ISSN 2415-7740 (Print)
ISSN 2415-7074 (Online)

Belarusian State University of Informatics and Radioelectronics

Open Semantic Technologies for Intelligent Systems

Research Papers Collection

Founded in 2017

Issue 4

Minsk
2020

UDC 004.822+004.89-027.31

The collection includes peer-reviewed articles approved by the Editorial board. Designed for university professors, researchers, students, graduate students, undergraduates, as well as for specialists of enterprises in the field of intelligent systems design.

Editorial address:
220005, Minsk, Platonova str., 39, rm. 606 b

Phone: +375 (17) 293-80-92

E-mail: ostisconf@gmail.com

Web-site: ***http://proc.ostis.net***

The collection is included in the Russian Science Citation Index

ISSN 2415-7740 (Print)
ISSN 2415-7074 (Online)

Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Открытые семантические технологии проектирования интеллектуальных систем

Сборник научных трудов

Основан в 2017 году

Выпуск 4

Минск
2020

УДК 004.822+004.89-027.31

Сборник включает прошедшие рецензирование и утвержденные Редакционной коллегией статьи.

Предназначен для преподавателей высших учебных заведений, научных сотрудников, студентов, аспирантов, магистрантов, а также для специалистов предприятий в сфере проектирования интеллектуальных систем.

Адрес редакции:

220005, г. Минск, ул. Платонова 39, каб. 606 б

Телефон: +375(17)293-80-92

Электронный адрес: ostisconf@gmail.com

Сайт: **http://proc.ostis.net**

Сборник включён в Российский индекс научного цитирования

OSTIS-2020 CONTRIBUTORS

ТЕХНИЧЕСКАЯ И ИНФОРМАЦИОННАЯ ПОДДЕРЖКА

TABLE OF CONTENTS

СОДЕРЖАНИЕ

FOREWORD

Research papers collection “Open Semantic Technology for Intelligent Systems” is devoted to the flexible and compatible technologies development that ensure the rapid and high-quality design of intelligent systems for various purposes.

The collection reflects research in the field of artificial intelligence in the following areas:

- Hybrid intelligent computer systems;
- Intelligent human-machine systems;
- Computer vision;
- Fuzzy computing ;
- Intelligent agents ;
- Intelligent automation;
- Knowledge management;
- Knowledge engineering;
- Ontological design;
- Semantic networks;
- Machine learning;
- Neural networks;
- Natural-language interface

The first article in this issue is dedicated to the memory of Dmitry Alexandrovich Pospelov, doctor of technical sciences, professor, academician of the Russian Academy of Natural Sciences, founder of the Soviet Association of Artificial Intelligence.

The main focus of this issue is on the standardization of intelligent systems.

In total, the collection contains 57 articles. The editors are thankful for all authors who sent their articles. Scientific experts selected for publication the best of the submitted works, many of them were revised in accordance with the comments of reviewers.

We are grateful our scientific experts for their great job in reviewing the articles in close cooperation with the authors. Their work allowed to raise the level of scientific results presentation, and also created a platform for further scientific discussions.

We hope that, as before, the collection will perform its main function — to promote active cooperation between business, science and education in the field of artificial intelligence.

Editor-in-chief
Golenkov Vladimir

ПРЕДИСЛОВИЕ

Сборник научных трудов «Открытые семантические технологии проектирования интеллектуальных систем» посвящен вопросам разработки гибких и совместимых технологий, обеспечивающих быстрое и качественное построение интеллектуальных систем различного назначения.

В сборнике отражены исследования в сфере искусственного интеллекта по следующим направлениям:

- Гибридные интеллектуальные компьютерные системы;
- Интеллектуальные человеко-машинные системы;
- Компьютерное зрение;
- Нечеткие вычисления;
- Интеллектуальные агенты;
- Интеллектуальная автоматизация;
- Управление знаниями;
- Инженерия знаний;
- Онтологическое проектирование;
- Семантические сети;
- Машинное обучение;
- Искусственные нейронные сети;
- Естественно-языковой интерфейс.

Первая статья в данном выпуске посвящена памяти Дмитрия Александровича Поспелова, доктора технических наук, профессора, академика РАЕН, основателя Советской Ассоциации искусственного интеллекта.

Основной акцент в этом выпуске сборника сделан на стандартизацию интеллектуальных систем.

В общей сложности сборник содержит 57 статей. Редакция сборника благодарит всех авторов, представивших свои статьи. Для публикации научными экспертами были отобраны лучшие из представленных работ, многие из них были переработаны в соответствии с замечаниями рецензентов.

Мы также благодарим экспертов за большой труд по рецензированию статей в тесном взаимодействии с авторами, который позволил повысить уровень изложения научных результатов, а также создал платформу для дальнейших научных дискуссий.

Надеемся, что, как и прежде, сборник будет выполнять свою основную функцию — способствовать активному сотрудничеству между бизнесом, наукой и образованием в области искусственного интеллекта.

Главный редактор
Голенков Владимир Васильевич

Искусственный интеллект: от прошлого к будущему. О жизненном пути и научном наследии профессора Д.А.Поспелова

В. Б. Тарасов
МГТУ им. Н.Э.Баумана
Москва, Росссия
Vbulbov@yahoo.com

***Аннотация*—В работе прослежены этапы жизненного пути и рассмотрены главные составляющие научного наследия выдающегося советского и российского учёного, основателя научного направления «Искусственный интеллект» и ведущей школы по искусственному интеллекту в СССР, инициатора создания и первого президента Советской (затем Российской) ассоциации искусственного интеллекта, доктора технических наук, профессора, академика Российской академии естественных наук Дмитрия Александровича Поспелова. Им открыты новые междисциплинарные научные области, такие как ситуационное управление и прикладная семиотика, псевдофизические логики, нечёткие и кольцевые шкалы, построены формальные модели индивидуального и коллективного поведения, разработаны элементы теории искусственных агентов и многоагентных (децентрализованных) систем, предложены оригинальные методы описания рассуждений «здравого смысла, введены уровни понимания для искусственных систем**

***Ключевые слова*—Искусственный интеллект; Психоника; Ситуационное управление; Семиотическое моделирование; Прикладная семиотика; Псевдофизическая логика; Модель внешнего мира; Модель поведения; Агент; Многоагентная система; Когнитивное моделирование: Оппозиционная шкала.**

I. Введение

19 декабря 2019 года исполнилось 87 лет со дня рождения выдающегося учёного, основоположника искусственного интеллекта в СССР и России, доктора технических наук, профессора, действительного члена Российской академии естественных наук, основателя и первого президента Советской (впоследствии Российской) ассоциации искусственного интеллекта (ИИ) **Дмитрия Александровича Поспелова**.

С его именем связана целая эпоха формирования и развития, официального признания и расцвета ИИ в нашей стране на протяжении почти четырёх десятилетий XX-го века. Для новой тогда науки это было романтическое время.

Дмитрий Александрович был прирождённым лидером, учёным с широчайшим кругозором и удивительной способностью к научному синтезу, успешно работавшим на стыке различных и ранее казавшихся далёкими друг от друга дисциплин. Он был мощным и никогда не терявшим жизненной энергии генератором идей, многие из которых обращены в будущее ИИ. При этом у Д. А. Поспелова был настоящий дар научного общения: вокруг него всё время собирались группы талантливых и интересных людей, приходивших из самых разных областей науки, техники и культуры, а он успешно находил им нужные «экологические ниши» для творческого развития и применения их профессионального опыта в искусственном интеллекте.

Профессор Д. А. Поспелов ушёл из жизни совсем недавно, 30 октября 2019 года, после тяжёлой и продолжительной болезни, с которой мужественно боролся на протяжении 21 года. Даже долгая жизнь, полная творческих достижений, скоротечна: за первые два десятилетия XXI-го века выросло новое поколение научных работников, аспирантов, студентов, живо интересующихся проблемами ИИ, которые никогда в жизни не видели Дмитрия Александровича, не подпадали под обаяние его яркой личности, и ничего не знают о его выдающихся научных достижениях. Им не довелось слышать его великолепные, глубокие научные доклады и увлекательные популярные лекции, читать полные новыми мыслями, богатыми ассоциациями и неожиданными метафорами книги, просто встречаться с ним и ездить на конференции, беседуя с удивительно эрудированным человеком на самые разные и необычные темы -– от архитектурных канонов русских церквей (дед Димы Поспелова был священником) до природы Фанских гор, куда он, будучи инструктором по альпинизму, не раз ходил в молодые годы.

Трудно серьёзно заниматься наукой, не ведая её истории и совсем не зная её творцов. Эта статья адресована заинтересованным представителям научной молодёжи. Автор будет рад, если знакомство с личностью, научными результатами и идеями профессора Д. А. Поспелова вызовет отклик, пробудит интерес, а, главное, желание понять, развить и воплотить в прикладных системах хотя бы часть нереализованных замыслов этого выдающегося учёного.

Всё же, в первую очередь, эта статья – для его

Рис. 1. Посвящается участникам учредительного съезда Советской ассоциации искусственного интеллекта

Рис. 2. Академик РАЕН, профессор Дмитрий Александрович Поспелов -– основоположник искусственного интеллекта в СССР и России

учеников и соратников, тех, кто был долгое время рядом с Д. А. Поспеловым, участников учредительного съезда нашей Ассоциации искусственного интеллекта. Для тех, кого не перестанут удивлять неизвестные грани его творчества и неисчерпаемость идей.

II. ЭТАПЫ НАУЧНОЙ БИОГРАФИИ Д. А. ПОСПЕЛОВА

Дмитрий Поспелов родился 19 декабря 1932 года в Москве. В 1956 году окончил механико-математический факультет МГУ им. М.В.Ломоносова по специальности «Вычислительная математика» и был распределён в Московский энергетический институт (МЭИ) на кафедру вычислительной техники (ВТ).

A. Работа в МЭИ, семинар по психонике и защита докторской диссертации

Уже в первые годы работы в МЭИ, на кафедре ВТ, наряду с преподавательской деятельностью, Дмитрий Поспелов активно занимался научными исследованиями по теории вычислительных систем, теории автоматов, теории игр, многозначным логикам. Важными результатами этих исследований стали его ранние монографии, написанные в в 1960-е годы и получившие международный резонанс: «Логические методы анализа и синтеза схем» [1] (выдержала три издания в СССР в 1964, 1968 и 1974 гг., переведена на болгарский и немецкий языки), «Игры и автоматы» [2] (переведена

Рис. 3. Дмитрий Поспелов в студенческие годы

Рис. 4. Школа по вероятностным автоматам в Казани (1964 г.). Слева направо: Д. А. Поспелов, В. Н. Захаров, О. П. Кузнецов.

на испанский и польский языки), «Вероятностные автоматы» [3].

В январе 1968 г. Д. А. Поспелов защитил на заседании совета факультета автоматики и вычислительной техники МЭИ докторскую диссертацию на тему «Проблемы теории вычислительных систем» [4]. При подготовке докторской диссертации им был создан аппарат ярусно-параллельных форм, позволивший ставить и решать ряд проблем, связанных с организацией и проведением параллельных вычислений в вычислительных комплексах и сетях. На его основе ещё в 1960-1970-е годы успешно решались такие проблемы, как синхронное и асинхронное распределение программ по машинам компьютерной системы, задачи оптимальной сегментации программ, а также важные оптимизационные задачи, связанные с распределением информационных обменов.

Уже к концу 1960-х годов в МЭИ возникла научная школа Д. А. Поспелова [5]: его первые аспиранты Ю. И. Клыков [6] и В. Н. Вагин [7] защитили кандидатские диссертации в 1967 и 1968 годах; впоследствии они стали докторами наук, известными учёными по ситуационному управлению и прикладной логике для ИИ соответственно. В создании этой школы и формировании творческих коллективов будущих междисциплинарных исследований по ИИ в СССР большую роль сыграл научный семинар по психонике [8,9], который работал в МЭИ на протяжении почти 7 лет с 1964 по 1970 г. За эти годы было проведено 92 заседания семинара [8]. Инициатором и душой этого семинара был Д. А. Поспелов.

В заседаниях семинара принимали участие видный психолог Б. В. Зейгарник, известные учёные-кибернетики М. Г. Гаазе-Рапопорт и Е. Т. Семёнова, и молодые учёные, многие из которых стали впоследствии знаменитыми. Это — психологи В. П. Зинченко и В. Н. Пушкин, психолог и математик В. А. Лефевр, кибернетик и будущий известный специалист по искусственному интеллекту Л.И.Микулич, будущий «отец машинной музыки» Р. Х. Зарипов и, конечно, аспиранты и молодые сотрудники МЭИ -– А. В. Бутрименко, В. П. Кутепов, В. В. Железнов, и др.

Предыстория семинара такова. 12 лет работы Дмитрия Александровича в МЭИ с 1956 по 1968 г. пришлись на первый период кибернетического бума, когда только рождались новые научные направления и дисциплины, связанные с отображением и имитацией важнейших характеристик сложных естественных систем в искусственных системах.

Одна из таких дисциплин -– *бионика* -– провозгласила своей целью практическое применение в технических системах тех биологических механизмов и принципов действия, которые природа «отработала» в ходе эволюции живых организмов. В сферу интересов бионики попадают автономные и гомеостатические системы, искусственные нейроны и эволюционные системы, искусственные конечности, управляемые биотоками, и биотехнические системы, где в качестве одной или нескольких подсистем используются живые организмы.

Вариантом реализации имитационной, бионической программы в ИИ стал *нейробионический* подход, приверженцы которого считают, что воспроизведение интеллектуальных процессов практически невозможно без воспроизведения их материальных носителей, т.е. тех процессов, которые протекают в специальным

Рис. 5. Д.А.Поспелов с любимым учеником В.Н.Вагиным (1968 г.)

образом организованной биологической ткани. Другими словами, основной целью этого подхода является создание искусственного мозга, структура и функционирование которого идентичны биологическому мозгу. В рамках нейробионического подхода ведутся исследования по разработке искусственных нейронов и построению на их основе структур, подобных биологическим.

Ещё на заре ИИ стало ясно, что бионика в ее исходном понимании в большей степени охватывает проблемы «тела» технических систем, а вопросы их «одушевления» оказываются как бы на периферии. Бионика того времени в основном оставалась на уровне рефлекторных процессов, моделей самосохранения, обмена и адаптации. Однако, для понимания естественных интеллектуальных систем важное место занимает исследование процессов целеобразования, формирования и принятия решений в сложных ситуациях, классификации и оценки ситуаций и многое другое, что традиционно связано с психологией, а не с биологией или физиологией. Основным отличием новых технических устройств от обычных компьютеров является наличие в них модели внешнего мира, которая отражает соотношения этого мира с помощью системы базовых понятий и отношений различного типа между этими понятиями (онтологии в современных терминах).

Так с «лёгкой руки» Д. А. Поспелова родилась психоника -– научная область, основной задачей которой стало изучение и использование в интеллектуальных системах результатов, добытых психологами при изучении психики человека и способов организации человеческой деятельности [9]. На первом заседании семинара был прочитан доклад Д. А. Поспелова «О задачах психоники» (см. [9]), в котором было отмечено появление круга проблем, которые с одной стороны являются предметом исследований психологов и лингвистов, а с другой стороны, представляют большой интерес для инженеров, специалистов в области технических наук, математиков, программистов, которые занимаются проектированием устройств, предназначенных для целенаправленной, «логической» деятельности. К таким проблемам относятся вопросы целенаправленного поведения, мотивации поступков, выработки абстрактных понятий, проведения индуктивных выводов, и т.д.

Автором были указаны три направления в психонике: 1) создание в машине внутренней модели внешней среды; 2) принятие целесообразных решений; 3) моделирование личности и коллективов.

Было отмечено, что наделение технических систем аналогами личностных характеристик позволило бы строить управляющие устройства, обладающие некоторой «субъективностью» подхода к решению задачи. Например, при моделировании игровой ситуации следует учитывать такие поведенческие аспекты как обман и блеф.

Рис. 6. Д.А.Поспелов выступает на первом семинаре по психонике в МЭИ: «модель внешнего мира – основа технических устройств нового типа»

Сам термин «гиромат» и идея подстройки структуры модели к особенностям решаемой задачи были заимствованы из произведений польского писателя фантаста Ст.Лема. Процедура принятия решений в гироматах опиралась на соотнесение текущей ситуации с некоторым типовым классом событий или ситуаций, для которого предполагалось известным априорное решение. При этом гиромат уже содержал «агентообразующие» модули: блок мотивации; блок селекции (рецепторы); блок построения внутренней модели внешней среды; блок выдвижения гипотез; блок модельного опыта; блок выработки решений; блок активного опыта; блок времени. Общая идея работы гиромата представлена на рис. 7.

Информация о текущей ситуации в дискретной ситуационной сети поступает в блок селекции. Из него она идет по двум направлениям: в блок построения модели, где происходит классификация ситуаций, и в блок гипотез, в котором вырабатываются различные гипотезы о связи ситуаций между собой. В эти же блоки из блока времени поступает информация о времени, когда была зафиксирована данная ситуация. Она используется для оценки повторяемости конкретных ситуаций и изменений во времени. В результате блок построения модели изменяет внутреннюю модель мира, хранящуюся в гиромате. В модели внешнего мира хранится весь опыт гиромата, накопленный им в процессе функционирования. На основе этой информации блок модельного опыта извлекает информацию, требуемую для принятия решений, и передает её в блок выработки решений.

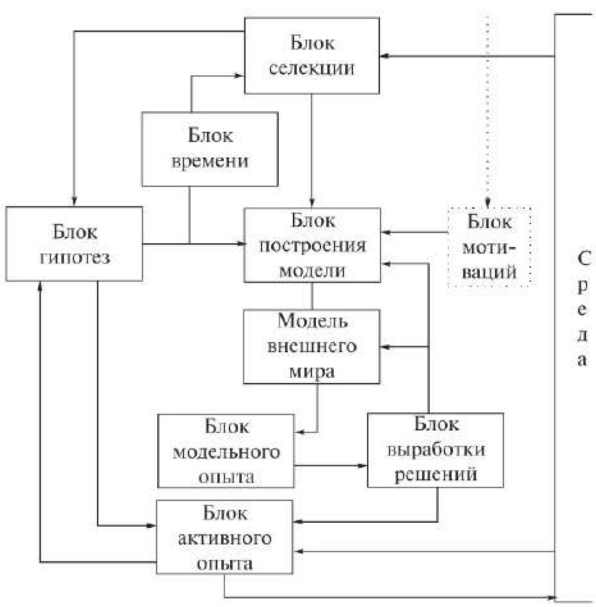


Рис. 7. Общая схема гиромата

Если рассматриваемая ситуация оказалось типичной, уже знакомой гиромату, то блок выработки решений формирует задание блоку активного опыта на выдачу воздействий на среду. В противном случае требуется дополнительное обращение к блоку гипотез. Блок активного опыта может непосредственно анализировать реакцию среды и передавать корректирующую информацию в блок гипотез.

Следует напомнить, что пионерская работа К. Хьюитта в области теории агентов «Viewing Control Structures as Patterns of Message Passing» [11] вышла в свет в 1977 году тогда как первые публикации Д.А.Поспелова по психонике и модели агента-гиромата, появились на десять лет раньше.

На заседаниях семинара, проходивших в 1967-1968 г. сформировались задачи, подходы и методы, составившие основу ситуационного управления. В нём нашли прямое отражение идеи гироматов.

Сам Д. А. Поспелов отмечал [8], что семинар по психонике в МЭИ стал одним из ключевых мероприятий, связанных с зарождением ИИ в нашей стране. Хотя в отличие от бионики, сам термин «психоника» так и не прижился в научной среде, сами исследования в этой области привели к формированию междисциплинарного научного сообщества, необходимого для развёртывания широкого фронта работ, относящихся к искусственному интеллекту. К этому можно добавить, что, начиная с середины 1960-х годов, сам Дмитрий Александрович выполнил немало экспериментальных исследований совместно с психологами. В первую очередь, искались альтернативы чисто бихевиористским подходам к решению задач, опиравшимся на лабиринтную модель. Именно результаты таких совместных исследований способствовали формированию широкого менталистского воззрения Д.А.Поспелова на предмет искусственного интеллекта, отнюдь не сводящегося к классической схеме инженерии знаний. В соответствии с идеями ментализма, адекватная характеристика сложного, целенаправленного человеческого поведения и самой жизни невозможна без привлечения сознания или психики в целом как «средств объяснения».

B. Переход в Вычислительный Центр, развитие ситуационного управления, семиотический пролог

В 1968 году Д. А. Поспелов перешёл на работу в Вычислительный центр АН СССР, где главной областью его профессиональной деятельности стало только возникшее в ту пору новое научное направление – искусственный интеллект (ИИ). Он проработал в ВЦ АН СССР (затем ВЦ РАН) 30 лет, с 1968 по 1998 год: сначала в должности заведующего сектором в Лаборатории теории и программирования больших систем, во главе которой стоял академик (тогда ещё член-корр. АН СССР) Г. С. Поспелов, а затем -– в должности заведующего отделом проблем искусственного интеллекта [12].

Также в 1968 году он стал профессором Московского физико-технического института.

В 1970-е годы вышли в свет новые книги Д. А. Поспелова -– «Введение в теорию вычислительных систем» [13], «Мышление и автоматы» (в соавторстве с В.Н.Пушкиным) [14] (переведена на чешский язык), «Системы управления» [15] (в соавторстве с В. Н. Захаровым и В. Е. Хазацким), а также «Большие системы. Ситуационное управление» [16]. Так в книге

«Мышление и автоматы» две области исследования – психология и теория автоматов – были объединены в одно общее русло. Авторы убедительно показали, что кибернетическая реализация методов психологических исследований может принести серьезный практический эффект.

Уже с начала 1970-х годов в Лаборатории больших систем стали активно проводиться исследования по ИИ. Впоследствии были сформированы научные группы по разработке интеллектуальных диалоговых систем, интеллектуальных решателей и планировщиков, интегрированных интеллектуальных систем.

В секторе Д. А. Поспелова первоначально основное внимание уделялось проблемам и методам *ситуационного управления сложными системами* [17,18]. 1-я Всесоюзная конференция по ситуационному управлению была проведена в июне 1972 г., а её избранные доклады вошли в сборник «Вопросы кибернетики. Вып. 13. Ситуационное управление: теория и практика», часть 1, изданный в 1974 году.

Ситуационное управление требуется для класса больших (или сложных) систем, таких как город, морской порт, транснациональное предприятие, где невозможна или нецелесообразна формализация процесса управления в виде математических уравнений, а доступно лишь его описание в виде последовательности предложений естественного языка с помощью логико-лингвистических моделей. На основе экспертной информации строится классификатор, позволяющий разделять все наблюдаемые ситуации на нечёткие классы (образующие покрытие, но не разбиение). Для описания ситуаций используются семантические сети и близкие к ним модели знаний.

Рис. 8. Д.А.Поспелов с участниками III-го Всесоюзного симпозиума по ситуационному управлению большими системами в Одессе

Обычно под семиотикой понимают гуманитарную, описательную науку о знаковых системах. При этом считается, что семиотика — удел лингвистов. Высокий потенциал Д. А. Поспелова в сфере междисциплинарного синтеза проявился при создании логико-лингвистических моделей, в которых логические средства применяются для преобразования данных, описанных в лингвистической форме. В результате им были разработаны основы *прикладной семиотики* -— новой синтетической научной дисциплины, рассматривающей вопросы использования знаков и знаковых систем в системах представления, обработки и использования знаний при решении различных практических задач.

Хотя официально этот термин появился лишь в 1995 г. в ходе дискуссий на российско-американских семинарах «Российское ситуационное управление и кибернетика» и «Архитектуры для семиотического моделирования и ситуационный анализ в больших сложных системах», одна из первых, фундаментальных работ Д. А. Поспелова по *прикладной семиотике* датируется 1970-м годом [19]. В её начале, аргументируя необходимость применения семиотических моделей в ИИ, он писал, что все технические устройства работают на досемиотическом уровне, поэтому они способны моделировать лишь простейшие формы поведения и имеют существенные ограничения в плане решения творческих задач. В отличие от технических систем высшие животные и человек решают подобные задачи на семиотическом уровне, что позволяет им находить такие способы решения, которые невозможно реализовать на досемиотическом уровне.

При этом в [19] знак трактовался им как единство сигнала и его значения, а знаковая система понималась как совокупность простых знаков и правил образования сложных знаков из совокупности простых. Структура знаковой системы в некотором смысле изоморфна системе отношений между объектами реального мира. Такую систему знаков можно назвать системой знаков первого уровня (или псевдофизической семиотической системой).

Но, кроме знаков, значениями которых выступают предметы или явления реального (или моделируемого) мира, можно рассматривать знаки знаков (метазнаки), значениями которых служат знаки семиотической системы первого уровня. Такую систему следует отнести к знаковой системе второго уровня. Путем индукции нетрудно ввести системы знаков любого k-го уровня.

Иерархия знаковых систем имеется у человека (например, система знаков естественного языка служит для знаковой системы математики системой нижнего, предшествующего уровня).

Там же в [19] была выдвинута следующая гипотеза Д. А. Поспелова: условием синтеза технических устройств, способных к решению творческих задач, является возможность формирования внутри такого устройства системы знаковых систем. По сути, эта статья явилась «манифестом будущей семиотической революции» в искусственных системах, вызывая и теперь большой интерес в плане приложения её идей к развитию интернета вещей и когнитивных роботов.

Комплекс методов управления, предложенных Д. А. Поспеловым, его учениками и последователями, опирается на гибридные логико-лингвистические (семиотические) модели, где логические средства используются для преобразования данных, выраженных в лингвистической форме. В отличие от классической формальной системы, которая моделирует замкнутый, полностью заданный мир, семиотическая модель позволяет описывать процессы, протекающие в открытых или динамических системах.

В СССР ещё в 1970-е годы с помощью методов ситуационного управления и средств семиотического моделирования были созданы эффективные системы оперативного диспетчерского управления такими сложными объектами, как грузовой морской порт, атомная электростанция, комплекс трубопроводов, автокомбинат, и др. Методы этого типа находят широкое применение и сегодня, особенно в области организационно-технического управления.

Также ещё в 1970-е годы у Д. А. Поспелова возникла идея *псевдофизических логик*, где пропозициональными переменными являются лингвистические переменные Л.Заде, имеющие в качестве значений слова естественного языка, которые выражаются в виде нечетких множеств на универсальной шкале. Появление нечётких классов в ситуационном управлении, а также использование лингвистических и нечётких переменных в создаваемых псевдофизических логиках предопределили его живой интерес к теории нечётких множеств и её приложениям.

Здесь «первой ласточкой» стал рижский семинар «Применение теории нечётких множеств в задачах управления сложными системами».

C. Официальное признание искусственного интеллекта в СССР и установление зарубежных контактов на «высшем научном уровне»

Профессор Дмитрий Александрович Поспелов стоял у истоков организации первых академических структур по ИИ. Переход Д. А. Поспелова в ВЦ АН СССР привёл к возникновению «дуэта Поспеловых», ставшего ключевым фактором становления искусственного интеллекта в Советском Союзе. В 2019 году исполнилось 45 лет со дня (точнее, двух дней) официального признания ИИ в СССР. Здесь выдающуюся организационную роль сыграл академик Гермоген Сергеевич Поспелов. Будучи главным инициатором и убеждённым сторонником развития ИИ в СССР, академик Г. С. Поспелов посвятил почти два десятилетия своей жизни тому, чтобы наши исследования в этой области получили признание Академии наук СССР и вышли на мировой уровень [20]. Его верным соратником в этом деле, а затем и «главным локомотивом» отечественного ИИ был Дмитрий Александрович Поспелов.

В январе 1974 г. был образован Совет по проблеме «Искусственный интеллект» в рамках Комитета по системному анализу Президиума АН СССР. Председателем совета стал Г. С. Поспелов, а его заместителем -– Д. А. Поспелов. Первым крупным научным мероприятием, организованным новым советом, было проведение в Тбилиси 4-6 июня 1974 г. 7-го всесоюзного симпозиума по кибернетике -– первого в нашей стране представительного форума, целиком посвященного проблемам ИИ. Главным докладом, прочитанным на его пленарном заседании, стал совместный доклад Г. С и Д. А. Поспеловых «Основные проблемы искусственного интеллекта», переработанная версия которого затем была опубликована в Вестнике АН СССР [21].

В декабре того же 1974 г. в рамках Научного совета по комплексной проблеме «Кибернетика» при Президиуме АН СССР была создана Секция по проблеме «Искусственный интеллект», во главу которой также стал «дуэт Поспеловых». С формирования двух указанных организационных структур в системе Академии наук началась официальная история развития ИИ в СССР.

В результате образования этих научных структур появились первые отечественные комплексные проекты, охватывавшие прорывные области исследований по созданию прикладных интеллектуальных систем и автономных роботов, в частности, проект «Ситуация» по главе с Д. А. Поспеловым, в рамках которого были сконцентрированы все работы по ситуационному управлению.

Кроме того, под эгидой этих организаций в СССР было проведено два важнейших международных мероприятия по ИИ: IV-я Международная объединённая конференция по искусственному интеллекту IJCAI-75 (Тбилиси, 3-8 сентября 1975 г.) и спустя полтора года Международное совещание по искусственному интеллекту в Репино под Ленинградом (18-24 апреля 1977 г.)

В работе этих форумов принимал участие «весь цвет» мирового искусственного интеллекта: Дж. Маккарти, М. Арбиб, Л. Заде, Д. Ленат, Д. Мики, Дж. Мур, Н. Нильсон, П. Уинстон, К. Хьюитт, Э. Фредкин, Э. Сандевал, Ж.-К. Симон и др. (см.[22]).

По материалам конференции в Репино был издан том 9 серии Machine Intelligence (главный редактор серии -– Д. Мики) [23]. Этот сборник стал «местом встречи», где первый раз были широко представлены работы как крупнейших западных, так и ведущих советских учёных в области искусственного интеллекта. В нём была опубликована статья Г. С. и Д. А. Поспеловых «Influence of Artificial Intelligence Methods on the Solution of Traditional Control Problems», в которой

Рис. 9. Академик Г. С. Поспелов и профессор Д. А. Поспелов на Международном совещании по искусственному интеллекту в Репино (1977 г.).

Рис. 10. На Международном совещании по искусственному интеллекту в Репино. Справа налево: вопрос задаёт Л. Заде, рядом с ним сидят: Дж. Маккарти, В. И. Варшавский, Д. А. Поспелов.

впервые на английском языке были подробно описаны недостатки классической теории автоматического управления как предпосылки появления ситуационного управления, сам метод ситуационного управления и его приложения.

По свидетельству В. Л. Стефанюка [22], именно после двух упомянутых конференций оригинальные работы Д. А. Поспелова приобрели международную известность, а сам он выдвинулся на позиции основоположника ряда научных направлений, тесно связанных с искусственным интеллектом.

III. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ПАРАДИГМА Д. А. ПОСПЕЛОВА

Таким образом, к концу 1970-х годов у Д. А. Поспелова уже сложился общий взгляд на искусственный интеллект как новую «науку-перекресток», предполагающую синтез подходов, методов и моделей естественных, технических и гуманитарных наук. Сам он не раз подчеркивал, что теоретические проблемы ИИ возникают и исследуются на стыке философии, психологии, лингвистики, семиотики, логики, этики, а необходимым инструментом построения формальных моделей и прикладных интеллектуальных систем являются методы и средства прикладной математики (включая прикладную логику), теории систем, теории управления, информатики и вычислительной техники, программирования [24]. На рис.11 приведена общая иллюстрация представлений Д. А. Поспелова об ИИ как «системе наук» (см. также [25]).

Рис. 11. Искусственный интеллект -– синтетическая «наука-перекресток»

Эта иллюстрация наглядно подтверждает, что научное творчество Д. А. Поспелова носит междисциплинарный характер. Недаром среди его близких друзей, соавторов, организаторов совместных секций на конференциях по ИИ, членов научного совета Ассоциации искусственного интеллекта были известные философы (Б. В. Бирюков, Ю. А. Шрейдер, Д. Б. Юдин), специалисты по системному анализу (Г. С. Поспелов, В. Н. Садовский, Ю. М. Горский), психологи (В. Н. Пушкин, В. П. Зинченко, О. К. Тихомиров, Б. М. Величковский, В. Ф. Петренко), лингвисты (И. Н. Горелов, Б. Ю. Городецкий, А. Е. Кибрик, В. В. Мартынов, Е. В. Падучева), логики (В. К. Финн, В. А. Горбатов, Н. Н. Непейвода). Своими учителями в науке Дмитрий Александрович называл выдающихся учёных в области кибернетики А. И. Берга и М. А. Гаврилова.

К числу открытых им новых научных направлений можно отнести *ситуационное управление* и *прикладную* (когнитивную) *семиотику*, *логико-лингвистические модели* и псевдофизические логики, когнитивную семантику, включая нечёткие и кольцевые шкалы, модели диалога и понимания, и формализованные *модели поведения*, в том числе фреймы поступков и структуры волшебных сказок. Трудно переоценить роль Д. А. Поспелова в исследовании таких ключевых теоретических проблем ИИ, как представление и организация знаний и моделирование рассуждений «здравого смысла», методы когнитивной графики и отражения образного мышления специалиста в искусственных системах. Нельзя не отметить его существенный, но пока явно недооценённый вклад в развитие теории агентов и многоагентных систем. На наш взгляд, Дмитрий Александрович Поспелов, наряду с К. Хьюиттом и М. Л. Цетлиным, должен считаться одним из основоположников теории искусственных агентов, хотя сам термин «искусственный агент» появился позже, и никто из этой тройки его не использовал (у Хьюитта был «актор»). Его идеи психоники как «психологии искусственных агентов» в целом и структура гиромата в частности, модели индивидуального и коллективного поведения, общая архитектура организма как агента, ранние результаты в области семиотических и децентрализованных систем являются неоспоримым свидетельством приоритета в данной области.

«Золотым веком» научного и литературного творчества Д. А. Поспелова стали 1980-е годы. Кипящая в ту пору научная жизнь, удивительно широкая эрудиция Дмитрия Александровича, его частые и увлекательные выступления на лекциях в Политехническом музее привели к большой популярности автора новых, необычных, междисциплинарных, полных метафор книг. Выход в свет каждой монографии профессора Д. А. Поспелова по ИИ, инженерии знаний, ситуационному и децентрализованному управлению, моделированию поведения был важным событием для многочисленных читателей; на протяжении десятилетия автор радовал их почти каждый год.

В период с 1981 по 1990 годы им было опубликовано восемь блестящих книг: «Логико-лингвистические модели в системах управления» [26] «Фантазия или наука: на пути к искусственному интеллекту» [27], «Оркестр играет без дирижера. Размышления об эволюции некоторых технических систем и управлении ими» (в соавторстве с В. И. Варшавским)[28], «Ситуационное управление: теория и практика»[29], «От амёбы до робота: модели поведения» (с М. Г. Гаазе-Рапопортом) [30], «Представление знаний о времени и пространстве в интеллектуальных системах» (совместно с Е. Ю. Кандрашиной и Л. В. Литвинцевой) [31], «Моделирование рассуждений» [32], «Нормативное поведение в мире людей и машин» (в соавторстве с В. А. Шустер) [33]. Одна из первых отечественных монографий по теории и приложениям нечётких множеств «Нечёткие множества в моделях управления и искусственного интеллекта» [34], написание которой было инициировано Д. А. Поспеловым, издана под его редакцией.

A. От психоники к Artificial General Intelligence

В заключительной части своей книги «Фантазия или наука: на пути к ИИ» Дмитрий Александрович Поспелов выдвинул программу будущих исследований в области искусственного интеллекта. Эти исследования должны быть нацелены на *«изучение психики человека* с целью ее имитации в технических системах, решающих определенный набор практических задач, традиционно считающихся интеллектуальными» [27, с. 211]. Такое представление целей ИИ является чрезвычайно широким и намного опередило свое время.

Чтобы обосновать это утверждение, вначале обратимся к определениям, предложенным «отцами-основателями» на заре ИИ. Так П. Уинстон [35, с. 11] определял ИИ весьма размыто как «науку о концепциях, позволяющих компьютерам делать такие вещи, которые у людей выглядят разумными». Методология, разрабатываемая для того, чтобы сделать разумнее машины, также может быть использована и для того, чтобы сделать разумнее самих людей. Схожей позиции придерживался А. Хоффман: ИИ стремится открыть общие принципы работы интеллекта, которые могут эффективно применяться даже без помощи программных средств [36].

Согласно Н. Нильссону, задача ИИ – создание *теории интеллекта*, базирующейся на обработке информации [37]. По Г. Саймону, теоретические единицы в ИИ могут быть трёх типов [38]: а) единицы описания самой интеллектуальной системы; б) единицы описания её среды (например, условия задачи или ограничения, влияющие на решение); в) единицы описания поведения системы в среде.

Наконец, патриарх искусственного интеллекта и автор самого этого термина Дж. Маккарти считал, что главной проблемой ИИ является *построение искусственных систем*, способных работать на уровне *«здравого смысла»*.Общая проблематика ИИ включает две ключевые составляющие: проблему представления (эпистемология ИИ) и проблему эвристик (какими процедурами и какими средствами решаются задачи, в частности, как строить гипотезы на основе имеющихся данных и как проводить рассуждения) [39].

В 2005-м году в своём приветственном обращении к читателям выпуска журнала AI Magazine, посвящённом 50-летию искусственного интеллекта, — «The Future of AI -– a Manifesto» и затем в 2007-м году в статье [40] он подчёркивал, что долгосрочная цель искусственного интеллекта как науки заключается в создании ИИ человеческого уровня. При этом Маккарти полагал, что основной надеждой здесь является развитие логического подхода к формализации повседневных знаний и рассуждений.

В целом, начиная с исторического Дартмутского семинара 1956 года, первопроходцы ИИ Дж. Маккарти и М. Минский, авторы общего решателя задач GPS Г. Саймон и А. Ньюэлл, как и многие другие, на протяжении всей научной карьеры мечтали разработать компьютерные системы, способные подобно человеку решать разные задачи из различных предметных областей. Однако это проблема оказалась слишком сложной и не разрешимой в рамках классической парадигмы инженерии знаний.

В середине 2000-х годов появилась новая волна исследователей в сфере «целостного интеллекта». Стали популярными такие направления как: «искусственный общий интеллект» AGI (от Artificial General Intelligence) [41,42], «сильный ИИ», «искусственное сознание», развиваемое, в частности, в рамках ассоциации BICA (в переводе: биологически инспирированные когнитивные архитектуры) и её конференций (см., например, [43]). Недавно возникла оригинальная концепция «открытого и неоканчиваемого ИИ» (Open-Ended AI [44].

В России также появились новые стратегии разработки интеллектуальных технологий, в частности, «Поведение, управляемое сознанием» [45] и синергетический ИИ [46]. Общим пунктом требований к любому проекту в сфере AGI является разработка концепции и модели *целостного интеллекта* и её инженерная реализация с помощью программно-аппаратных средств.

Начальным этапом формирования модели целостного интеллекта может быть выделение различных подмножеств в совокупности способностей естественного интеллекта человека. Согласно В.К.Финну [47, с.37], эта совокупность включает: 1) выделение существенного в знании; 2) способность к рассуждениям; 3) способность к рефлексии; 4) выдвижение цели и выбор средств ее достижения, т.е. способность к целеполаганию и планированию поведения; 5) познавательную активность; 6) адаптацию к ситуации; 7) формирование обобщений; 8) синтез познавательных процедур (т.е. взаимодействие индукции, аналогии и абдукции; 9) способность к отбору знаний; 10) способность к обучению и использованию памяти; 11) способность к аргументированному принятию решений, использующему представление знаний и результаты рассуждений,, соответствующие поставленной цели; 12) способность к рационализации идей, стремление уточнить их как понятия; 13) способность к созданию целостной картины предметной области.

Однако в первых проектах AGI гипертрофированное внимание уделяется способности к обучению в ущерб остальным.

Всё это демонстрирует актуальность психонической парадигмы Д. А. Поспелова в плане синтеза будущей концепции целостного интеллекта. В самом деле, имитация психики означает моделирование как сознания, так и бессознательного, а сознание, мыслимое как единство интеллектуальной, волевой и чувственной сфер, является главным регулятором поведения. Таким образом, интеллект рассматривается как открытая, неоднородная динамическая система.

В методологическом плане данную позицию можно обозначить как «умеренный функционализм», который предполагает возможность абстрагировать характерные свойства некоторого явления и воспроизвести их на других носителях. Речь идет о воспроизведении в искусственном агенте основных функций человеческого интеллекта (а, в более широком плане, психики человека) без учета лежащих за ними физиологических явлений.

Моделирование психики в контексте проектирования искусственного агента может опираться на общую классификацию психических явлений, среди которых выделяются психические процессы, состояния и свойства. *Психический процесс* характеризуется достаточно однородной структурой и сравнительно кратковременным протеканием. Его длительность колеблется от долей секунды (время сенсомоторной реакции) до десятков минут и нескольких часов (при чтении или слушании музыки). Психические процессы делятся на *когнитивные* (ощущения, восприятие, память, внимание, мышление, воображение, речь) и *эмоционально-волевые*.

Таким образом, «психология искусственных систем» опирается не только на когнитивное моделирование, но и на имитацию функций эмоций, в особенности, оценочной функции. Так распознавание роботом эмоций человека в партнёрской системе «человек – робот» может определять дальнейшее поведение робота. Сам Д. А. Поспелов уже в [19] рассмотрел вариант построения уровня эмоций в семиотической системе.

В то же время, *психическое состояние* есть «спутник» деятельности естественного агента, который проявляется в его повышенной или пониженной активности. С одной стороны, состояние агента является прямым эффектом его деятельности, а с другой стороны – фоном, на котором она протекает. По сравнению с процессами состояния естественных агентов имеют большую длительность (дни, недели).

Психические состояния делятся на *мотивационные*, т.е. основанные на потребностях (желания, интересы, стремления, влечения); *эмоциональные* (радость, горе, удовлетворение, стресс, фрустрация); *волевые* (решимость, настойчивость). В основе реактивного агента лежит простое стремление к состоянию удовлетворения (гедонизм), а исполнение желания интеллектуального агента обычно предполагает его настойчивость, принося радость.

Психические свойства -– это устойчивые образования, обеспечивающие определенный уровень деятельности или тип поведения. К ним относятся темперамент, характер, способности. Например, такие черты характера как эгоизм и альтруизм определяют диаметрально противоположные типы социального поведения агентов, возможности формирования и пути развития многоагентных систем.

В монографии [27] Д. А. Поспелов также обратился к проблеме понимания (с. 183-189). Кроме классификации элементов внешнего мира, у интеллектуальной системы должны быть и более развитые метапроцедуры, суть которых определяется термином «понимание». Термин этот многозначен. Первое из его уточнений принадлежит лингвистам, которые занимаются машинным переводом. В контексте понимания у них быстро возникла идея перехода от обычной схемы «язык–язык» к схеме «язык — действительность -– язык». Отсюда появилось следующее определение метапроцедуры понимания. Понимание -— это процесс соотнесения языкового описания с внеязыковой ситуацией. Другими словами, понимание является процедурой связи между семиотической системой естественного языка с теми образами, которые формирует в нас система чувственного восприятия мира.

B. Техноценозы и модели поведения: на пути к гибридному интеллекту в смешанных сообществах людей и машин

Затем вышла в свет монография «Оркестр играет без дирижёра» [28], посвящённая проблемам коллективного поведения, децентрализованного управления и эволюции сложных распределённых систем. В децентрализованных системах управление происходит за счёт локальных взаимодействий, а распределённые системы могут иметь единый орган управления.

Авторы книги — В. И. Варшавский и Д. А. Поспелов -— задаются вопросом «зачем нужна децентрализация» в сложных системах?» и приводят следующие аргументы в её пользу: 1) не все технические и, тем более, организационные и экономические системы появились на свет благодаря единому проектного замысла; многие возникли из более простых систем в результате технической эволюции (классический современный пример – сеть Интернет); 2) сложность технических систем достигла такого уровня, что централизованное управление в них становится невозможным или неэффективным из-за наличия огромных потоков информации, когда слишком много времени тратится на ее передачу в центр и принятие им решений; 3) с ростом сложности больших систем падаёт их надёжность, и наиболее эффективный способ борьбы с этим — децентрализация, обеспечивающая избыточность; 4) в ряде случаев трудно сформулировать в формальном виде с необходимой точностью цель работы и критерий управления сложным объектом; 5) как правило, децентрализации нет альтернативы при формировании сложных межгосударственных и межнациональных систем.

Мир, создаваемый человеком в технических системах, во многом похож на тот, что окружает человека в природе. В этом искусственном мире происходят процессы, подобные эволюции живых организмов, возникают колонии и сообщества технических систем. На первых же страницах книги, показывая, как возникает децентрализованное управление, авторы используют бионический подход и, по аналогии с биоценозом, вводят термин «техноценоз». Техноценозы представляют собой популяции различных технических систем, устройств, приборов, связанных между собой тесными отношениями. Подобно биоценозам, они заставляют входящих в них искусственных агентов жить по законам, диктуемым всем сообществом. Характерными примерами техноценозов служат аэропорты, крупные промышленные предприятия, отрасли, инфраструктура городского хозяйства и т.д.

Отличительным признаком техноценоза является тот факт, что полную документацию на него собрать невозможно, что означает необходимость эволюционного пути развития. После постановки проблемы управления в рамках техноценоза или группы техноценозов описана сложная техническая система, которую в полном объёме никто не проектировал (телефонная сеть земного шара), создаваемая эволюционным путём. В результате, делается вывод, что подобные системы не могут управляться одним органом управления, а предполагают согласование действий различных взаимодействующих подсистем. Далее рассмотрены варианты взаимодействия, кооперации и самоорганизации подсистем сложной системы при децентрализованном управлении. Обсуждаются ставшие классическими автоматные и игровые модели поведения в коллективе. В частности, описаны модели рефлексивного поведения, рассмотрены модели саморегулирования числа агентов в коллективе, проведён анализ поведения коллектива во времени.

В заключительной части книги проанализированы различные стратегии эволюционной адаптации: арогенез (расширение адаптационных возможностей системы); аллогенез (смена одной экологической ниши на другую); телогенез (глубинная адаптация к определённому состоянию среды за счёт специализации).

Книга «От амёбы до робота: модели поведения» [30], стала одной из первых (если не первой) монографий в мировой литературе, посвящённой основам единой теории поведения природных и искусственных систем. Само ее название прекрасно отражает смелый замысел авторов: провести междисциплинарное исследование общих принципов и механизмов поведения, последовательно рассмотрев основные ступени его эволюции. Эта эволюция простирается от элементарных форм раздражимости, свойственных простейшим организмам (таким как амёба), до весьма сложных форм нормативного и ситуационного поведения.

В монографии изложен широкий спектр моделей поведения: от моделей рефлекторного поведения, изучавшихся физиологами и биологами, до моделей поступков, личности и социального поведения, предложенных в различных областях психологии. Описаны варианты формальной и программной реализации многих моделей, в частности гиромат, фреймы поступков, планирование поведения на сетях. Необходимыми условиями реализации искусственным агентом некоторого поведения выступают специальные устройства, непосредственно воспринимающие сигналы внешней среды (рецепторы), и исполнительные органы, воздействующие на среду (эффекторы), а также процессор (блок переработки информации) и память. Под памятью понимается способность агента хранить информацию о своем состоянии и состоянии среды. Таким образом, исходное представление о простейшем агенте сводится к хорошо известной модели «организм – среда», описанной в монографии «От амебы до робота: модели поведения» (рис.3).

В предисловии к книге «От моделей поведения к искусственному интеллекту» [48], в которой продолжено исследование проблем, затронутых М. Г. Гаазе-Рапопортом и Д. А. Поспеловым, отмечено, что монография [30] сыграла существенную роль в развитии наук о поведении, став предвестником интересного и перспективного научного направления «Моделирование адаптивного поведения».

Рис. 12. Модель «организм – среда» как структура простейшего агента

Были введены фреймы поступков и дана их классификация, описана оригнальная модель личности. Заключительный раздел книги был посвящён проблемам взаимодействия людей и машин, включая проблемы объяснения и планирования поведения.

C. Предчувствие новых когнитивных наук

В первой главе книги «Моделирование рассуждений» [32], названной «У истоков формальных рассуждений» профессор Д. А. Поспелов развивает очень важную идею зависимости рассуждений от онтологических допущений о мире. Для её обоснования предварительно были рассмотрены суждения и оценки на биполярных шкалах. Главное положение традиционной теории оппозиционных (биполярных) шкал со времён Ч. Осгуда заключается в том, что мир для человека устроен в виде системы шкал, где края каждой шкалы связаны между собой чем-то вроде операции отрицания. Например, берутся шкалы «мы -– они», «друг -– враг», «добро -– зло» и пр. Всякий объект или агент, все их свойства, деяния или эмоции отображаются на подобные шкалы, где середина нейтральна, а далее могут быть градации. Но всегда есть два конца и середина, которая очень важна, поскольку делит всю шкалу на две половины – положительную и отрицательную. Именно середина как бы переключает нас с одного типа оценок на другой.

Различные виды оппозиционных шкал можно определить с помощью двух базовых критериев: а) сила оппозиции между полюсами; б) статус нейтрального значения. В случае сильной оппозиции между полюсами отрицательная часть шкалы является зеркальным отражением положительной части, и эти две области считаются взаимно исключающими. Ослабление оппозиции соответствует возможности одновременного сосуществования положительных и отрицательных оценок, и в дальнейшем возникают новые связи между ними.

Согласно гипотезе Д. А. Поспелова, семантика операций над экспертными оценками на шкалах сильно зависит от контекста. Для подтверждения этого им было исследовано, как изменяется толкование операции отрицания на оппозиционных шкалах «мы–они» и «друг-враг» (во втором случае были предварительно проведены психологические эксперименты). В результате был показан естественный переход от отрицания как противопоставлению (антагонизму) к отрицанию как различию. Были выделены следующие интерпретации отрицания: а) жёсткая конфронтация; б) гибкая оппозиция; в) учет различий в условиях сосуществования (см. также [49]).

По сути, в этом разделе была очерчены контуры новой области исследований на стыке психологии и логики – *когнитивной семантики*, возникшей из анализа экспериментов по восприятию людьми друг друга и попыток осмысления их с помощью нетрадиционных динамических шкал и многозначных логик.

Во второй главе монографии [32] Д. А. Поспелов напоминает читателю о роли забытых в XX-м веке наук о рассуждениях. На пути развития человеческих знаний о внешнем мире возникали и исчезали целые науки. Одни из них, например, *диалектика* (в смысле Сократа) или *риторика*, известны современным учёным и педагогам хотя бы понаслышке, о других, например, *герменевтике* или *экзегетике* знает весьма узкий круг специалистов, занимающихся историей науки и культуры. В последние годы герменевтические схемы становятся предметом тщательного изучения специалистами, работающими в области понимания текстов.

В ИИ нас интересуют возможные схемы человеческих рассуждений. Эти схемы распадаются на три следующих типа: герменевтические, экзегетические, гомилетические. В *герменевтических схемах* заключения выводятся на основании, главным образом, содержания текста. Два других типа рассуждений для построения заключения используют вне текстовую информацию.

Для получения *экзегетических выводов* привлекается информация, связанная с контекстом, в котором был порожден данный текст. Это может быть информация об исторических условиях создания текста, о его авторе, о принятых во времена написания текста условностях при использовании конкретных выражений, и т.п.

Наконец, *гомилетические рассуждения* основаны на получении заключений, опирающихся на морально-этические и нравственные посылки, связанные с текстом и его создателями. Рассуждения такого типа порождают собственное поведение человека на основе истолкования текста или оценку на этой основе поведения других лиц.

Несколько основополагающих работ были написаны профессором Д. А. Поспеловым в 1990-е годы. В первую очередь, к ним относятся статьи «Серые» и/или «черно-белые» [50] и «Знания и шкалы в модели мира» [51]. В них была продолжена начатая в [32] разработка основ когнитивной семантики для ИИ. Так в [50] были введены два типа биполярных шкал -— «серые» и «чёрно-белые шкалы». Поясним их различие на следующем примере. Пусть биполярная шкала строится на базе пары антонимов «положительный (+) -— отрицательный (-)». В случае серой шкалы наблюдаем плавный переход от положительных оценок a^+ к отрицательным оценкам a^- и наоборот. В нейтральной точке имеем противоречивую оценку «и большой, и малый», т.е. в ней оба антонима представлены в равной степени.

Будем обозначать через $\uparrow$ и $\downarrow$ соответственно увеличение и уменьшение оценки a на шкале. Для серой шкалы имеем:

1) $a^+\uparrow \Rightarrow a^-\downarrow$, $a^-\downarrow \Rightarrow a^+$ (принцип взаимной компенсации положительных и отрицательных оценок: чем больше a^+, тем меньше a^- и, наоборот, чем больше a^-, тем меньше a^+;

2) нейтральная точка *c* есть точка наибольшего противоречия, где a^+ и a^- равны между собой;

3) $a^+ = a^-$.

Напротив, для чёрно-белой шкалы нейтральная точка понимается как точка разрыва (полная неопределённость: «ни малый, ни большой»); её окрестность может трактоваться как область хаоса, где возможно формирование нового смысла путём перескока с одной шкалы на другую. Шкала разделяется на две независимые половинки. Здесь не происходит взаимной компенсации положительных a^+ и отрицательных a^- оценок. При этом, чем ближе обе оценки a+ и a- к точке разрыва «чёрнобелой» шкалы, тем неопределённее становятся их значения.

В [50] были также введены неклассические круговые (или кольцевые) шкалы и предложены двухосновные оценки объектов на таких шкалах, отражающие динамику экспертных суждений, оценок и рассуждений.

В [51] оппозиционные шкалы были рассмотрены как образующие «модели мира» (в смысле А. Н. Леонтьева). Было развито представление об обобщённой шкале. В отличие от обычных шкал, где каждой точке соответствует один объект, на обобщённых шкалах одной точке может с разными степенями соответствовать множество объектов. Базовые категории в модели мира проецируются на систему обобщённых шкал, задающих отношение частичного или нестрогого порядка, толерантности или доминирования.

Таким образом, когнитивная семантика Д. А. Поспелова связана с конструированием смысла и представлением знаний на нетрадиционных (нечётких, обобщённых, кольцевых, динамических) шкалах с целью формирования целостного образа мира.

Монография [31] посвящена представлению знаний о времени и пространстве на основе псевдофизических логик (ПФЛ). В ней приведены формы представления знаний на ограниченном естественном языке, описаны проблемы шкалирования, показана ключевая роль отношений в представлении знаний и изложены основы формальной теории отношений. Центральное место в книге занимают логики времени в интеллектуальных системах и модели времени в базах знаний, а также логики пространства, включая логики расстояний на прямой и на плоскости, логику направлений на плоскости, пространственную логику для нормативных фактов.

Псевдофизическая логика -— это логика, отражающая восприятие субъектом или искусственной системой закономерностей внешней физической среды. Особенностью ПФЛ является наличие нечётких шкал, на которые проецируются объекты. Примерами ПФЛ являются временные логики, пространственные логики, логики действий и т.п. [52, с. 45-46].

Этот класс логических систем имеет следующие особенности:

1) В качестве пропозициональных переменных используются лингвистические переменные (ЛП) с термами, представленными нечёткими множествами. Например, в частотной логике И.В.Ежковой и Д.А.Поспелова [53] в качестве ЛП берётся «Частота события» с терм-множеством никогда, чрезвычайно редко, редко, ни часто, ни редко, часто, очень часто, почти всегда, всегда, и базовыми значениями числового универсального множества 0,1/5, 2/5, 3/5, 4/5, 1.
2) На множестве значений для всех переменных имеются порядковые шкалы с отношением стро-

гого порядка. Точнее для ЛП существуют порядковые шкалы, а для числовых переменных — метрические шкалы.

3) Выводы в ПФЛ учитывают порядковые и метрические шкалы, а также расположение событий на них.

По аналогии с современной психофизической схемой и в отличие от классической аристотелевской логики ПФЛ описывают не идеальный платоновский мир, а восприятие реального физического мира конкретным агентом. Суть псевдофизических логик составляет работа с событиями, т.е. с формулами, которые соотнесены с отметками на шкалах. Взаимное положение событий на множестве шкал, возможные перемещения по шкалам и связь этих перемещений с изменениями на других шкалах позволяют описать те процессы вывода, которые характерны для псевдофизических систем.

Псевдофизическая логическая система представляет собой семейство взаимосвязанных логических подсистем, которые следует отнести к двум основным уровням. На первом уровне находятся *временная*, *пространственная*, *каузальная логика*, а также *логика действий*. На втором, более высоком уровне находятся *логика оценок*, *логика норм*, *логика мнений*, и пр.

Псевдофизические логики непосредственно связаны с познанием и являются прямыми предшественниками современной «когнитивной логики». Первые работы Д.А.Поспелова по временным ПФЛ были написаны в середине 1970-х годов [54,55], тогда как «когнитивная логика» возникла уже в XXI-м веке [56].

Псевдофизические логики можно положить в основу логического моделирования когнитивных агентов. В частности, логики первого уровня непосредственно связаны с взаимодействием агентов (например, роботов) с внешней средой. В 1990-е годы профессор Д. А. Поспелов опубликовал две фундаментальные работы, посвященные основам теории агентов и многоагентных систем (где понятие «агент» уже рассматривалось им в явном виде): «От коллектива автоматов к мультиагентным системам» [57], а затем «Многоагентные системы -– настоящее и будущее» [58]. В них он описал исторический переход от моделей коллективного поведения к теории агентов, ввел классификацию агентов с помощью тройки критериев «тип среды – уровень свободы воли -– степень развития социальных отношений», выделил и проанализировал различные типы сред функционирования агентов, указал ключевые интенциональные характеристики агентов. В ходе 1-го Международного семинара «Распределенный искусственный интеллект и многоагентные системы» в Санкт-Петербурге в 1997 году состоялась чрезвычайно интересная дискуссия Д. А. Поспелова с В. Л. Стефанюком (их противоположные исходные позиции описаны в [57] и [59]) о взаимосвязях между моделями коллективного поведения и многоагентными системами.

Всё же главное место в научном поиске 1990-х годов у Д. А. Поспелова занимала *прикладная семиотика* [60-62]. Здесь катализатором послужил уже упомянутый ранее совместный российско-американский семиотический проект, в частности, семинар по архитектурам для семиотического моделирования и ситуационному анализу в больших сложных системах, прошедший в Монтерее в 1995 году. На нём был сделан доклад Д. А. Поспелова и др. «Семиотическое моделирование и ситуационное управление» [60].

В прикладной семиотике центральное место занимает понятие *семиотического моделирования*, характеризующее динамику интеллектуальной системы при изменении её знаний об окружающем мире и способах поведения в нем. Если классическая формальная система описывает замкнутый, полностью определённый мир, то введённое Д. А. Поспеловым представление семиотической системы позволяет моделировать процессы, протекающие в открытых или динамических системах.

Стратегия перехода от формальных к семиотическим системам, по Д. А. Поспелову, заключается в изменении различных параметров формальной системы: алфавита, синтаксических правил, аксиом, правил вывода, ценностных ориентаций, стратегий поиска решений, ограничений и т.п. Эти изменения производятся с помощью некоторых правил или моделей. Таким образом, семиотические системы являются открытыми, активными, ориентированными на работу с динамическими базами знаний, реализацию как дедуктивных, так и индуктивных и абдуктивных рассуждений, сосуществование разных логик.

Существенную роль в прикладной семиотике играет введенная Д. А. Поспеловым модификация классического треугольника Фреге -– наглядного представления знака как триединства имени, концепта и денотата. Здесь речь идет о введении метауровня в знаковых представлениях (метазнаков), с которым связывается активность знаковых систем и возникновение в них свойства рефлексии. Впоследствии ученик Д. А. Поспелова, ныне президент РАИИ, д.ф.-м.н., профессор Г. С. Осипов предложил назвать наглядное представление метазнака (в русле «динамической интерпретации знака») квадратом Поспелова [63].

Точно так же, как и его когнитивная концепция нетрадиционных шкал, прикладная семиотика у Д. А. Поспелова, по существу, является *когнитивной семиотикой*, поскольку она опирается на определение соответствий между знаковыми структурами и структурами знаний в виде фреймов [61,62] (рис. 12).

Так, например, паре процедур 1 знака соответствуют поиск информации по адресу и ассоциативный поиск информации по содержанию, а паре процедур 2 отвечают приобретение знаний и построение конкретных

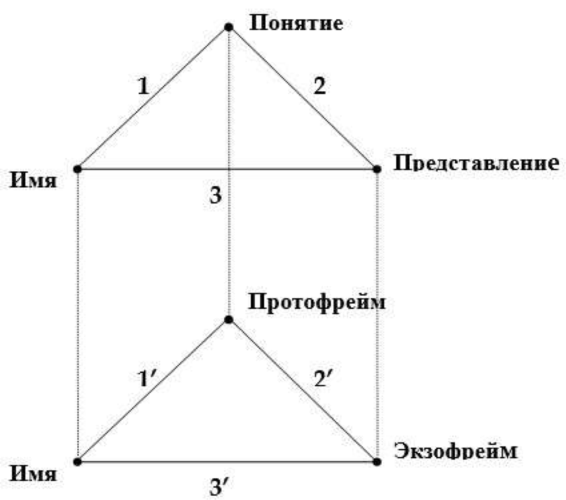


Рис. 13. К определению знака-фрейма

представлений на основе понятия, т.е. порождение экзофреймов на базе имеющегося протофрейма.

Проведённый анализ показывает, что столь разные, на первый взгляд, достижения и открытия Д. А. Поспелова, как прикладная семиотика, псевдофизические логики и нетрадиционные оппозиционные шкалы связаны воедино методологией когнитивных наук.

IV. БУДУЩЕЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

На формирование синтетического и синергетического научного мировоззрения Д. А. Поспелова несомненное влияние оказали идеи всеединства и всечеловечности, лежащие в основе философии и искусства русского космизма (в первую очередь, концепция ноосферы В. И. Вернадского, образуемой благодаря росту научного знания, а также концепция гармонии и синтезирующей роли культуры Н. К. Рериха). Русский космизм подхватил эстафету человеческих исканий в плане осмысления идеи единства Человека и Космоса и поднял эти искания на удивительную высоту прозрений человеческого духа. Именно русский космизм выступил как предтеча ныне формирующихся представлений о мировом общественном интеллекте. С этим миропониманием тесно связаны грани уникального дарования Д. А. Поспелова как учёного, писателя и искусствоведа.

Одной из важнейших тем русского космизма является предвидение и управление будущим. Соответственно, футурологические очерки и научные прогнозы занимают немаловажное место в творческом наследии Дмитрия Александровича.

A. Некоторые прогнозы Д. А. Поспелова

В наше время мало кто знает, что ещё в 1970-е годы профессор Д. А. Поспелов высказал ряд удивительно точных и прозорливых прогнозов развития компьютерной техники, робототехники, теории управления. Так в 1975 году в небольшом интервью, опубликованном в газете МФТИ «За науку» (12 сентября 1975 г., № 24 (542)), он ***предсказал появление глобальной информационной сети Интернет***: «Сети вычислительных машин сейчас эволюционируют столь же быстро, как когда-то эволюционировала телефонная сеть. И скоро они сольются в межнациональную сеть обработки данных, что позволит двум людям, выросшим в разных точках земного шара, общаться и понимать друг друга».

Сегодня только начинает сбываться другой его давний прогноз о том, что теория управления станет центральным звеном прогресса науки и общества в целом. Ещё в 1973 году им отмечалось, что «в ближайшие десятилетия самые странные и неожиданные открытия будут сделаны именно в теории управления, на её стыке с вычислительной техникой, экономикой, социологией и биологией. Пальма первенства от физиков, долгие годы будораживших человечество великими открытиями, перейдет к теории управления. Великие физики первой половины XX-го века уступят свое место великим управленцам»... Будет развиваться новый взгляд на суть управления благодаря учёту взаимосвязей между сложностью и эффективностью управления, выбору нужной степени централизации и иерархичности управления, организации сложных систем за счет локального управления их подсистемами» (статья Д. А. Поспелова в газете «За науку», 6 апреля 1973 г., № 11 (452)).

Современное развитие концепций управления в эпоху цифровой экономики и стратегии Индустрия 4.0, появление сетевых, виртуальных, интеллектуальных организаций, успешная практика стратегического менеджмента как управления корпоративными знаниями демонстрируют актуальность этого прогноза.

Несколько позже, в монографии [27, с. 204] Дмитрий Александрович отметил, что «в области промышленных роботов появятся ***роботы нового поколения***, способные выполнять набор близких по своей специфике работ и ***адаптироваться*** к новым условиям производства. Эти роботы будут способны к ***групповой деятельности*** в рамках предусмотренных для этого сценариев». Сегодня развитие коллективной и коллаборативной робототехники делает данный прогноз реальностью.

B. Искусственный интеллект в новом тысячелетии

«Будущее искусственного интеллекта». Так назывался составленный Д. А. Поспеловым и К. Е. Левитиным сборник [64], изданный в год заката СССР, в котором ведущие отечественные и зарубежные учёные обсуждали различные аспекты ИИ, а также перспективы этой междисциплинарной науки. Первая часть книги написана в форме актуальных вопросов ИИ и ответов на них академика Г. С. Поспелова, а во второй части собраны статьи и высказывания об ИИ учёных и специалистов из разных областей знаний: пионеров информатики и ИИ в СССР академика А. П. Ершова, Л. Т. Кузина, М. М. Ботвинника, британского патриарха ИИ Д. Мики, ведущего логика в области ИИ В. К. Финна, математика и родоначальника когнитивной графики А. А. Зенкина, видных психологов В. П. Зинченко, О. К. Тихомирова, и других специалистов.

В [64] отмечено, что на XI-й Международной конференции по ИИ были предсказаны три новых направления в этой области в 1990 годах: 1) компьютеры параллельной архитектуры, в том числе, реализующие асинхронно протекающие волновые процессы; нейронные сети, нейроструктуры и нейрокомпьютеры; 2) гибридные интеллектуальные системы; 3) компьютерная когнитивная графика.

Среди ключевых проблем рассуждений были названы: формализация повседневных рассуждений «здравого смысла», переход от вывода к обоснованию, а от обоснования — к оправданию, построение ценностных рассуждений, связанных с формированием поведения.

Через пять лет в 1996 году была опубликована статья Д. А. Поспелова «Десять «горячих точек» в исследованиях по ИИ» [65]. В ней Дмитрий Александрович вынес на обсуждение «горячие точки» ИИ -– прорывные направления, на которых будут сосредоточены основные усилия специалистов в начале нового тысячелетия: 1) переход от вывода к аргументации; 2) проблема оправдания; 3) порождение объяснений; 4) поиск релевантных знаний; 5) понимание текстов; 6) синтез текстов: 7) когнитивная графика; 8) многоагентные системы; 9) сетевые модели; 10) метазнания. При этом в пункте 9 он предсказал разрыв между адептами нейронных сетей и основным ядром специалистов в области ИИ. Профессор Д. А. Поспелов считал, что нарастание этого разрыва, по–видимому, приведёт к становлению двух разных наук, связанных с построением интеллектуальных систем. Одна из них будет по-прежнему опираться на уровень информационных (или ментальных) представлений, а другая — на уровень структурной организации, порождающей нужные решения.

Почти четверть века спустя этот прогноз полностью подтвердился. Впечатляющие успехи свёрточных нейронных сетей глубокого обучения и капсульных нейронных сетей вдохновляют неофитов ИИ на фактическое сведение его к нейросетевой парадигме и алгоритмам машинного обучения. Но даже простой перечень интеллектуальных способностей человека [47], не говоря уж о целостных структурах сознания или психики, показывает ограниченность такого подхода, поскольку в его рамках проблемы аргументации, оправдания, объяснения и понимания не решаются, а становятся только острее.

Здесь, очевидно, требуются новые гибридные модели, развиваемые на основе комплексных подходов типа мягких вычислений (Soft Computing), вычислительного интеллекта (Computational Intelligence) и т.п.

Рассмотрим некоторые собственные работы Дмитрия Александровича по проблемам обоснования и оправдания, понимания и когнитивной графики.

В работе [66], посвящённой первой из выделенных «горячих точек», выявлены различия между процедурами обоснования и оправдания, а также предложена интересная классификация видов рассуждений в зависимости от уровня самосознания (состояния своего «Я») у рассуждателя.

Классическое *обоснование* есть погружение нового факта в систему уже имеющихся фактов, такое, что добавление этого факта не порождает противоречий. В то же время *оправдание* означает погружение нового факта в систему ценностей, при условии, что положительная оценка данного факта ей не противоречит, Иными словами, реальное обоснование связано с трактовкой истинности как когерентности — степени согласованности факта с располагаемым знанием, а степень оправдания сильно зависит от системы ценностей.

Предложенная классификация рассуждений опирается на психологический структурный анализ и выделение трёх ипостасей (состояний) «Я» агента (по Э. Берну): Родитель, Взрослый и Ребёнок. Эти типы рассуждений являются совершенно различными. Система знаний Родителя замкнута и неизменна, его рассуждения основаны на жёсткой системе аксиом, содержащих незыблемые истины и нормы, причём все шаги вывода считаются априори правильными и не подверженными пересмотру. Таким образом, Родитель осуществляет строгий рациональный вывод при полном отсутствии неопределенности.

Рассуждения Взрослого являются более гибкими. Это состояние самосознания подразумевает критическое осмысление поступающей информации, а также её проверку на соответствие реальному миру. Знания Взрослого объективированы в том смысле, что, кроме данного агента, они принимаются и другими членами социума. Работа с такими знаниями соответствует процедурам, используемым в открытых базах знаний для интеллектуальных систем. Здесь вывод является скорее правдоподобным, чем достоверным, а порой и немонотонным. Нередко он заменяется аргументацией или простым погружением нового факта в базу знаний, если это не приводит к противоречию.

Наконец, схема рассуждений, относящаяся к состоянию Ребёнок, тесно связана с эмоциональной сферой. Здесь могут быть существенные отклонения от нормативного вывода и совсем неожиданные заключения. При этом мы имеем дело с оправданиями, а не с обоснованиями. Если на уровне обоснования возникает задача синтеза нормативного поведения, то на уровне оправдания формируется ситуативное поведение.

Эта классификация рассуждений представляет интерес и в плане определения уровня понимания собеседника.

Изучение проблемы понимания текстов, затронутой уже в книге [27],привело Д. А. Поспелова к идее выделения уровней понимания у компьютера [67, 68]. Им было введено 7 уровней понимания. При этом понимание трактуется прагматически, как способность отвечать на вопросы.

На нулевом уровне понимания компьютерная система способна отвечать на сообщения пользователя безо всякого анализа их сути. На этом уровне понимание как таковое у системы отсутствует. В общении людей между собой нулевому уровню понимания соответствует так называемый *фактический диалог*, когда разговор ведется без анализа сути высказываний собеседника за счет чисто внешних форм поддержки диалога.

На первом уровне понимания система должна быть способной отвечать на все вопросы, ответы A на которые есть во введенном в нее тексте T, т.е. имеем простую схему: $T \Rightarrow A$.

На втором уровне понимания добавляются *база знаний* и *машина вывода I (блок пополнения текста)*. В его функции входит автоматическое пополнение текста за счет хранящихся в памяти компьютера процедур пополнения. Имеем I: $T \rightarrow T^*$, и исходная схема приобретает вид $T^* \Rightarrow A$.

На третьем уровне понимания к средствам второго уровня добавляются правила пополнения текста знаниями системы о среде. Отличие третьего уровня понимания от второго заключается в процедурах, реализуемых блоком вывода ответа I. Формируя ответы, этот блок использует теперь не только информацию, хранящуюся в базе знаний, куда введено расширенное представление T^* исходного текста $T_{,,}$ но и дополнительные знания о типовых сценариях Sc развития ситуаций и процессов, характерных для той предметной области, с которой работает компьютерная система.

Четвёртый уровень понимания достижим при наличии двух каналов получения информации. По одному в систему передается текст T^*, а по другому – дополнительная информация D, отсутствующая в тексте. Так два канала коммуникации имеют интеллектуальные роботы, обладающие зрением.

Пятый уровень понимания предполагает наличие информации о конкретном источнике текста, а также сведений о системе общения (знания об организации общения, целях участников общения, нормах общения, и пр.). Его идейной основой является теория речевых актов. Здесь широко используются средства правдоподобного вывода.

На шестом уровне понимания происходит порождение метафорических знаний с помощью специальных процедур, опирающихся на ассоциации и вывод по аналогии.

По аналогии с уровнями понимания компьютера, в [69] были описаны уровни понимания интеллектуального робота.

Следует отметить, что в XXI-м веке резко возрос интерес к таким объектам понимания как результаты измерений, намерения, поступки, действия, поведение, ситуации, и пр. (см. [69]).

Ещё две проблемы будущего ИИ, над которыми Д. А. Поспелов активно работал в 1990-е годы, это –– *когнитивная графика* [70, 71] и перспективы изучения *правополушарного мышления* [72,73].

Накопление знаний и решение задач, возникающих перед человеком возможно двумя путями: алгебраическим и геометрическим. В XX-м веке первый путь стал преобладающим и даже почти единственным. Широкое распространение компьютеров лишь усилило доминанту алгебраической модели. Развитие высшего образования создало убеждение, что алгебраизация — суть единственный подлинно научный подход к познанию мира и принятию решений в нём. Зато у геометрического или графического подхода есть одно главное, неоспоримое преимущество. Апеллируя к образу, рисунку, геометрическому узору, этот подход генерирует у человека пучки ассоциаций, с помощью которых формируются интеллектуальные подсказки [71].

Графика позволяет представить знания в наиболее удобной для человека и компактной форме. Её важнейшей характеристикой является способность непосредственно влиять, даже направленно воздействовать на интуитивно образное мышление человека. В книге [70] выделены иллюстративная и когнитивная функции графических изображений. Здесь иллюстративная функция связана с визуализацией уже известного знания и обеспечением узнаваемости изображаемого объекта, а когнитивная функция графики, направленная на визуализацию внутреннего содержания, смысла научных абстракций, способствует порождению нового знания. Между иллюстративной и когнитивной графикой нет чёткой границы: концептуальное сжатие и наглядное представление известного знания может подсказать новую идею или гипотезу, для подтверждения которой могут также применяться диаграммы или картинки.

Когнитивная графика есть, прежде всего, средство визуализации идей, которые ещё не получили какого-либо точного выражения. Известными примерами когнитивных образов являются лица Чернова, используемые при анализе и управлении деятельностью компаний. Широкое распространение мультимедиа и гипермедиа-технологий также стимулирует развитие когнитивной графики. Она наиболее востребована на этапе исследовательского проектирования (формализации проблемы и выдвижения гипотез).

В итоге, Д. А. Поспелов делает общий вывод, что сплав алгебраического и геометрического подходов позволит создать полноценные интеллектуальные системы с куда большими возможностями, чем у современных систем ИИ [71]. По-видимому, это и будет решающий шаг на пути к «ИИ человеческого уровня», о котором так мечтал Дж. Маккарти.

В статье [72] была затронута проблема моделирования в компьютерных системах «правополушарных процедур». Человеческое сознание и человеческий способ познания мира отличает от компьютера наличие двух параллельных систем познания. Компьютер обладает одной системой познания, а именно символьной. Человек же обладает двумя совместно работающими системами познания: одна есть рассудок или символьно-логическое мышление, а другая -– система восприятия и образного мышления. Неоднократно формулировалась гипотеза о том, что в основе человеческого механизма познания лежит интегрированная система, в которой образная и символьно-логическая компоненты слиты воедино [72,73].

Важным средством изучения и формирования взаимосвязей и организации совместной работы левополушарных и правополушарных процессов, символьно-логического и образного мышления служат системы виртуальной и дополненной реальности. Термин *«виртуальная реальность»* подразумевает погружение и навигацию человека в киберпространстве (искусственном, чаще всего трёхмерном мире), создаваемом с помощью компьютера (смартфона) и воспринимаемом посредством специальных устройств (шлема, очков, перчатки данных, и пр.).

В то же время *дополненная реальность* (ДР) есть результат добавления к объектам реального мира мнимых объектов, применяемых для получения новой информации. Она предполагает совмещение компонентов реального и виртуального миров при работе в реальном времени и трехмерном пространстве. Например, в поле восприятия человека на производстве оказываются дополнительные сенсорные данные, в первую очередь, зрительные и тактильные, которые позволяют улучшить представление и понимание задачи. Таким образом физическим объектам добавляют виртуальные свойства, и формируется их индивидуализированное представление под конкретного пользователя.

3-4 марта 1998 г. в Переславле-Залесском состоялся первый (и до сих пор единственный по этой тематике) семинар РАИИ «Отражение образного мышления и интуиции специалиста в системах искусственного интеллекта». На нём с первым пленарным докладом «Метафора, образ и символ в познании мира» [73] выступил профессор Д. А. Поспелов. Опираясь на таблицу Шиклоши, содержащую сравнение возможностей человека и компьютера в генетическом плане, он обратил внимание на противоположность путей развития людей и классических систем ИИ: у человека вначале формируется образная система, а потом понятийно-логическая, а у компьютера всё наоборот. Так в логике мы, в первую очередь, формируем общую систему понятий за счёт сходства и различия, а затем уже используем эту систему понятий для вывода, например, о принадлежности конкретного индивидуального понятия к общему понятию. Между тем, по утверждениям психологов, у человека сначала возникает единый образ -– гештальт, а в дальнейшем всё остальное. Пока ещё нет искусственных систем, функционирующих по такому принципу.

Докладчик остановился на принципах гештальта и работы образной системы познания, предложенных в гештальт-психологии. Прежде всего, это принципы равновесия и простоты. Зрительные, кинестетические, слуховые и другие образы очень тесно связаны с устойчивостью и ощущением равновесия. Нарушения равновесия образов обычно обусловлены двумя вещами: 1) отклонением от симметрии (положением осей симметрии, плоскости симметрии); 2) понятиями «верх-низ», «слева-направо» и т.л. Например, в картинах живописи, фигуры, расположенные в правой половине картины, выглядят «тяжелее» точно таких же фигур в левой половине. В сознании человека равновесный образ является аналогом истины в логике. Неравновесный образ требует либо до осмысления, либо поиска дополнительной информации, либо включения в такую систему, где его неравновесность будет как то скомпенсирована. Подобные смутные образы быстрее вытесняются из зоны активного внимания.

Наличие эталонных образов в плане простоты легко пояснить на геометрических фигурах. У человека есть врождённые эталоны простых форм – это квадрат и круг. Многоугольник сложнее треугольника, незамкнутая фигура сложнее замкнутой. Диагональное направление слева направо и вверх — стандартный эталон возрастания, а справа налево и вниз–убывания. Цвета также упорядочены по предпочтениям: у художников простые или главные цвета -– это синий, красный и жёлтый.

Соответственно, базовыми преобразованиями образов («операциями над образами») являются уравновешивание и упрощение (или сведение к эталонам). Ритмы и орнаменты также играют в образах особую роль. Например, образ может быть неуравновешенным, но ритмичным, «возвращающимся через равные промежутки времени», и нас это устраивает как компенсация его неустойчивости.

Композиции, свёртки простых образов также играют ключевую роль. Так сочетания из простых форм, красок и теней образуют узоры. Повторяющиеся и чередующиеся узоры создают орнамент. Орнаментальность, ощущаемая предсказуемость «застывшего движения» способствует устойчивости зрительного образа. Поэтому операции создания ритмической структуры и орнаментализации также следует отнести к числу базовых операций образного представления мира, образного мышления.

Ещё одна динамическая группа образов, которые связываются с восприятием и процессами воображения –– это «топляки». Название «топляки» предложено по аналогии с брёвнами, которые ушли при сплаве под воду, а затем всплывают. Речь идёт об образах, которые в какое-то время вытесняются в подсознание, а потом «всплывают». Весьма интересны происхождение и условия применения операций «потопления» и «всплывания» образов. Почему они топятся? Топятся они точно так же, как и знания, когда те нас не устраивают. Даже знания логического типа часто убираются из активной зоны нашего внимания потому, что они нам не нравятся в силу своей необъяснимости, неприемлемости, неполноты, противоречивости и других НЕ-факторов.

Итак, если образ кажется непонятным, раздражает или никак не достраивается до уравновешенного состояния, то он вытесняется. И будет находиться в подсознании, пока у человека не сработает интуиция, не произойдёт озарение или инсайт. Возникает «ощущение истинности».

Эквивалентом истинности в образах является красота. Красота в образе, как и красота теории, есть удовлетворение всё тем же исходным принципам равновесия, простоты, объяснимости, и т.д.

Целостность восприятия и образного мышления, наличие эталонных образов, ощущения равновесия, симметрии, простоты и завершённости формы являются главными характеристиками «правополушарного мира». Сначала всё воспринимается целиком, потом раскладывается на части, но не собирается целое из частей.

В [73] Д. А. Поспелов также проследил связи между образами и метафорами в интеллектуальных системах.

Создание ИИ «человеческого уровня» означает наделение искусственной системы правополушарными способностями человека и реализацию в ней совместной работы символьно-логического и образного мышления. Насколько, это окажется возможным, покажет время. Более реальной перспективой видится создание целостного гибридного, синергетического интеллекта в условиях партнёрства человека и робота, а, в более общем плане, смешанных коллективов людей и машин. Тогда на смену «третьей волны» развития цивилизации (по О. Тоффлеру) придёт четвертая волна — «волна единого всемирного партнёрского интеллекта людей и машин». Но уже сегодня надо думать о том, чтобы она не захлестнула наш естественный мир!

V. НАСЛЕДИЕ Д. А. ПОСПЕЛОВА

Дмитрий Александрович Поспелов внёс огромный вклад в становление и развитие искусственного интеллекта в СССР, а затем в России и СНГ. Можно выделить четыре основных измерения его многогранной деятельности.

Во-первых, это научное измерение, его собственные выдающиеся научные результаты на стыке разных дисциплин, Создание новых научных направлений, прямо относящихся или близких к ИИ, включая не только хорошо разработанные подходы, методы и модели ситуационного управления, прикладной семиотики, псевдофизических логик, но и базовые идеи новых когнитивных наук, как «ростки будущего», в частности, модели когнитивной семантики на нестандартных биполярных и кольцевых шкалах, модели рассуждений «здравого смысла», триаду «объяснение-обоснование-оправдание» и её связи с типами рассуждений и уровнями понимания. Формирование основ психонической парадигмы целостного ИИ и «психологии искусственных агентов», которая становится очень востребованной в наши дни.

Во-вторых, блестящий педагогический и, особенно, лекторский талант Д. А. Поспелова привели к созданию одной из самых сильных научных школ по искусственному интеллекту –– междисциплинарной, меж организационной, межрегиональной, и, без преувеличения, международной. Важнейшая миссия Дмитрия Александровича как учителя заключалась в воспитании достойных учеников и производстве профессионалов высокого уровня, способных не только и не столько брать на вооружение готовые программные продукты и реализовывать чужие подходы и алгоритмы, сколько генерировать свои идеи, методы, модели и воплощать их в жизнь.

Официально, среди его непосредственных учеников 5 докторов наук, в том числе, крупные учёные в области ИИ В. Н. Вагин, Г. С. Осипов, В. Ф. Хорошевский, и более 50 кандидатов наук, включая таких известных специалистов, как А. Н. Аверкин, А. Ф. Блишун, И. В. Ежкова, Л. В. Литвинцева и др., но это только «верхушка айсберга». Немало видных учёных не только в России, но и в странах СНГ и дальнего зарубежья с гордостью называют себя выходцами из школы Д. А. Поспелова.

В-третьих, он без видимого напряжения выполнял очень большую научно-организационную работу. Вместе с академиком Г. С. Поспеловым Дмитрий Александрович в 1974 году стоял у истоков официального признания искусственного интеллекта как научного направления в СССР (см. выше).

Яркими страницами международного сотрудничества в области искусственного интеллекта, связанными с Д. А. Поспеловым, были мероприятия по координации исследований по ИИ в социалистических странах,проведённые в 1980-е –1990-е годы под эгидой Комиссии «Научные вопросы вычислительной техники» Совета Экономической Взаимопомощи (СЭВ). Была создана международная базовая лаборатория стран СЭВ по ИИ, работавшая в Братиславе, одним из руководителей которой стал профессор Д. А. Поспелов. Вскоре после этого были образованы международные рабочие группы: сначала РГ18[74], а затем РГ-22 [75] как её продолжение. Итогом этой деятельности стало издание трёхтомного справочника по искусственному интеллекту, причём том 2 «Модели и методы» вышел под редакцией Д. А. Поспелова [68].

Дмитрий Александрович много работал также в секции «Прикладные проблемы кибернетики» при Московском доме научно-технической пропаганды (общество «Знание»), был заведующим Международной лабораторией ЮНЕСКО по ИИ при Институте программных систем РАН, руководителем ВНТК «Интеллектуальные системы» РАН, членом редсоветов издательств «Энергоатомиздат», «Радио и связь», главным редактором созданного им в 1991 году журнала «Новости искусственного интеллекта». На протяжении более 20 лет профессор Д. А. Поспелов был заместителем главного редактора журнала «Известия Академии наук: «Теория и системы управления».

В-четвертых, важнейшим достижением Д. А. Поспелова стало создание открытой, междисциплинарной научной среды, связанной с интенсивным формированием новых научных идей и свободным обменом ими в процессе взаимодействия между различными научными, учебными, промышленными сообществами и структурами.

С 21 по 25 ноября 1988 года в Переславле-Залесском на базе Института программных систем (ИПС) состоялась 1-я Всесоюзная конференция по искусственному интеллекту. Сопредседателями программного комитета были академики Е. П. Велихов и Г. С. Поспелов, а их заместителями — директор ИПС А. К. Айламазян и профессор Д. А. Поспелов. Это была весьма представительная конференция, которая собрала более 500 участников из многих республик Советского Союза и была разбита на 13 секций.

В мае 1989 года Д. А. Поспелов выступил главным инициатором объединения ведущих учёных и специалистов по ИИ в официальное сообщество, названное Советской (а с 1986 года -– Российской) ассоциацией искусственного интеллекта. Он более семи лет был президентом Ассоциации, а в 2000-е годы являлся почётным членом её научного совета.

Вот уже более 20 лет Ассоциация живёт и успешно функционирует. Ею регулярно проводятся Национальные конференции по ИИ: за период с 1988 г. по 2019 г. было проведено 17 Национальных конференций. Ранее они проводились 1 раз в два года, а, начиная с XVII-й конференции КИИ-2019 в Ульяновске, ежегодно.

Творческая натура Дмитрия Александровича Поспелова не ограничивалась только профессиональной научной и преподавательской деятельностью. Стихи и проза, история и археология, нумизматика и живопись – вот далеко не полный перечень его увлечений. В 2000-е годы, даже будучи тяжело больным, он работал над изданием сборников своих стихов «Размышления» (2005 г.) и рассказов «Знак Водолея» (1998 г.), «Чужое пространство» (2005 г.), продолживших своеобразное реально-виртуальное, полумистическое направление в современной российской литературе. Близкая ему философия русского космизма нашла своё отражение в литературных образах.

Рис. 14. Профессор Д. А. Поспелов в своём домашнем кабинете среди картин П. Фатеева (2007г.).

В 2007 году вышла в свет замечательная книга Д. А. Поспелова-искусствоведа о творческом руководителе группы русских художников-интуитивистов «Амаравелла: мистическая живопись Петра Фатеева» [76]. В аннотации на обложке автор написал: «эти картины для тех, в ком живет творческий дух, кому тесно в этом Старом тяжеловесном Мире, кто хочет поднять завесы ложной реальности и разложить вещи, чтобы найти в них то, что он никогда не видел, кто жаждет прохлады горных высот и чистого воздуха, кто, змеёй извиваясь, хочет сбросить свою ветхую шкуру и взглянуть новыми глазами на Новый Мир».

Именно так можно оценить жизнь и творчество самого Дмитрия Александровича Поспелова. Это *амаравелла* -– «ростки бессмертия» в задуманном им Новом Мире Искусственного Интеллекта.

VI. ВМЕСТО ЗАКЛЮЧЕНИЯ

Вместо заключения автору хотелось бы привести высказывания о профессоре Д. А. Поспелове его коллег и друзей –– ведущих учёных и руководителей РАИИ на первых Поспеловских чтениях в Политехническом музее «Искусственный интеллект — проблемы и перспективы» (2003-й год).

Один из основоположников исследований в области кибернетики и искусственного интеллекта, автор известной, переведённой на русский язык монографии «Искусственный интеллект», профессор Алекс Эндрю (Великобритания)

В 1851–1852 годах леди Лавлейс стала обсуждать такие вопросы, как применение вычислительных устройств для выполнения символьных операций или для сочинения музыки, которые далеко выходили за пределы обычных вычислительных задач. Может быть, это и были первые идеи, положившие начало тому направлению, которое сегодня называется «искусственный интеллект». В данной области работы Д. А. Поспелова являются фундаментальными, вносящими существенный вклад в ее развитие.

Дмитрий Поспелов привнёс в ИИ и смежные науки совершенно новое качество. В частности, это фантазия и воображение, которые присущи многим его работам, например, работам по компьютерному моделированию волшебных сказок.

Применение компьютеров в искусственном интеллекте значительно отличается от их применений в вычислительной области. И Д. А. Поспелов одним из первых показал, что даже обычный фон-неймановский компьютер может эффективно использоваться не только для счёта и решения вычислительных задач, но и для имитации различных типов поведения, поддержки образного мышления, моделирования творческих процессов.

Член Научного совета Российской ассоциации искусственного интеллекта, д.т.н., профессор Виктор Константинович Финн.

Можно выделить 3 основные измерения многогранной деятельности профессора Д. А. Поспелова. Во-первых, Дмитрий Александрович –– основоположник когнитивного направления в отечественной науке -– в широком смысле этого слова, т.е. и в смысле математических формализмов, и в смысле связей с психологией и лингвистикой.

Второе измерение, второй аспект его творчества –– это формирование научной школы, воспитание учеников, производство специалистов, близких ему по духу и по идеям.

Наконец, третий аспект его творчества –– это создание открытой, междисциплинарной научной среды (ничуть не менее важный, чем два предыдущих). Сейчас многие политики (а, в особенности, политтехнологи) говорят о «гражданском обществе», смутно представляя себе, что это такое. Между тем, Д. А. Поспелов продемонстрировал пример создания ячейки «гражданского общества» – Ассоциации искусственного интеллекта, которая живет самостоятельной жизнью и является примером открытого сообщества, активно взаимодействующего с самыми разными научными, учебными, промышленными структурами. Относясь равнодушно и даже с некоторой брезгливостью к чинам, иерархии, официозу и всему прочему, Дмитрий Александрович продемонстрировал, что и без чинов высоких можно создавать, творить и иметь множество учеников.

В заключение, я хочу сказать, что все мы вышли «из двух шинелей» –– Гермогена Сергеевича Поспелова и Дмитрия Александровича Поспелова. Они были «тёплыми» и создали для всех нас удивительный научный микроклимат.

Председатель Научного совета РАИИ, д.т.н., профессор Олег Петрович Кузнецов.

Мне хотелось бы вспомнить ещё об одной «шинели», из которой в значительной степени вышел и сам Дмитрий Александрович. Это Михаил Александрович Гаврилов, которого он считает своим учителем, и теория автоматов. Сейчас теория автоматов –– один из многочисленных разделов информатики, а в 1950-е и 1960-е годы она была почти синонимом кибернетики. Теорией автоматов занимались одни из основателей кибернетики –– Джон фон Нейман и Алан Тьюринг. Львиную долю литературы по кибернетике того времени составляли работы по теории автоматов.

Одна из первых и наиболее ярких школ, воспитавших целое поколение отечественных исследователей, –– это знаменитые Гавриловские школы [77,78]. Если бесспорным лидером школы в целом был действительно сам М. А. Гаврилов, то неформальным лидером молодежи, которая собиралась на этих школах, стал, безусловно, Дмитрий Поспелов. Самые первые научные работы Дмитрия Александровича были в области теории автоматов. Первая его книга, знакомая не одному поколению студентов, –– «Логические методы анализа и синтеза схем» – ныне считается одним из двух–трех классических учебников по теории автоматов и логическим методам схемотехники.

По крайней мере, наше поколение, первые ученики и последователи Дмитрия Александровича хорошо знают такие его прекрасные книги, как «Игры и автоматы», «Вероятностные автоматы», «Мышление и автоматы». Именно под воздействием идей Д. А. Поспелова произошёл плавный переход от казавшихся скучноватыми формализмов классической теории автоматов к содержательным и интересным моделям ситуационного управления, которые впоследствии стали частью проблематики искусственного интеллекта.

Как раз из Гавриловских школ Д. А. Поспелов почерпнул и привил своей школе искусственного интеллекта удивительное чувство научной и человеческой солидарности, создав ту проникнутую неформальной атмосферой уникальную научную среду, в которой возможно по-настоящему свободное творчество, свободный поиск истины.

Действительный член Европейского координационного комитета по искусственному интеллекту, вице-президент Российской ассоциации искусственного интеллекта, д.т.н., профессор Вадим Львович Стефанюк.

То, что делал Дмитрий Александрович Поспелов в искусственном интеллекте, я обозначу термином «поиски смысла». Поиски смысла во всем. Поиски смысла в текстах, поиски смысла в картинах, которые написаны художниками, поиски смысла в репликах, которыми обмениваются диспетчеры в морском порту, поиски смысла в волшебных сказках... Наконец, поиск скрытого смысла в математике. Именно отсюда появились семиотические системы, при построении которых ставился вопрос: что же происходит при обучении в системах типа конечных автоматов?

Конечно, многие ищут смысл. Но Д. А. Поспелов отличался тем, что он не просто искал смысл, а ещё и разработал уникальный формализованный аппарат, позволяющий двигаться в этом направлении.

Руководитель тематического отделения Российской ассоциации искусственного интеллекта «История информатики», д.т.н., профессор Яков Ильич Фет.

Ещё одна, очень важная сторона многогранной деятельности Д. А. Поспелова -– это его работы по истории кибернетики, информатики и искусственного интеллекта. В начале 1980-х годов Дмитрий Александрович вместе с Модестом Георгиевичем Гаазе-Рапопортом организовали постоянно действующий семинар по истории кибернетики, который проводился в здании Политехнического музея. Первое заседание состоялось 22 декабря 1983 года. На этом семинаре выступали с воспоминаниями многие наши известные специалисты. Имеется около 30 кассет с записями этих семинаров.

В 1991 году вышел в свет первый номер официального журнала Ассоциации искусственного интеллекта «Новости искусственного интеллекта», главным редактором которого был Д. А. Поспелов. И уже в первом номере он ввел специальную рубрику «Из истории искусственного интеллекта». В ней были опубликованы воспоминания о семинаре по психонике. Позднее, в 1994 году, в журнале появились его статьи о решающей роли Г. С. Поспелова в признании у нас искусственного интеллекта как научного направления и развёртывании соответствующих исследований в нашей стране, а также об организации деятельности международных рабочих групп в области ИИ РГ-18 и РГ-22.

Затем Дмитрий Александрович Поспелов выступил как инициатор издания книг по истории информатики, в частности, в серии «Информатика: неограниченные возможности и возможные ограничения».

Первая книга под названием «Очерки истории информатики в России» вышла летом 1998 года. Он написал блестящую вводную статью под названием «Становление информатики в России» [79]. По существу, эта статья стала программной для всех наших дальнейших публикаций. Наши работы по истории информатики продолжаются. Вышло еще несколько книг: об А. Н. Колмогорове, А. А. Ляпунове, Л. В. Канторовиче, А. И. Берге.

В наших планах – продолжение серии изданий под общим названием «История отечественной информатики». История, в которой ярок след Д. А. Поспелова.

Работа выполнена при поддержке РФФИ, проект №20-07-00770.

Список литературы

[1] Поспелов Д. А. Логические методы анализа и синтеза схем. М.: Энергия, 1964.

[2] Поспелов Д. А. Игры и автоматы. М.: Энергия, 1965.

[3] Поспелов Д. А. Вероятностные автоматы. М.: Энергия, 1970.

[4] Поспелов Д. А. Проблемы теории вычислительных систем. Автореферат диссертации . . . д.т.н. М.: МЭИ, 1968.

[5] Вагин В. Н., Еремеев А. П. Научная школа искусственного интеллекта в Московском энергетическом институте на базе кафедры прикладной математики: становление и развитие. Вестник МЭИ. №2, 2015, с.29-37.

[6] Клыков Ю. И. Модельный метод управления динамическими ситуационными системами. Автореферат диссертации. . . к.т.н. М.: МЭИ, 1967.

[7] Вагин В. Н. Игровые методы управления сложными системами. Автореферат диссертации. . . к.т.н. М.: МЭИ, 1968.

[8] Поспелов Д. А. Семинар по психонике. Новости искусственного интеллекта, №1, 1991, с.31-36.

[9] Поспелов Д. А. О задачах психоники. Проблемы бионики. М.: Наука, 1967, с.294-297.

[10] Поспелов Д. А. Теория гироматов. Проблемы бионики. М.: Наука, 1973, с.397-402.

[11] Hewitt C. Viewing Control Structures as Patterns of Message Passing. Artificial Intelligence, vol.8, no 3, pp.323-364.

[12] Аверкин А. Н., Дулин С. К., Хорошевский В. Ф., Эрлих А. И. Отдел интеллектуальных прикладных систем// 50 лет ВЦ РАН: история, люди, достижения. М.: ВЦ РАН, 2005. С. 198-211

[13] Поспелов Д. А Введение в теорию вычислительных систем. М.: Сов.Радио, 1972.

[14] Поспелов Д. А., Пушкин В. Н. Мышление и автоматы. М.: Сов.Радио,1972.

[15] Захаров В. Н., Поспелов Д. А., Хазацкий В. Е. Системы управления. М.: Энергия, 1972

[16] Поспелов Д. А. Большие системы. Ситуационное управление. М.: Знание, 1975.

[17] Поспелов Д. А. Принципы ситуационного управления. Известия АН СССР: Техническая кибернетика, №2, 1971, с.10-17.

[18] Клыков Ю. И., Поспелов Д. А., Пушкин В. Н. Зачем нужно ситуационное управление? Вопросы кибернетики. Вып.13. Ситуационное управление: Теория и практика. Ч.1. М.: НСКП «Кибернетика» при Президиуме АН СССР, 1974. С.7-15.

[19] Поспелов Д. А.Системный подход к моделированию мыслительной деятельности. Проблемы методологии системного исследования. М.: Мысль, 1970. С.338-358.

[20] Поспелов Д. А. Новые информационные технологии -– это те ключи, которые откроют нам путь в новое общество. Новости искусственного интеллекта, №2, 1994, с.57-76.

[21] Поспелов Г. С., Поспелов Д. А. Искусственный интеллект. Вестник АН СССР, №10, 1975, с.26-36.

[22] Стефанюк В. Л. От автоматов М. Л. Цетлина к искусственному интеллекту. Новости искусственного интеллекта, №4,1995,с.56-92.

[23] Machine Intelligence. Vol.9. Ed. by J.E.Hayes, D.Michie, L.Mikulich. Chichester: Ellis Horwood Limited, 1979.

[24] Поспелов Д. А. Системы искусственного интеллекта: итоги и ближайшее будущее. Problems of |Control and Ibformation Theory, vol.9, no 1, 1980, pp.3-18.
[25] Тарасов В. Б. Д. А. Поспелов — основоположник искусственного интеллекта в СССР и России. Политехнические чтения. Вып.7. Искусственный интеллект: проблемы и перспективы. М.: РАИИ, 2004. С.34-47.
[26] Поспелов Д. А. Логико-лингвистические модели в системах управления. М.: Энергоатомиздат, 1981.
[27] Поспелов Д. А. Фантазия или наука: на пути к искусственному интеллекту. М.: Наука, 1982.
[28] Варшавский В. А., Поспелов Д. А. Оркестр играет без дирижера. Размышления об эволюции некоторых технических систем и управлении ими. М.: Наука, 1984.
[29] Поспелов Д. А. Ситуационное управление: теория и практика. М.: Наука, 1986.
[30] Гаазе-Рапопорт М. Г., Поспелов Д. А. От амёбы до робота: модели поведения. М.: Наука, 1987.
[31] Кандрашина Е. Ю.,Литвинцева Л. В., Поспелов Д. А. Представление знаний о пространстве и времени в системах искусственного интеллекта. М.: Наука, 1988.
[32] Поспелов Д. А. Моделирование рассуждений. Опыт анализа мыслительных актов. М.: Радио и связь, 1989.
[33] Поспелов Д. А., Шустер В. А. Нормативное поведение в мире людей и машин. Кишинев: Штиинца, 1990.
[34] Нечеткие множества в моделях управления и искусственного интеллекта. Под ред.Д. А. Поспелова. М.: Наука, 1986.
[35] Уинстон П. Искусственный интеллект: Пер. с англ. М.: Мир, 1980.
[36] Hoffman A.G. On the Principles of Intelligence. Proceedings of the First World Conference on the Fundamentals for Artificial Intelligence (Paris, France, July 1-5, 1991). Paris: Masson, 1991. P.257-266.
[37] Нильсон Н. Принципы искусственного интеллекта: Пер. с англ. М.: Радио и связь, 1985.
[38] Simon H.A. Artificial Intelligence: an Empirical Science. Artificial Intelligence, vol.77, 1995, pp.95-127.
[39] McCarthy J., Hayes P. J. Some Philosophical Problems from the Standpoint of Artificial Intelligence. Machine Intelligence. Vol 4. Ed. by B.Meltzer and D.Michie. Edinburgh: Edinburgh University Press, 1969, pp.463-502.
[40] McCarthy J. From Here to Human-Level AI. Artificial Intelligence, vol.171, 2007, pp.1174-1182.
[41] Artificial General Intelligence. Ed. by B.Goertzel and C.Pennachin. Berlin: Springer, 2007
[42] Theoretical Foundations of Artificial General Intelligence/ Ed. by Pei Wang and B.Goertzel. Beijing: Atlantis Press, 2012.
[43] Reggia J. The Rise of Machine Consciousness: Studying Consciousness with Computational Models .Neural Networks. 2013. Vol. 44. P.112-131.
[44] Stanley K. Open-Endedness: a New Grand Challenge for AI. O'Reilly Artificial Intelligence Conference, 2019.
[45] Осипов Г. С., Панов А. И, Чудова Н. В. Управление поведением как функция сознания. Известия РАН: Теория и системы управления, №4, 2014, с.49-62.
[46] Тарасов В. Б. От многоагентных систем к интеллектуальным организациям. М: Эдиториал УРСС, 2002.
[47] Финн В. К. Искусственный интеллект: методология, применения, философия. М.: КРАСАНД, 2011.
[48] От моделей поведения к искусственному интеллекту. Под ред. В.Г.Редько. М.: КомКнига, 2006.
[49] Тарасов В.Б. Оппозиционные шкалы в модели мира. К 70-летию Д.А.Поспелова. Программные продукты и системы, №2, 2003, с.44-47.
[50] Поспелов Д. А. «Серые» и/или «черно-белые». Прикладная эргономика. Специальный выпуск «Рефлексивные процессы», №1, 1994, с.29-33.
[51] Поспелов Д. А. Знания и шкалы в модели мира. Модели мира. М.: РАИИ, 1997. С.69-84.
[52] Толковый словарь по искусственному интеллекту. А. Н. Аверкин, М. Г. Гаазе-Рапопорт, Д. А. Поспелов (редакторы – составители). М:. Радио и связь, 1992.
[53] Ежкова И. В., Поспелов Д. А. Принятие решений при нечетких основаниях. I. Универсальная шкала. Известия АН СССР: Техническая кибернетика, №6, 1977, с.3-11.
[54] Поспелов Д. А. Элементы аксиоматики временных отношений. Вопросы кибернетики. Вып.14. Ситуационное управление: Теория и практика. Ч.2. М.: НСКП «Кибернетика» при Президиуме АН СССР, 1975. С.15-21.
[55] Литвинцева Л. В., Поспелов Д. А. Время в роботах и диалоговых системах. Вопросы кибернетики. Вып.61. Проблемы искусственного интеллекта. М.: НСКП «Кибернетика» при Президиуме АН СССР, 1980. С.61-70.
[56] Pei Wang, Cognitive Logic vs Mathematical Logic. Lecture Notes of the 3rd International Seminar on Logic and Cognition, Guangzhou, China, 2004.
[57] Поспелов Д. А. От коллектива автоматов к мультиагентным системам. Труды Международного семинара «Распределённый искусственный интеллект и многоагентные системы» (DAIMAS'97, Санкт-Петербург, Россия, 15-18 июня 1997). СПб.: СПИИРАН, 1997. С.319-325.
[58] Поспелов Д. А. Многоагентные системы -– настоящее и будущее. Информационные технологии и вычислительные системы, №1, 1998, с.14-21.
[59] Стефанюк В. Л. От многоагентных систем к коллективному поведению. Труды Международного семинара «Распределённый искусственный интеллект и многоагентные системы» (DAIMAS'97, Санкт-Петербург, Россия, 15-18 июня 1997). СПб.: СПИИРАН, 1997. С.327-338.
[60] Pospelov D. A., Ehrlich A. I., Osipov G. S. Semiotic Modeling and Situation Control. Proceedings of 1995 ISIC Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems (Monterey CA, USA, August 1995), p.127-129.
[61] Поспелов Д. А. Прикладная семиотика и искусственный интеллект. Программные продукты и системы, №3, 1996, с.10-13.
[62] Поспелов Д. А., Осипов Г. С. Прикладная семиотика. Новости искусственного интеллекта, №1, 1999, с.9-35.
[63] Осипов Г. С. От ситуационного управления к прикладной семиотике. Новости искусственного интеллекта, №6, 2002, с.3-7.
[64] Будущее искусственного интеллекта. К. Е. Левитин, Д. А. Поспелов (редакторы – составители). М.: Наука, 1991.
[65] Поспелов Д. А. Десять горячих точек в исследованиях по искусственному интеллекту. Интеллектуальные системы (МГУ), т.1, вып.1-4, 1996, с.47-56.
[66] Поспелов Д. А. Моделирование рассуждений: состояние и ближайшее будущее. Теория и применение искусственного интеллекта. Созополь, Болгария, 1989. С.46-53.
[67] Поспелов Д. А. Интеллектуальные интерфейсы для ЭВМ новых поколений. Электронная вычислительная техника. Сборник статей. Вып. 3. М.: Радио и связь, 1989. С.4-20.
[68] Искусственный интеллект – в 3-х кн. Кн.2. Модели и методы: Справочник Под ред. Д. А. Поспелова. М.: Радио и связь, 1990.
[69] Тарасов В. Б. Проблема понимания: настоящее и будущее искусственного интеллекта. Открытые семантические технологии проектирования интеллектуальных систем. Материалы V-й Международной научно-технической конференции (OSTIS-2015, Минск, БГУИР, 19-21 февраля 2015 г.). Минск: Изд-во БГУИР, 2015. С.25-42.
[70] Зенкин А. А. Когнитивная компьютерная графика. М: Наука, 1991.
[71] Поспелов Д. А. Когнитивная графика -– окно в новый мир. Программные продукты и системы, №3, 1992, с.4-6.
[72] Поспелов Д. А., Литвинцева Л. В. Как совместить левое и правое? Новости искусственного интеллекта, №2, 1996, с.66-71.
[73] Поспелов Д. А. Метафора, образ и символ в познании мира. Новости искусственного интеллекта, №1, 1998, с.94-114.
[74] Поспелов Д. А., Хорошевский В. Ф. Рабочая группа РГ-18. Новости искусственного интеллекта, №2, 1992, с.87-100.
[75] Поспелов Д. А., Хорошевский В. Ф. Можно ли повторить чудо вторично? (История РГ-22). Новости искусственного интеллекта, №2, 1994, с.77-95.

[76] Поспелов Д. А. Амаравелла: мистическая живопись Петра Фатеева. М.: Фантом пресс, 2007.
[77] Кузнецов О. П. Гавриловские школы: жизнь после смерти. Новости искусственного интеллекта, №2, 1996, с.88-92.
[78] Поспелов Д. А. Школа МАГа. Новости искусственного интеллекта, №3, 1997, с.80-129.
[79] Поспелов Д. А. Становление информатики в России. Очерки истории информатики в России. М.: Наука, 1998.

Artificial Intelligence: From the Past to the Future. About the Life Path and Scientific Heritage of Prof. Dmitry A. Pospelov

Valery B. Tarassov
Bauman Moscow State Technical University Moscow, Russia E-mail: Vbulbov@yahoo.com

The contribution of the outstanding scientist, the father of Artificial Intelligence in the USSR and Russia, the creator and first president of Soviet (then Russian) Association for AI, Prof. Dmitry Pospelov into various AI areas and related sciences has been considered. He was the founder of such new interdisciplinary research fields as Situational Control and Applied Semiotics, Lingua-Logical Modeling,Pseudo-Physical Logics and Non-Standard circular, fuzzy and general Scales. He also was the pioneer of Computers with Parallel and other Non-Traditional Architectures, the author of innovative approaches to individual and collective Behavior Modeling, including some original topics like computer-based behavior analysis of fairy tales heroes, as well as new techniques of Act Frames. Besides, Professor D.A.Pospelov has suggested some interesting approaches to common-sense reasoning with taking into account the results of psychological studies and proposed a classification of understanding levels. He should be viewed as a forerunner of agent theory and multi-agent systems with his early original contributions — Gyromata Theory and Decentralized Control in Collective Behavior.

By analogy with Bionics Dmitry Pospelov has coined the term Psychonics to specify a sort of Psychology for Artificial Systems. The objective of Psychonics is the investigation of main psychological approaches, methods, models and tools in order of their practical use in AI to create advanced intelligent systems. Here the main results, concerning the architecture of human psyche, organization of human activity and features of human personality, are of primary concern.

The paradigm of Psychonics may be viewed as a precursor of Artificial General Intelligence (AGI). Indeed, already in 1982 D.A.Pospelov formulated in his book «Fantasy or Science: on the Way to Artificial Intelligence» a very wide vision of Artificial Intelligence: «Modern investigations in AI should be aimed at studying and modeling of Human Psyche for the sake of its simulation in technical systems to solve a variety of practical problems conventionally viewed as intelligent ones» (p.211).

It is worth stressing that in psychology, the word Psyche stands for the totality of human mind, both conscious and unconscious. The investigation of consciousness makes appeal to a well-known psychological triad: «Cognition — Affect -– Conation». So D.A.Pospelov has envisioned Intelligence as a wholistic, open, heterogeneous dynamic system that perfectly corresponds to AGI paradigm.

Received 30.01.2020

Predicative representations of relations and links of the intellectual operational-technological control in complex dynamic systems

Alexander Kolesnikov, Eric Jasinski
Immanuel Kant Baltic Federal University
Kaliningrad, Russia
avkolesnikov@yandex.ru, ejasinski96@gmail.com

Sophiya Rumovskaya
Kaliningrad branch of FRC IC RAS
Kaliningrad, Russia
sophiyabr@gmail.com

Dmitry Alexandrovich Pospelov – the scientist, the author of the situational control and the mastermind of the artificial intelligence research, of blessed memory.

***Abstract*—The research belongs to the field of "intelligent control systems", where it is important to "learn", "understand", "remember", "evaluate the situation", "find a solution", "check the execution" and combines methods for solving intellectual problems, independently not solved by the human operator "in principle". The cognitive hybrid intellectual systems are proposed for computer simulation of cognitive formations and enhancing human intelligence in operational work. In respect of such systems we developed the language for the description of relations and links for predicative coding of verbal knowledge about resources, properties and actions of personnel, for coding of the grammar of identity substitution and the functional deformation of the cognitive image of the object state.**

***Keywords*—intelligent control systems, cognitive image, operational image, cognitive hybrid intellectual systems, language for describing relations and links, predictive coding**

I. Introduction

In the seminal work [1] is emphasized: "Thinking is primarily the establishment of relationships between objects. With the help of special thought mechanisms the environment, in which the subject lives and acts, is recreated in his head with the reflection of those signs and links that are found between objects. In contrast to the language of perception, in which, first of all, those properties of objects are fixed, which are manifested in influences on the senses, a special language of relations and links is characteristic of thinking. Through this specific language, a subject gets the opportunity for internal work with those objects and their properties that are not given in perception, which are outside the scope of his direct contacts". The first results of our researches on the semantics of relations in verbal-symbolic representations of knowledge of intelligent systems are reflected in [2]. Later they were developed in [3], [4] within the linguistic approach to hybridization and within the problem-structured methodology and technology of the functional hybrid intellectual systems (HyIS). In the paper we considered the development of the language describing relations and links (LDRL) for predictive coding of word and verbal knowledge about resources, properties and actions of the personnel, for coding the grammar of identical transformations and functional deformation of the cognitive image of the state of a control object.

The reported study was funded by RFBR according to the research project №19–07–00208.

II. Predicative model of cognitive and operational image of the object of operational-technological control in complex dynamic systems

The key aspect of the operator's performance is remote control of the object which is out of sight of the controlling operator. Then the operator by the actions physically transforms the control object, without observing it directly. In such conditions the image of the control object (CO) acquires the specific meaning — representation "turning an absent object into a present in the mind" [5]. The control would be effective if the image reflects the important for the operator aspects of the CO and if the image is relevant to his operational actions. The cognitive image (CI) of the object of control as per D.A. Oshanin — the result of the cognitive reflection function, an instrument for completing the knowledge of the CO. This is the worldview and the "depository" of information about the control object. CI is redundant, entropic and continuously developing, enriching and reducing, internally restructuring. CI is an instrument of "inventory" of potentially useful properties of CO. A cognitive image is an open information system for flexible switching from one structure to another depending on changes in relevant parameters.

The cognitive image of the object of control as per D.A. Oshanin — the result of the cognitive reflection function, an instrument for completing the knowledge of the CO. This is the worldview and the "depository" of information about the control object. CI is redundant, entropic and continuously developing, enriching and reducing, internally restructuring. CI is an instrument of "inventory" of potentially useful properties of CO. A cognitive image is an open information system for flexible switching from one structure to another depending on changes in relevant parameters.

The operational image of an object (OO) of control as per D.A. Oshanin – the image of the OC as a result of the regulatory function of reflection, which is formed by the human-operator (the subject of control) in the process of fulfilling the tasks of operational-technological control (OTC). Its a functional psychological system (model) that processes information about the sequential states of the object into appropriate action on him. The operator's reflections on the CO image and the changes (transformations) that the CO image undergoes under this are called the dynamics of the operational image according to D.A. Oshanin. The operational image is dynamic, changeable, fluid, contradictory and regulates the specific action in time and space. It is pragmatic, "subordinate" to the problem being solved and relevant to it, specific, suitable only for a specific task, concise, peculiarly limited, "distracted" from the CO features that are not currently used to solve the task.

The semantic of CI and OO is reflected by the language of the professional activity (LPA) of the human-operator. The traditional model of abstract knowledge of LPA is predicative model with "predicate–argument" as the base structure, which can be assigned the values "true" or "false" [6]. It had a success in linguistic due to the possibility of organizing calculations, and for many years predetermined the "linguistic turn" in the philosophy of knowledge [7]. In classical logic of Aristotel the subject was defined as the entity of statement. The subject (σ) can be associated either with a resource of the CO, or with the human-operator notion of it. The statement of resource is made in the form of a predicate (ρ), positive or negative statement. The structure of attributive statement is represented by the formula $\sigma - \rho$. A connective (the verb *is*) can be introduced into it as a component expressing the inherent or non-inherent character of the feature to the resource.

Afterwards, attributive logic was being displaced by the relational logic, defining the predicate not as a property of a resource (σ — ρ or σ *is* ρ), but as a relation between two (or three, etc.) resources (arguments). This was reflected in formulas of type $\rho\ (x, y)$, $\rho\ (x, y, z)$, where ρ symbolize monadic or dyadic predicate.

Predicate (lat. praedicatum, from praedicare – make known beforehand, declare) [8] – something which is affirmed or denied concerning an argument of a proposition; assert (something) about the subject of a sentence or an argument of proposition. Thus, predicate is the term of logic and linguistics, denoting the constitutive member of a statement. It is what one expresses (affirms or denies) about the subject. By the number of actants predicates are divided into monadic; dyadic; triadic; four-place, etc. At a syntactic level a predicate is a kernel structure with actants. A kernel is a verbal construction. Actants combine with a kernel by the system of relations. Nodes of this construction are names (noun, pronoun, numerals) in their attributive form. A syntax (formalized record) of a predicate for the atomic sentence is accepted in the following interpretation:

(ACTION) *to have as a subject* (SUBJECT), (ACTION) *to have as an object* (OBJECT1), (ACTION) *to exist* (OBJECT2), *to have a result* (OBJECT3), (OBJECT) *to have a mean* (OBJECT4), (ACTION) *to have a mean* (OBJECT5), *whence* (OBJECT6), (ACTION) *to where* (OBJECT7),

(ACTION) (*R0* (SUBJECT) *R1* (OBJECT1)*R2* (OBJECT2), *R3* (OBJECT3) *R4* (OBJECT4) *R4* (OBJECT5) *R5*(OBJECT 6) *R6* (OBJECT7)),

(XD) (R0 (XS) R1(XO1) R2 (XO2), R3 (XO3) R4 (XO4) R4 (XO5) R5(XO6) R6 (XO7)),

where ACTION, *XD* – the kernel of a predicate (predicator), in the general case, its a verb structure that may have additional attribute components; SUBJECT, *XS* – an object of the predicate, a material entity, an intangible object, "empty"; OBJECT1, *XO1,. . .* , OBJECT 7, *XO7* – actants, may be represented by special objects or in the form of concrete characteristics of the predicator, indicated by adverbs (*yesterday, today, there, here, etc.*); *R0* – the relation "*to be a subject*", *R1,. . . , R7* – relations of a predicate.

The size of the message decrease, it is coding by the replacement according to the qualified agreement names of ACTION, SUBJECT, OBJECT and relations by their code values, on which the quality and efficiency of using CI and processing OO in cognitive HyIS depends.

Such coding is named predicative RKX-coding. Codes of names of relations and concepts are formed by the rules in conformity with the next classification of relations: "resource-resource", "property-property", "action-action", "resource-property", "property-resource", "resource-action", "action-resource", "action-property", "property-action" and their formal-logical properties [9]

RKX -codes include not only features of the name (relation – *R* , concept *X*, their number in the dictionary), but also features of the formal-logical properties (symmetric, transitive, direct, reverse, etc.) and features of the membership to classes according to D.A. Pospelov (classifications, indicative, quantitative, comparisons, implementation, etc.) and to classes according to A.V.

Kolesnikov и V.F. Ponamarev – K.

The formation of syntagmatic, situational relations occurs in a different way. This relationship is “to be time”, “to be space”, “to be a cause”, “to be a consequence”, etc. Situational relationships are common in LPA because they recreate spatio-temporal and causal relationships that explicitly or implicitly always present in operational messages. Language practice has developed forms for the implementation of these relations – syntagmas and syntagmatic chains.

A syntagma according to P.M. Kolychev [10] – binomial juxtaposition of any language elements in a linear sequence: VALUE *equal* 50, SUBSTATION SEVERNAYA *near* KALININGRAD, TRANSFORMATOR *is at* SUBSTATION. At the abstractive level a syntagma has the next form:

$$(OBJECT1)R(OBJECT2)or(XO1)R(XO2) \quad (1)$$

where OBJECT1, *XO1* – abstraction from VALUE, SUBSTATION SEVERNAYA, TRANSFORMATOR; OBJECT2, *XO2* – abstraction from 50, KALININGRAD, SUBSTATION; *R* – abstraction from *equal, near, is at* and suchlike.

A syntagma (1), a “tripod” as simple kernel construction according to D.A. Pospelov. It is the basic unit of the language of situation control. There is a relation *R* in the center of the construction. On the ends of the construction there are the left (*XO1*) and the right (*XO2*) poles – concepts (parentheses are used as delimiters). Relations *R* connect together the resources of the control object, which appear in them as certain roles, the semantics of which is determined by the type of relationship. Such relations *R* are called role-based.

III. Lexis of the language of relations and links of cognitive and operational images in a predicative model

The classification of concepts interprets the three philosophical categories of A.I. Uemov (“thing”, “property” and “relation”) in terms of "intelligent control systems for electrical networks and electrical installations" (the subject area).

Each concept in LDRL is denoted by $x^n \in X^n$, where n = 0 – a feature of a basic concept, $n \in \{1, 2, \dots\}$ – a feature of a derivative, complex concept. Let’s represent a set of concepts of the LDRL as $X^0 = X^0_\alpha \bigcup X^0_\beta$, where X^0_α, X^0_β – the basis and the auxiliary subsets. The basis set is denoted by three categories: *resource, property and action*. In the LPA sentences, they are mainly expressed by nouns, adjectives, ordinal numbers and definitive adverbs. The auxiliary set of concepts is denoted by categories: *measure, value, characteristic, parameter, name, state, estimate, exotic*. To construct the classifier X^0, the corresponding definitions are adopted [3], which are one-to-one related to the following predicates: $\rho^1_1\left(x^0\right)$ - *to be a resource*, $\rho^1_2\left(x^0\right)$ - *to be a property*, $\rho^1_3\left(x^0\right)$ - *to be an action*, $\rho^1_4\left(x^0\right)$ - *to be a measure*, $\rho^1_5\left(x^0\right)$ - *to be a value*, $\rho^1_6\left(x^0\right)$ - *to be a character* (*characteristic property*), $\rho^1_7\left(x^0\right)$ - *to be a parameter* (*physical property*), $\rho^1_8\left(x^0\right)$ - *to be a name* (*name property*), $\rho^1_9\left(x^0\right)$ – *to be a state*, $\rho^1_{10}\left(x^0\right)$ - *to be an estimate*, $\rho^1_{11}\left(x^0\right)$ – *to be an exotic concepts*.

The introduced categories are defined through gender-specific differences, examples and relationships with parts of speech. This makes the class separation of X^0 understandable and outlines a certain formalism in the semantic decomposition of expressions of the LPA. Exotic concepts make the classification easily extensible by analysis of elements ${}^{11}X^0$ and introducing of new definitions and corresponding predicates.

If we set three of imposed above predicates over X^0

$$M_1 =< X^0, \rho^1_i\left(x^0\right) \mid i \in \overline{1;5}, i \in \{9, 10, 11\} >, \quad (2)$$

then we would get the separation of X^0 into eight subsets: ${}^1X^0$ – resources, ${}^2X^0$ – properties and ${}^3X^0$ – actions, ${}^4X^0$ – measures, ${}^5X^0$ – values, ${}^6X^0$ – states, ${}^7X^0$ – estimates и ${}^8X^0$ – exotic concepts. Meanwhile, there must be implemented the next: $X^0 = \bigcup_i {}^iX^0$, if $i \in \overline{1;8}$ and $\varnothing = \bigcup\left({}^iX^0 \bigcap {}^jX^0\right)$, if $i \neq j$ and $i, j \in \overline{1;8}$. Because of informality of the separation procedure of X^0 into subsets, two things can take place: ${}^ix^0_k \equiv^j x^0_m$ and ${}^ix^0_k, {}^j x^0_m \in {}^zX^0 | i = j = z$ – duplication of concepts, and also ${}^ix^0_k \equiv^j x^0_m$ and ${}^ix^0_k \in {}^iX^0, {}^j x^0_m \in {}^jX^0 | i \neq j$ – a disambiguation. Both of it can be detected formally, but if the first one is eliminated easily, the second demands the analysis by a subject.

If we set predicates $\rho^1_6\left(x^0\right)$, $\rho^1_7\left(x^0\right)$, $\rho^1_8\left(x^0\right)$ over ${}^2X^0$

$$M_2 =< {}^2X^0, \rho^1_i\left({}^2x^0\right) | i \in \overline{6;8} > \quad (3)$$

then we would get the separation of ${}^2X^0$ into three subsets: ${}^{21}X^0$ – parameters, ${}^{22}X^0$ – characteristics, ${}^{23}X^0$ – names. There must be implemented: ${}^2X^0 = \bigcup_i {}^{21i}X^0$ if $i \in \overline{1;3}$ and $\varnothing = \bigcup\left({}^{21i}X^0 \bigcap {}^{21j}X^0\right)$, if $i \neq j$ and $i, j \in \overline{1;3}$. If $x^0_k \in X^0$ and $\forall \rho^1_i\left(x^0_k\right) | i \in \overline{1;10}$ = false, then $x^0_k \in {}^8X^0$

The second group of LDRL lexical items includes relations of the next classes (by D.A. Pospelov): of definition (DEF15 – here and below this is the code of the class of a relation “definition”), of comprasion (DEF16), spatial (DEF17), temporal (DEF18), of inclusion (DEF19), of causality (DEF20) and of preference (DEF21). In LPA relations of resources is denoted by nouns (*part, identity, equality, object, subject, means, etc.*) and related to the them adjectives (*identical, equal, etc.*), verbs (*have, be, exist, characterize, etc.*), prepositions, unions or by their combination (*be over, be behind, be before, etc*).

Each relation (link) in LDRL is denoted by $r^0 \in R^0$ or $r^\pi \in R^\pi$ depending on the denoting in the LPA by a word or an expression. Let’s relate predicates to definitions DEF15–DEF21: $\rho^1_{12}\left(r^0\right)$ – *to be a relation of definintion*, $\rho^1_{13}\left(r^0\right)$ – *to be a relation of comprasion*, $\rho^1_{14}\left(r^0\right)$ – *to* *be a spatial relation*, $\rho_{15}^{1}\left(r^{0}\right)$ *– to be a temporal relation*, $\rho_{16}^{1}\left(r^{0}\right)$ *– to be a relation of inclusion*, $\rho_{17}^{1}\left(r^{0}\right)$ *– to be a relation of causality*, $\rho_{18}^{1}\left(r^{0}\right)$ *– to be a relation of preference.*

Then, having denoted imposed predicates over R^0

$$M_3 =< R^0,\ \rho_i^1\left(r^0\right)|i \in \overline{12;18} > \quad (4)$$

we would get the separation of R^0 into seven subsets of relations: ${}^1R^0$ – of definition, ${}^2R^0$ – of comprasion, ${}^3R^0$ – spatial, ${}^4R^0$ – temporal, ${}^5R^0$ – of inclusion, ${}^6R^0$ – of causality, ${}^7R^0$ – of preference. Meanwhile, there must be implemented: $R^0 = \bigcup_i {}^iR^0$, if $i \in \overline{1;7}$ and, because r_1^0 *– to be* $\in$ ${}^2R^0$, ${}^3R^0$, ${}^4R^0$, to be r_2^0 *– to have* $\in$ ${}^1R^0$, ${}^5R^0$, ${}^6R^0$, ${}^7R^0$, then generally $\varnothing = \bigcup\left({}^iR^0 \bigcap {}^jR^0\right)$, if $i \neq j$ and $i, j \in \overline{1;7}$.

The model M_3 follows the classification of D.A. Pospelov and gives the opportunity to allocate in the LPA relations and links of seven classes. It also allows to fill the corresponding dictionaries, but it doesn't denote the structure of signs of the LDRL.

Let's build the complete graph G_1

$$G_1 = \Big({}^1X^0, {}^2X^0, {}^3X^0; \\ R_{11}, R_{22}, R_{33}, R_{12}, R_{21}, R_{13}, R_{31}, R_{23}, R_{32}\Big), \quad (5)$$

where $R_{11} = \{R_{11}^0, R_{11}^\pi\}$, $R_{22} = \{R_{22}^0, R_{22}^\pi\}$, $\ldots$, $R_{32} = \{R_{32}^0, R_{32}^\pi\}$. Let us define relations R_{ij}^0, without going beyond the classification of concepts M_1 and M_2, the classification of relations M_3. Meanwhile, $\frac{1}{R}$ – inverse ***R*** – this is such relation in X^0, that for every ${}^iX^0$ and ${}^jX^0$ from X^0 the settlement ${}^iX^0\ \frac{1}{R}\ {}^jX^0$ equivalents to ${}^jX^0\ \boldsymbol{R}\ {}^iX^0$. R30(R_{11}^0 – the relation between concepts ${}^1x^0 \in^1 X^0$, including ${}^1R_{11}^0$, ${}^3R_{11}^0$, ${}^5R_{11}^0$) . R31(R_{22}^0 – ${}^2x^0 \in {}^2X^0$; ${}^1R_{22}^0$, ${}^3R_{22}^0$, ${}^5R_{22}^0$). R32(R_{33}^0 – ${}^3x^0 \in {}^3X^0$; ${}^1R_{33}^0$, ${}^4R_{33}^0$, ${}^5R_{33}^0$, ${}^6R_{33}^0$). R33(${}^1X^0\ R_{12}^0\ {}^2X^0 = \{{}^1X^0\ {}^1R_{12}^0\ {}^2X^0\}$). R34(${}^2X^0 R_{21}^0\ {}^1X^0 = \{{}^2X^{0 1}R_{21}^0\ {}^1X^0 | {}^1R_{21}^0 = \frac{1}{{}^1R_{12}^0}\}$). R35(${}^1X^0 R_{13}^0\ {}^3X^0 = \{{}^1X^{01}R_{13}^0\ {}^3X^0, {}^1X^{06}R_{13}^0\ {}^3X^0\}$). R36(${}^3X^0\ R_{13}^0\ {}^1X^0 = \{{}^3X^{01}R_{31}^0\ {}^1X^0, {}^3X^{06}R_{31}^0\ {}^1X^0 | {}^6R_{31}^0 = \frac{1}{{}^6R_{13}^0}\}$). R37(${}^3X^0\ R_{32}^0\ {}^2X^0 = \{{}^3X^{01}R_{32}^0\ {}^2X^0\}$). R38(${}^2X^0\ R_{23}^0\ {}^3X^0 = \{{}^2X^{01}R_{23}^0\ {}^3X^0 | {}^1R_{23}^0 = \frac{1}{{}^1R_{32}^0}\}$).

Consider the graph G_2, whose vertexes are indicated by the sets ${}^2X^0$, ${}^4X^0$, ${}^5X^0$ – are values. Then edges of the graph define the system of relations, from which we allocates the next: R_{24} – *property-measure*, R_{25} – *property-value*, R_{45} – *measure-value*:

$$G_2 = \left({}^2X^0, {}^4X^0, {}^5X^0, R_{44}, R_{55}, R_{25}, R_{45}\right), \quad (6)$$

where $R_{24} = \{R_{24}^0, R_{24}^\pi\}$, $R_{25} = \{R_{25}^0, R_{25}^\pi\}$, $\ldots$, $R_{45} = \{R_{45}^0, R_{45}^\pi\}$. The last relations that doesn't contradict the next: R_{44} – *measure–measure*, for example, a km is more than a m, a kV is more than a V; R_{55} - *value–value*, for example, 10 > 2, high voltage is more than low voltage, we'll consider, denoting R_{22}. R39(${}^2X^0\ R_{24}^0\ {}^4X^0 = \{{}^2X^0\ {}^1R_{24}^0\ {}^4X^0\}$). R40(${}^2X^0\ R_{25}^0\ {}^5X^0 = \{{}^2X^0\ {}^1R_{25}^0\ {}^5X^0\}$). R41(${}^4X^0\ R_{45}^0\ {}^5X^0 = \{{}^2X^0\ {}^1R_{24}^0\ {}^4X^0\}$).

Consider the graph G_3, whose vertexes are indicated by the sets ${}^1X^0$, ${}^3X^0$, ${}^6X^0$, ${}^7X^0$.

$$G_3 = \left({}^1X^0, {}^3X^0, {}^6X^0, {}^7X^0, R_{66}, R_{77}, R_{16}, R_{37}\right), \quad (7)$$

where $R_{66} = \{R_{66}^0, R_{66}^\pi\}$, $R_{77} = \{R_{77}^0, R_{77}^\pi\}$, $R_{16} = \{R_{16}^0, R_{16}^\pi\}$, $R_{37} = \{R_{37}^0, R_{37}^\pi\}$. Then edges define the system of relations, from which we allocates: R_{66} – *state-state*, R_{77} – *estimate-estimate*, R_{16} – *resource-state*, R_{37} – *action-estimate*. R42(${}^1X^0\ R_{16}^0\ {}^6X^0 = \{{}^1X^0\ {}^1R_{16}^0\ {}^6X^0\}$). R43(${}^3X^0\ R_{37}^0\ {}^7X^0 = \{{}^3X^{01}R_{37}^0\ {}^7X^0\}$). R44(${}^6X^0 R_{66}^0\ {}^6X^0 = \{{}^6X^0\ {}^4R_{66}^0\ {}^6X^0, {}^6X^0\ {}^6R_{66}^0\ {}^6X^0\}$). R45(${}^7X^0\ R_{77}^0\ {}^7X^0 = \{{}^7X^0\ {}^7R_{77}^0\ {}^7X^0, {}^7X^0\ {}^2R_{77}^0\ {}^7X^0\}$).

We establish a one-to-one correspondence between definitions R30 – R45 and predicates: $\rho_{20}^1\left(r^0\right)$ *– to be* R_{11}^0 , $\rho_{21}^1\left(r^0\right)$ *– to be* R_{22}^0 , $\rho_{22}^1\left(r^0\right)$ *– to be* R_{33}^0 , $\rho_{23}^1\left(r^0\right)$ *– to be* R_{12}^0 , $\rho_{24}^1\left(r^0\right)$ *– to be* R_{21}^0 , $\rho_{25}^1\left(r^0\right)$ *– to be* R_{13}^0 , $\rho_{26}^1\left(r^0\right)$ *– to be* R_{31}^0 , $\rho_{27}^1\left(r^0\right)$ *– to be* R_{23}^0 , $\rho_{28}^1\left(r^0\right)$ *– to be* R_{32}^0 , $\rho_{29}^1\left(r^0\right)$ *– to be* R_{24}^0 , $\rho_{30}^1\left(r^0\right)$ *– to be* R_{25}^0 , $\rho_{31}^1\left(r^0\right)$ *– to be* R_{45}^0 , $\rho_{32}^1\left(r^0\right)$ *– to be* R_{16}^0 , $\rho_{33}^1\left(r^0\right)$ *– to be* R_{37}^0, $\rho_{34}^1\left(r^0\right)$ *– to be* R_{66}^0, $\rho_{35}^1\left(r^0\right)$ *– to be* R_{77}^0 . Then having devoted imposed predicates over R^0

$$M_4 =< R^0, \rho_i^1\left(r^0\right)|i \in \overline{20;31} > \quad (8)$$

we would get the separation of R^0 into subsets R_{11}^0, R_{22}^0, $\ldots$, R_{23}^0, R_{24}^0, R_{25}^0, R_{45}^0, R_{16}^0, R_{37}^0, R_{66}^0, R_{77}^0. Meanwhile, there must be implemented the next: $R^0 = \bigcup_{ij} R_{ij}^0$, if $i \in \overline{1;7}$, $j \in \overline{1;7}$, and generally $\varnothing = \bigcup\left(R_{ij}^0 \bigcap R_{km}^0\right)$, if $i \neq k, j \neq m$ and $i, k \in \overline{1;7}$, $j, m \in \overline{1;7}$. For example: R_{11}^0 = {*to be, in, near, a part, to have a composition*}, R_{22}^0 = {*to be, more, less, equal*}, R_{33}^0 = {*to be simultaneously, earlier*}, R_{31}^0 = {*to have, a subject, an object, a means*}, R_{13}^0 = {*to be, condition, an object, a means, a place*}, R_{32}^0 = {*to have, a duration, a result*}. Relations from the classes R_{12}^0, R_{21}^0, R_{23}^0, R_{24}^0, R_{25}^0, R_{45}^0, R_{16}^0, R_{37}^0 include verbs *to be*, *to have*, and devoted in M_1 M_2 concepts *parameter*, *characteristic, name*, *measure, etc.*

After definition of the base relations R_{ij}^0 we'll consider the auxiliary ones R_{ij}^π. We define the language of the first level as $L_1(X^0, R_{ij}^0, P^1)$ = $\{r_{ij_l}^\pi\}$ – the set of auxiliary relations $\{\ r_{11_1}^\pi, r_{11_2}^\pi, \ldots, r_{11_\delta}^\pi, r_{22_1}^\pi, r_{22_2}^\pi, \ldots, r_{22_\varepsilon}^\pi, \ldots, r_{45_1}^\pi, r_{45_2}^\pi, \ldots, r_{45_\varrho}^\pi\ \}$ = R_{ij}^π derived from target relationships r_1^0 *– to have* and r_2^0 *– to be* by the rules $P^1 = \{p_1^1,\ p_2^1\}$.

Rule p_1^1, which we devote inductively. It's applied to the classes R_{11}^0, R_{22}^0, R_{33}^0*: 1) If* r_k^0*–* the target relationship and $r_{\mathrm{ij}_m}^\pi \in R_{\mathrm{ij}}^0 | i = j = \{1, 2, 3\}$*, then* $r_{\mathrm{ij}_l}^\pi = r_k^0 r_{\mathrm{ij}_m}^0 | k \in \overline{1;2}$ – the *auxiliary* relation in R_{ij}^π; 2) If $r_{\mathrm{ij}_l}^\pi$– the *auxiliary* relation, then $r_{\mathrm{ij}_l}^\pi r_{\mathrm{ij}_m}^0$ is also the *auxiliary* relation; 3) There are no others *auxiliary* relations in R_{ij}^π .

Rule p_2^1, which is applied to the classes R_{12}^0, R_{21}^0, R_{23}^0, R_{32}^0, R_{24}^0, R_{25}^0, R_{45}^0, R_{16}^0, R_{37}^0, ${}^{211}X^0, {}^{212}X^0$: 1) *If* r_k^0– the target relationship and $r_{\mathrm{ij}_m}^0 \in R_{\mathrm{ij}}^0 | i = \{1,2,4\}, j \in \overline{1;5}, i \neq j,\ {}^{21n}x^0 \in {}^{21n}X^0 | n = \{1,2\}$, then $r_{\mathrm{ij}_l}^\pi = r_k^0 r_{\mathrm{ij}_m}^0 | k \in \{1,2\}$ or $r_{\mathrm{ij}_l}^\pi = r_k^0\ {}^{21n}x^0$ – the auxiliary relation in R_{ij}^π; 2) There are no others auxiliary relations in R_{ij}^π .

IV. RKX – coding of the semantics of relations and links of cognitive and operational images within the predicative model of the professional activity language

Thus, basis and auxiliary relations are defined (3)–(7) over ${}^1X^0,\ {}^2X^0,\ \ldots, {}^7X^0$ and we can build the graph $G = G_1 \bigcup G_2 \bigcup G_3$. Models M_1 – M_4 (1)–(4) define of components of G, which define the structure of strings of the LDRL. To show this we'll expand the system by the next model:

$$M_5 = <\widehat{X},\ \{\rho_{i,j}^1({}^vx^0)\},\ \{\rho_{i,j}^2({}^dx^0, {}^\varepsilon x^0)\},$$
$$\{\rho_{i,j}^3({}^gx^0, {}^qx^0, {}^hx^0)\},\ |\ i = \{12, \ldots, 18\},$$
$$j = \{1,2,3,20,\ldots,31\} >,$$

where $\rho_{i,j}^1\left({}^vx^0\right) = \begin{cases} t, \text{if } r_{\mathrm{ij}_l}^\pi\, {}^vx^0, \\ f, \text{otherwise} \end{cases}$,

$\rho_{i,j}^2({}^dx^0, {}^\varepsilon x^0) = \begin{cases} t, \text{if } {}^dx^0 r_{\mathrm{ij}_m}^\pi\, {}^\varepsilon x^0, \\ f, \text{otherwise} \end{cases}$,

$\rho_{i,j}^3\left({}^gx^0, {}^qx^0, {}^hx^0\right) = \begin{cases} t, \text{if } {}^gx^0\ r_{\mathrm{ij}_k}^\pi\ {}^qx^0\ \ r_{\mathrm{ij}_k}^\pi\, {}^hx^0, \\ f, \text{thewise} \end{cases}$,

$\widehat{X} = X^0 \bigcup \varnothing$; $r_{\mathrm{ij}_l}^\pi,\ r_{\mathrm{ij}_m}^\pi,\ r_{\mathrm{ij}_k}^\pi \in R_{\mathrm{ij}}^\pi$; ${}^fx^0 \in {}^fX^0 | f = v,\ d,\ \varepsilon,\ g,\ q,\ h.$

If to devote concatenation operator " $\circ$ " over the signature $P_5 = \{\{\rho_{i,j}^1\left({}^vx^0\right)\}, \left\{\rho_{i,j}^2\left({}^dx^0, {}^\varepsilon x^0\right)\right\}, \{\rho_{i,j}^3({}^gx^0, {}^qx^0, {}^hx^0)\}\}$, then the following five axioms are possible. Structures of strings denoting resources, properties, actions and links between resources and actions are defined only by expressions A1 – A5 correspondingly:

A1. $\rho_{11,1}^1\left({}^1x_a^0\right) \circ \rho_{11,20}^2\left({}^1x_a^0, {}^2x_b^0\right) \circ \rho_{11,22}^2\left({}^1x_a^0, {}^3x_c^0\right) \circ \rho_{15,17}^2\left({}^1x_a^0, {}^1x_d^0\right)$,

A2. $\rho_{11,2}^1\left({}^2x_a^0\right) \circ \rho_{11,21}^2\left({}^2x_a^0, {}^1x_b^0\right) \circ \rho_{11,26}^2\left({}^2x_a^0, {}^3x_c^0\right) \circ \rho_{11,26}^2\left({}^2x_a^0, {}^4x_d^0\right) \circ \rho_{11,27}^2\left({}^2x_a^0, {}^5x_c^0\right) \circ \rho_{15,18}^2\left({}^2x_a^0, {}^2x_t^0\right)$,

A3. $\rho_{11,3}^2\left({}^3x_a^0\right) \circ \rho_{11,25}^2\left({}^3x_a^0, {}^2x_b^0\right) \circ \rho_{11,23}^2\left({}^3x_a^0, {}^1x_c^0\right) \circ \rho_{15,19}^2\left({}^3x_a^0, {}^3x_d^0\right)$,

A4. $\rho_{13,1}^2\left({}^1x_a^0, {}^1x_b^0\right) \circ \rho_{13,1}^3\left({}^1x_a^0, {}^1x_b^0, {}^2x_c^0\right)$,

A5. $\rho_{16,3}^2\left({}^3x_a^0, {}^3x_b^0\right) \circ \rho_{14,3}^2\left({}^3x_a^0, {}^3x_b^0\right) \circ \rho_{14,3}^3\left({}^3x_a^0, {}^3x_b^0, {}^2x_c^0\right)$.

A1–A3 are clear. In A4 – A5 dyadic predicates determine qualitative spatial relations, and triadic predicates determine quantitative spatial and temporal. Considered system M_1– M_5 is the semantic, predicative model of expressions of the LPA. From one side, it allows to analyze expressions (M_1– M_4), and, from the other side, it allows to synthesize strings of the LDRL.

Concepts ${}^1x^0$, ${}^3x^0$ devote classes of resources and actions in LDRL whithout taking into account any distinctive features of the latter and correspond, basically, to the words of the LPA. Modelling of resources and actions consists in building of expressions-strings, devoting elements of classes and subclasses of resources and actions, related to phrase of the LPA and consisting of concepts ${}^1x^0,\ {}^2x^0, \ldots,\ {}^7x^0$ and auxiliary relations $r_{\mathrm{ij}_l}^\pi$. Denotation is not only substitution of reality with a sign in LDRL, but also an indication when a resource, action or their links can be associated with any string. Denotation of simple and complex resources, properties and actions in LDRL is implemented at the second and the third levels, correspondingly. Consider the language of the second level as

$$L_2\left(X^0, {}^1R_{\mathrm{ij}}^\pi,\ (,),\ \mathcal{P}^2\right) = \{{}^ix_k\}$$

a set of auxiliary concepts $\{{}^1x_1, {}^1x_2, \ldots,\ {}^1x_\eta, {}^2x_1, {}^2x_2, \ldots,\ {}^2x_\varphi, {}^3x_1, {}^3x_2, \ldots,\ {}^3x_\psi\}$ = X, obtained from $X^0, {}^1R_{\mathrm{ij}}^\pi, (,)$ within rules $\mathcal{P}^2$ and devoting simple resources, properties and actions. Consideration of rules $\mathcal{P}^2 = \{p_1^2, p_2^2, p_3^2\}$, interpreting correspondingly axioms A1–A3, more convenient to start from p_2^2.

Denotations: BF – a basis feature, AF – an additional feature, AC – an auxiliary concept.

Rule p_2^2. 1) If $r_{22_l}^\pi \in {}^1R_{22}^\pi$ and ${}^2x_m^0 \in {}^2X^0$, then $r_{22_l}^\pi\, {}^2x_m^0$ – CO in ${}^2X$; 2) If $r_{\mathrm{ij}_l}^\pi \in {}^1R_{21}^\pi, {}^1R_{23}^\pi, {}^1R_{24}^\pi, {}^1R_{25}^\pi$ and i = 2, j = {1,3,4,5}, ${}^jA = {}^jx_k^0,\ {}^jA = {}^jA\ {}^jx_n^0$, then $({}^2xk^0 r_{\mathrm{ij}_l}^\pi\, {}^jA)$ or $r_{\mathrm{ij}_l}^\pi\left({}^jA\right)$ is AF in ${}^2X$; 3) If ${}^2B$ is BF and ${}^2C$ is AF, then ${}^2B{}^2C$ is AC in ${}^2X$; 4) If ${}^2C$ is BF and ${}^2D$ is AF, then ${}^2C{}^2D$ is AC in ${}^2X$; 5) There are no others AC in ${}^2X$.

Rule p_1^2. 1) If $r_{11_l}^\pi \in {}^1R_{11}^\pi$ and ${}^1x_m^0 \in {}^1X^0$, then $r_{11_l}^\pi\ {}^1x_m^0$ is BF in ${}^1X$; 2) If $r_{\mathrm{ij}_f}^\pi \in {}^1R_{12}^\pi, {}^1R_{13}^\pi$ and i = 1, j = {2,3}, ${}^jA = {}^jx_\pi^0, {}^jA = {}^jx_k,\ {}^jA = {}^jA\ {}^jx_a^0, {}^jA = {}^jA\ {}^jx_b$, then $({}^1xk^0 r_{\mathrm{ij}_f}^\pi\, {}^jA)$ or $r_{\mathrm{ij}_f}^\pi\left({}^jA\right)$ is AF in ${}^1X$; If ${}^1B$ is BF and ${}^1C$ is AF, then ${}^1B{}^1C$ is AC in ${}^1X$; 4) If ${}^jD$ is AC in ${}^2X$ or ${}^3X$, then $\left({}^1x_m^0 r_{\mathrm{ij}_f}^\pi\left({}^jD\right)\right)$ is AF in ${}^1X$; 5) If ${}^1D$ is AC and ${}^1C$ is AF, then ${}^1D{}^1C$ is AC in ${}^1X$; 6) There are no others AC in ${}^1X$.

Rule p_3^2. 1) If $r_{33_l}^\pi \in {}^1R_{33}^\pi$ and ${}^3x_m^0 \in {}^3X^0$, then $r_{33_l}^\pi\ {}^3x_m^0$ is BF in ${}^3X$;
2) If $r_{\mathrm{ij}_f}^\pi \in {}^1R_{31}^\pi, {}^1R_{32}^\pi$ and i = 3, j = {1,2}, ${}^jA = {}^jx_n^0, {}^jA = {}^jx_k,\ {}^jA = {}^jA\ {}^jx_a^0, {}^jA = {}^jA\ {}^jx_b$, then

$({}^3x^0_m r^\pi_{\text{ij}_f}\, {}^jA)$ or $r^\pi_{\text{ij}_f}({}^jA)$ is AF in ${}^3X$; 3) $If\ {}^3B$ is BF and ${}^3C\ is\ AF,\ then\ {}^3B^3C$ is AC in L_2; 4) If ${}^jD$ is AC in ${}^1X$ or${}^2X\ ,then\ \left({}^3x^0_m r^\pi_{\text{ij}_f}\ ({}^jD)\right)$ is AF in ${}^3X$; 5) If ${}^3D$ is AC and ${}^3C$ is AF, then ${}^3D^3C$ is AC in ${}^3X$; 6) There are no others AC in ${}^3X$.

Finally, we can define the auxiliary concept of the L_2. It is AC in ${}^1X$ or AC in ${}^2X$ or AC in ${}^3X$. There are no others AC in L_2.

From the definition of L_2 and rules $\mathcal{P}^2$ we can see, that an auxiliary concept, from one side, is a symbol ${}^1x_k$ and from the other side, it is a string of features. If one establishes a one-to-one correspondence ${}^1x_k$=${}^jD_k$ between ${}^1x_1, {}^1x_2, \ldots\ ,\ {}^1x_\eta$, ${}^2x_1, {}^2x_2, \ldots\ ,\ {}^2x_\varphi, {}^3x_1, {}^3x_2, \ldots\ ,\ {}^3x_\psi$ and strings ${}^jD_k | k = \{1, 2, \ldots, \eta, \ldots, \varphi, \ldots, \psi\ \}$, then ${}^1x_k$ is the defined concept and ${}^jD_k$ is its definition. Such a record, when the derived concept is represented in the unity of the definable and the definition is used in identical transformations.

Consider the language of the third level as

$$L_3\left(X^0, X, {}^5R^\pi_{\text{ij}},\ (,)\ ,\ \mathcal{P}^3\right) = \{{}^ix^n_k\}$$

a set of AC $\{{}^1x^n_1, {}^1x^n_2, \ldots\ ,\ {}^1x^n_\eta, {}^2x^n_1, {}^2x^n_2, \ldots\ ,\ {}^2x^n_\varphi$, ${}^3x^n_1, {}^3x^n_2, \ldots\ ,\ {}^3x^n_\psi\} = X^n$, obtained from $X^0, X, {}^5R^\pi_{\text{ij}},\ (,)$ within the rule $p^3_1 \in \mathcal{P}^3$, common for resources, properties and actions: 1) If $r^\pi_{\text{ij}_l} \in {}^5R^\pi_{\text{ij}}$ and $i = j = \{1{,}3\}$, ${}^iA = {}^ix^0_f$, ${}^iA = {}^ix_k$, ${}^iA = {}^iA\, {}^jx^0_a$, ${}^iA = {}^iA\, {}^jx^0_b$, then $r^\pi_{\text{ij}_l}\left({}^iA\right)$ is a feature in ${}^iX^n$ (generally can be ${}^iA = \varnothing$); 2) If ${}^iC$ is a feature in ${}^iX^n$ and ${}^iD$ is AF in ${}^iX^n$, $then\ {}^iD^iC$ is AC in ${}^iX^n$; 3) There are no others AC in ${}^iX^n$.

Essentially, the rule p^3_1 of the language L_3 sequences auxiliary concepts of L_2 concerning the relation of inclusion (to have in composition or to be a part of) and forms a hierarchical structure. Thus, if ${}^1x^n_k$ are concrete, i.e. each of them responds to the only one resource, then ${}^3x^n_k$ is an abstraction, devoting the plan of actions and defining its properties and a list of features of resources involved for realization. If this list contain a name property, then the plan comes true. This leads to the establishing of relations $r^\pi_{\text{ij}_l} \in {}^3R^\pi_{\text{ij}}$ over resources and it also forms the situation in CO.

Modelling of spatial, productive structure, P- and O-situations is performed at the fourth level of LDRL. Consider the language of the fourth level as

$$L_4\left(X^0, X^n, \{{}^fR^\pi_{\text{ij}}\},\ (,)\ ,\ \mathcal{P}^4\right) = \{{}^i\pi_k\}$$

A set of AC $\{{}^1\pi_1, {}^1\pi_2, \ldots\ ,\ {}^1\pi_\lambda, {}^3\pi_1, {}^3\pi_2, \ldots\ ,\ {}^3\pi_\nu\}$ = Π, obtaining from $X^0, X^n, {}^3R^\pi_{\text{ij}}$, ${}^4R^\pi_{\text{ij}}$,${}^6R^\pi_{\text{ij}}$,$(,)$ within the rule $p^4_1 \in \mathcal{P}^4$: 1) If $r^\pi_{\text{ij}_l} \in {}^3R^\pi_{11}$, ${}^1D_a$ and ${}^1D_m$ are AC in ${}^1X^n$, then $({}^1D_a\ r^\pi_{11_l}\ ({}^1D_m))$ is AC in ${}^1\Pi \subset \Pi$; 2) If $r^\pi_{33_l} \in {}^4R^\pi_{33}$, ${}^6R^\pi_{33}$, ${}^3D_a$ and ${}^3D_m$ are AC in ${}^3X^n$, then $({}^3D_a\ r^\pi_{33_l}\ ({}^3D_m))$ is AC in ${}^3\Pi \subset \Pi$; 3) If ${}^i\pi_k$ and${}^i\pi_m$ is AC in ${}^i\Pi$, then $\pi_k\pi_m$ is AC in ${}^i\Pi$; 4) If $r^\pi_{11_l} \in {}^3R^\pi_{11}$, $r^\pi_{33_l} \in {}^4R^\pi_{33}$, ${}^6R^\pi_{33}$, ${}^3D_k$ and ${}^3D_m$ are AC in ${}^1X^n, {}^3X^n$, then $r^\pi_{11_l}\ {}^1D_k$ and $r^\pi_{33_l}\ {}^3D_m$ are features in ${}^1\Pi$, ${}^3\Pi$4; 5) If ${}^iA$ and ${}^iB$ are features, then ${}^iA^iB$ is also a feature in ${}^i\Pi$; 6) If ${}^iD_b$ is AC in ${}^iX^n$ and ${}^iA$ is a feature in ${}^i\Pi$, then $({}^iD_b\ \ ({}^iA))$ is AC in ${}^i\Pi$; 7) There are no others AC in Π.

Obvious that ${}^1\Pi$ can be described as spatial structure of CO ${}^1\Pi_C \subset^1 \Pi$. The current P-situation ${}^1\Pi_P \subset^1 \Pi$ similarly in ${}^3\Pi$ productive structure corresponds to ${}^3\Pi_C \subset^3 \Pi$ and O-situations to ${}^3\Pi_O \subset^3 \Pi$.

The formalism of the fifth level language allows to define the concept of a "cognitive image of the state of the control object". Consider the fifth level of LDRL as

$$L_5\left(\diamond,\ \mathcal{P}^5\right) = \{s_k\}$$

a set of AC – CI of states $\{s_1, s_2, \ldots, s_\omega\}$ = S, formally deduced within the rule $p^5_1 \in \mathcal{P}^5$. Before giving the inductive definition of CI of a state, let's consider the substitution rule: if ${}^1x_m$=${}^1D_m$, ${}^3x_l$=${}^3D_l$ are correspondingly auxiliary concepts in ${}^1X^n$, ${}^3X^n$ and ${}^1x^0_f \in {}^1D_m$, ${}^3D_l$, then the changing of ${}^1x^0_f$ in ${}^3\ D_l$ to ${}^1x_m$ or${}^1D_m$ is called substitution. After substitution the plan of actions ${}^3x_l$=${}^3D_l$ becomes true ${}^3\widehat{x}_l = {}^3\ \widehat{D}_l$.

Rule p^5_1 : *1) If* ${}^1\Pi_P \subset^1 \Pi$, ${}^3\Pi_O \subset^3 \Pi$ and $\{{}^1x_P = {}^1D_P\}$, $\{{}^3x_O = {}^3D_O\}$ are subsets of *auxiliary concepts* in ${}^1X^n$ and ${}^3X^n$, correspondingly, and ${}^{212}x^n = {}^{212}D | (r^\pi_{22_3}\, {}^{212}x^0_\mu\ r^\pi_{25_1}({}^5x^0_\omega)) \in {}^{212}D$, ${}^{212}x^0_\mu$ - *mode*, ${}^5x^0_\omega \in$\{*normal*, *emergency*\} are AC in ${}^{212}X^0$, then ${}^3\widehat{\Pi}_O{}^1\Pi_P{}^{212}x^n_\mu$ (where $\{{}^3\widehat{x}_O = {}^3\ \widehat{D}_O\}$ is obtained as a result of substitution) is AC in *S*; 2) There are no others AC in *S*.

Consider the example of the expression $s \in S$. Suppose that at time *t* two actions are simultaneously performed
${}^3x^1_1 = r^\pi_{33_1}\, {}^3x^0_1 r^\pi_{31_1}\left({}^1x^0_1\right) r^\pi_{31_2}\left({}^1x^0_2\right) r^\pi_{31_3}\left({}^2x^0_3\right)$,
${}^3x^1_2 = r^\pi_{33_1}\, {}^3x^0_2 r^\pi_{31_1}\left({}^1x^0_4\right) r^\pi_{31_2}\left({}^1x^0_5\right)$, then ${}^3\pi_O = \left({}^3x^1_1\ r^\pi_{33_2}\, {}^3x^1_2\right)$.

If there were involved resources for realization

$$\begin{aligned}{}^1x^1_1 = r^\pi_{11_1}\, {}^1x^0_1 r^\pi_{12_1}(r^\pi_{22_1}\, {}^{21^3}x^0_1 r^\pi_{25_1}({}^5x^0_1))\\ r^\pi_{12_1}(r^\pi_{22_2}\, {}^{21^1}x^0_1 r^\pi_{24_1}\left({}^4x^0_1\right) r^\pi_{25_1}({}^5x^0_2)),\end{aligned}$$

$$\begin{aligned}{}^1x^1_2 = r^\pi_{11_1}\, {}^1x^0_2 r^\pi_{12_1}(r^\pi_{22_1}\, {}^{21^3}x^0_1 r^\pi_{25_1}({}^5x^0_3))\\ r^\pi_{12_1}(r^\pi_{22_2}\, {}^{21^1}x^0_1 r^\pi_{24_1}\left({}^4x^0_1\right) r^\pi_{25_1}({}^5x^0_4)),\end{aligned}$$

$$
\begin{aligned}
{}^{1}x_{3}^{1} &= r_{11_1}^{\pi}\,{}^{1}x_{3}^{0} r_{12_1}^{\pi} (r_{22_1}^{\pi}\,{}^{21^3}x_{1}^{0} r_{25_1}^{\pi} ({}^{5}x_{5}^{0}))\\
&\quad r_{12_1}^{\pi} (r_{22_3}^{\pi}\,{}^{21^2}x_{1}^{0} r_{25_1}^{\pi} ({}^{5}x_{6}^{0})),
\end{aligned}
$$

$$
\begin{aligned}
{}^{1}x_{4}^{1} &= r_{11_1}^{\pi}\,{}^{1}x_{4}^{0} r_{12_1}^{\pi} (r_{22_1}^{\pi}\,{}^{21^3}x_{1}^{0} r_{25_1}^{\pi} ({}^{5}x_{7}^{0}))\\
&\quad r_{12_1}^{\pi} (r_{22_3}^{\pi}\,{}^{21^2}x_{2}^{0} r_{25_1}^{\pi} ({}^{5}x_{8}^{0})),
\end{aligned}
$$

$$
\begin{aligned}
{}^{1}x_{5}^{1} &= r_{11_1}^{\pi}\,{}^{1}x_{4}^{0} r_{12_1}^{\pi} (r_{22_1}^{\pi}\,{}^{213}x_{2}^{0} r_{25_1}^{\pi} ({}^{5}x_{9}^{0}))\\
&\quad r_{12_1}^{\pi} (r_{22_3}^{\pi}\,{}^{212}x_{3}^{0} r_{25_1}^{\pi} ({}^{5}x_{10}^{0})),
\end{aligned}
$$

then as a result of substitution we'll have ${}^{3}\widehat{x}_{1}^{1} = r_{33_1}^{\pi}\,{}^{3}x_{1}^{0} r_{31_1}^{\pi} \left({}^{1}x_{1}^{1}\right) r_{31_2}^{\pi} \left({}^{1}x_{2}^{1}\right) r_{31_3}^{\pi} \left({}^{2}x_{3}^{1}\right)$,

${}^{3}\widehat{x}_{2}^{1} = r_{33_1}^{\pi}\,{}^{3}x_{2}^{0} r_{31_1}^{\pi} \left({}^{1}x_{4}^{1}\right) r_{31_2}^{\pi} \left({}^{1}x_{5}^{1}\right)$, ${}^{3}\widehat{\pi}_{O}= {}^{3}\widehat{x}_{1}^{1} r_{33_2}^{\pi}\,{}^{3}\widehat{x}_{2}^{1}$.

The obtained O-situation forms some P-situation ${}^{1}\Pi_{\ P} = ({}^{1}x_{1}^{1} r_{11_2}^{\pi}\,{}^{1}x_{2}^{1})\ ({}^{1}x_{3}^{1} r_{11_3}^{\pi}\,{}^{1}x_{4}^{1})\ ({}^{1}x_{3}^{1} r_{11_4}^{\pi}\,{}^{1}x_{5}^{1})$. This allows to compute *value* ${}^{5}x_{\omega}^{0}$ of characteristic *mode* ${}^{212}x_{\mu}^{\pi}$. Then CI of the state of CO is $s\left(t\right) = {}^{3}\widehat{\Pi}_{O}{}^{1}\Pi_{\ P}({}^{212}x_{\mu}^{0} r_{25_1}^{\pi} ({}^{5}x_{\omega}^{0})) = {}^{3}\widehat{\pi}_{O}{}^{1}\Pi_{\ P} = ({}^{3}\widehat{x}_{1}^{1} r_{33_2}^{\pi}\,{}^{3}\widehat{x}_{2}^{1})\ ({}^{1}x_{1}^{1} r_{11_2}^{\pi}\,{}^{1}x_{2}^{1})\ ({}^{1}x_{3}^{1} r_{11_3}^{\pi}\,{}^{1}x_{4}^{1})\ ({}^{1}x_{3}^{1} r_{11_4}^{\pi}\,{}^{1}x_{5}^{1})({}^{212}x_{\mu}^{0} r_{25_1}^{\pi} ({}^{5}x_{\omega}^{0}))$.

So, an expression about CI of a state includes: ${}^{3}\widehat{\Pi}_{O}$– AC "O-situation" over relations "action-action", ${}^{1}\Pi_{\ P}$ – AC "P-situation" over relations "resource-resource", $({}^{212}x_{\mu}^{0} r_{25_1}^{\pi} ({}^{5}x_{\omega}^{0}))$ – characteristic *mode* ${}^{212}x_{\mu}^{0}$ with computed *value* ${}^{5}x_{\omega}^{0}$ over relations "resource-property" for resources of P-situation. If we take into account, that ${}^{3}\pi_{0}$ is considered within the context of productive structure, then knowing $s\left(t_{O}\right)$ we can predict the behavior of CO in the next moment of the time t_1, i.e. $s\left(t_0\right) \longrightarrow s\left(t_1\right)$, where $t_1 > t_0$.

V. The grammar of identical transformations and functional deformation of a co-cognitive image of a state

R1. If AC in ${}^{i}X$ or ${}^{i}X^{n}$ has AF of view $\left({}^{i}x_{m}^{0} r_{\mathrm{ij}_l}^{\pi}\,{}^{j}A\right) | i \neq j,$ then it can be converted to the view $r_{\mathrm{ij}_l}^{\pi}\left({}^{j}A\right)$ and on the contrary, if AC contains BF $r_{\mathrm{ij}_l}^{\pi}\,{}^{i}x_{m}^{0}$ and AF $r_{\mathrm{ij}_k}^{\pi}\left({}^{j}A\right)$, then the last one can be transformed as follows $\left({}^{i}x_{m}^{0} r_{\mathrm{ij}_k}^{\pi}\,{}^{j}A\right)$. For example:

$$
\begin{aligned}
{}^{1}x_{1}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{1}^{0}({}^{1}x_{1}^{0} r_{12_1}^{\pi} (r_{22_1}^{\pi}\,{}^{213}x_{1}^{0}({}^{213}x_{1}^{0}\\
\left({}^{213}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{1}^{0}\right)))
\end{aligned}
$$

$$
({}^{1}x_{1}^{0} r_{12_1}^{\pi} (r_{22_2}^{\pi}\,{}^{211}x_{1}^{0} \left({}^{211}x_{1}^{0} r_{24_1}^{\pi}\,{}^{4}x_{1}^{0}\right) \left({}^{211}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{2}^{0}\right)))
$$

it can be replaced by

$$
\begin{aligned}
{}^{1}x_{1}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{1}^{0} r_{12_1}^{\pi}((r_{22_1}^{\pi}\,{}^{21^3}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{1}^{0}) r_{12_1}^{\pi}\\
\left(r_{22_2}^{\pi}\,{}^{211}x_{1}^{0} r_{24_1}^{\pi}\,{}^{4}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{2}^{0}\right)
\end{aligned}
$$

. R2. If AF of the view $r_{\mathrm{ij}_l}^{\pi}\left({}^{j}A\right)$ contain the same relation, then it can be put out of the bracket (the reverse is also true):

$$
\begin{aligned}
{}^{1}x_{1}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{1}^{0} r_{12_1}^{\pi}(\left(r_{22_1}^{\pi}\,{}^{21^3}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{1}^{0}\right)\\
\left(r_{22_2}^{\pi}\,{}^{21^1}x_{1}^{0} r_{24_1}^{\pi}\,{}^{4}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{2}^{0}\right)).
\end{aligned}
$$

R3. If in AC ${}^{i}x_{m}^{n}$ there is a string of defining concepts $({}^{j}D)$ and $i \neq j$, then ${}^{j}x_{k} = \left({}^{j}D\right)$ can be imposed and $\left({}^{j}D\right)$ can be replaced by ${}^{j}x_{k}$. The reverse is also true. For example:

${}^{2}x_{1}^{1} = \left(r_{22_1}^{\pi}\,{}^{213}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{1}^{0}\right)$. If we impose ${}^{2}x_{2}^{1} = \left(r_{22_2}^{\pi}\,{}^{211}x_{1}^{0} r_{24_1}^{\pi}\,{}^{4}x_{1}^{0} r_{25_1}^{\pi}\,{}^{5}x_{2}^{0}\right)$,

then ${}^{1}x_{1}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{1}^{0} r_{12_1}^{\pi} \left({}^{2}x_{1}^{1}{}^{2}x_{2}^{1}\right)$.

R4. If in ${}^{i}x_{m}^{\pi}$ there is subsidiary feature $r_{\mathrm{ij}_l}^{\pi}\left({}^{j}A\right)$, where $r_{\mathrm{ij}_l}^{\pi} \in {}^{5}R_{\mathrm{ij}}^{\pi}$ and in ${}^{j}A$ there is a string of defining concepts ${}^{j}D$, then we can impose ${}^{j}x_{k} ={}^{j} D$ and replace ${}^{j}D$ by ${}^{j}x_{k}$, and also replace ${}^{j}x_{m}^{n}$ by ${}^{j}x_{m}^{n+1}$. The reverse is also true. For example:

$$
\begin{aligned}
{}^{1}x_{1}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{1}^{0} r_{12_1}^{\pi}({}^{2}x_{1}^{1}{}^{2}x_{2}^{1}{}^{2}x_{3}^{1})\\
r_{11_2}^{\pi}(r_{11_1}^{\pi}\,{}^{1}x_{2}^{0} r_{12_1}^{\pi}({}^{2}x_{3}^{1}{}^{2}x_{4}^{1}{}^{2}x_{5}^{1})\\
r_{11_1}^{\pi}\,{}^{1}x_{3}^{0} r_{12_1}^{\pi}({}^{2}x_{6}^{1}{}^{2}x_{7}^{1}),
\end{aligned}
$$

Impose

$$
\begin{aligned}
&{}^{1}x_{1}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{2}^{0} r_{12_1}^{\pi}({}^{2}x_{3}^{1}{}^{2}x_{4}^{1}{}^{2}x_{7}^{1}),\\
&{}^{1}x_{2}^{1} = r_{11_1}^{\pi}\,{}^{1}x_{3}^{0} r_{12_1}^{\pi}({}^{2}x_{6}^{1}{}^{2}x_{7}^{1}),\\
&{}^{1}x_{1}^{2} = r_{11_1}^{\pi}\,{}^{1}x_{1}^{0} r_{12_1}^{\pi}({}^{2}x_{1}^{1}{}^{2}x_{2}^{1}{}^{2}x_{3}^{1}) r_{11_2}^{\pi}({}^{1}x_{1}^{1}{}^{1}x_{2}^{1})
\end{aligned}
$$

R5. If $r_{\mathrm{ij}_l}^{\pi}$ is symmetric, i.e. $r_{\mathrm{ij}_l}^{\pi} = \frac{1}{r_{\mathrm{ij}_l}^{\pi},}$ then $({}^{1}D_{k}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ ({}^{j}D_{m})) \in {}^{i}\Pi$ can be replaced by $({}^{1}D_{m}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ ({}^{j}D_{k}))$. If $r_{\mathrm{ij}_l}^{\pi}$ is asymmetric, i.e. from two rations $({}^{1}D_{k}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ {}^{j}D_{m})$, $({}^{1}D_{m}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ {}^{j}D_{k})$ at least one is not completed and $\overline{r}_{\mathrm{ij}_l}^{\pi} = \frac{1}{r_{\mathrm{ij}_l}^{\pi},}$ then $({}^{1}D_{m}\ \ \overline{r}_{\mathrm{ij}_l}^{\pi}\ \ ({}^{j}D_{k})) \in {}^{i}\Pi$. For example: if we have $({}^{1}x_{1}^{1} r_{11_5}^{\pi}({}^{1}x_{2}^{1}))\left({}^{1}x_{3}^{1} r_{11_6}^{\pi}({}^{1}x_{4}^{1}\right) \in {}^{1}\Pi$ and $r_{11_5}^{\pi}$ is symmetric and $r_{11_6}^{\pi}$ is asymmetric, then $({}^{1}x_{2}^{1} r_{11_5}^{\pi}({}^{1}x_{1}^{1}))\left({}^{1}x_{4}^{1}\overline{r}_{11_6}^{\pi}({}^{1}x_{3}^{1}\right) \in {}^{1}\Pi$.

R6. If $({}^{i}D_{n}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ ({}^{i}D_{m}))$, $({}^{i}D_{n}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ ({}^{i}D_{f})) \in {}^{i}\Pi,$ then they can be replaced by

$({}^{i}D_{n}\ \ r_{\mathrm{ij}_l}^{\pi}\ \ ({}^{i}D_{m}{}^{i}D_{f})) \in {}^{i}\Pi.$ The reverse is also true. For example: if we have $({}^{1}x_{1}^{1} r_{11_5}^{\pi}({}^{1}x_{2}^{1}))\left({}^{1}x_{1}^{1} r_{11_5}^{\pi}({}^{1}x_{3}^{1}\right)\ \left({}^{1}x_{1}^{1} r_{11_5}^{\pi}({}^{1}x_{4}^{1}\right))$, it can be replaced by$({}^{1}x_{1}^{1} r_{11_5}^{\pi}({}^{1}x_{2}^{1}{}^{1}x_{3}^{1}\ {}^{1}x_{4}^{1}))$.

R7. If left poles of AC in ${}^{i}\Pi$ identical, then they can be put out of the bracket. The reverse is also true. For example: $({}^{1}x_{1}^{1} r_{11_5}^{\pi}({}^{1}x_{2}^{1}))({}^{1}x_{1}^{1} r_{11_6}^{\pi}({}^{1}x_{3}^{1}))$ can be replaced by $({}^{1}x_{1}^{1}(r_{11_5}^{\pi}({}^{1}x_{2}^{1}) r_{11_6}^{\pi}({}^{1}x_{3}^{1}))$. Putting out of right poles comes down to the R5,7.

The following group of rules is used for the functional deformation of the semantics of the entropy, cognitive image of the CO state into the regulatory, pragmatic information of a concise and localized operational image. The moment of display to the human-operator of the operational image is calculated by the "resource-property" relations, based on the establishment of a "threshold" of operation. If the threshold value is reached, then the cognitive image is deformed as a sign of the operational image. Since according to relations “property-resource” we know those resources in ${}^{1}\Pi_{\ P}$, properties of which are out-of-range (emergency resources; let it can be only one), then first of all, ${}^{1}x_{k}^{n}{=}{}^{j}D_{k}$ is reduced by extraction from ${}^{j}D_{k}$ irrelevant features from AF (by the rule R8). After that the spatial structure ${}^{1}\Pi_{\ C}$ of CO is reduced to a fragment, containing information only about the position of emergency resource in its real nearest environment. In addition, because an emergency resource performs a particular role by relations “resource-action” within an action in O-situation${}^{3}\Pi_{\ O}$, it is connected by relations “resource-resource” with other resources from ${}^{1}\Pi_{\ P}$. That’s why a sign ${}^{1}\Pi_{\ P}$ is reduced within the rule R9.

R8. If for an operation ${}^{3}x_{f}^{n}$ one involve resources ${}^{1}\widetilde{x}_{1}^{n}, {}^{1}\widetilde{x}_{2}^{n}, \ldots, {}^{1}\widetilde{x}_{m}^{n},$ then from ${}^{1}\Pi_{\ c}$ a fragment ${}^{1}\ \widetilde{\Pi}_{c}$ can be cut. The last contains links only of ${}^{1}\widetilde{x}_{1}^{n}, {}^{1}\widetilde{x}_{2}^{n}, \ldots, {}^{1}\widetilde{x}_{m}^{n},$ between each other. All other links ${}^{1}\Pi_{\ c}$ are not significant. For example, if ${}^{1}\ \widetilde{\Pi}_{c}$ = $\left({}^{1}x_{1}^{1}r_{11_5}^{\pi}\,{}^{1}x_{2}^{1}\right)\left({}^{1}x_{3}^{1}r_{11_5}^{\pi}\,{}^{1}x_{4}^{1}\right),$ then taking into account the previous example for ${}^{3}x_{1}^{1}$ we get ${}^{1}\ \widetilde{\Pi}_{c} = \left({}^{1}x_{1}^{1}r_{11_5}^{\pi}\,{}^{1}x_{2}^{1}\right).$ Because resources ${}^{1}x_{1}^{n}, {}^{1}x_{2}^{n}, \ldots,\ {}^{1}x_{m}^{n}$ form a P-situation ${}^{1}\Pi_{\ p}$, then exactly this P-situation defines the cut fragment ${}^{1}\ \widetilde{\Pi}_{c}$.

R9. Every AC of a cognitive image $s\left(t\right)\ =^{3} \widehat{\Pi}_{O}{}^{1}\Pi_{\ P}\ \left({}^{212}x_{\mu}^{0}r_{25_1}^{\pi}\left({}^{5}x_{\omega}^{0}\right)\right)$ can be distorted into AC of an operational image ${}^{3}\widetilde{\widehat{\Pi}}_{o}{}^{1}\ \widetilde{\Pi}_{p}({}^{212}x_{\mu}^{0}r_{25_1}^{\pi}({}^{5}x_{\omega}^{0})),$ if $\{{}^{1}D_{\ p}\}$ has been fragmented within the rule R8.

VI. Conclusion

Cognitive hybrid intellectual systems are proposed for computer simulation of cognitive formations and enhancing human intelligence in operational work, by supplementing the natural abilities of the operator (in the work with operational-technological information) with software and hardware that expand the human mental processes. In relation to them, a language has been developed for describing relations and links for predictive coding of word-verbal knowledge about resources, properties and actions of personnel. There also has been developed a grammar of identical transformations and functional deformation of the semantics of the entropy, cognitive image of the state of a control object into regulatory, pragmatic information of concise and localized operational image.

References

[1] Pospelov D. A., Pushkin V. N. Myshlenie i avtomaty [Thinking and Automata], Moskow, Izdatel’stvo «Sovetskoe radio», 1972. 224 p.

[2] Kolesnikov A.V., Ponomarev V.F. Klassifikatsiya ponyatii modeli staticheskikh otnoshenii [Classification of concepts of the model of static relations]. Sbornik “Voprosy kibernetiki” [Collection "Questions of Cybernetics"], 1974, Vol. 13., pp.154–159.

[3] Kolesnikov A.V. Gibridnye intellektual’nye sistemy. Teoriya i tekhnologiya razrabotki. [Hybrid artificial systems. Theory and development technology], Saint-Petersburg: SPbGTU, 2001.-711 p.

[4] Kolesnikov A.V., Kirikov I.A., Listopad S.V., etc. Reshenie slozhnykh zadach kommivoyazhera metodami funktsional’nykh gibridnykh intellektual’nykh system [Solving of the complex tasks of a travelling salesman using functional hybrid intellectual systems]. Moskow, IPI RAN, 2011. 295 p.

[5] Oshanin D.A. Predmetnoe deistvie i operativnyi obraz: Izbrannye psikhologicheskie trudy [The objective act and operational pattern: The elected psychological works], Voronezh, SPA “MODEK” Publs., 1999. 512 p.

[6] Rishar Zh.F. Mental’naya aktivnost’. Ponimanie, rassuzhdenie, nakhozhdenie reshenii [Mental activity. Understanding, reasoning, finding solutions], Moskow, Izdatel’stvo “Institut psikhologii RAN”, 1998. 232 p.

[7] Mikeshina L.A. Filosofiya nauki: Sovremennaya epistemologiya. Nauchnoe znanie v dinamike kul’tury. Metodologiya nauchnogo issledovaniya: ucheb. Posobie [Philosophy of Science: Modern Epistemology. Scientific knowledge in the dynamics of culture. Research Methodology: Handbook], Moskow, Izdatel’stvo “Progress-Traditsiya”: MPSI : Flinta, 2005. 464 p.

[8] English terms dictionary. 2015. Predicate. Available at: https://terms_en.academic.ru/32514/predicate (accessed 2020, Jan)

[9] Osipov G.S. Priobretenie znanii intellektual’nymi sistemami. Osnovy teorii i tekhnologii [Knowledge acquisition by intelligent systems. Fundamentals of theory and technology]. Moskow, Izdatel’stvo Nauka. Fizmatlit, 1997. 112 p.

[10] Kolychev P.M. Kategoriya sootnosheniya [Correlation category], Saint Petersburg, Izdatel’stvo Sankt-Peterburskogo Universiteta, 2006. 259 p.

Предикативная модель когнитивного и оперативного образа объекта оперативно-технологического управления (ОТУ) в системах с высокой динамикой

Колесников А. В., Румовская С. Б., Ясинский Э. В.

Для компьютерной имитации когнитивных образований и усиления человеческого интеллекта в оперативной работе, путем дополнения естественных способности оператора к работе с оперативно-технологической информацией, программно-аппаратными средствами, расширяющими мыслительные процессы человека предложены когнитивные гибридные интеллектуальные системы. Применительно к ним разработан язык описания отношений и связей для предикативного кодирования словесно-вербальных знаний о ресурсах, свойствах и действиях персонала, грамматика тождественных преобразований и функциональной деформации семантики энтропийного, когнитивного образа состояния объекта управления в регуляритивную, прагматическую информацию лаконичного и локализованного оперативного образа.

Received 01.02.2020

Agent Interaction Protocol of Hybrid Intelligent Multi-Agent System of Heterogeneous Thinking

Listopad Sergey
Kaliningrad Branch of the Federal Research Center
“Computer Science and Control“ of the Russian Academy of Sciences
Kaliningrad, Russia
ser-list-post@yandex.ru

***Abstract*—The paper aims at modeling the collective decision-making processes in dynamic environments. The need for such modeling is caused by the irrelevance of traditional abstract-mathematical methods to such problems, as well as the limited time in a dynamic environment to organize a comprehensive discussion of the problem and make decisions by a real team of experts. To provide information support for solving problems, a new class of intelligent systems is proposed that simulate teamwork under the guidance of a facilitator, namely hybrid intelligent multi-agent systems of heterogeneous thinking. The paper proposes the agent interaction protocol of the hybrid intelligent multi-agent system of heterogeneous thinking for solving the problem of regional power grid restoration after large-scale accidents.**

***Keywords*—heterogeneous thinking, hybrid intelligent multi-agent system, agent interaction protocol**

I. Introduction

Problems in dynamic environments are characterized, firstly by a whole set of non-factors (in the sense of A.S. Narinyani [1]) that do not allow to use traditional abstract mathematical models, and secondly by the limited time for making decisions that make it difficult for traditional collective problem-solving in horizontal structures that integrate heterogeneous knowledge, such as “round table”, staff meeting, operational meeting, etc. In this regard, computer simulation of collective problem solving processes using a new class of intelligent systems, namely hybrid intelligent multiagent systems of heterogeneous thinking (HIMSHT), is relevant. They combine the apparatus of multi-agent systems in the sense of V.B. Tarasov [2], and methods of heterogeneous thinking: divergent, non-judgmental perception of the problem situation and convergent thinking of S. Kaner [3], parallel thinking, designing solutions from the field parallel to the existing capabilities of E. de Bono [4]. This make it possible to model the process of solving problems under insufficiently identified conditions and with an uncertain purpose. The paper proposes the agent interaction protocol of the HIMSHT for solving the problem of regional power grid restoration after large-scale accidents.

II. The problem of regional power grid restoration after large-scale accidents

Planning of the power grid (PG) restoration is a combinatorial problem that requires extensive knowledge, includes many restrictions and conditions on which operator evaluations are necessary, which further complicates its comprehensive solution [5]. Three main features make this problem interesting for modern planners: partial observability, the dimension of the state space, which makes a complete enumeration of states absolutely impossible, the consequences of actions are difficult to model [6].

Based on works [6]–[9] devoted to the planning of the distribution power grid restoration after accidents the problem of power grid restoration (PPGR) was formulated. It’s in the development of a plan consisting of a sequence of switching events, the order of repair crews’ trips to perform switching and rehabilitation. Baseline data for PPGR are elements of the power grid and following sets:

- incidence relations between elements of the power grid;
- locations;
- routes between locations;
- repair crews;
- vehicles;
- resources for restoration of the power grid;
- actions for restoration of the power grid.

Optimality criteria of the plan are following:

- minimizing the time of disconnection of the priority load;
- maximizing the total recovered load power;
- maximizing the reliability of the power system (resistance to subsequent accidents).

The following restrictions apply to the plan:

- the preservation of the radial network structure of the powered lines;
- for each line, the total value of loads that are fed from a source of distributed generation through it should not exceed its carrying capacity;
- compliance with the balance of active and reactive power;

- voltage and frequency values must be within acceptable limits;
- consumers not affected by the initial outage should not be turned off as a result of switching;
- restoration and switching must be carried out by repair crews with appropriate admission if the necessary resources are available in their vehicle;
- vehicle capacity;
- working time of crews;
- vehicles must return to base;
- when forcibly dividing the grid to islands, the power lines between the islands must have synchronization equipment for the subsequent merging of the islands.

In accordance with the problem-structural methodology of A.V. Kolesnikov [10] the problem is reduced to decomposition of the following tasks (Fig. 1):

- localization of the accident site, taking into account the possibility of sensors' failure, which requires relay protection and automation engineer's knowledge;
- operational and short-term forecasting of active and reactive power consumed by customers after connecting to the network, for solving which the engineer's in analyzing and forecasting energy consumption modes knowledge is required;
- assessment of requirements for restoration actions of the power grid elements, for solving which power equipment repair engineer's knowledge is relevant;
- building routes for field crews to perform switching and repairing damaged equipment, which needs electrical network area head's knowledge;
- determining the switching order, for solving which the engineer's on power operational modes knowledge is needed;
- recovery plan development that is the task of coordinating intermediate solutions and integrating private solutions of tasks, for solving which the regional operational and technological dispatcher's knowledge is required.

To solve the PPGR it is proposed to model collective decision making by the operational personnel of the energy supplying organization, power engineers, logisticians, and labor protection specialists using the HIMSHT methods.

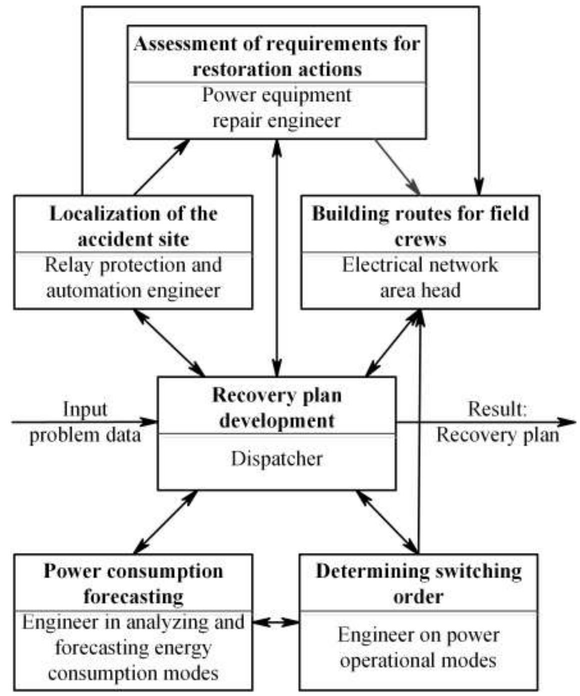


Figure 1. Decomposition of the power grid restoration problem.

III. Formal model of the hybrid intelligent multi-agent system of heterogeneous thinking

Formally HIMSHT is defined as follows:

$$himsht = < AG^*, env, INT^*, ORG, \{ht\} >,$$
$$act_{himsht} = \left(\bigcup_{ag \in AG^*} act_{ag} \right) \cup$$
$$\cup act_{dmsa} \cup act_{htmc} \cup act_{col},$$
$$act_{ag} = (MET_{ag}, IT_{ag}), \left| \bigcup_{ag \in AG^*} IT_{ag} \right| \geqslant 2,$$

where $AG^* = \{ag_1, ..., ag_n, ag^{dm}, ag^{fc}\}$ is the set of agents, including expert agents (EA) ag_i, $i \in \mathbb{N}$, $1 \leqslant i \leqslant n$, decision-making agent (DMA) ag^{dm}, and facilitator agent (FA) ag^{fc}; n is the number of EA; env is the conceptual model of the external environment of HIMSHT; $INT^* = \{prot_{gm}, lang, ont, dmscl\}$ are the elements for structuring of agent interactions: $prot_{gm}$ is the interaction protocol, allowing to organize their collective heterogeneous thinking; $lang$ is the message language; ont is the domain model; $dmscl$ is the classifier of collective solving problem situations, identifying the stages of this process (Fig. 2); ORG is the set of HIMSHT architectures; $\{ht\}$ is the set of conceptual models of macro-level processes in the HIMSHT: ht is the model of the collective problem solving process with heterogeneous thinking methods, that is the "diamond of participatory decision-making" model by S. Kaner, L. Lind, C. Toldi, S. Fisk and D. Berger (Fig. 2) [3]; act_{himsht} is the function of the HIMSHT as a whole;

act_{ag} is the function of EA from the set AG^*; act_{dmsa} is the FA's function "analysis of the collective problem solving situation", which provides identification of the heterogeneous thinking process stage based on the private solutions offered by the EA, the intensity of the conflict between the EA and the previous stage of the problem solving process; act_{htmc} is FA's function "choice of heterogeneous thinking method", which is implemented using a fuzzy knowledge base about the effectiveness of heterogeneous thinking methods depending on the characteristics of the problem, the stage of its solution process, and the current situation of the solution in the HIMSHT; $act_{col} =< met_{ma}, it_{ma} >$ is the collective dynamically constructed function of HIMSHT with multi-agent method met_{ma} and intelligent technology it_{ma}; $met_{ag} \in MET_{ag}$ is the problem solving method; $it_{ag} \in IT_{ag}$ is the intelligent technology, with which the method met_{ag} is implemented.

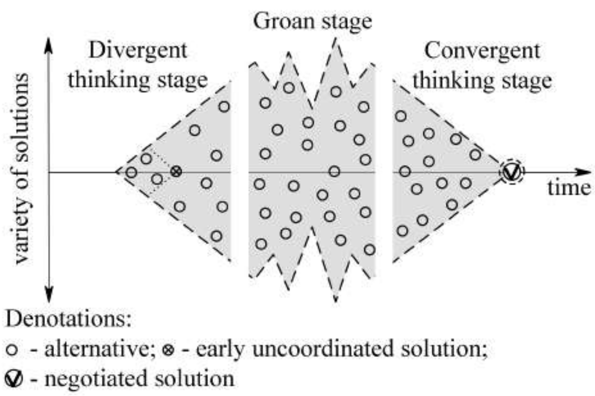


Figure 2. The "diamond of participatory decision-making" model by S. Kaner et al.

According to the model presented in Fig. 2, the problem solving process in HIMSHT consist of three stages: divergent, groan and convergent. At the stage of divergent thinking, expert agents generate a variety of solutions to the problem, and the facilitating agent stimulates their production by appropriate methods [11]. If there are no contradictions, that is, the problem has an obvious solution, the process is completed. Otherwise, agents of the HIMSHT conflict with knowledge, belief, opinions, that is, they participate in a kind of cognitive conflicts [11], [12]. Conflict is a distinctive feature of the groan stage, which allows the facilitator agent to take measures to bring agents' points of view together. At the convergent thinking stage agents will jointly reformulate and refine the solutions until they receive a collective decision relevant to the diversity of the HIMSHT expert models.

The functional structure of HIMSHT for solving problem of power grid restoration after large-scale accidents is presented in [12]. Consider the agent interaction protocol of HIMSHT in the process of collective heterogeneous thinking.

IV. Heterogeneous thinking protocol

The main purpose of development of agent interaction protocol of hybrid intelligent multi-agent system is to encapsulate allowed interactions. It defines the schemes (distributed algorithm) of information and knowledge exchange, coordination of agents during problem solving [13]. Such protocol specification unambiguously determines whether a specific implementation of agent interaction satisfies the specified protocol and whether specific agent is compatible with HIMSHT [14].

The scheme of the hybrid intelligent multi-agent system operation by the heterogeneous thinking protocol, based on the theory of speech acts, is presented in Fig. 3.

As shown in Fig. 3, the standard speech act protocol [15] is extended with the following message types: "request-ch-tt", "commit-ch-tt", "request-start-ps", "request-stop-ps", "request-task", "report-decision", used for organization of interaction of FA, DMA and EA. EAs interact with each other and with other agents with speech act protocol [15].

According to proposed protocol, the process of collective heterogeneous thinking begins with sending of "request-ch-tt" type message by FA to DMA and EA, the body of which indicates the method of heterogeneous thinking used at this stage of the HIMSHT operation. FA suspends its work in anticipation of responses, confirmations from the EA and the DMA. DMA and EA having received the "request-ch-tt" message from the FA, select the appropriate algorithm and go to the waiting mode for the signal to start solving the problem in accordance with the given algorithm. To confirm receipt of the message and readiness for operation according to the established algorithm DMA and EA send to FA a confirmation-commit "commit-ch-tt". When FA have got confirmations from the DMA and all EAs, it sends a "request-start-ps" message to the DMA, indicating that all agents have switched to the appropriate method of heterogeneous thinking, and the system is ready for further work. After that FA awaits solutions from the EA.

When DMA receives the "request-start-ps" message, it generates and sends EA tasks using a "request-task" type message, in the body of which the initial data of the task is also described. Then it proceeds to collecting solutions derived from the EA and works with them in accordance with the previously established heterogeneous thinking algorithm. EA after receiving a task from the DMA begin to solve tasks in accordance with the established algorithm of heterogeneous thinking. Depending on the algorithm, they can generate one or several solutions to the problem. EA sends all received alternative solutions in the body of a "report-solution" message to DMA and FA simultaneously.

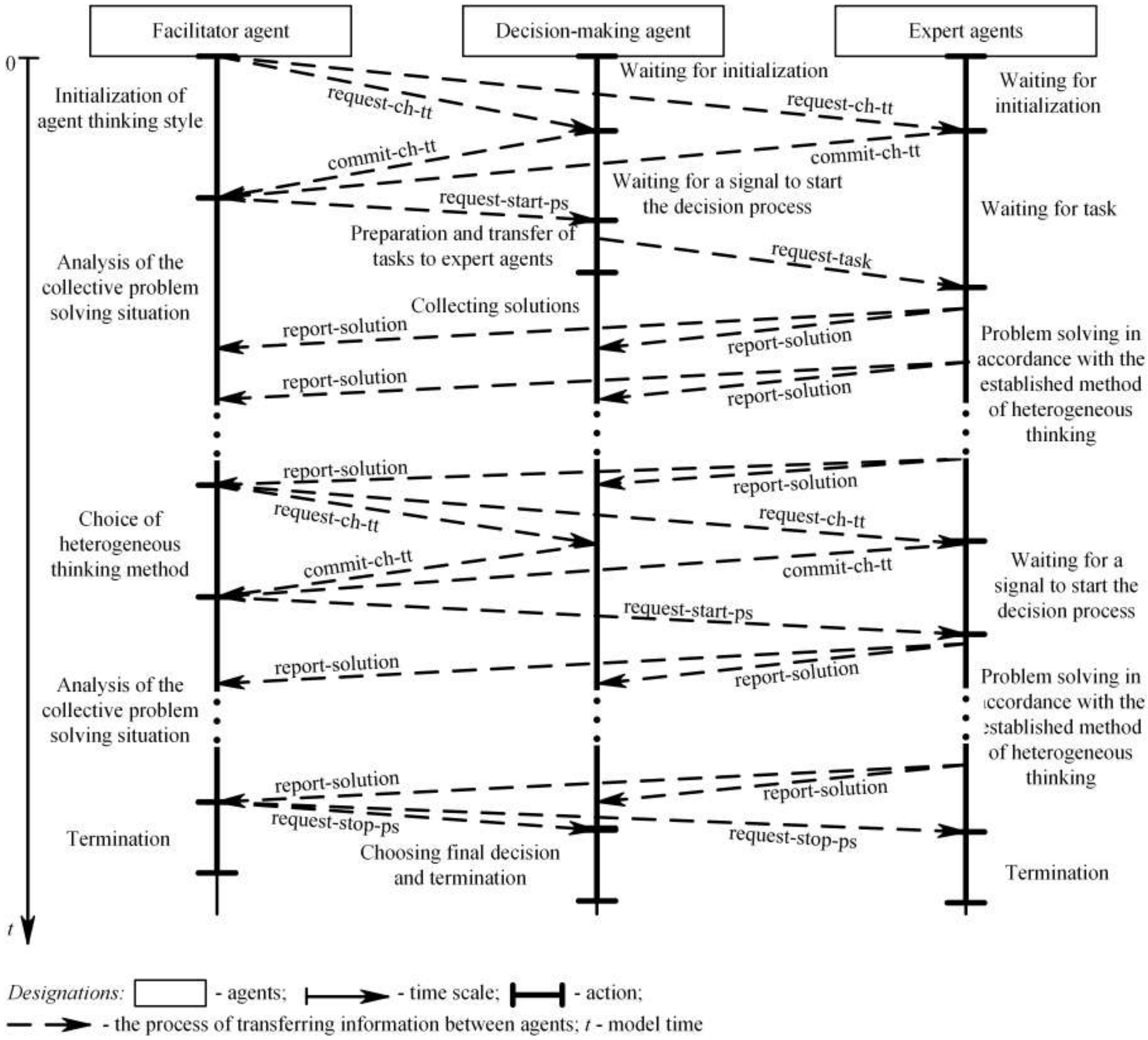


Figure 3. The scheme of the heterogeneous thinking protocol.

When FA receives problem solution from any of EAs it launches the function “analysis of the collective problem solving situation” , and determines with its help the intensity of the conflict in each pair of EAs and in the HIMSHT as a whole [16]. When a certain level of conflict intensity is reached in accordance with the fuzzy knowledge base, the FA launches the function “choice of heterogeneous thinking method” , which allows it to choose a method of heterogeneous thinking relevant to the collective decision-making situation. To implement this function, the FA has a fuzzy knowledge base about the relevance of agents’ “thinking styles” to different decision-making situations in the HIMSHT, and the methods to different features of problems and stages of collective decision-making. To form such a knowledge base, it is necessary to conduct a series of computational experiments to solve problems from different classes [10] and establish the correspondence between the class of problems and relevant approaches to the organization of heterogeneous thinking. By selecting the relay method, the FA sends to EA and DMA “request-ch-tt” type messages, the body of which indicates the method of heterogeneous thinking for EA and DMA. Then it waits for confirmations from the DMA and all EAs, after which it sends to EAs “request-start-ps” message indicating that all agents have switched to the appropriate method of heterogeneous thinking and the EAs can continue to work. The FA begins to receive solutions from the EAs, analyzes the situation, and chooses the method of heterogeneous thinking, until the convergent thinking stage is completed in the HIMSHT (Fig. 2). After its completion, the FA sends a “request-stop-ps” signal to stop the problem solving process to DMA and EAs and finishes its work. Having received such a signal, the EA interrupts the execution of tasks and finishes its work, and the DMA chooses solution in accordance with the established algorithm of heterogeneous thinking, which became the final collective solution. For example, it could be solution on which a consensus was reached or which got a majority of agent votes during the convergent thinking stage. After that, it passes this solution to the interface agent and also terminates.

The number of "switchings" of thinking methods by the protocol is not determined in advance, since the groan stage may be absent, and different methods at the divergent and convergent thinking stages can be used consistently. Thus, due to the presence in FA of a fuzzy knowledge base, as well as to the ability of representation of the heterogeneous functional structure of a complex task and heterogeneous collective thinking of intelligent agents HIMSHT interacting in accordance with the proposed protocol develops a relevant problem solving method for each problem without simplification and idealization in a dynamic environment.

V. System's effectiveness estimation

To evaluate accurately the effectiveness of the proposed HIMSHT architecture, it is necessary to accomplish its software implementation and conduct a series of computational experiments with various models of power grids. At the moment, a rough estimate of the HIMSHT effectiveness can be given by comparing its capabilities with other implemented systems designed to solve problems in various areas of the economy. For comparative analysis, two intelligent system is used: 1) hybrid intelligent system AGRO [10] for crop forecasting and planning of agricultural events, which allowed to increase the planning quality by 7-14%, and the planning speed by four times; 2) hybrid multi-agent intelligent system TRANSMAR [17], designed to solve complex transport and logistics problems and provided an increase in the efficiency of routing by more than 7%, and routing speed by 23% compared to methods existed at the moment of its creation.

Table I
Comparative analysis of the features of intelligent systems for solving heterogeneous problems

Features	AGRO	TRANSMAR	HIMSHT
Handling problem heterogeneity	+	+	+
Handling tool heterogeneity	+	+	+
Modelling expert reasoning	+	–	+
Autonomy of elements / agents	–	+	+
Ontology-based reasoning	–	+	+
Modelling collective heterogeneous thinking	–	–	+
Self-organization type	–	Weak	Strong

Designations: + - feature present; – - no feature.

As shown in Table I the proposed class of HIMSHT combines the representation of the heterogeneous functional structure of the problem with heterogeneous structure of the expert team and heterogeneous collective thinking methods, creating conditions for solving problems in dynamic environments without simplification and idealization. Due to the implementation of the system's heterogeneous elements as autonomous intelligent agents and ontology-based reasoning, the HIMSHT can effectively adapt to changing conditions of the problem, including modifying its structure and parameters, and develop a new relevant method during each problem solving process, showing signs of "strong" self-organization. Thus, HIMSHT has advantages over AGRO and TRANSMAR and more relevant to real expert teams solving problems in dynamic environments, therefore, as result of its software implementation, performance indicators could be no worse than those of reviewed intelligent systems could.

VI. Conclusion

The features of the problem of regional power grid restoration after large-scale accidents are considered and a new class of intelligent systems for its solution, namely HIMSHT, is proposed. A formalized description of the HIMSHT, and its main components are presented. The proposed HIMSHT moves the imitation of the collective development of operational actions to the field of synergetic informatics, when interaction between the elements of an intelligent system is no less important than their composition and quantity. This leads to self-organizing, social management models, each element of which is developing, obtaining data and knowledge from other elements. This reduces the cost of developing and operating the system. Modelling the methods of heterogeneous thinking by the agents of the system allows it to adapt to the dynamically changing conditions of the problems, each time re-establish connections between the agents, choosing the interaction style and developing a new decision-making method relevant to the situation.The protocol for organizing collective heterogeneous thinking of agents based on the theory of speech acts is proposed. Its use in hybrid intelligent multi-agent systems containing heterogeneous intelligent self-organizing agents allow to relevantly model effective practices of collective problem solving. The use of HIMSHT will allow dispatching personnel of power supply organizations to make relevant decisions on the restoration of the power grid in the shortage of time.

Acknowledgment

The reported study was funded by RFBR according to the research project No. 18-07-00448A.

References

[1] A.S. Narinyani, “Inzheneriya znaniy i NE-faktory: kratkiy obzor-08 [Knowledge Engineering and Non-Factors: A Brief Overview-08]" in *Voprosy iskusstvennogo intellekta* [*Artificial Intelligence Issues*], 2008, №1, pp.61–77. (in Russian)

[2] V.B. Tarasov, *Ot mnogoagentnykh sistem k intellektualnym organizatsiyam: filosofiya, psikhologiya, informatika* [*From multiagent systems to intelligent organizations: philosophy, psychology, and informatics*]. Moscow: Editorial URSS, 2002. (in Russian)

[3] S. Kaner, L. Lind, C. Toldi, S. Fisk, and D. Beger, *The Facilitator's Guide to Participatory Decision-Making*, San Francisco, CA: Jossey-Bass, 2011.

[4] E. de Bono, *Parallel Thinking: From Socratic to De Bono Thinking*, Melbourne: Penguin Books, 1994.

[5] M.I. Uspenskiy, and I.V. Kyzrodev, *Metody vosstanovleniya elektrosnabzheniya v raspredelitel'nykh setyakh* [*Methods for restoring electrical supply in distribution networks]*, Syktyvkar, 2010. (in Russian)

[6] S. Thiebaux, and M.-O. Cordier, “Supply Restoration in Power Distribution Systems — A Benchmark for Planning under Uncertainty" in *Proceedings of the 6th European Conference on Planning (ECP-01)*, Toledo, 2001, pp. 85-96.

[7] I.V. Kyzrodev, and M.I. Uspensky, “Avtomatizatsiya vosstanovleniya elektrosnabzheniya v raspredelitel'nykh setyakh [Automating the restoration of power supply in distribution networks]" in *Izvestiya Komi nauchnogo tsentra UrO RAN* [*Izvestiya Komi Scientific Center, Ural Branch of the Russian Academy of Sciences*], 2010, № 2, pp. 84-91. (in Russian)

[8] V. Kumar, H.C.R. Kumar, I. Gupta, and H.O. Gupta, “DG integrated approach for service restoration under cold load pickup" in *IEEE Transactions on Power Delivery*, 2010, Vol. 25(1), pp. 398-406.

[9] P. Hentenryck, and C. Coffrin, “Transmission system repair and restoration" in *Mathematical Programming*. Ser. B, 2015, Vol. 151, pp. 347–373.

[10] A.V. Kolesnikov, *Gibridnye intellektual'nye sistemy. Teoriya i tekhnologiya razrabotki* [*Hybrid intelligent systems: theory and technology of development*], Saint Petersburg: Izdatel'stvo SPbGTU, 2001. (in Russian)

[11] A.Y.C. Tang, G.S. Basheer, “A Conflict Resolution Strategy Selection Method (ConfRSSM) in Multi-Agent Systems" in *International Journal of Advanced Computer Science and Applications*, 2017, Vol. 8(5), pp. 398–404.

[12] A.V. Kolesnikov, S.V. Listopad, “Funktsional'naya struktura gibridnoy intellektual'noy mnogoagentnoy sistemy geterogennogo myshleniya dlya resheniya problemy vosstanovleniya raspredelitel'noy elektroseti [Functional structure of the hybrid intelligent multiagent system of heterogeneous thinking for solving the problem of restoring the distribution power grid]" in *Sistemy i sredstva informatiki* [*Systems and Means of Informatics*], 2019, Vol. 29(1), pp, 41–52. (in Russian).

[13] V.I. Gorodetskiy, M.S. Grushinskiy, A.V. Khabalov, “Mnogoagentnyye sistemy (obzor) [Multiagent Systems (Review)]" in *Novosti iskusstvennogo intellekta* [*Artificial Intelligence News*], 1998, № 2, pp. 64–116. (in Russian).

[14] M.P. Singh, A.K. Chopra, “Programming Multiagent Systems without Programming Agents" in *Lecture Notes in Computer Science*, 2010, Vol. 5919, pp. 1–14.

[15] D. Weerasooriya, A.S. Rao, K. Ramamohanarao, “Design of a concurrent agent-oriented language" in *Lecture Notes in Computer Science*, 1994, vol. 890, pp. 386–401.

[16] A.V. Kolesnikov, S.V. Listopad, “Hybrid intelligent multiagent system of heterogeneous thinking for solving the problem of restoring the distribution power grid after failures" in *Open Semantic Technologies for Intelligent Systems (OSTIS-2019): Research Papers Collection*, Minsk, 2019, pp. 133–138.

[17] Kolesnikov, A.V., Kirikov, I.A., Listopad, S.V., Rumovskaya, S.B., Domanitskiy, A.A. *Reshenie slozhnykh zadach kommivoyazhera metodami funktsional'nykh gibridnykh intellektual'nykh system* [*Complex travelling salesman problems solving by the methods of the functional hybrid intelligent systems*], Moscow: IPI RAN, 2011. (in Russian).

Протокол взаимодействия агентов гибридной интеллектуальной многоагентной системы гетерогенного мышления

Листопад С.В.

В статье рассматриваются вопросы моделирования процессов коллективного решения проблем в динамических средах. Необходимость такого моделирования обусловлена нерелевантностью традиционных абстрактно-математических методов к решению практических проблем, а также ограниченным временем в динамичной среде для организации всестороннего обсуждения проблемы и принятия решений реальной командой экспертов. Для обеспечения информационной поддержки решения проблем предлагается новый класс интеллектуальных систем, имитирующих коллективную работу под руководством фасилитатора — гибридные интеллектуальные многоагентные системы гетерогенного мышления. В статье предложен протокол взаимодействия агентов гибридной интеллектуальной многоагентной системы гетерогенного мышления для решения проблемы восстановления региональных энергосистем после крупных аварий.

Received 13.01.2020

What technologies should be used to create artificial intelligence systems? Subjective view at the problem

Valeriya Gribova
Institute of Automation and Control Processes FEBRAS, Far Eastern Federal University)
Vladivostok, Russia
gribova@iacp.dvo.ru

***Abstract*—This paper provides an overview of the current state in the development of artificial intelligence systems. An attempt is being made to answer the question: what should be the perfect artificial intelligence system? The requirements to the tools for creating artificial intelligence systems are identified and substantiated.**

***Keywords*—artificial intelligence, shells, machine learning, knowledge bases**

I. Introduction

Artificial intelligence as a scientific direction was created in the mid-50s, and the main research conducted by scientists were related to knowledge representation models and methods for solving intellectual tasks. First of all, it became obvious that problem-solving technologies based on the algorithmic approach are not suitable for solving creative or intellectual problems, so specialized technologies are needed to create systems of this class.

During the lifetime of artificial intelligence (AI) as a scientific direction, a number of models of knowledge representation and methods for solving intellectual problems were created. The development of AI systems was supported first by specialized programming languages, then by universal tool, as well as problem-independent and specialized shells. Typically, AI system development tools and shells were intended for knowledge-based AI system design, while AI system development based on other methods were much more poorly supported by specialized tools. However, the situation has changed dramatically in the past few years. Numerous frameworks and libraries have entered the market that support the neural network approach to creating systems of this class. They are convenient, well implemented (they are developed by the leaders of the IT industry), and have a low enough barrier to entry for developers. As a result, AI has often been reduced exclusively to the neural network approach, and the approach based on knowledge is often characterized as outdated, lost its relevance. Is that so? What should AI technologies be, what properties should they have? The article contains the author's opinion on all these issues.

The research is carried out with the partial financial support of the Russian Foundation for Basic Research (projects 19-07-00244 and 18-07-01079).

II. Current state of artificial intelligence system development technologies

The first scientific research in the AI field led to the creation of a number of specialized programming languages, such as PROLOG, LISP, SMALLTALK, FRL, Interlisp, etc. These languages have many advantages, however, their use has shown a high complexity of creating software systems, which, as noted in [1] has led to fact, the development of intelligent systems has become almost inaccessible.

Together with the development of knowledge representation models and methods for solving intellectual problems, universal tools and specialized software shells were actively developed, mainly for creating systems with knowledge bases. Typical representatives of this class of systems are: IACPaaS, OSTIS, AT-TECHNOLOGY, Level5, Object, G2, Clips, Loops, VITAL, KEATS, etc. [1], [2], [3], [4], [5]. Significant differences between various tools are related to the level of instrumental support, development technologies, the formalisms used to represent knowledge, the methods for their formation and debugging, the used inference machine, and user interface design tools. Specialized shells designed to solve problems of a certain class are also widely used. The main advantage of this class of tools is significant reduction of development time and maintenance of software. The main disadvantage is limitations of the area of use.

Approximately from the last decade of the 20th century and the first of the current one, active development has received another direction of AI, - machine learning and big data analysis. This is due to the development of new computing platforms, a significant increase in their performance, as well as the creation of new theoretical models and architectures, primarily based on neural networks. To date, we can see a boom in neural networks, that foster a number of tools - frameworks (for example, Deeplearning4j, TensorFlow and Theano, PyTorch), specialized libraries (e.g., Keras, PuzzleLib, Caffe, etc.) allows developers to create, train, test, and deploy a network of any complexity [6], [7], [8]. However, it is very important that the use of libraries for machine learning significantly save development time, thereby eliminating the most difficult stages of its development. Many machine learning libraries are free and available, which has led to significant progress in this area in recent years. In addition to tools, developers have access to ready-made, already trained, neural networks for solving various problems.

Unfortunately, the boom in the field of neural networks, as well as a significantly increased interest in AI, has generated not only a huge number of "experts" in this field, who, taking a ready-made model, data and running training, get perhaps even a good solution for simple cases, consider themselves experts in this field and reduce AI exclusively to a neural network approach. Moreover, they actively promote, and not without success, the assertion that the approach based on knowledge bases (while they usually mention only the production model of knowledge representation), hopelessly outdated, in modern AI systems it is used less and less and tends to complete extinction.

Estimating the current state of AI technologies, the following can be stated:

1. AI technologies based on knowledge bases continue to be developed and used. These technologies have evolved significantly. According to a study published in [2], the most popular production model of knowledge representation in the 80-90 years is losing its popularity (according to estimates in [2], the number of systems with this model of knowledge representation has decreased by about 4 times), giving way to systems with other models of knowledge representation. The trend to decrease production can be explained by the expansion of the available tools that support other representation models that are most appropriate to the domain and tasks for which the production representation is not convenient. First of all, it is an ontological approach to the representation of knowledge. The most well-known technology is Protege, which supports the object-oriented representation model [9], as well as the relatively new (compared to Protege), IACPaaS [10] and OSTIS [5] technologies, which support the semantic model of knowledge representation. These three technologies are developed by research teams, which, on the one hand, ensures that the proposed technologies correspond to the current level of scientific research in this area and their continuous improvement, on the other, these technologies, admittedly, lack a "beautiful wrapper", context assistance systems, and other marketing elements, which are important for promoting technologies "to the masses". Research teams have neither the experience of promoting software products, nor the human resources.

2. Machine learning technologies, primarily neural networks, are currently leading. It is neural network technologies that are often declared as modern artificial intelligence technologies in contrast to knowledge-based technologies. Yes, in solving some problems primarily related to the analysis of images, text, they demonstrate impressive results, their use is undoubtedly justified for a number of applications and, importantly, world leaders of the IT industry made it easier for developers to use them. However, it is still difficult to agree that the neural network approach is a modern AI that simulates, albeit simplistically, the work of the human brain. Do not forget that a neural network is a fitting of a function implemented by a neural network for training data by selecting the parameters of this function. This is still very little like how the human brain learns.

Accepting and understanding the importance and need of all available technologies today, however, it is worth noting that the two types of technologies work very poorly together, despite some attempts to integrate them and imitate only certain functions of the natural intelligence. Therefore, the scientific community faces an important task of determining the future of artificial intelligence technologies and their development directions.

III. What is the perfect artificial intelligence system

In order to answer the question of what the perfect AI system is, it is necessary to understand what properties, abilities the natural AI has (the author understands that the perfect systems do not exist, in this case, the perfect requirement should be understood as a softer requirement - as close as possible to the perfect one). V.K. Finn listed 13 such abilities [11]. Here's their compressed enumeration: the discovery of the essential in the data; the creation of "goal-plan-action" sequences search for assumptions relevant to the purpose of reasoning; ability to reason: conclusions as consequences from assumptions; conclusions that do not follow directly from assumptions: conclusions by analogy, inductive generalizations of data, conclusions by explaining data (abduction), solving problems through various cognitive procedures; the ability to estimate knowledge and action; the ability to explain - the answer to the question "why?"; argumentation when making a decision; recognition as the answer to the question "what is this?"; the ability to learn and use memory; the ability to integrate knowledge, to form concepts and theories; the ability to refine unclear ideas and transform them into precise concepts; ability to change the knowledge system while learning and changing the situation.

The perfect AI system should have all these properties, and as V.K. Finn notes, that not all natural intelligence functions can be automated, some of them can be implemented only through "man-computer" interaction. Thus, if we really want to create an AI system, it must have at least a subset of these properties. It is important to note: in natural intelligence the result of one (or several) cognitive activities is always (!) integrated (consistent) with other types of cognitive procedures.

For example, if we have knowledge in an area and have an object with a set of attributes, then we, on the basis of our knowledge system, can make some logical inference (depends on the problem being solved). It is possible that the attributes of this object are not consistent with our knowledge system. We have a precedent and need to understand what to do. Next, we make some decisions (possibly wrong). When it is necessary to draw a logical inference about a certain object again, we proceed as follows: draw a conclusion based on our knowledge system, or compare an object with an existing precedent to make a decision (a reasoning by analogy). The accumulation of precedents with a known result (learning) creates a situation where a person needs to correct his or her knowledge system, etc. , so it is important that both knowledge (in the computer system is a knowledge base), and precedents (data) are clear to the person. Thus, terminology must be defined, which is understandable and accepted by a certain community, and the connection between cognitive procedures should be "seamless". Such system can claim to be called an AI system. In this case, the knowledge obtained either from a person or as a result of machine learning will be integrated into the knowledge base, which will be continuously improved. Knowledge itself will be available to both the AI system and the person who will be able get new knowledge (using the updated knowledge base) in a way that is not accessible to the computer system. As in real life, a person, using a computer, makes calculations, receives information from external devices, respectively, the AI system must "understand and accept" information from external devices.

Thus, if we are talking about the AI system, it should support a subset of cognitive procedures: the inference of consequences from assumptions - working with the existing system of knowledge, reasoning by analogy (this function is often overlooked and underestimated, although in real life, natural intelligence, it plays a significant role), learning as a way of correction the knowledge system and obtaining new knowledge, the ability to explain and justify their actions. At the same time, all procedures should be understandable to a person (described in his or her system of concepts), and the connection between different types of cognitive procedures should be a natural "seamless" way to carry out their "circle" with the participation of a person.

IV. Requirements for artificial intelligence system tools

Here are the basic requirements for AI tools.

Support the development of several intellectual activities with "seamless" integration.

This requirement is discussed in the paragraph above. It is undoubtedly fundamental and means that the tools for creating AI systems should be much richer in their functionality than frameworks and tools designed for systems of other classes, because they should allow developers to create knowledge bases, ontologies, solvers, focused on logical inference from the knowledge base. Using them it is possible to describe the algorithms of learning, also oriented to the knowledge bases, reasoning on precedents, generation of explanations and other cognitive activities. Is it hopefully to develop different types of cognitive activity with a set of tools? The answer is yes, you can, but in this case it is difficult to integrate technologies, you need to create additional software modules that provide a link to them into a single whole. And this link is not reduced to a set of formats. It is important that the results obtained are embedded in all other cognitive procedures.

Availability of means to support the viability of the created systems and their development tools.

This requirement is vast and, in fact, includes many other requirements. Its implementation, like the previous requirement, indicates the maturity of the development. This requirement should be given key attention both when creating tools and AI systems based on them. Unfortunately, it is common practice for most implementations to declare the functionality of a software to demonstrate results. However, developers find it difficult to list the mechanisms providing the viability of the created system. But modern statistics state that the process of evolution, maintenance of the system after implementation of its release in terms of labor costs is many times higher than the stage of actual development.

In general, this requirement should be provided by the following solutions:

- declarative representation of components, shifting the efforts of developers from programming to designing software components;
- automatic generation of software components;
- creation of technology that separates competencies among developers with sequential or parallel work on the formation of models and components, by providing the architectural integrity of the project;
- availability of tools for intellectual support of development of software component, a visual and understandable representation of components to software developers.

In the implementation of these requirements key role belongs to the tools with which the system is created. The longer the life cycle of any system, the more obvious the need for developers to improve their adaptation mechanisms due to changes of user requirements, domain, and operating conditions. It is important that users of the tool can also improve it, so the means for expanding the tool should be as close as possible to the means for developing the AI systems themselves, and the complexity of this process is comparable to the complexity of developing the corresponding components of the AI systems.

Integration with external software and components.

The AI system should be able to interact with third-party systems based on different interaction models.

Collective development.

The AI system development requires the participation of many specialists, so the tool must have the means to ensure the collective development of system components, including collective development one component (for example, a knowledge base can be developed by several experts). To do this, a hierarchical system should be provided for managing access rights to components, monitoring changes, coordinating between components through user accounts and cloud tools for creating and managing components.

Support for the creation of evolving shells, knowledge portals and/ or ecosystems in various domains based on universal tools.

It is known that there are two approaches to creating tools: universal tool complexes designed to create a wide class of software and specialized tools for development of either a class of tasks or a set of tasks in a particular domain. The evolution of the development of two opposite approaches has shown that these approaches do not conflict with each other, if there are specialized tools, this is preferable for developers due to a significant reduction in the complexity of creating systems of this class, which is supported by specialized tools. It is important that development of such shells or ecosystems is supported by universal tools for permanent modification and evolution of systems based on it.

V. Discussion

The development of AI systems is a complex and time-consuming work that requires the participation of specialists in various fields and domains - programmers, cognitive scientists, interface designers and domain experts.

We have to admit that the scientific community has not yet given a clear and understandable definition of AI systems, which has given rise to many interpretations and definitions of a wide range of "experts".

The situation requires a serious redefining and consolidation, first of all, of the scientific community.

1. It is important to determine what an AI system is. It is justified to give this definition, based on the range of cognitive abilities of the natural intelligence. As a basis I propose to take the list proposed by V.K. Finn (perhaps a little modified). In this case, it is easy to understand the "level" of the intelligence of the AI system. For example, a system based on neural networks implements (imitates) one function of natural intelligence - learning, a system based on knowledge base, implementing the search and criticism of hypotheses with the generation of explanations - three functions of natural intelligence, etc. In this case, it is clear what functions the AI system lacks and what needs to be further implemented to increase its intelligence.

2. Technologies for AI system development. It is obvious that they should develop in two directions - the creation of universal tools and specialized, taking into account the specifics of the architecture, methods and models of AI, as well as having a set of tools for automating development. In this case, the scientific community faces a significant challenge in the development of universal and specialized methods of imitation of cognitive abilities providing coherence with other cognitive abilities, and technology developers have to provide instrumental support for their implementation. An important task is the transition from particular models to their generalization (see Fig. 1 Evolution of development of diagnosis process ontologies of Intelligent System Laboratory of Institute of Automation and Control Process FEB RAS).

3. Moving from demonstration prototypes to viable implementations. This requirement is important for all types of software systems, but it is especially important for AI systems. It is obvious that neither the knowledge base, nor the learning process, nor the realization of other intellectual functions can be complete and require permanent revision. Therefore, it is proposed to evaluate the implementation of AI systems not only from the point of view of listing simulated cognitive procedures, but also from the point of view of mechanisms implementing their viability. There is undoubtedly a significant role of tools for the creation of the AI system.

4. Use terminology. Today, there is a wide range of terminology used. The same concepts are called and interpreted differently, and vice versa. It is necessary to bring it (at least within the scientific community) to a common interpretation and designation.

5. Development of tools (universal and specialized). In my opinion, different implementations, different tools and approaches to creating tools for implementing AI is a positive factor in the development of AI. It is unlikely that in the near future we can expect "Babel", if we do not take into account the libraries and frameworks for creating systems based on neural networks, implementing, as discussed in this article, only one of the cognitive abilities. Creating tools is an extremely difficult and, unfortunately, often very thankless job that requires understanding both your own vast experience in this

field, and analysis of the literature to generalize methods and models in order to create universal or specialized tools that implement them. Thus, sharing the experience of 30 years of development of the Protege system, M. Musen stated that new scientific results, the availability of alternative / competing systems and tools, feedback and user requests - these are the main drivers of the development of the Protege system [9]. Another issue is that today tool systems, even those created within the same professional community (IACPaaS and OSTIS platforms), although in different countries, supporting a very similar list of basic requirements and key principles of development, do not support compatibility with each other. At the same time, both platforms have already accumulated a large range of reusable modules, components and implemented systems. For example, the IACPaaS platform has more than 200 active developers, and for example, an ecosystem created on its basis for solving a complex of intellectual problems in practical medicine and education [12], see Fig. 2, contains a terminological base of more than 27 thousand concepts, their properties and values, ontologies for diagnosis, monitoring, treatment of diseases (which is important independent of the section of medicine), knowledge bases only for the diagnosis of diseases containing more than 100 thousand vertices (knowledge bases have a semantic representation), as well as knowledge bases for drug and rehabilitation treatment, problem solvers for decision support systems, computer simulators, and other reusable components. Therefore, it is important to create an interested community of developers of AI tools and applications, which would develop a set of requirements for compatibility and integration of various types of tools and application systems (and, perhaps, a list of the main properties for which you need to compare tools with each other).

6. Intellectual property. This issue is also quite relevant and topical. It is important to preserve the intellectual property of knowledge in knowledge bases, services, and their components.

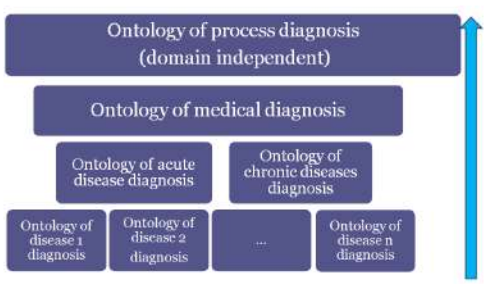


Figure 1. Evolution of diagnosis process ontologies

REFERENCES

[1] Rybina G. V. Intellektual'nye sistemy: ot A do YA. Seriya monografij v trekh knigah. Kn. 3. Problemno-specializirovannye intellektual'nye sistemy. Instrumental'nye sredstva postroeniya intellektual'nyh system (Intelligent systems: A to Z. A series of monographs in three books. Book 3. Problem-specialized intelligent systems. Tools for building intelligent systems), M.: Nauchtekhlitizdat, 2015, 180 p. (in Russian).

[2] Emmanuel C. Ogu, , Adekunle, Y.A. Basic Concepts of Expert System Shells and an Efficient Model for Knowledge Acquisition // Intern. J. of Science and Research Intern. Journal of Science and Research (IJSR), India Online ISSN: 2319-7064. 2013, vol. 2, iss. 4, pp. 554–559.

[3] Valeria Gribova, Alexander Kleschev, Philip Moskalenko, Vadim Timchenko, Leonid Fedorischev, Elena Shalfeeva. The IACPaaS Cloud Platform: Features and Perspectives // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC) (Vladivostok, Russia, 25-29 sept. 2017). IEEE. 2017, pp. 80–84. DOI: 10.1109/RPC.2017.8168076.

[4] Gensym G2. The World's Leading Software Platform for Real-Time Expert System Application. URL: http://www.gensym.com/wp-content/uploads/Gensym-l-G2.pdf (accessed 14.01.2020)

[5] Vladimir Golenkov, Daniil Shunkevich, Irina Davydenko. Principles of organization and automation of the semantic computer systems development. Open semantic technologies for intelligent systems, 2019, iss. 3, pp.53–91.

[6] S. Markidis, S. W. D. Chien, E. Laure, I. B. Peng, J. S. Vetter, "NVIDIA Tensor Core Programmability Performance and Precision", 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 522–531, May 2018.

[7] A. A. Awan, K. Hamidouche, J. M. Hashmi, D. K. Panda, "S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters", Proceedings of the 22Nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming ser. PPoPP'17, 2017, pp. 193–205.

[8] Machine learning // Wikipedia, the free encyclopedia: internet-portal. URL: https://en.wikipedia.org/wiki/Machine_learning (accessed 14.01.2020).

[9] Musen M. A. The Protégé Project: A Look Back and a Look Forward, AI Matters, 2015, vol. 1, issue 4, pp. 4–12.

[10] Valeriya Gribova, Alexander Kleschev, Philip Moskalenko, Vadim Timchenko, Elena Shalfeeva. The Technology for Development of Decision-Making Support Services with Components Reuse // Advances in Intelligent Systems and Computing. Vol. 902, pp. 3–13. ISSN 2194-5357. ISBN 978-3-030-12081-8. https://doi.org/10.1007/978-3-030-12082-5_1. Springer, Cham. https://link.springer.com/content/pdf/bfm

[11] Y.M. Arskiy, V.K. Finn Principi konstruirovaniya intellektualnih sistem. Informacionnie tehnologii i vichislitelnie sistemi, 2008, no.4. s. 4–37 (Principals of intelligent system design // Information technology and computer systems, 2008, no. 4, pp. 4–37 (in Russian))

[12] Valeria V. Gribova, Margaret V. Petryaeva, Dmitry B. Okun, Alexey V. Tarasov, Software Toolkit for Creating Intelligent Systems in Practical and Educational Medicine // IEEE Xplore. 2018. DOI: 10.1109/RPC.2018.8482130.

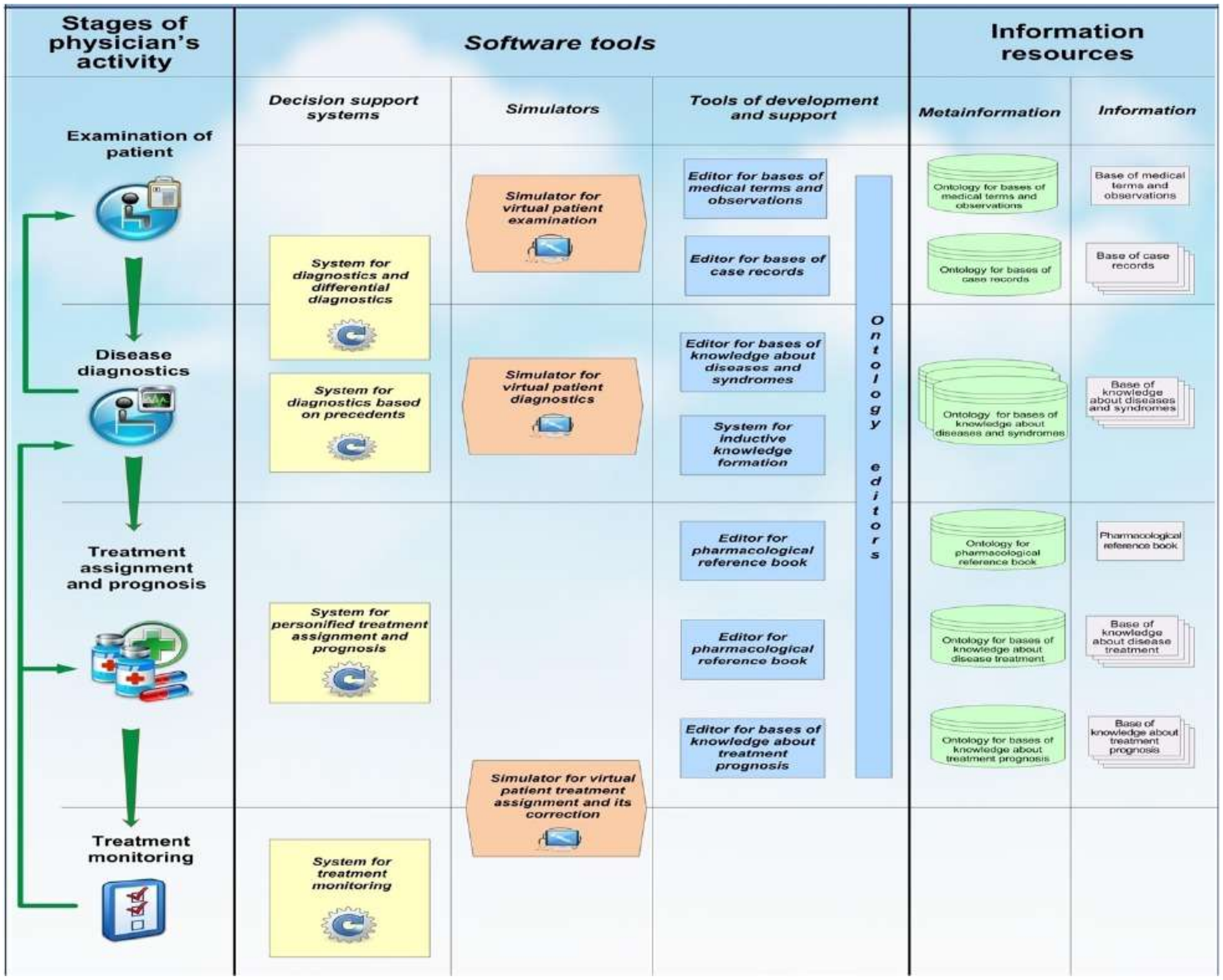


Figure 2. Cloud infrastructure for creating AI systems in medicine

Какими должны быть технологии для создания систем искусственного интеллекта? Субъективный взгляд на проблему

Грибова В.В.

Искусственный интеллект как научное направление возник в середине 50-х годов, и основные исследования, проводимые учеными, были связаны с моделями представления знаний и методами решения интеллектуальных задач. Прежде всего, стало очевидно, что технологии решения задач на основе алгоритмического подхода плохо подходят для решения творческих или интеллектуальных задач, поэтому необходимы специализированные технологии для создания систем такого класса.

За время существования искусственного интеллекта (ИИ) как направления, был создан ряд моделей представления знаний, методов решения интеллектуальных задач. Технологически это поддерживалось сначала специализированными языками программирования, затем универсальными инструментальными системами и проблемно-независимыми и специализированными оболочками. Как правило, инструментарий для разработки систем ИИ поддерживал проектирование систем ИИ на основе знаний, разработка систем ИИ, основанная на других методах, гораздо более слабо поддерживались инструментально. Однако в последние несколько лет ситуация коренным образом изменилась. На рынок вышли многочисленные фреймворки и библиотеки, поддерживающие нейросетевой подход к созданию систем такого класса. Они удобны, хорошо реализованы (их разработкой занимаются лидеры IT-индустрии), имеют достаточно низкий порог вхождения для разработчиков. В результате ИИ зачастую стал сводиться исключительно к нейросетевому подходу, а подход, основанный на знаниях, многие характеризуют как устаревший, потерявший свою актуальность.

Так ли это? Какими должны быть технологии ИИ, какими свойствами обладать? Данная работа содержит мнение автора по всем указанным вопросам.

Received 15.01.2020

The Technology of Temporal Knowledge Obtaining from Various Sources: the Experimental Research of Combined Knowledge Acquisition Method

Galina Rybina, Aleksandr Slinkov, Dmitriy Buyanov
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow,Russian Federation
Email: gvrybina@mephi.ru, galina@ailab.mephi.ru, sanch971@mail.ru

***Abstract*—This paper analyzes the results of the experimental research of automated knowledge base construction for dynamic intelligent systems, in particular dynamic integrated expert systems on the basis of the so-called original combined knowledge acquisition method with temporal extensions.**

The focus of this work is on some aspects of the application and development of technologies of knowledge acquisition from various sources (experts, NL-texts, data bases) in order to create new applied intellectual technologies that can be used, for example, in the field of healthcare (personalized medicine, "smart" hospital, etc.). The paper describes general description of basic technologies of knowledge acquisition from various sources, features of the combined knowledge acquisition method and means of its implementation, features of a language experiment and the knowledge acquisition from a temporal database. Analysis of experimental results of software modeling is given.

***Keywords*—artificial intelligence, intelligent systems, automated temporal knowledge acquisition, temporal inference.**

I. Introduction

An important place among the priority directions of the development and use of artificial intelligence (AI) technology, determined by Decree of the President of the Russian Federation (No. 490), is given to the development of software that uses such basic AI technologies as: technologies of knowledge acquisition from various sources and intellectual analysis of Big Data; technologies of forecasting and decision support; technologies of planning and multi-agent control of targeted behavior in unstructured environments; technologies of processing natural languages (NL), etc.

Nowadays, based on these basic technologies, it is necessary to develop applied AI technologies that will be used in various fields of life, production and economics. A particular role among the basic AI technologies is currently given to automated technologies of knowledge acquisition from various sources, such as experts, NL-texts and databases, which can be effectively used to create new applied technologies, for example , in the field of healthcare (personalized medicine, «smart» hospital, etc.) and other problem domains. In this regard, the focus of research and development in recent years (domestic and foreign) turned out to be issues related to solving a set of scientific and technical problems of obtaining, presenting and processing of temporal knowledge for building dynamic intelligent systems, in particular, dynamic integrated expert systems [1], [2], used to solve a wide class of unformalized and formalized tasks of various practical significance and complexity. The aim of this work is to discuss some new results in this domain in the context of the development of a problem-oriented methodology of IES construction for static and dynamic problem domains.

II. General description of basic technologies of knowledge acquisition from various sources

As shown in [1], [2], the problems of obtaining knowledge from sources of knowledge of various typologies (experts, NL-texts, DB), as well as the issues of creating effective technologies of automated knowledge acquisition are still in the focus of modern intellectual systems developers' attention, in particular, the most popular integrated expert systems (IES) with scalable architecture and extensible functionality [1].

The analysis conducted in [3]–[5] has shown that the most acute problem of knowledge acquisition arises when solving complex practical problems in such areas as energy, space, ecology, etc., as well as in the social field, for example, in the field of healthcare, where significant amounts of data have been accumulated, for which various Big Data technologies are used today [6], and when building knowledge bases (KBs) in intelligent systems such as the IES, several experts or groups of experts are required, which significantly increases the cost and time parameters of system development in the absence of automated means to support the process of obtaining knowledge from the expert / experts, who are the main knowledge source.

Currently, the typology of knowledge sources is no longer limited only to experts because significant amounts of expert knowledge are accumulated in the NL-texts and in the information accumulated in the databases (DB) of modern business information systems. Problems of obtaining (revealing) knowledge from NL-texts are related to the rapidly advancing technology Text Mining [7], [8], and various methods and algorithms of automated knowledge acquisition from DB are included in the technology Data Mining / Deep Data Mining, Knowledge Discovery in Databases (KDD), etc. [9]–[12]. Success of Text Mining technology is connected with different aspects of application of textual methods of obtaining knowledge from NL-texts, which have received the greatest development in three types of modern web-oriented NL-systems - information retrieval, information extraction and understanding of NL-text (Text / Message Understanding) [2], [5].

It should be noted that for solving the problems of knowledge acquisition from NL-texts, various methods and approaches are currently used: machine learning of neural networks [13]–[15], ontology construction and search in the concept graph [16], Markov logic [17], consideration the syntax and semantics of a particular language [18], the use of a large body of problem-oriented texts annotated using the TimeML language [19], and the obtaining of temporal information from time series [20], [21].

However each of the above-mentioned technologies has emerged and developed independently and today such autonomy and distribution does not allow to carry out effective monitoring of such information resources as KB, DB, and in recent years ontologies possessed by intellectual systems, in particular, IES [2]. Moreover, at present, except for the works [2]–[5], [12], [22], there is practically no research in the field of creation of tools and technologies of distributed knowledge acquisition from various sources.

Experience of practical use of a number of applied IESs, including the most complex dynamic IESs [2], [3], [5], etc, developed on the basis of problem-oriented methodology (author G.V. Rybina) [1] and supporting its software - AT-TECHNOLOGY workbench, for such problem domains (PD) as express-diagnostics of blood, diagnostics of complex technical systems, design of unique mechanical engineering objects, complex environmental problems, etc., showed the effectiveness of the joint use of three sources of knowledge - experts, NL-texts and databases. For example, the analysis of experimental data obtained in the course of creation of KB of several applied IESs using the combined method of knowledge acquisition (CKAM) [1]–[5], [12], [13], which is an integral part of this methodology, has shown that the local use of the database as an additional source of knowledge can supplement the volume of developed BDs by 10-20 percent, depending on the specifics of PD.

Now on the basis of the problem-oriented methodology the intellectual program technology and the automated workplace of the knowledge engineer - the complex AT-TECHNOLOGY on which basis, including with use of three versions of means of support CKAM (local, distributed, dynamic), more than 20 applied IES for static PD are developed, and also prototypes of the most difficult dynamic IES, i.e. the IES using dynamic representations of a subject domain and solving dynamic problems [1], [2] are created.

The modern stage of development of the technology of knowledge acquisition from various sources in the context of creation of a dynamic version of the CKAM and its support tools, functioning as part of the new generation of tools such as WorkBench - AT-TECHNOLOGY complex - is associated with the automation of the processes of obtaining, presenting and processing of temporal knowledge for the construction of KB in dynamic IESs. Relevance of the research is related to the fact that, despite the existence of a significant number of approaches to the representation of temporal dependencies in the context of automatic processing of NL-texts, the issues of obtaining temporal knowledge (both using manual techniques and automated methods) for constructing temporal KB in dynamic intelligent systems, in particular, in dynamic IES [5], are practically not considered.

The aim of this work is to present new results of experimental software modeling of the processes of temporal knowledge acquisition for the automated construction of KB in dynamic IES (for example, medical diagnostics).

III. Features of the combined knowledge acquisition method and means of its implementation

The basic version of CKAM [1], [2], [12] and its support facilities, created in the end of 1990s, are constantly being developed and successfully used to automate the processes of KB development in static PD, forming the core of the automated workstation of the knowledge engineer on the basis of the software platform AT-TECHNOLOGY. Today a distributed variant of computer knowledge acquisition is supported [12], [13], which provides integration of three types of knowledge sources (experts, NL-texts, DB) taking into account their geographical distribution within the client-server architecture. Proceeding from the context of this work, we will focus only on those features of CKAM, which are most important from the point of view of studying the possibilities of CKAM development for the purpose of automated construction of temporal databases for dynamic IES.

The first peculiarity of CKAM is the way of organizing the process of direct obtaining of knowledge from experts by means of computer interviewing at all stages of the life cycle of the IES construction on the basis of the author's approach "orientation to the model of solving a typical problem" [1], according to which the controlling knowledge about strategies (methods) of solving specific classes of problems solved in a similar way is formed in the form of some heuristic model of a typical problem [1] (diagnostics, design, planning, etc.). Therefore, the processes of obtaining knowledge are controlled by means of sets of models for solving typical tasks, for which a number of methods and approaches have been developed and are constantly being developed, which allow to create scenarios of dialogues with experts, reflecting both the thematic structure of the dialog (i.e. the scheme of solving a typical task [1], [2]) and the local structure of the dialog (steps of the dialog [1], [2]), i.e. a set of specific actions and reactions between the expert and the system.

Thus, the processes of knowledge acquiring from experts and NL-texts are computer modeling, allowing in the process of dialogue with experts to build and denote all the components of the model of solving a typical problem and form both fragments of the knowledge field [1], [2] (intermediate representation of structured knowledge used to verify information obtained from various sources), and the corresponding fragments of KB. To build an "action-response" scheme of partners, several implementation techniques are used, in particular, the "simulation of consultation" method, etc.

Expert interviewing processes are supported by a dialogue script interpreter, with each scenario corresponding to a specific type of task. In addition, special screen forms are provided for entering unreliable knowledge [1] (uncertainty, inaccuracy, fuzziness) and connecting the means of implementing the adaptive method of repertoire grids [1] (for example, for the implementation of procedures of differentiation of diagnoses in the case of activation of the scenario for the task of medical diagnosis). A specialized linguistic processor and a set of dynamically updated dictionaries occupy an important place in the software for supporting basic and distributed CKAM (linguistic aspects of CKAM are described in detail in [1], [2], [12]).

Another important feature of the CKAM is the integration of closely interrelated processes of expert computer interviewing with methods of NL-texts processing (entered both during the interview session and after the end, in the form of expert interviewing protocols), as well as methods of knowledge acquisition from the DB [1], [12], [13].

In order to use temporal DB [9], [10] as an additional source of knowledge, the basic functionality of the means of supporting the distributed version of the CKAM has been significantly expanded by developing new algorithms of knowledge acquisition from temporal DB and means of integration of various sources of knowledge (experts, NL-texts, DB) [1], [12], [13]. Instead of the CART algorithm used in the local version of the basic CKAM [13], the well-known Random Forest algorithm [11] was implemented, modified to support the work with temporal DB.

The essence of the modification was the use of multivariate feature space, one of which is a time stamp. The ensemble of solution trees is constructed in accordance with the basic algorithm; however, the calculation of the value of the partition criterion has undergone changes due to the use of the multidimensional space of characters (the partition criterion will be the arithmetic mean of the calculated values of information entropy). In addition, unlike the decisive trees based on modified CART and C4.5 algorithms [1], [12], [13], the tree construction is performed until all the elements of the sub-sample are processed without the application of the branch cutting procedure. The solution tree algorithm is executed as many times as necessary to minimize the error of classification of objects from the test sample (classification of objects is done by voting by analogy with basic version of the Random Forest algorithm [11]).

Thus, the actual problem of the current stage of research is the further evolution of the CKAM, in order to develop methods and means of automated construction of temporal KB in dynamic IES. To date, models, methods and software for representation and processing of temporal knowledge have already been developed and tested in the creation of several prototypes of dynamic IES models [2]–[5], [14]. Below is a description of the current results of the experimental program modeling of the temporal version of the CKAM.

IV. Features of a language experiment and the knowledge acquisition from a temporal database

For modeling of processes of direct knowledge acquisition from experts and NL-texts (sublanguage of business prose [1], [2]) the typical problem - medical diagnostics - was used, and as PD the complex diagnostics of diseases of a mammary gland and diagnostics of traumas of a knee joint was considered. Model dialogues were conducted in the form of a "language experiment" [2], [4], [5] related to the search for temporal information, i.e. temporal relations both within each NL-proposal coming from the expert and/or in neighboring sentences (taking into account the current state of the local structure of the dialog) and with the search for relations indicating the time of text creation.

For these purposes, the dictionary of temporal lexemes developed on the basis of the works [15], [16], a specialized linguistic processor and interviewing support tools functioning as a part of the AT-TECHNOLOGY complex were used. Scenarios and corresponding screen forms were developed and tested with the help of model dialogues.

In total, several hundred modeling sessions of interviewing were implemented with the participation of about 80 students who, according to the principle of "doctor to himself", introduced lexemes (temporary pretexts, target pretexts, causal pretexts, particles, adverbs of time, etc.) into the corresponding screen forms to build fragments of the knowledge field. On the basis of the experiments carried out, a set of modified scenarios describing the thematic and local structure of the dialogue when solving a typical problem of medical diagnostics was obtained, which made it possible to implement the elements of the "through" technology of direct acquisition and representation (in terms of an extended language of knowledge representation [2]) of fragments of temporal KB, ready to implement temporal output on the production rules [2], [4], [17].

Thus, the use of a set of model dialogues made it possible to experimentally determine which temporal entities (markers) [5] can be detected on the basis of algorithms and software of the temporal version of the CKAM and significantly add to the current temporal lexeme vocabulary.

Another complex of experiments was carried out with the modified Random Forest algorithm, which is a part of the CKAM and is the core of knowledge acquisition from temporal DB. As an input to this algorithm, some medical temporal DB containing data in a certain format was used, and the set of medical data was exported to the database under the control of SqLite 3, and then to a separate table with the allocation of identifiers with assigned classes. Thus, a table with objects is formed, which contains their attributes at each moment of time, and a table with classes. The Random Forest algorithm builds an ensemble of trees according to this temporal DB, where each committee tree allocates the classified object to one of the classes, i.e. it votes, and wins the class for which the largest number of trees voted. A fragment of the knowledge field containing the rules of tree voting in intervals and rules in the extended language of knowledge representation is built on the trees.

The obtained from temporal DB fragment of knowledge field is suitable for further verification and integration [1], [2], [12], [13] with fragments of the knowledge field obtained as a result of expert interviewing sessions [1], [2], [12], [13].

V. Analysis of experimental results of software modeling

Below are the results of experimental research of algorithms and software tools for the formation and verification of knowledge field elements with temporal entities. For these purposes, modal dialogues were used for medical ultrasound diagnostics and diagnostics of the knee joint. Experiment scenarios were used, including:

• addition of new events and intervals to the model dialogues, due to which the vocabulary of temporal lexemes, previously not detected in experiments carried out with the support of the algorithm for identifying temporal markers, was replenished (positive testing);

• inclusion of events and intervals in the knowledge fields without references and/or with incorrect values, to test the reaction of means of supporting the verification of the knowledge field to anomalies (negative testing);

• the use of synonymous events and intervals for subsequent experimental research of means for combining elements of the knowledge field obtained from sources of various typologies.

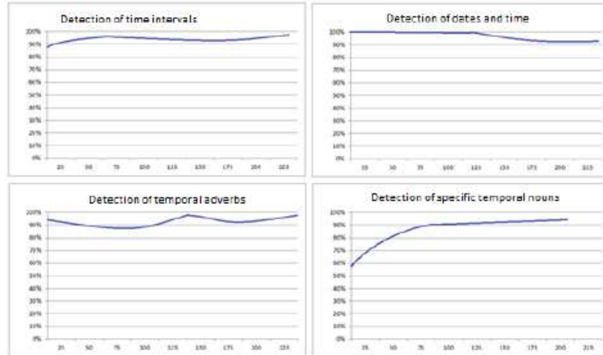


Figure 1. The results of experiments to detect temporal markers (the Oy axis is the percentage of detected markers of the total number, the Ox axis is the number of temporal markers)

Figure 1 shows some results in the form of statistical data obtained as a result of "language experiments".

Figure 2 shows examples of pie charts that display the quantitative result of these experiments.

Now the results of experiments are presented with other sources of knowledge - temporal databases. The experimental scenario included:

• registration of the temporal database in the dynamic version of the AT-TECHNOLOGY workbench (registration means adding a file containing the temporal database to the directory where the executable file is located);

• opening the database and reading data stored in the database, namely: identifiers of objects, classes and timestamps;

• creation of files with serialization of the ensemble of trees, a description of the knowledge field in the extended language of knowledge representation, as well as a description of the knowledge field in the internal representation.

It should be noted that the Random Forest algorithm was tested on several temporal databases that have the same structure, but a different number of objects, classes, timestamps, and also with different formats of times-

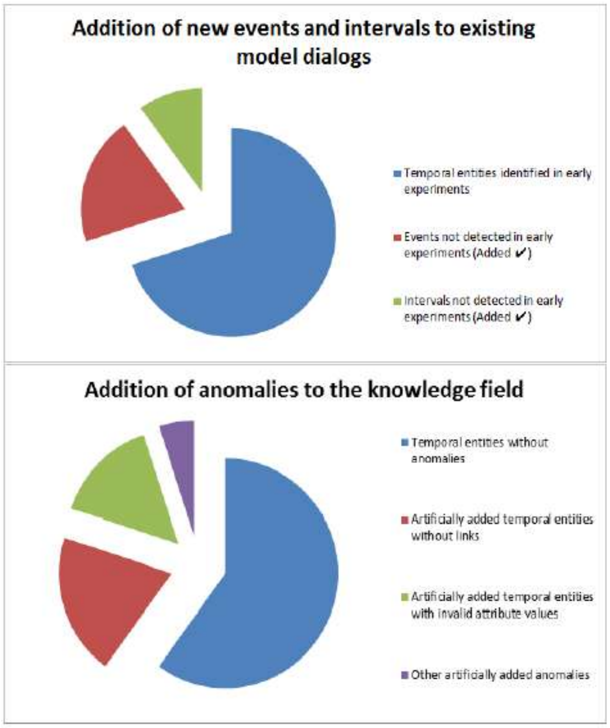


Figure 2. Pie charts with experimental results of addition new temporal entities

	Echostructure	Echogenicity	Circuit	Inclusions	The size
1	Cystic	Anechogenic	Smooth	No inclusions	1cm
2	Almost completely cystic	Hyperechoic	Unable to determine	Large comet tail artifacts	1 cm to 1.5 cm
3	Spongy	Isoechogenic	Lobed	Macrocalcinates available	From 1.5 cm
4	Mixed solid cystic	Hypoechoic	The appearance of extra-thyroid distribution	Peripheral calcification	
5	Solid				
6	Almost completely solid				

Figure 3. The name of the classifying attributes with possible variants of their values (Class attribute: 0/1 (biopsy required / not necessary). Number of objects: 140)

tamps. In addition, during testing, ensembles of 1, 50, and 100 trees were built.

Consider in more detail some of the features of temporal databases used in experiments. The temporal database is represented as a set of two tables (Objects, Classes), in turn Objects (id, attrN, timestamp), where id is the identifier column, attrN is the column with the attributes of the object, timestamp is the column for storing the timestamp, and Classes (id, class), where id is the identifier column, class are the classes to which objects can belong. Figure 3 shows a description of the model temporal database fragment.

Figure 4 shows the example of mapping of a temporal database into elements of a knowledge field in an extended language for representing knowledge.

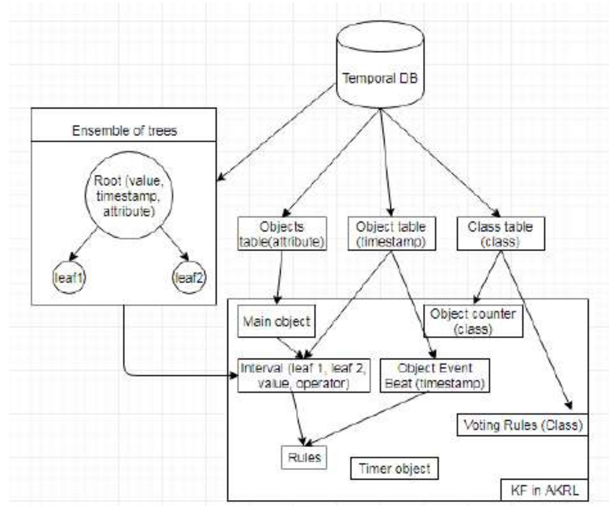


Figure 4. Mapping of the temporal database to the elements of the knowledge field in the extended language of knowledge representation

VI. Conclusion

The developed methods, algorithms and technologies of temporal knowledge acquisition from various sources (experts, NL-texts, DB) are especially important for medical PD, where significant volumes of temporal information are accumulated even about one patient, including all his previous conditions and diseases in a wide time range. This is a crucial task, which is necessary to improve the quality of healthcare in our country.

Acknowledgment

This work was supported by RFBR (project No. 18-01-00457).

References

[1] Rybina G.V. Theoriya i tekhnologiya postroeniya integrirovannykh expertnikh sistem. Monographiya. M .: Nauchtekhlitizdat. 2008, – 482 p.

[2] Rybina G.V. Intellektualnye systemy ot A do YA. Seriya monographiy v 3 knigakh. Kniga 2. Intellektualnye dialogovie systemy. Dynamicheskie intellektualnye systemy. M .: Nauchtekhlitizdat. 2015, – 160 p.

[3] Rybina G.V. Sovremennye arkhitekturi dynamicheskikh intellektualnykh system: problem integratcii i sovremennye tendentsii // Pribory i Systemi. Upravlenie, Kontrol, Diagnostika, No.2, 2017, p. 1-12.

[4] Rybina G.V., Danyakin I.D. The combined method of automated temporary collection of information for the development of knowledge bases of intelligent systems. Materials of the 2nd International Conference of 2017 on the Development and Application of Knowledge, 2017, - London: IEEE, p. 117-123.

[5] Rybina G.V. Dynamicheskye integrirovannye expertnye systemi: tekhnologiya avtomatizirovannogo polucheniya, predstavlniya i obrabotki temporalnykh znanij // Informatcionnye izmeritelnye i upravlyayuschye systemi. 2018, V.16, No.7, p. 20-31.

[6] Tsvetkova L.A., Cherchenko O.V. Vnedrenye tekhnologiy Big Data v zdravookhranenie: otzenka tekhnologicheskikh I kommercheskikh perspektiv // Ekonomica nauki. 2016, V. 2, No.2, p. 139-150.

[7] Aggarwal S. S., Zhai S. Mountain text data. - Springer, 2012, – 535 p.

[8] Pusteyovsky J., Castano J., Ingria R., Sauri R., Gauzauskas R., Setzer A., Katz G. TimeML: Reliable specification of events and temporary expressions in the text // Materials on new directions in answering the question , 2003, – p. 28-34.

[9] Kaufmann M., Manjili A., Wagenas P., Fisher P., Kossmann D., Fireber F., May N. Timeline index: Undefined data structure for processing requests for temporary data in SAP HANA. In the book: SIGMIOD, 2013.

[10] Van Ishaq, Van Hussein and Ku Mahamud, Ku Ruhana and Mary Norawi, Nortia. Extraction of temporary formation data using sliding window technology // CiiT International Journal of Mining and Knowledge Engineering, 2011, Vol. 3, No.8., p. 473-478.

[11] Tsacheva A.A., Bhagavati A., Ganesan P.D. MR is a random forest algorithm for detecting distributed action rules. // International Journal of Data Mining and Knowledge Management Process, 2016, Vol. 6, No.5., p. 15-30.

[12] Rybina G.V. Kominirovanniy metod priobreteniya znanij dlya postroeniya baz znanij integrirovannikh expertnikh system // Pribory i Systemi. Upravlenie, Kontrol, Diagnostika. 2011, No.8., p. 19-41.

[13] Verhagen M., Gaizauskas R., Schilder F., Hepple M., Katz G., Pustejovsky J. SemEval-2007 Task 15: TempEval Temporal Re-lation Identification // SemEval 07 Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg: Asso-ciation for Computational Linguistics. 2007. p. 75–80.

[14] Pan E. Learning Temporal Information from Text / Encyclopedia of Data Warehousing and Mining, Second Edition, Montclair State University. USA. 2009. p. 1146–1149.

[15] Mani I., Verhagen M., Wellner B., Lee Ch. M., Pustejovsky J. Machine learning of temporal relations // In ACL-44:Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Morristown, NJ, USA. Association for Computational Linguistics. 2006. p. 753–760.

[16] Schilder F. Temporal Relations in English and German Narrative Discourse. University of Edinburgh. College of Science and Engineering. School of Informatics. 1997. 216 p.

[17] Yoshikawa K., Riedel S., Asahara M., Y. Matsumoto Y. Jointly Identifying Temporal Relations with Markov Logic // In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec. Singapore. 2009. p. 405–413.

[18] Hagege C., Tannier X. XTM: A Robust Temporal Text Processor // CICLing 08 Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, Springer-Verlag Berlin, Heidelberg. 2008. p. 231–240.

[19] Pustejovsky J., Castano J., Ingria R., Sauri R., Gauzauskas R., Setzer A., Katz G. TimeML: Robust Specification of Event and Temporal Expression in Text // In Proceedings of New Directions in Question Answering. 2003. p. 28–34.

[20] Yankovskaya A.E., Kolesnikova S.I., Bukreev V.G. Vyyavlenie zakonomernostey vo vremennykh ryadakh v zadachakh raspoznava-niya sostoyaniy dinamicheskikh obektov. Tomsk: TPU. 2010. 254 p.

[21] Zaboleeva-Zotova A.V., Famkhyng D.K., Zakharov S.S. Gibridnyy podkhod k obrabotke vremennoy informatsii v tekste na russkom yazyke // Trudy 11-y natsionalnoy konfer. po iskusstvennomu intellektu s mezhdunarodnym uchastiem. M.: URSS. 2008, p. 228–235.

[22] Rybina G.V., Dejnenko A.O. Raspredelennoe priobretenie znanij dlya avtomatizirovannogo postroeniya integrirovannikh expertnikh system // Iskusstvennij intellect I prinyatie reshenij. 2010, No.4., p. 55-62.

[23] Rybina G.V., Mozgachev A.V. Realizatsiya temporalnogo vivoda v dinamicheskikh integrirovannikh expertnikh systemakh // Iskusstvennij intellect I prinyatie reshenij. 2014, No.1., p. 34-45.

[24] Efimenko I.V. Semantika vremeni modeli, metodi i algoritmi identifikatcii v systemakh avtomaticheskoj obrabotki estestvennogo yazyka // Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta Seriya "Lingvistika". No.2, - M.: Publishing House of Moscow State University, 2007.

[25] Arutyunva N.D., Yanko T. E. Logical analysis of the language: Language and time / Edited by: N.D. Arutyunva, T. E. Yanko. - M.: Indrik, 1997.

[26] Rybina G.V., Sorokin I.A., Sorokin D.O. Dynamic integrated expert systems: automated construction features of temporal knowledge bases using problem-oriented methodology // Open Semantic Technologies for Intelligent Systems Research Papers Collection. Issue 3. Minsk:BSUIR. 2019, p. 129-132.

Технология получения темпоральных знаний из различных источников: экспериментальное исследование комбинированного метода приобретения знаний

Рыбина Г.В., Слиньков А.А., Буянов Д.Р.

В данной статье анализируются результаты экспериментального исследования автоматизированной технологии построения баз знаний для динамических интеллектуальных систем, а именно динамических интегрированных экспертных систем, на основе оригинального комбинированного метода приобретения знаний с темпоральными сущностями.

Рассматриваемый метод является неотъемлемой частью задачно-ориентированной методологии построения интегрированных экспертных систем, обладающих масштабируемой архитектурой и расширяемой функциональностью.

В фокусе внимания данной работы находятся некоторые аспекты применения и развития технологий извлечения знаний из различных источников (эксперты, ЕЯ-тексты, базы данных) с целью создания новых прикладных интеллектуальных технологий, которые могут использоваться, например, в сфере здравоохранения (персонализированная медицина, «умная» больница и др.).

Приводится общая характеристика базовых технологий извлечения знаний из различных источников, описываются особенности комбинированного метода приобретения знаний и средств его реализации, а также особенности языковых экспериментов и приобретения знаний из темпоральной базы данных. Проанализированы результаты экспериментального программного моделирования.

Received 13.12.2019

Some Aspects of Intelligent Technology of Construction of the Tutoring Integrated Expert Systems

Galina Rybina, Ilya Sorokin, Anton Korsachev, Maxim Dvorchik
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow, Russian Federation
Email: gvrybina@mephi.ru, galina@ailab.mephi.ru

***Abstract*—We analyze the experience of development and use of tutoring integrated expert systems in MEPhI educational process. These systems were created on the basis of problem-oriented methodology and intelligent software environment of AT-TECHNOLOGY workbench. The emphasis is on the peculiarities of the implementation of certain tasks of intellectual training, related to the identification of students' knowledge and skills to solve non-formalized problems.**

***Keywords*— artificial intelligence, integrated expert systems, problem-oriented methodology, AT-TECHNOLOGY workbench, intelligent software environment, automated planning, tutoring integrated expert systems, intelligent training.**

I. INTRODUCTION

Methods of intelligent planning, and their integration with knowledge engineering, proposed and described in detail in [1], [2]. and in other works, are the basis of a new technology for building one of the most common classes of intelligent systems - integrated expert systems (IES) [1]–[3] , the demand for which in modern conditions of extraordinary attention to the use of methods and technologies of artificial intelligence (AI) has increased significantly, especially in the light of priority areas of development and use of technologies defined by the Decree of the President of the Russian Federation (No. 490 of 10.10.2019).

For the construction of IES is created, actively used and constantly evolving problem-oriented methodology [1], the essence of which consists in modeling the conceptual architecture of IES at all levels of consideration of the integration processes in IES and focus on modeling specific types of non-formalized problems (NF-problems), relevant technology,knowledge based systems(KBS). AT-TECHNOLOGY workbench [1]–[4] provide intelligent support for the development of the IES at all stages of the life cycle (LC) of building and maintenance of the IES (a detailed description of the intelligent software environment of the AT-TECHNOLOGY workbench can be found in [1]–[4]and other works).

The problem of *tutoring* is the least formalized among the "typical tasks" considered in the problem-oriented methodology [1] , which is associated with the insufficient development of pedagogical and psychological theories of knowledge acquisition, formation of concepts, construction of conclusions and other problems. However, as shown in [1] ,this problem is easily decomposed into a sequence of simpler tasks (*diagnosis, interpretation, planning, design*, etc.), which in general allowed us to develop a new approach to the construction of intelligent tutoring systems (ITS) based on the architecture of *tutoring* IES, linking the solution of these problems with the construction of appropriate models - *tutoring* (diagnosis), *trainee* (planning, design), *explanation* (interpretation), etc.

Currently, considerable experience has been accumulated in the use of tutoring IES developed using the basic tools of the AT-TECHNOLOGY workbench in the educational process of the MEPhI. *Applied* ontologies are implemented and dynamically supported for all basic courses / disciplines, which together form a *generalized* ontology "Intelligent systems and technologies". A unified ontological space of *knowledge* and *skills* was created for the automated construction of *competence-oriented* models of specialists in the field of AI and software engineering.

A significant place in the framework of problem-oriented methodology is given to methods and means of intelligent support of the most labor-intensive all stages of LC construction of IES - analysis of system requirements and design. Here, as a conceptual basis of intelligent technology is the concept of the *intellectual environment model* [1], [2] , the main components of which are the *technological knowledge base* (KB) and *the intelligent planner*.

The accumulation of experience associated with the development of various architectures of the IES for specific problem domains (PrD) and classes of solved problems showed that the *prototyping* processes of tutoring IES [1]–[3] have the greatest complexity due to the great complexity of building a *model of the architecture* of tutoring IES, performed, as a rule, by a subject teacher with the help of a knowledge engineer, as well as the need to use a large number of separate software and information components of the AT-TECHNOLOGY workbench, implementing the basic functionality of tutoring IES (*building a model of the trainee, tutoring model, ontology of courses/disciplines*, etc.).

Therefore, the focus of this work is on the issues related to the experimental program research of prototyping processes of tutoring IES based on the use of intelligent technology for the development and maintenance of applied IES.

II. Features of intellectualization of prototyping processes of tutoring IES

Let us consider the features of the new technology for supporting prototyping processes of applied IES using the main components of the intelligent software environment [2]–[4].

The basic declarative component of the intelligent software environment model in accordance with [1] [1] is the *technological* KB, which contains knowledge about the accumulated experience of building the IES in the form of a set of *standard-design procedures* (SDP) and *reusable components* (RUC). An important operational component of the model of intelligent software environment is the means of intelligent planning of actions of knowledge engineers, which provide generation and execution of plans for building prototypes of IES, i.e. *intelligent planner* developed on the basis of integration of models and methods of intelligent planning with knowledge engineering methods applied in the field of IES [1], [4].

As input to generate plans for the development of prototypes of IES are an *architectural model* of a prototype IES described by a hierarchy of *extended data flow diagrams* (EDFD [1]), and technological KB contains set of SDP and RUC. Accordingly, the model of prototyping processes of IES [3], [4] includes the function of planning the actions of knowledge engineers to obtain the current prototype of ES for a specific PrD. The main task of the intelligent planner is to automatically generate plans (*global and detailed* [4]) based on the model of the architecture of the IES and a set of RUC from the technological KB, which significantly reduces the risks of erroneous actions of knowledge engineers.

Execution of the plan tasks is carried out with the help of operational (instrumental) RUC. In terms of SDP as a key element of algorithmic, intelligent planner at each moment of time produces a detailed building plan for the development of IES depending on the current state of the project (type of solving NF-problems [1], reflected on the architectural model) , features of PrD, the availability of the architectural model unprojected drives, etc..

In accordance with [1], [2] the general architecture of the AT-TECHNOLOGY workbench is built in such a way that all functionality is distributed, i.e. "spread" on the components registered in the environment of the workbench and operating under the control of the intelligent development support environment. In other words, these components are the RUC of the workbench and are implemented according to the results defined for the RUC [1].Currently, within the intelligent technology of building IES used two groups of RUC – components implementing the procedural capabilities of *procedure* RUC, and components that implement the capabilities of *information* RUC [1], [2].

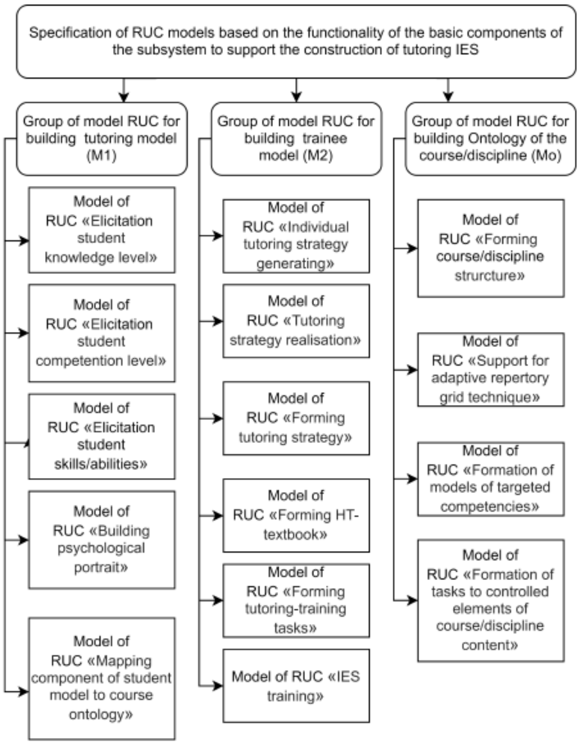


Figure 1. Specification of RUC models based on the functionality of the basic components of the subsystem to support the construction of tutoring IES

It is important to note that in the course of the research conducted, the basic components of the AT-TECHNOLOGY workbench (information and software) that implement the technology of building tutoring IES (Fig. 1) were specified on the basis of what the ontology of RUC is built. Ontology model RUC is a development of the basic ontology model proposed in the framework of problem-oriented methodology [1], [2], and is a semantic network where the vertices represent the different elements of the model tutoring objectives (tutoring model (M1), trainee model (M2), a model ontology (Mo), etc.), as well as a set of models of different training impact on the identification of knowledge and skills of trainees, and arcs show the variouslink types (relations between RUC, training impacts, educational-training tasks, fragments of of hypertext textbooks (HT-textbooks)). Currently, the RUC ontology has more than 50 peaks.

In the process of prototyping of IES intelligent planner, having knowledge of all SDP forms the set of tasks for the development of any prototype of IES (in accordance with LC of development), and then based on the requirements to the prototype of IES generated in the stage of analysis of system requirements, decompositum the development plan into small tasks (subtasks). All SDP are classified as follows [1]:SDP that do not depend on the type of task, for example, related to the processes of acquiring knowledge from different sources (experts, EY-texts, databases); SDP, depending on the type of task, for example, the construction of components of tutoring ICS; SDP, associated with the RUC, i.e. procedures that contain information about the LC of RUC from the beginning of its configuration to inclusion in the layout of the prototype ICS, as well as information about the tasks solved by this RUC, the necessary settings and, possibly, their values.

Complication of architecture of IES, and the emergence of the technological KB of a large number of SDP and RUC led to an increase of the complexity of the search increased the negative effect of nonoptimality of the solutions. Therefore, several years ago there was a need to improve the methods and algorithms of planning [1]used by the intelligent planner. The results of the complex analysis of modern methods of intellegent planning and the conducted experimental studies [4] have shown the expediency of using to solve the problem a fairly well-known approach associated with planning in the space of states [5]–[9].

The method of generation of the action plan of the knowledge engineer is to perform a sequence of transformations of the model architecture the current prototype of IES performed in 4 stages [4]: receiving generalized EDFD graph; generate an exact *cover* (i.e. a set of instances of the SDP from mutually disjoint fragments containing all vertices of the graph) using a heuristic search; generate plan of knowledge engineer based on the detailed coverage; generation of a presentation plan.

A detailed description of methods and algorithms for the implementation of intelligent planning is given in [1], [4] and others, and below in the context of this work, the results of experimental modeling for the SDP “Construction of tutoring IES”are considered.

III. Research of possibilities of application of intelligent technology of prototyping of tutoring IES (on the example of plan generation)

All of the above steps are performed by intelligent planner that fully implements the functionality associated with the planning of processes of prototyping IES. With the help of the EDFD hierarchy preprocessor, the DPD hierarchy is preprocessed by converting it into a single generalized maximum detail diagram (Fig.1). The task of covering the detailed EDFD with the existing SDP is implemented using the global plan generator, which on the basis of the technological KB and the built generalized EDFD provides the task, as a result of which the exact coverage is built, which is later converted into a global development plan.

The generator of the detailed plan is carried out detailing each element of the coating, i.e. on the basis of the obtained coating EDFD and technological KB is made detailing each element of the coating, thus forming a preliminary detailed plan.

Then, based on the analysis of available RUC and their versions (data about which are requested from the development process management component), the plan interpretation component generates a detailed plan, where each task is associated with a specific RUC and can be performed by a knowledge engineer. With the help of the building component of the final plan, the necessary representation of the plan is formed for its use by other components of the intelligent software environment (the visualization component of the plan (Fig.3), etc.).

The following example describes the representation method

First, we describe how to store the RUC ontology that the planner works with. The RUC ontology is stored as an XML-document [10] that defines how all elements are described using tags. First, the description of 4 types of ontology elements is presented (Fig. 2): operation - function, unformalized operation-nffunction, entity-entity, storage-store. The following is a description of the relationships between the elements-flow (Fig. 3), followed by a description of a group of strongly connected RUC - fragment (Fig. 4), where document is the file created as a result of the RUC work, and extension - its type. The end describes the chronology of the execution of RUC groups - network (Fig. 5), where the task - description of the work performed by the group RUC (input - description files necessary for the correct operation, output the description of the files created as a result of work, predecessor - a description of what the RUC group should interact to perform his current job, executor - name operational RUC, which will carry out the ongoing work).

```
<procedure name="Построение обучающей ИЭС">
 <nodes>
   <node id="41" name="Преподаватель" type="Entity" />
   <node id="31" name="создать &quot;тренинг с ИЭС&quot;" type="NFFunction" />
   <node id="21" name="создать УТЗ" type="Function" />
   <node id="22" name="создать ГТ-учебник" type="Function" />
   <node id="11" name="БД онтологий" type="Storage" />
```

Figure 2. A fragment of an XML document describing the PIC ontology (elements)

```xml
<flows>
  <flow data="задание" src_id="41" dest_id="21"/>
  <flow data="знания" src_id="41" dest_id="22"/>
  <flow data="знания" src_id="41" dest_id="31"/>
  <flow data="знания, опыт" src_id="41" dest_id="24"/>
  <flow data="знания, опыт" src_id="41" dest_id="23"/>
  <flow data="знания" src_id="41" dest_id="25"/>
```

Figure 3. A fragment of an XML document describing the PIC ontology (links)

```xml
<fragments>
  <fragment id="0" name="Основной" primary="1">
    <nodes>
          <node_ref id="32"/>
          <node_ref id="42"/>
          <node_ref id="12"/>
    </nodes>
    <documents>
      <document id="0" name="Поле знаний" extension="log"/>
      <document id="1" name="Протокол верификации" extension="log"/>
      <document id="2" name="База знаний"  extension="kbs"/>
```

Figure 4. A fragment of an XML document describing the PIC ontology (group of strongly connected RUC)

```xml
<network>
  <task id="0" fragment_id="0" name="Приобретение знаний"
        stage="Анализ системных требований" executor="KMPZ">
    <output>
      <document id="0"/>
      <document id="4"/>
    </output>
  </task>
```

Figure 5. A fragment of an XML document describing the PIC ontology (chronology of the execution of RUC groups)

IV. Conclusion

Thus, the experimental base in the form of accumulated information and software tutoring of IES for individual courses/disciplines that are used in the educational process of the department of Cybernetics at MEPhI since 2008 proved to be a good "testing ground" for the continuation and development of studies in developing elements of a new intelligent planning and control processes of building intelligent systems, including tutoring of IES.

It is important that a single conceptual basis for their development is a problem-oriented methodology, and as a tool - AT-TECHNOLOGY workbench.

In fact, there was a technological transition from "automation" to "intellectualization" of labor-intensive processes of design and maintenance of information and software tutoring IES, by creating conditions for effective use of the intelligent planner, in particular, to create a technological knowledge base (SDP, RUC information and operational nature), and then conduct full-fledged research on the creation of elements of new technology.

Acknowledgment

The work was done with the Russian Foundation for Basic Research support (project No 18-01-00457)

References

[1] 1. Rybina G.V. Teoriya i tekhnologiya postroeniya integrirovannyh ekspertnyh sistem. Monografiya. M.: Nauchtekhlitizdat, 2008, 482p.

[2] Rybina G.V. Intellektual'nye sistemy: ot A do YA. Seriya monografij v 3 knigah. Kniga 1. Sistemy osnovannye na znaniyah. Integrirovannye ekspertnye sistemy. - M.:Nauchtekhlitizdat. 2014. - 224 p.

[3] Rybina G.V. Intellektualnaya tekhnologiya postroeniya obuchayushchih integrirovannyh ekspertnyh sistem: novye vozmozhnosti. // Otkrytoe obrazovanie. Tom 21. No. 4. 2017 p.43-57

[4] Rybina G.V., Blokhin Y. M. Methods and Software Implementation of Intelligent Planning for Integrated Expert System Design // Scientific and Technical Information Processing. 2019. Vol. 46. No. 6. pp. 1–12.

[5] Nau D.S. Current trends in automated planning // AI Magazine, 2007, vol.28 No4 P.43–58

[6] Ghallab M., Nau D. S., Traverso P. Automated planning — theory and practice. Elsevier, 2004.

[7] Ghallab M., Nau D. S., Traverso P. Automated Planning and Acting, 1st ed. New York, NY, USA: Cambridge University Press, 2016.

[8] Geffner H., Bonet B. A Concise Introduction to Models and Methods for Automated Planning, ser. Online access: IEEE (Institute of Electrical and Electronics Engineers) IEEE Morgan &

[9] Claypool Synthesis eBooks Library. Morgan & Claypool, 2013. Gerevini A., Haslum P., Long D. et al. Deterministic planning in the fifth internation– al planning competition: PDDL3 and experimental evaluation of the planners // Artif. Intell.. 2009. Vol. 173. No5-6. Pp. 619–668.

[10] Furche T., Bry F., Bolzer O. XML perspectives on RDF querying: Towards integrated access to data and metadata on the web // Tagungsband zum 17. GI-Workshop u¨ber
Grundlagen von Datenbanken (17th GI-Workshop on the Foundations of Databases), Wo¨rlitz, Sachsen-Anhalt, 17.-20. Mai 2005. — 2005. Pp. 43–47.

Некоторые аспекты интеллектуальной технологии построения обучающих интегрированных экспертных систем

Рыбина Г.В., Сорокин И.А., Корсачев А.А., Дворчик М.А.

Мы анализировали опыт разработки и использования интегрированных экспертных систем обучения в учебном процессе НИЯУ МИФИ. Эти системы были созданы на основе основы проблемно-ориентированной методологии и интеллектуального программного обеспечения среды инструментального комплекса АТ-ТЕХНОЛОГИЯ. Выразительность речь идет об особенностях реализации тех или иных задач организации интеллектуальное обучение, связанное с идентификацией личности обучающихся знания и умения решать неформализованные задачи.

Received 13.12.2019

The standardization of intelligent computer systems as a key challenge of the current stage of development of artificial intelligence technologies

Vladimir Golenkov
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus
golen@bsuir.by

Vladimir Golovko
Brest State Technical University
Brest, Belarus
vladimir.golovko@gmail.com

Natalia Guliakina
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus
guliakina@bsuir.by

Viktor Krasnoproshin
Belarusian State University
Minsk, Belarus
krasnoproshin@bsu.by

***Abstract*—This work is devoted to the consideration of the most important factor providing semantic compatibility of intelligent computer systems and their components – standardization of intelligent computer systems, as well as standardization of methods and tools of their design.**

***Keywords*—intelligent computer system; integrated technology for the development of intelligent computer systems; knowledge base of an intelligent computer system; educational activities in the field of artificial intelligence; research activities in the field of artificial intelligence; development of artificial intelligence technologies; engineering activities in the field of artificial intelligence; convergence of scientific disciplines in the field of artificial intelligence; convergence of models for the representation and processing of knowledge in intelligent computer systems; general theory of intelligent computer systems; convergence of various methods and means of developing intelligent computer systems; semantic compatibility of intelligent computer systems and their components; hybrid intelligent computer system; semantic representation of knowledge; standard of intelligent computer systems.**

I. Introduction

The report at the OSTIS-2019 conference [1] examined the key problem for the current stage of development of artificial intelligence technologies: to ensure ***semantic compatibility*** of intelligent computer systems and *semantic compatibility* of various components of such systems (various types of knowledge that are part of knowledge bases, various problem solving models, various components of a multimodal interface).

This work is devoted to the consideration of the most important factor providing *semantic compatibility* of intelligent computer systems and their components – ***standardization of intelligent computer systems***, as well as standardization of methods and tools of their design.

The basis of the proposed approach to ensuring a high level of *learnability* and *semantic compatibility* of intelligent computer systems, as well as to the development of ***standard of intelligent computer systems*** is the unification of ***sense representation of knowledge*** in the memory of such systems and the construction of global ***sense space*** of knowledge.

II. The current state of artificial intelligence as a field of scientific and technical activities

artificial Intelligence
:= [An interdisciplinary (transdisciplinary) field of scientific and technical activity aimed at the development and application of ***intelligent computer systems***, which provide automation of various spheres of human activity]
:= [The field of human activity aimed at (1) building the theory of *intelligent computer systems* (2) development of technologies (methods and tools) for their design, (3) development of applied *intelligent computer systems*]
⇐ *decomposition**:
{
- *educational activities in the field of artificial intelligence*

- *research activities in the field of artificial intelligence*
- *development of integrated technologies in the field of artificial intelligence*
- *development of a set of intelligent computer systems*
- *business in the field of artificial intelligence*

}

⇒ *feature**:

[A feature of artificial intelligence as a field of scientific and technical activity is that it has a pronounced interdisciplinary, interdisciplinary, integration, collective character. Success here is largely determined by the consistency of actions and the level of compatibility of the results of these actions.]

Despite the presence of ***serious scientific results*** in the field of *artificial intelligence*, the pace of development of ***market of intelligent computer systems*** is not so impressive. There are several reasons for this:

- there is a big gap between scientific research in the field of *artificial intelligence* and the creation of industrial complex technologies for the development of *intelligent computer systems*. Scientific research in the field of *artificial intelligence* is mainly concentrated on the development of new methods for intelligent problems solving;
- these studies are fragmented and the need for their integration and the creation of a general formal theory of intelligent systems is not recognized, that is, there is a "Babylonian crowd" of various models, methods and tools used in *artificial intelligence* in the absence of awareness of the problem of providing their compatibility. Without a solution to this problem, neither a general theory of intelligent systems can be created, nor, therefore, an integrated technology for *intelligent computer systems* development, available to both engineers and ***experts***;
- the specified integration of models and methods of artificial intelligence is very complicated, because it is interdisciplinary in nature;
- intelligent systems as design objects have a significantly higher level of complexity compared to all the technical systems that humanity has dealt with;
- as a consequence of the above, there is a big gap between scientific research and engineering practice in this area. This gap can only be filled by creating an evolving technology for *intelligent computer systems* development, the evolution of which is carried out through the active cooperation of scientists and engineers;
- the quality of the development of applied intelligent systems to a large extent depends on the mutual understanding of experts and knowledge engineers. Knowledge engineers, not knowing the intricacies of the application field, can make serious mistakes in the developed knowledge base. The mediation of knowledge engineers between experts and the developed knowledge base significantly reduces the quality of the developed intelligent computer systems. To solve this problem, it is necessary that the language of knowledge representation in *knowledge base* be "convenient" not only for intelligent systems and knowledge engineers, ***but also for experts***.

The current state of ***artificial intelligence technologies*** can be characterized as follows [2]–[7]:

- There is a large set of particular *artificial intelligence technologies* with appropriate tools, but there is no general *theory of intelligent systems* and, as a result, there is no general *complex design technology for intelligent computer systems*;
- The compatibility of particular artificial intelligence technologies is practically not realized and, moreover, there is no awareness of such a need.

The development of *artificial intelligence technologies* is substantially hampered by the following socio-methodological circumstances:

- The high social interest in the results of work in the field of *artificial intelligence* and the great complexity of this science give rise to superficiality and dishonesty in the development and advertising of various applications. Serious science is mixed up with irresponsible marketing, conceptual and terminological sloppiness and illiteracy, throwing in new absolutely unnecessary effective terms, confusing the essence of the matter, but creating the illusion of fundamental novelty.
- The interdisciplinary nature of research in the field of *artificial intelligence* substantially complicates this research, since work at the intersections of scientific disciplines requires high culture and qualifications.

To solve the above problems of the development of *artificial intelligence technologies*:

- Continuing to develop new formal models for *intelligent problems* solving and to improve existing models (logical, neural networks, production), it is necessary to ensure **compatibility** of these models both among themselves and with traditional models for solving problems that are not among the intellectual tasks. In other words, we are talking about the development of principles for organizing *hybrid intelligent computer systems* that provide solutions to **complex tasks** that require sharing and in unpredictable combinations of a wide variety of types of knowledge and a wide variety of problem solving models.
- A transition is needed from the eclectic construction of complex intelligent computer systems using different types of knowledge and various types of problem solving models to their deep integration, when the same presentation models and knowledge processing

models are implemented in different systems and subsystems in the same way.
- It is necessary to reduce the distance between the modern level of *theory of intelligent systems* and the practice of their development.

artificial intelligence
⇒ *development trends**:
{
- [Erasing interdisciplinary barriers between different areas of research in *artificial intelligence*.]
- [Transferring the emphasis from scientific research aimed at studying the phenomenon of intelligence and building formal models of intelligent processes to creating of industrial complex technology for *intelligent computer systems* design.]

}
⇒ *development problems**:
{
- [Lack of motivation among scientists to integrate their research results in the field of *artificial intelligence* within the framework of the general theory of intelligent systems.]
- [Insufficient level of motivation and consistency for the transition from the theory of intelligent systems to ***integrated technology for intelligent computer systems design***, ensuring their semantic compatibility.]
- [Lack of effective interaction between various activities that provide development of *artificial intelligence* (educational activities, research activities, technology development, engineering, the business in the field of artificial intelligence).]

}
⇒ *consequence**:
[The consequence of these problems is that the current state of artificial intelligence technologies does not provide the required development of the market for artificial intelligent systems.]
⇒ *what to do**:
[For the development of artificial intelligence technologies, close interaction between practical engineers, developers of new technologies and scientists in the field of artificial intelligence is necessary.]

artificial intelligence
⇒ *key development tasks**:
{
- [the convergence of various areas of scientific research in the field of artificial intelligence in order to create *general theory of intelligent systems*]
- [integration of the existing variety of models, methods and tools for developing various components of intelligent systems into *unified integrated technology*, providing intelligent systems development for automating various fields of activity]
- [ensuring of *semantic compatibility* of the developed intelligent computer systems]
- [integration and coordination of various activities that ensure the sustainable development of artificial intelligence:
  – educational activities aimed at training of specialists capable of effectively participating in the development of artificial intelligence;
  – artificial intelligence research activity;
  – activities aimed at the development of artificial intelligence technologies;
  – applied intelligent systems engineering;
  – of a business aimed at organizing and financially supporting all the above types of activities and, first of all, at introducing the developed systems.

  ]

}

The problem of creating a fast-growing market for semantically compatible intelligent systems is a challenge addressed to specialists in the field of artificial intelligence, requiring overcoming the "Babylonian crowd" in all its manifestations, the formation of a high culture of negotiability and a unified, consistent form of representation of collectively accumulated, improved and used knowledge.

Scientists working in the field of artificial intelligence should ensure the convergence of the results of different areas of artificial intelligence and build a general theory of intelligent computer systems and *integrated technology for semantically compatible intelligent computer systems design*, including the appropriate ***standards of intelligent computer systems*** and their components.

Engineers developing intelligent computer systems should collaborate with scientists and participate in the development of integrated technology for intelligent computer systems design.

III. The convergence of different activities in the field of artificial intelligence

The further development of artificial intelligence affects all forms and directions of activity in this area. We list the main areas of convergence in the field of artificial intelligence [8]–[10]:

- The convergence of various disciplines in the training of specialists in the field of artificial intelligence in order to form a holistic picture of **problems** in the field of *artificial intelligence*;
- The convergence of various scientific studies in the field of *artificial intelligence* in order to build ***general theory of intelligent computer systems***;
- Convergence of development methods and tools for various *intelligent computer systems* to create ***integrated technology for intelligent computer systems*** development, available to a wide range of engineers;

- Convergence of engineering activities in order to build an ecosystem of semantically compatible and effectively interacting intelligent computer systems that support semantic compatibility in the process of operation and evolution;
- The convergence of all these types of activities among themselves, aimed at ensuring their consistency.

As an example of the integration of these types of activities, we can mention the organization of training for a modern engineer of intelligent systems.

How to teach an engineer to feel confident in such conditions of rapidly developing technologies? To do this, there is only one way – to immerse him in a problem field, teach him to see problems in the current state of technology, give him the opportunity to participate in eliminating these problems, to form appropriate skills and a sense of ownership in the development of technologies. If engineers who are users of relevant technologies do not participate in its development and provide feedback, the pace and quality of technology development will be significantly reduced.

But for this technology should be open so that anyone can contribute to their development, and should be protected from unskilled or malicious actions. Therefore, we need the organization of free examination, the development of harmonization rules and the possibility of the existence of points of view that are not universally recognized.

In fact, the convergence between different types of activities in the field of artificial intelligence will really be realized only when every specialist in this field will participate and sufficiently orient in all these types of activities, i.e. at the same time will be a teacher, a scientist, and a participant in the development of integrated technology, and an engineer, and a businessman. The idea of the impossibility of such a combination is clearly exaggerated and is the strongest brake on the implementation of convergence processes in the field of artificial intelligence.

IV. Semantic compatibility of intelligent computer systems, semantic representation of knowledge and its standardization

The key problem of the current stage of development of the general theory of intelligent computer systems and the technology for their development is the problem of ensuring *semantic compatibility* (1) of the various types of knowledge that make up the knowledge bases of intelligent computer systems (2) of various types of problem solving models (3) of various intelligent computer systems in general [11]–[13].

To solve this problem, unification (standardization) of the form of knowledge representation in the memory of intelligent computer systems is required. A logical approach for such unification is the orientation to ***sense representation of knowledge*** in the memory of intelligent computer systems.

In the development of knowledge-driven computer systems, a key role is played by the way knowledge is represented internally. The quality of knowledge representation of a computer system is determined by how close this representation is to sense (semantic) representation. Our approach to formalizing the *sense* of the knowledge presented is based on the following principles [14].

The sense of any information construction is a structure that includes:

- described objects that can be entities of any nature;
- links between the described objects, which can also be considered as these links themselves (it follows from this that there may be links between links, as well as links between links and objects that are not links);
- typology of relationships and non-relationship objects. This typology can be reduced to links that connect the described objects (including relations) with the classes to which the described objects belong. Moreover, these classes are also one of the types of described objects.

It is necessary to distinguish (1) sense as a sense structure of the above kind, and (2) ***sense*** as a semantic representation of information, as an information construct in which objects (entities) of the above sense structure are replaced by unique signs of these objects and which, respectively, is isomorphic to the sense structure indicated above, (3) ***sense**** as a relation connecting informational constructions with the corresponding sense structures.

sense
:= [sense model of information construction]
:= [sense representation of the information construct]
:= [sense structure]
:= [sense construct]
:= [sense representation of information]

sense*
:= [Binary oriented relation connecting informational constructions with semantic equivalent semantic structures]

Thus, in the final analysis, the sense of any informational construction is a configuration of links between the described objects (entities). In other words, the information is contained not in the signs (in particular, in the names) of the described entities, but in the configuration of the revealed links between these entities. Moreover, the links between the signs of the described entities are information models of the links between these entities themselves.

At the same time, we emphasize that entities are different (material, abstract). The described entities include links, and classes of entities (concepts), and structures.

Following the principle of Occam's razor, it is necessary to rid the form of internal representation of information in the memory of an intelligent computer system of parts that are not directly related to the *sense* of the information presented.

"It is necessary to bridge the gap between the syntactic data structures, on the one hand, and the *sense* (meaning) expressed by them, on the other" [15].

If the signs of the described entities have an internal structure (for example, are words or phrases), then the configuration of the links between the entities being described is not explicitly displayed, but is "camouflaged" against the background of explicitly and implicitly defined links between fragments of the signs used (between letters, words, phrases).

All known languages perform two functions – communicative (as a means of exchanging messages between subjects) and cognitive (as a means of representing the information model of the described World).

The language of the internal representation of knowledge in the memory of a computer system is not required to perform a communicative function. The language of the internal representation of knowledge is only required to ensure that knowledge is stored in a form convenient for processing. Convenience of processing knowledge stored in memory is determined by:

- simplicity of information retrieval procedures for fragments of a stored knowledge base that satisfy specified requirements;
- simplicity of the procedures for integrating new knowledge added to the knowledge base;
- ease of implementation of inference procedures.

Thus, everything that provides only the communicative functions of the language can be excluded from the language of the internal representation of knowledge. The language of the internal representation of knowledge in the memory of a computer system, based on the formalization of ***sense*** of this knowledge, should fulfill only a cognitive function – be a means of representing *internal information model* of some described World (including the external environment of the corresponding computer system).

The signs that make up the internal representation of knowledge should not have an internal structure, in particular, should not be represented in the form of some name of the corresponding (designated) entity. *Sense* of each sign is determined solely by its links with other signs that make up the internal representation of the knowledge base of a computer system. In contrast, semantic analysis and understanding of messages (external texts) requires structuredness and easy recognition of signs. By the similarity of structures representing signs (for example, character strings), the syntactic equivalence of signs is determined, although in informal languages it does not always coincide with their semantic equivalence (i.e. with their synonymy).

Within the framework of the internal semantic representation of the knowledge base of a computer system, *synonymy* (duplication) of signs is excluded. Internal signs denoting the same entity must be "glued", identified. As a consequence of this, within the framework of each knowledge base, *semantic equivalence* (duplication) of its constituent fragments is excluded, i.e. fragments that carry the same information. At the same time, the possibility of the existence of *logically equivalent* fragments of knowledge bases when one fragment is a logical consequence of the second and vice versa remains.

The texts of the language of the internal sense representation of knowledge should be non-linear in contrast to the usual texts, because the configuration of links between the entities of the described World in the general case is not linear. Each entity described can be associated with an unlimited number of links with other entities. Moreover, for any group of entities there always exists a link connecting them – everything in the World is interconnected. The question is which links are appropriate and which are impractical to explicitly represent in the knowledge base. The linearity of familiar texts is the result of projecting a nonlinear World onto a linear (one-dimensional) space, which requires additional special language tools that provide not a description of this World itself, but its projection onto a linear space. It should be noted that ***specialized non-linear languages*** are widely used, for example:

- circuit diagrams;
- flowcharts;
- blueprints.

Unlike these languages, the language of the sense representation of knowledge should be a universal nonlinear language.

Powerful and simple means of transition from information to meta-information (in particular, from poorly structured data to related data) are introduced within the language of the internal sense representation of knowledge. For this, the texts that are part of the knowledge base are also considered as described entities, for the designation of which the corresponding signs are introduced, each of which is interpreted as a signs denoting the set of all the signs that make up the designated text, including the signs of all kinds of links that are included in it.

Since the language of the internal sense representation of knowledge should be universal, it should provide the representation of all kinds of knowledge:

- specifications of various entities;
- documentation of various technical systems;
- various subject domains (both static and dynamic);
- various types of domain ontologies;
- utterance texts;

- texts of proof of theorems;
- statement of tasks;
- formulations of task classes;
- texts of solutions to specific problems;
- ways to solve various classes of problems;
- descriptions of the evolutionary histories of various systems;
- descriptions of projects aimed at creating and improving various technical systems.

The use of the universal language of the internal representation of knowledge with the possibility of its unlimited expansion, if the need arises for the presentation of new types of knowledge, creates the conditions for the unlimited expansion of the fields of application of computer systems based on such an internal language.

Within the framework of the language of the internal sense representation of knowledge, the names, terms, designations used in the transmission and reception of external messages are also independent described entities that have their own internal signs, which are associated with the internal links of the internal signs of those entities that are called these external designations. All external languages are part of the external World described by it for the knowledge base.

Atomic fragments of the internal sense representation of the knowledge base are only signs. Moreover, each inner sign itself can be a described entity (meaning the sign itself, and not the entity denoted by this sign). In addition, each link between the described entities is also itself a described entity, which in its internal representation has its own internal sign and which is interpreted as a set, the elements of which are signs of entities linked by the described link. Thus, everything that is not related to the presentation of sense, but to the form of representation used, is excluded from the internal representation of knowledge. So, for example, in the internal representation of knowledge there are absent not only letters, words, phrases, but also delimiters, limiters, prepositions, conjunctions, pronouns, declensions, conjugations, etc.

The language of the internal sense representation of knowledge should not only be convenient for processing knowledge in a computer system, but should be understandable and "transparent" for both the developer of the computer system and its end user. For this, along with the development of the language of the internal semantic representation of knowledge, external languages close to it should be developed, which should be based on simple and quickly acquired procedures for translating texts from internal to external forms of their presentation. We emphasize that the principles of organizing the memory ***modern computers*** do not correspond to the principles of the semantic representation of information in their memory. Therefore, the level of compatibility of modern computer systems and their users, the level of their "mutual understanding" is clearly insufficient. The computer system should not be **"black box"** for users. To solve this problem, it is necessary to make the principles of organization, storage and processing of information in computer systems understandable and convenient for users by using the sense representation of information in the memory of a computer system.

The typology of signs that make up the internal sense representation of knowledge is completely determined by the typology of entities denoted by these signs. In this case, the basic typology of the described entities is distinguished, which defines the syntactic typology (alphabet) of internal signs.

Our proposed standard for the internal sense representation of knowledge in the memory of an intelligent computer system is called ***SC-code*** (Semantic Code) [16]. The signs included in the texts of *SC-code* are called ***sc-elements***. Each *sc-element* can be considered an invariant of the whole variety of presentation forms (in various languages and sign constructions) of the entity that is denoted by this *sc-element*. Such an invariant is only that the indicated sc-element denotes the corresponding entity. Therefore, *sc-element* has no form. In this sense, he abstracts from the form of his presentation within the framework of a particular sign construction.

SC-code
:= [Language of a unified sense representation of knowledge in the memory of an intelligent computer system]

Syntax of ***SC-code*** is determined by
- typology (alphabet) of ***sc-elements*** (atomic fragments of ***SC-code*** texts);
- rules for connecting (incidence) of sc-elements (for example, what types of *sc-elements* cannot be incident to each other);
- typology of configurations of *sc-elements* (links, classes, structures), links between configurations of *sc-elements* (in particular, set-theoretic)

SC-code denotational semantics is specified by
- semantic interpretation of *sc-elements* and their configurations;
- semantic interpretation of incidence of sc-elements;
- hierarchical system of subject domains;
- the structure of the concepts used in each subject domain (studied classes of objects, studied relations, studied classes of objects of relations from related subject domains, key instances of studied classes of objects);
- subject domain ontologies.

It should be emphasized that unification and the maximum possible simplification of ***syntax*** and ***denotational semantics*** of the internal language of intelligent computer systems are necessary because the overwhelming volume of ***knowledge*** stored in the knowledge base of an intelligent computer system are ***meta-knowledge*** describing the properties of other knowledge. Moreover, for the indicated reason, constructive (formal) development of the theory of intelligent computer systems is impossible without clarification (unification, standardization) and ensuring semantic compatibility of various types of knowledge stored in the knowledge base of an intelligent computer system. It is obvious that the variety of forms of representing semantically equivalent knowledge makes the development of a general theory of intelligent computer systems practically impossible. *meta-knowledge*, in particular, should include various kinds of logical statements and all kinds of programs, descriptions of methods (skills). Providing the solution of various classes of information problems.

We list the basic principles underlying ***SC-code***:

- The signs (notation) of all entities described in *sc-texts*, (SC-code texts) are represented in the form of syntactically elementary (atomic) fragments of *sc-texts* and, therefore, not having an internal structure, not consisting from simpler fragments of text, such as names (terms) that represent the signs of the described entities in familiar languages.
- Names (terms), natural-language texts and other informational constructs that are not *sc-texts* can be included in *sc-text*, but only as files described (specified) by *sc- texts*. Thus, the knowledge base of an intelligent computer system based on *SC-code* can include names (terms) that denote some described entities and are represented by the corresponding files. Each sc-element will be called internal (to indicate some entity, and the name of this entity represented by the corresponding file will be called the external designation of this entity. Moreover, each named (identifiable) *sc-element* is linked by an arc belonging to the relation "***be an external identifier***", with a node whose contents are an identifier file (in particular, a name) denoting the same entity as the above *sc-element*. An external designation may be not just a name (term), but also a hieroglyph, pictogram, voiced name, gesture. We especially note that the external designations of the described entities in an intelligent computer system based on *SC-code* are used only (1) to analyze the information received in this system from outside from various sources, and input (understanding and immersion) of this information into the knowledge base, as well as (2) for the synthesis of various messages addressed to various subjects (including to users).
- Texts of *SC-code* (*sc-texts*) generally have a non-linear (graph) structure, since (1) the sign of each described entity is included in the sc-text once and (2) each such sign may be incident to an unlimited number of other signs, since each described entity can be linked by an unlimited number of links with other described entities.
- The knowledge base, represented by the text of *SC-code*, is a graph structure of a special kind, the alphabet of elements of which includes set of nodes, set of edges, set of arcs, set of basic arcs – arcs of a specially selected type that provide structuring of knowledge bases, and there are also many special nodes, each of which has content, which is a file stored in the memory of an intelligent computer system. The structural feature of this graph structure is that its arcs and edges can connect not only a node with a node, but also a node with an edge or arc, an edge or arc with another edge or arc.
- All elements of the above graph structure are signs that are part of the text of *SC-code*. Those. all its nodes, edges and arcs are the designations of various entities. Moreover, an edge is a designation of a binary undirected link between two entities, each of which is either represented in the graph structure under consideration by a corresponding sign, or is this sign itself. An arc is a designation of a binary oriented link between two entities. A special-type arc (***base arc***) is a sign of link between a node denoting a certain set of elements of the considered graph structure and one of the elements of this graph structure that belongs to the specified set. A node that has content (a node for which the content exists but might not be currently known) is a sign of the file that is the content of this node. A node that is not a sign of a file can denote a material object, a primary abstract object (for example, a number, a point in some abstract space), some binary link, some set (in particular, a concept, structure, situation, event, process). At the same time, entities denoted by the elements of the graph structure under consideration can be permanent (always existing) and temporary (entities to which the period of their existence corresponds). In addition, entities denoted by the elements of the considered graph structure can be constant (specific) entities and variable (arbitrary) entities. Each element of the considered graph structure, which is a designation of a variable entity, is assigned a range of possible values of this designation. The range of possible values of each variable edge is a subset of the set of all kinds of constant edges, the range of possible values of each variable arc is a subset of the set of all possible constant arcs, the range of possible values of each variable node is a subset of the set of all possible constant nodes.
- In the considered graph structure, which is a representation of the knowledge base in *SC-code*, different elements of the graph structure can be, but should not exist, denoting the same entity. If a pair of such elements is detected, then these elements are glued together (identified). Thus, the synonymy of

internal designations in the knowledge base of an intelligent computer system based on *SC-code* is prohibited. In this case, the synonymy of external signs is considered normal. Formally, this means that several arcs belonging to the relation "***be an external identifier***" come out of some elements of the graph structure under consideration. Of all the indicated arcs belonging to the relation "***be an external identifier***" and emerging from one element of the considered graph structure, one (very rarely two) is selected by including them in the number of arcs belonging to the relation "***be the primary external identifier***". This means that the external identifier indicated in this way is not homonymous, i.e. cannot be used as an external identifier corresponding to another element of the considered graph structure.
- In addition to files representing various external symbols (names, characters, pictograms), files of various texts (books, articles, documents, notes, comments, explanations, can be stored in the memory of an intelligent computer system built on the basis of *SC-code* drawings, drawings, diagrams, photographs, video materials, audio materials).
- Any entity that requires a description can be designated as an element of the considered graph structure. We emphasize that the elements of the graph structure under consideration are not just the designations of the various entities described, but designations that are elementary (atomic) fragments of the sign structure, i.e. fragments, the detailed structure of which is not required for the "reading" and understanding of this sign structure.
- The text of *SC-code*, like any other graph structure, is an abstract mathematical object that does not require detailing (refinement) of its encoding in the memory of a computer system (for example, as an adjacency matrix, incidence matrix, list structure). But such detail will be required for the technical implementation of the memory in which sc-constructions are stored and processed.
- On the other hand, for the operation of intelligent computer systems based on *SC-code*, in addition to the method of abstract internal representation of knowledge bases (*SC-code*), several methods of external image of abstract sc-constructions convenient for users will be required and used in the design of the source texts of the knowledge bases of the indicated intelligent computer systems and source texts of fragments of these knowledge bases, as well as used to display to users various fragments of the knowledge bases user queries. *SCg-code* and *SCn-code* are proposed as such methods for external image of sc-constructions.
- The most important additional property of *SC-code* is that it is convenient not only for the internal representation of knowledge in the memory of an intelligent computer system, but also for the internal representation of information in the memory of computers specially designed for interpreting semantic models of intelligent computer systems. That is, *SC-code* defines the syntactic, semantic and functional principles of organizing the memory of new generation computers, oriented to the implementation of intelligent computer systems – the principles of organizing graphodynamic associative semantic memory.

V. From standardization of the semantic representation of knowledge in the memory of intelligent computer systems to the standardization of intelligent computer systems

After we have defined the standard of *universal* (!) internal language of the sense representation of knowledge in the memory of an intelligent computer system, we can proceed to refine the standard of intelligent computer systems based on the specified language. Since this language is universal, it is possible to describe the intellectual computer system itself with a sufficient degree of detail and completeness, in the memory of which the specified description is stored. The integrated set of knowledge stored in the memory of an intelligent computer system and *sufficient* (!) For the functioning of this system is called ***knowledge base*** of the specified system. *Knowledge Base* of an intelligent computer system includes:
- a description of the facts and laws of the external environment in which the intelligent computer system functions ("dwells") and, in particular:
  - a description of external actors (for example, users) with whom the system interacts (a description of their properties, a description of the situations and events associated with them, a description of protocols for direct interaction with them);
  - a description of the syntax and semantics of the languages of communication with external entities;
  - a description of the facts of the behavior of an intelligent computer system in the external environment;
  - a description of the rules (skills) of the behavior of an intelligent computer system in the external environment;
- a description of the rules of behavior of internal entities of an intelligent computer system that perform actions (information processes) in the memory of an intelligent computer system (we will call such entities internal agents of an intelligent computer system);
- a description of the information processes themselves, planned, executed, or executed in the memory

of an intelligent computer system (a description of the behavior of an intelligent computer system in the internal environment);
- a description of methods (skills) that provide solutions to the corresponding classes of information problems – a description of the rules of behavior of internal agents of an intelligent computer system in its memory, which is the internal environment of an intelligent computer system.

Thus, the standardization of intelligent computer systems is determined by language means of knowledge bases structuring, means of systematizing the various types of knowledge that are part of knowledge bases.

At the same time, the intelligent computer system itself is considered as a system consisting of two main components:
- knowledge base, which is a complete description of this intelligent computer system;
- universal interpreter of knowledge bases of intelligent computer systems, consisting of:
  - memory in which the processed knowledge base is loaded and stored;
  - a processor that provides a direct interpretation of the knowledge base stored in the above memory.

Note that in the hardware implementation of a universal interpreter of knowledge bases that have a sense representation, the line between memory and the processor of the interpreter can be blurred. Those. the knowledge base interpreter can be implemented as a processor-memory in which processor elements will be connected to memory elements.

The above *completeness* description of an intelligent computer system in the knowledge base of this system itself is determined by the following properties and capabilities of intelligent computer systems.

Intelligent computer system is ***knowledge-based computer system*** and controlled by its knowledge. Those, the basis of an intelligent computer system is its knowledge base, which is a systematic information picture (information model) of the world (environment) in which an intelligent computer system operates. This environment is understood as ***external environment*** of an intelligent computer system, and its ***internal environment***, which is the knowledge base stored in memory of intelligent computer system.

An intelligent computer system is a computer system that has a high degree of ***learnability***, which boils down to expanding the *external environment* of one's "habitat" (functioning), to expanding one's *internal environment* (of one's knowledge base) with new declarative knowledge and new skills, to an improved quality of one's knowledge base (improving the quality of structuring a knowledge base, observing the principle of Occam's razor, minimizing contradictions, information holes, information garbage).

The high degree of *learnability* of intelligent computer systems is determined by the high speed of almost unlimited expansion of knowledge and skills of the system (and, including, the expansion of the variety of types of acquired knowledge and the variety of types of acquired skills – types of problem solving models). Such learnability is provided by:
- semantic compatibility of the knowledge used (including knowledge of various types) – the presence of an automated method for various knowledge integration;
- semantic compatibility of the skills used – the possibility of associative use of any required knowledge (including the same) in interpreting (performing) any skills (including recently acquired);
- the high level of flexibility of an intelligent computer system —- the low complexity of intelligent computer systems modifying at all levels while maintaining the integrity of the system (in particular, the complexity of modifying stored knowledge and skills used);
- high level of stratification;
- a high level of reflectivity – the ability to introspection and, including, to reasoning.

Intelligent computer system is a social subject that is able to exchange information and coordinate with other intelligent systems (including people) in the direction of achieving corporate goals, as well as maintain a sufficient level of semantic compatibility (mutual understanding) with other entities to prevent the "Babylonian crowd" syndrome.

An intelligent computer system is a computer system capable of solving a *integral* (!) complex of tasks that ensure the effective functioning of a computer system in the corresponding "habitat" environment. This includes
- means of *perception* of the current state (situations and events) of the "habitat" environment,
- means of *analysis* of this state,
- means of *goal-setting* (generation of tasks to be solved, with the specification of their priority),
- means of solving initiated tasks (relevant skills and interpreters of these skills),
- tools of targeted impact on the environment "habitat", i.e. tools of changing the state of this environment.

The basis of our approach to ensuring semantic and logical (functional) compatibility of knowledge representation and processing models is based on the following principles:
- *universal* way of *sense* representation of knowledge – *SC-code*;
- hierarchical system of formal ontologies represented in *SC-code* and ensuring compatibility (consistency) of the concepts used;
- general abstract graphodynamic associative memory, integrating all the knowledge used by the intelligent

system;
- *hierarchical* system of agents over the specified memory;
- programming tools for these agents – tools for decomposing (reducing) agents to agents of a lower level (to interpreting agents).

We emphasize that the development of an intelligent computer system standard is a prerequisite not only for ensuring the semantic compatibility of intelligent computer systems, but also for the formation of a developing market of *industrial* (!) intelligent computer systems – i.e. for mass industrial development of intelligent computer systems in various fields [17], [18].

The standardization of intelligent computer systems that we propose is the basis for the technology of semantically compatible intelligent computer systems design that focus on the use of new generation computers specially designed for this purpose.

VI. Methodology for the development and continuous improvement of the standard of intelligent computer systems as a knowledge base of intelligent computer metasystem

The description of the standard of intelligent computer systems based on the semantic representation of knowledge, presented as a section of the knowledge base ***Intelligent Computer Metasystem*** (IMS.ostis) has the following structure:

- The subject domain and ontology of the internal language of the semantic representation of knowledge in the memory of an intelligent computer system (syntax and denotational semantics of the internal language).
- The subject domain and ontology of the external languages of the intelligent computer system, close to the internal sense language (syntax and denotational semantics of external languages).
- The subject domain and ontology of structuring the knowledge bases of intelligent computer systems.
- The subject domain and ontology of integrated problem solvers of intelligent computer systems.
- The subject domain and ontology of verbal and non-verbal interfaces of intelligent computer systems.

It is important not only to develop a standard of intelligent computer systems, but also to organize the process of permanent and rapid improvement and harmonization of this standard. If the standard of intelligent computer systems is presented in the form of a knowledge base of intelligent computer systems, which is specifically designed to automate the use and improvement of this standard and which *itself is built in strict accordance with this standard*, then the process of permanent improvement of the standard of intelligent computer systems becomes a process of permanent collective improvement of the knowledge base of the specified intelligent computer system.

This process is based on the following principles:

- According to the specified knowledge base, any user can carry out free *navigation*, asking a wide range of questions to the corresponding intelligent computer system.
- Any section of this knowledge base and even its entirety can be represented in the form of a semantically structured and ”readable” source text.
- Anyone can become a co-author of the next version of the standard. To do this, he needs to register accordingly and comply with the relevant rules of interaction of the authors of the standard.
- Direct editing of the current *version of the standard* is carried out *automatically* by special internal agents for knowledge bases processing on the basis of a fully completed review process and coordination of the editorial editing proposed by one of the authors. A sufficiently representative group of authors should participate in each such procedure.
- Any proposal aimed at improving the current version of the standard is subject to review. Such a proposal may be a new fragment of the standard, an indication of a contradiction, an information hole, a garbage fragment, a dubious fragment, the proposed editorial revision.
- If the proposed new fragment of the standard is the result of eliminating a previously identified contradiction, or an information hole, or the result of processing a previously identified doubtful fragment, then in the specialization of the specified new fragment, this should be clearly indicated.
- All actions of each author are recorded in the knowledge base in the section of the stories of its evolution with automatic indication of the author and the moment the action was performed.
- Based on the history of evolution of the standard, the level of activity of each author and the level of value of his contribution to the development of the standard are automatically determined.
- The types of activities of the author of the standard include:
  - the construction of the proposed new fragment of the standard, together with the corresponding explanations, notes, examples, as well as with reference to some signs used by the current version of the standard (primarily to the concepts used);
  - building a specification of the contradiction (in particular, errors) found in the current state of the standard;
  - building a specification of the information hole detected in the current state of the standard;
  - building a specification of the garbage (excess) fragment detected in the current state of the

standard and expected to be deleted;
- building a specification requiring revision of a dubious fragment of the current state of the standard indicating the directions of such revision;
- building specifications of the proposed editorial revision in the current state (current version) of the standard (deletion, replacement);
- building a review of the proposal made by another author, indicating his opinion ("I agree", "I agree in case the elimination of the relevant comments", "I disagree for the reasons indicated");
- construction of a repeated review of the revised proposal made by other authors (after the elimination of comments);
- building a revised version of your assumption after removing the comments made by reviewers.

- In the process of collective improvement of the standard, special attention should be paid to achieving consensus (consensus) on issues such as:
  - distinguished entities (primarily concepts) and their unambiguous basic formal **specifications** (for concepts – definitions);
  - *basic* terms attributed to selected entities (including concepts);
  - distinguished subject domains and their structural specifications (classes of objects of study, studied relations defined on objects of study).

To organize the coordination of work on the development and continuous improvement of the standard of intelligent computer systems and to build the appropriate infrastructure, it is necessary to create a Consortium for the standardization of intelligent computer systems, ensuring their semantic compatibility.

VII. Standards as a kind of knowledge

Standards in various fields are the most important type of knowledge, ensuring the coherence of various types of mass activity. But so that the standards do not inhibit progress, they must be constantly improved.

Standards must be used effectively and competently. Therefore, the design of standards in the form of text documents does not meet modern requirements.

Standards should be in the form of intelligent help systems that are able to answer a variety of questions. Thus, it is advisable to formalize the standards in the form of knowledge bases, corresponding intelligent reference systems. Moreover, these intelligent reference systems can coordinate the activities of standards developers aimed at improving these standards [19]–[21].

From a semantic point of view, each standard is a hierarchical ontology, clarifying the structure and concept systems of their respective subject areas, which describes the structure and functioning of either a certain class of technical or other artificial systems, or a certain class of organizations, or a certain type of activity.

Obviously, to build an intelligent reference system for the standard and an intelligent system for supporting collective improvement of the standard, formalization of the standard is necessary in the form of an appropriate formal ontology.

The convergence of various activities, as well as the convergence of the results of these activities, requires deep semantic convergence (semantic compatibility of the relevant standards), which also urgently requires formalization of standards.

It should also be noted that the most important methodological basis for formalizing standards and ensuring their semantic compatibility and convergence is the construction of a hierarchical system of formal ontologies and observance of the Occam razor principle.

VIII. Experience in integrating various types of knowledge and various problem solving models on the basis of their standardization

On the basis of the proposed standard for the semantic representation of knowledge, means have been developed that provide coordinated integration (the possibility of sharing when solving problems within the same intelligent system) of various types of knowledge and various problem solving models.

Integration of various types of knowledge is ensured by the presence of a model of hybrid knowledge bases, considered in [22] and, in turn, including a hierarchical family of top-level ontologies that provide semantic compatibility of various types of knowledge. Based on this model, it is supposed to integrate such types of knowledge as facts, specifications of various objects, logical statements (definitions, axioms, theorems), events, situations, programs and algorithms, processes, problem formulations, domain models, ontologies and others.

Integration of various problem solving models (knowledge processing models) is provided within the framework of the hybrid problem solver model considered in [23]. In the framework of this model, the solver is interpreted as a hierarchical model of agents interacting with each other by specifying the actions they perform in the general semantic memory. Based on this approach, the integration of such problem-solving models as logical (including clear, fuzzy, reliable, plausible, etc.), neural network models, genetic algorithms, various strategies for finding ways to solve problems is supposed. One of the components of the hybrid task solver model is a basic programming language oriented to the processing of semantic networks, on the basis of which it is proposed to develop programming languages of a higher level.

The considered approaches to the integration of various types of knowledge and problem-solving models have found application in a number of works, including joint ones:

- The paper [24] proposed an approach to the integration of neural network image processing models and logical inference models to build a decision support system in production;
- In the works [25] and other works of the authors, types of knowledge are considered, the presentation of which is necessary for building a comprehensive integrated knowledge base of an enterprise, examples of formalization of specific types of knowledge are given;
- The paper [26] proposed an approach to the integration of neural network and semantic models to improve the quality of image recognition;
- The paper [27] and other works of the authors considered approaches to improving the quality of recognition of voice messages due to the context described in the knowledge base;
- A number of prototypes of reference and training systems have been developed, within the framework of which various types of knowledge are integrated (facts, logical statements, programs, illustrations, examples, etc.) and various problem solving models within the framework of the corresponding subject domains are applied.

IX. From standardization of intelligent computer systems to standardization of their design technology and to the construction of an intelligent knowledge portal that ensures the efficient use and improvement of these standards

From the standard of intelligent computer systems built on the basis of the semantic representation of knowledge, from the intelligent computer metasystem discussed above that supports the operation and improvement of this standard, it is easy to move to the standard of integrated technology for designing an intelligent computer system based on a semantic representation of knowledge, which includes the specified standard of intelligent computer system, and the standard of methods for their development, and the standard of their development tools. Moreover, the standard for the specified integrated technology can be brought into line with the metasystem that supports the operation and improvement of the standard of this technology. We called this metasystem ***IMS.ostis Metasystem*** (Intelligent MetaSystem for Open Semantic Technology for Intelligent Systems).

IMS.ostis Metasystem
:= [Intelligent computer metasystem for design support of intelligent computer systems based on the semantic representation of knowledge]
:= [Intelligent knowledge portal that ensures the efficient use and improvement of the standard of intelligent computer systems, built on the basis of the sense representation of knowledge, as well as the standard of methods and design tools for these intelligent computer systems]

IMS.ostis metasystem knowledge base has the following structure:

- Description of the standard of intelligent computer systems based on the sense representation of knowledge (the structure of this standard was discussed above).
- Description of the standard design methods for intelligent computer systems based on the semantic representation of knowledge.
- Description of the standard design tools for intelligent computer systems based on the semantic representation of knowledge.

The considered knowledge base will be published in the form of the source text of this knowledge base, which, we hope, will be quite "readable" for a wide range of specialists due to the "transparent" formalized structuring of this text and the inclusion of informal information structures (explanations, notes, images) [1], [28].

X. Standardization of the global semantic space and prospects for the development of science, education, industry and the knowledge market

The structuring of *semantic space* of knowledge is determined by a system of interconnected *subject domains* and their corresponding *ontologies*.

Subject domains and, accordingly, ontologies specifying them are different.

The subject domain and ontology of a certain class of technical systems is one thing and the subject domain and ontology of project activities (including techniques) aimed at developing technical systems of a specified class, as well as the subject domain and ontology of tools used in the development of these technical systems, are another thing.

In addition, each class of technical systems can be associated with a subject domain and an ontology of actions that ensure the efficient operation of technical systems of this class.

Thus, on the set of subject domains and ontologies corresponding to them, a whole family of relations connecting them is defined. We list some of them:

- be a particular subject domain or ontology in the set theoretic sense (for example, the subject domain of planar figures and the subject domain of polygons and the subject domain of triangles are connected this way);
- to be a particular subject domain or ontology in a spatial sense, this provides a detailed description of the *spatial parts* of objects studied in a given subject domain;

- be a particular subject domain or ontology in the temporal sense, which provides a detailed description of the temporal parts (stages of existence) of objects studied in a given subject domain.

The list of types of links between subject domains and ontologies can be continued. We emphasize that

- each scientific-technical or educational discipline and each standard in the semantic space is represented in the form of a hierarchical system of interconnected subject domains and ontologies;
- the convergence of different disciplines ultimately boils down to *increase in the number of connections* between subject domains and ontologies that are part of different disciplines.

The experience of convergence of various activities in the field of artificial intelligence can be used to implement a transdisciplinary approach in the development of science as a whole.

Moreover, the design of knowledge bases of intelligent computer systems should be based on the results of scientific research of the whole complex of scientific and technical disciplines. The knowledge bases of intelligent computer systems should make full use of the results of scientific research, and not distort them with the hands of knowledge base engineers.

Of fundamental importance for the development of transdisciplinary research and the deepening convergence of various disciplines is the concept of *formalized* (!) ***global semantic spaces***, which is the result of the abstract integration of knowledge bases designed in the standard of sense representation of knowledge and included in the portals of scientific and technical knowledge corresponding to *all kinds* (!) of scientific and technical disciplines.

If each scientific and technical discipline will evolve not only within the framework of its formalized local sense space, focusing attention only on its class of objects of research and on its subject of research, but also simultaneously within the framework of the formalized *global semantic spaces*, then obviously interdisciplinary research will develop more intensively, constructively and accurately.

The system of evolving semantically compatible portals of scientific and technical knowledge is the basis of a new level of development of the organization of scientific and technical information and the organization of scientific activity in general. From a system of scientific books and articles, a transition to a system of semantically compatible databases of scientific and technical knowledge is necessary, in which duplication of knowledge is fundamentally excluded and in which the meaning (essence) of the contribution of each scientist to the development of the global semantic space of knowledge is clearly defined and localized.

Convergence, integration of the results of various scientific disciplines, transdisciplinarity of scientific research is the most important trend in the development of modern science:

"...the era of analyticism and its inherent differentiation of science and closed scientific theories is already behind us. It became apparent that the real problems facing human society are much more complicated than scientific ones, and science is not able to fundamentally solve them due to the fragmentation of scientific disciplines and their specialization, poor coordination of research teams and their subjects, lack of system monitoring and a common formalized language for representing knowledge." [8].

The most urgent is the need for the convergence of various disciplines:

- in the field of artificial intelligence to build a general theory of intelligent computer systems;
- in medicine for creating intelligent computer systems for integrated transdisciplinary diagnostics;
- in education for the formation and learners of a comprehensive picture of the world;
- in the integrated automation of enterprises to ensure semantic compatibility of all levels of enterprise management;
- in the complex informatization of scientific activity.

Problems that need to be solved at the stage of increasing the level of informatization of various areas of human activity, require the use of modern artificial intelligence technologies, are fundamental in nature and, therefore, for different areas of human activity are largely the same in the informatization of scientific activity and the informatization of education (Univercity 3.0), industry (Industry 4.0), healthcare.

Based on the use of the global semantic space of knowledge, it becomes possible to create a rapidly developing ***knowledge market***, within which

- the authorship of each action aimed at the development of global semantic space will be clearly fixed;
- will automatically (and therefore objectively) based on the analysis of the opinions of reviewers establish the fact of recognition (consistency) of each proposal aimed at the development of a global sense space;
- will automatically (and therefore objectively) evaluate the significance of the contribution of each author to the development of global sense space;
- will automatically (and therefore objectively) evaluate the practical relevance of knowledge created by each author recognized by the scientific community, based on an analysis of the frequency of use of this knowledge in problems solving, in the development of various intelligent computer systems;
- will be recorded even those points of view that are correctly framed, but did not receive the recognition of reviewers (over time, such recognition can be obtained).

XI. Prospects for the development of the market of intelligent computer systems based on their standardization

The main product of *OSTIS Technology*, focused on the design of semantically compatible intelligent computer systems based on the sense representation of knowledge in the memory of these systems, is not the intelligent computer systems themselves, but the sociotechnical ***OSTIS Ecosystem***, which is a self-organizing system consisting of interacting computer systems built using *OSTIS Technology* (*ostis-systems*), as well as their users (end users and developers), and constantly supporting compatibility and, as a result, a high level of "mutual understanding" between the indicated *ostis-systems*, as well as between these systems and their users.

Thus, the basis of *OSTIS Technology* is the constantly evolving standards that provide the specified compatibility and "understanding".

OSTIS Technology is a technology for the development of semantically compatible hybrid intelligent systems that have a high level of learnability and constantly maintain the specified compatibility both among themselves and with their users.

We emphasize that *OSTIS Ecosystem* is a *hierarchical* multi-agent system, since the components (agents) of this multi-agent system [29] can be not only individual *ostis-systems* and their users, but and teams consisting of *ostis-systems*, users, as well as other such teams. Moreover, each *ostis-system* or user can be a member of several teams at once.

The transition from individual intelligent computer systems to an ecosystem of intelligent computer systems requires the development of

- ***standards*** of intelligent computer systems;
- ***standards evolution support tools*** of intelligent computer systems;
- ***compatibility support tools*** of intelligent computer systems during their evolution and operation;
- ***compatibility support tools*** of intelligent computer systems in the process of changing (evolution) of intelligent computer systems standards;
- ***integration automation tools*** of intelligent computer systems;
- ***means of coordination*** the activities of intelligent computer systems within various teams.

XII. Conclusion

The creation of *OSTIS Technology* will allow to solve a number of significant problems considered in this paper, which, in turn, will radically change the capabilities of a developer of intelligent computer systems and expand the range of problems solved by such systems [30], [31].

To reduce the time needed to improve the quality of developed intelligent computer systems, it is necessary to organize *component* design of intelligent computer systems with a powerful and constantly expanding library of reusable components of intelligent computer systems, but for this it is necessary to provide ***semantic compatibility*** of the specified components.

For the implementation of cooperative purposeful and adaptive interaction of intelligent computer systems within the framework of automatically formed teams of intelligent computer systems, ***semantic compatibility*** of intelligent computer systems is required, and this, in turn, requires unification of intelligent computer systems. The unification of an intelligent computer system is possible only on the basis of ***general formal theory of intelligent computer systems*** and the corresponding ***standard of intelligent computer systems***, and this requires a deep convergence of different areas of research in the field of artificial intelligence.

Since the result of the development of artificial intelligence as a scientific discipline is the permanent evolution of the general theory of intelligent computer systems and the corresponding standard of intelligent computer systems, in order to increase the pace of development of artificial intelligence and, accordingly, the technology for intelligent computer systems development, it is necessary to create a portal of scientific and technical knowledge on artificial intelligence that provides coordination of the activities of specialists, as well as coordination and inter radio results of this activity.

To switch to a new technological structure in the field of artificial intelligence, which is based on ensuring the semantic compatibility of intelligent computer systems, a prerequisite for which is the standardization (unification) of intelligent computer systems, certain socio-psychological prerequisites are required – an appropriate motivation and a fairly high level of determination, as well as mathematical and systemic culture.

The key points of *OSTIS Technology* are:

- targeting at a new generation of computers specifically designed for the production of semantically compatible intelligent computer systems;
- compliance with the principle of Occam's Razor at all levels of intelligent computer systems and, first of all, on the verge between the software and hardware of computer systems using *semantic representation of information* in the memory of intelligent computer systems;
- convergence of various types of components of intelligent computer systems, as well as methods and means of their development;
- orientation to solving the problem of semantic compatibility of intelligent computer systems *at all levels* of the organization of their activities.

Very promising object of mathematical research is the concept of the global sense space represented in *SC-code*, which includes the signs of *all kinds* of entities and, accordingly, the signs of all kinds of links between these entities and/or their signs. The peculiarity of this space is that it has a unique combination of properties:

- property of objectivity (independence from the point of view of a particular subject);
- with topological properties, which allows it to be investigated using topology methods by specifying, in particular, the concept of semantic proximity;
- algebraic properties, which allows you to explore it using the theory of algebraic systems, category theory;
- graph-theoretic properties, which allows you to explore its structure using the methods and tools of graph theory.

REFERENCES

[1] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, "Methods and tools for ensuring compatibility of computer systems," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[2] L. S. Bolotova, *Sistemy iskusstvennogo intelekta: modeli i tekhnologii, osnovannye na znaniyakh [Artificial intelligence systems: knowledge-based models and technologies]*. Moscow: Finansy i statistika [Finance and statistics], 2012, (In Russ).

[3] T. A. Gavrilova, D. V. Kudryavtsev, and D. I. Muromtsev, *Knowledge Engineering. Models and methods: A Textbook*. Saint Petersburg: Lan', 2016.

[4] G. S. Osipov, *Metody iskusstvennogo intellekta [Artificial intelligence methods]*. Moscow: Fizmatlit [Fizmatlit], 2015, (In Russ).

[5] A. V. Kolesnikov, *Gibridnye intellektual'nye sistemy: Teoriya i tekhnologiya razrabotki [Hybrid intelligent systems: theory and development technology]*. Saint Petersburg: SPbGTU [SPsSIT], 2001, (In Russ).

[6] A. L. Yalovets, *Predstavlenie i obrabotka znanii s tochki zreniya matematicheskogo modelirovaniya: problemy i resheniya [Presentation and processing of knowledge from the point of view of mathematical modeling: problems and solutions]*. Kiev: Naukova dumka, 2011, (In Russ).

[7] V. A. Golovko and V. V. Krasnoproshin, *Neirosetevye tekhnologii obrabotki dannykh [Neural network data processing technologies]*. Minsk: Minsk BGU [Minsk BSU], 2017, (In Russ).

[8] A. V. Palagin, "Problemy transdistsiplinarnosti i rol' informatiki [the problems of transddisciplinary and the role of informatics]," *Kibernetika i sistemnyi analiz [Cybernetics and systems analysis]*, no. 5, pp. 3–13, 2013.

[9] A. E. Yankowska, "Analiz dannykh i znanii na osnove konvergentsii neskol'kikh nauk i nauchnykh napravlenii [analysis of data and knowledge based on convergence of several sciences and scientific directions]," in *materialy Mezhdunarodnoi konferentsii «IOI-8» [materials of the International conference «IOI-8»]*. Tomsk State University of Architecture and Construction, Paphos, 2010, p. 1.

[10] A. E. Yankovskaya, A. A. Shelupanov, A. N. Kornetov, N. N. Ilinskaya, and V. B. Obukhovskaya, "Gibridnaya intellektual'naya sistema ekspress-diagnostiki organizatsionnogo stressa, depressii, deviantnogo povedeniya i trevogi narushitelei na osnove konvergentsii neskol'kikh nauk i nauchnykh napravlenii [hybrid intelligent system of express diagnostics of organizational stress, depression, deviant behavior and anxiety of violators based on convergence of several sciences and scientific directions]," in *Trudy kongressa po intellektual'nym sistemam i informatsionnym tekhnologiyam «ISIT'17». Nauchnoe izdanie v 3-kh tomakh. [Works of congress on intelligent 17 scientific publication in 3 volumes]*, ser. T. 1. Stupin A. S. publishing House, Taganrog, 2017, pp. 323–329.

[11] G. V. Rybina, "Sovremennye arkhitektury dinamicheskikh intellektual'nykh sistem: problemy integratsii i osnovnye tendentsii [modern architectures of dynamic intelligent systems: integration problems and main trends]," *Pribory i sistemy. Upravlenie, kontrol', diagnostika [Instruments and systems. Management, control, diagnostics]*, no. 2, pp. 7–12, 2017.

[12] I. Z. Batyrshin, A. O. Nedosekin, A. A. Stetsko, V. B. Tarasov, A. V. Yazenin, and N. G. Yarushkina, *Nechetkie gibridnye sistemy [Fuzzy hybrid systems]*. Moscow: Fizmatlit [Fizmatlit], 2007, (In Russ).

[13] A. V. Kolesnikov and I. A. Kirikov, *Metodologiya i tekhnologiya resheniya slozhnykh zadach metodami funktsional'nykh gibridnykh intellektual'nykh sistem [Methodology and technology of the solution of difficult tasks by methods of functional hybrid intellectual systems]*. Moscow: Moskva IPI RAN [Moscow IIP RAS], 2007, (In Russ).

[14] V. V. Golenkov and N. A. Guliakina, "Structurization of sense space," in *Open semantic technologies for intelligent systems*. BSUIR, Minsk, 2014, pp. 65–78.

[15] M. S. Tsalenko, *Modelirovanie semantiki v bazakh dannykh [Simulating semantics in databases]*. Moscow: Fizmatlit [Fizmatlit], 1989, (In Russ).

[16] V. V. Golenkov and N. A. Guliakina, "Principles of building mass semantic technologycomponent design of intelligent systems," in *Open semantic technologies for intelligent systems*. BSUIR, Minsk, 2011, pp. 21–58.

[17] N. M. Borgest, "Strategii intellekta i ego ontologii: popytka razobrat'sya [strategies of intelligence and its ontology: an attempt to understand]," *Ontologija proektirovanija. [Ontology of designing]*, vol. 9, no. 9 (34), pp. 407–425, 2019.

[18] G. B. Evgenev, "Ekspertopediya kak sredstvo sozdaniya ontologicheskogo interneta znanii [expetopedia as a means of creating ontological internet of knowledge]," *Ontologija proektirovanija. [Ontology of designing]*, vol. 9, no. 3 (33), pp. 307–319, 2019.

[19] "IT/APKIT professional standards [Electronic resource]," Available at: http://www.apkit.webtm.ru/committees/education/meetings/standarts.php, (accessed 7.05.2012).

[20] A. I. Volkov, L. A. Reingold, and E. A. Reingold, "Professional'nye standarty v oblasti it kak faktor tekhnologicheskogo i sotsial'nogo razvitiya [professional standards in the field of it as a factor of technological and social development]," *Prikladnaya informatika [Journal of applied informatics]*, vol. 10, no. 2, pp. 37–48, 2015.

[21] P. Serenkov, V. Solomaho, V. Nifagin, and A. Minova, "Koncepcija infrastruktury standartizacii kak bazy znanij na osnove ontologij [the concept of a standardization infrastructure as an ontology-based knowledge base]," *Novosti. Standartizacija i sertifikacija. [News. Standardization and certification.]*, no. 5, pp. 25–29, 2004.

[22] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 99–118.

[23] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.

[24] V. Golovko, A. Kroshchanka, V. Ivashenko, M. Kovalev, V. Taberko, and D. Ivaniuk, "Principles of decision-making systems building based on the integration of neural networks and semantic models," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 91–102.

[25] V. Taberko, D. Ivaniuk, V. Kasyanik, V. Golovko, K. Rusetski, D. Shunkevich, and N. Grakova, "Design principles of integrated information services for batch manufacturing enterprise employees," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 215–224.

[26] N. Iskra, V. Iskra, and M. Lukashevich, "Neural network based image understanding with ontological approach," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system*

[Open semantic technologies for intelligent systems], BSUIR, Minsk, 2019, pp. 113–122.
[27] V. Zahariev, T. Lyahor, N. Hubarevich, and E. Azarov, "Semantic analysis of voice messages based on a formalized context," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, BSUIR, Minsk, 2019, pp. 103–112.
[28] V. V. Golenkov, N. A. Guliakina, N. V. Grakova, I. T. Davydenko, V. Y. Nikulenka, A. P. Eremeev, and V. B. Tarasov, "From training intelligent systems to training their development tools," in *Open semantic technologies for intelligent systems*, BSUIR, Minsk, 2018, pp. 81–98.
[29] V. B. Tarasov, *Ot mnogoagentnykh sistem k intellektual'nym organizatsiyam: filosofiya, psikhologiya, informatika [From multi-agent systems to the intellectual organizations: philosophy, psychology, informatics]*. Moscow: Moscow URSS, 2002, (In Russ).
[30] D. I. Spivak and R. E. Kent, "Ologs: A categorical framework for knowledge representation," *PLoS ONE*, vol. 7, no. 1, p. e24274, Jan. 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0024274
[31] Y. Manin and M. Marcolli, "Semantic spaces," 2016.

Стандартизация интеллектуальных компьютерных систем – ключевая задача текущего этапа развития технологий искусственного интеллекта

Голенков В.В., Головко В.А.,
Гулякина Н.А., Краснопрошин В.В.

Данная работа посвящена рассмотрению важнейшего фактора, обеспечивающего *семантическую совместимость* интеллектуальных компьютерных систем и их компонентов – ***стандартизации интеллектуальных компьютерных систем***, а также стандартизации методов и средств их проектирования.

Основой предлагаемого подхода к обеспечению высокого уровня *обучаемости* и *семантической совместимости* интеллектуальных компьютерных систем, а также к разработке ***стандарта интеллектуальных компьютерных систем*** является унификация ***смыслового представления знаний*** в памяти таких систем и построение глобального ***смыслового пространства*** знаний.

В работе рассмотрены текущие проблемы в области развития технологий искусственного интеллекта, в области автоматизации научно-технической и образовательной деятельности, а также в области развития информационных технологий в целом.

Описаны принципы, лежащие в основе ***стандарта интеллектуальных компьютерных систем*** и соответствующей технологии их разработки, рассмотрена структура базы знаний портала научных знаний, в рамках которой формально описывается указанный стандарт.

Received 21.12.19

OSTIS technology overview in the context of fuzzy systems development

1st Anton Romanov
Department of Information Systems
Ulyanovsk State Technical University
Ulyanovsk, Russia
romanov73@gmail.com

2nd Aleksey Filippov
Department of Information Systems
Ulyanovsk State Technical University
Ulyanovsk, Russia
al.filippov@ulstu.ru

3rd Gleb Guskov
Department of Information Systems
Ulyanovsk State Technical University
Ulyanovsk, Russia
guskovgleb@gmail.com

4th Anton Vasin
Department of Information Systems
Ulyanovsk State Technical University
Ulyanovsk, Russia
vasin.anton404@gmail.com

5th Anton Skalkin
Department of Information Systems
Ulyanovsk State Technical University
Ulyanovsk, Russia
anton.skalkin73@gmail.com

6th Yulia Stroeva
Department of Information Systems
Ulyanovsk State Technical University
Ulyanovsk, Russia
stroeva95@mail.ru

***Abstract*—This article provides an overview of OSTIS technology. The prospects of using OSTIS technology in the field of data analysis and the development of fuzzy systems are considered. An approach to modeling time series based on type 2 fuzzy sets is described. An approach to constructing a fuzzy knowledge base is presented.**

***Keywords*—ostis, overwiev, type 2 fuzzy sets, fuzzy knowledge base**

I. Introduction

The implementation and use of knowledge-based intelligent systems are relevant in all problem areas nowadays [1], [2]. A huge amount of information produces the need to develop new models, methods, and algorithms for storing, presenting, and operating with knowledge. However, a serious problem in the evolution of information technology is ensuring information compatibility and integration of various computer systems.

This problem is especially important in the domain of artificial intelligence researches in studies related to solving problems based on machine learning methods. The isolation of groups that develop various technologies in this direction determines the slow growth of the market for intelligent systems. The significant results have been obtained in various researches in the domain of artificial intelligence, but integrating these results is a difficult task.

There is a need to create a technology that would integrate various models of problem areas and various types of knowledge. The relevance of solving this problem is determined by a number of reasons [1], [2]:

1) A single technology for data and knowledge representation will allow increasing the level of learning, and the degree of diversity of knowledge. Such knowledge can be used in solving various problems without the need to develop specialized software and ensure their information compatibility.
2) Simplification of the computer systems development process. The development and continuous expansion of the set of components, solutions, and typically embedded knowledge can significantly reduce development time.
3) Simplification of the maintaining computer systems process and the compatibility of information stored in them. The form of data storage in information systems can change significantly during the life cycle. This situation produces the need to modify the software to ensure compatibility with other information systems.

These problems partially solved in the concept of open semantic technology for intelligent systems (OSTIS).

II. Important aspects of the open semantic technology for intelligent systems

The problems of unification of the principles for constructing various components of computer systems are solved in the OSTIS project [1], [2]. The OSTIS project aims to create an open semantic technology for designing knowledge-driven systems.

Important aspects of OSTIS technology are [1], [2]:

1) Applying modern experience in the field of information technology. Language tools have been created for a unified description of the developed intelligent systems as part of the OSTIS technology.
   Each OSTIS-system uses semantic memory (sc-memory). Each action in sc-memory denotes a transformation performed by some subjects. The problem area of abstract sc-agents has been developed within the framework of OSTIS technology. A set of ontologies is used to describe sc-agents. These ontologies describe a specification of the

sc-agent concept and related to sc-agent concepts, and also includes formal tools that ensure synchronization of actions produces by sc-agents in sc-memory.

The SCP language is considered as the base language for programs that describe some activity of sc-agents with sc-memory. The SCP language is a graph procedural programming language designed to efficiently process homogeneous semantic networks with set-theoretic interpretation encoded using SC-code.

SC-code is a unified format for representing knowledge in the form of homogeneous semantic networks with a set-theoretic interpretation in OSTIS technology.

2) Separation of a unified description of the developed intelligent system and various interpretations of formal descriptions of intelligent systems.

   Clear separation of the design process of a formal description of the semantic model of the developed knowledge base from the process of interpreting this model is performed in the process of developing knowledge bases using OSTIS technology.

3) Common library for a set of components, solutions, and typically embedded knowledge that will be available to all developers.

   The problem arises of systematization and structuring of knowledge with the accumulation of large data volumes. Each of the concepts of the current domain has a role. This role can either be a class of the research object or a relation defined on the set of objects of study, etc.

III. Perspectives for the development of OSTIS technology in the field of fuzziness

The previously discussed advantages of OSTIS technology can be used to develop high-level abstractions of data mining models and methods. Such high-level abstractions can form a higher level of intelligent computing machine by analogy with high-level programming languages and machine instructions. Thus, you can create sets of functions for solving problems in the field of data mining using the OSTIS technology. Libraries of these functions can significantly reduce the time to development of intelligent systems.

A. Dynamic data analysis based on fuzzy sets of type 2

It will be possible to model time series with more quality if create a component for working with type 2 fuzzy sets for the OSTIS platform.

There is also a need to form type 2 fuzzy sets based on the features of the current problem area.

It is necessary to create tools for the automated building of an ontology based on the analysis of enterprise data to solve such a problem.

Let see an example of an approach to solving the problem of modeling time series using type 2 fuzzy sets.

The task of forecasting is always relevant in the context of management. Prediction is made using a variety of methods and approaches. If the deterministic model of the system is absence then such methods based on the study of the evolutionary history of processes and indicators. One of the popular tools for studying the evolutionary history of processes and indicators are time series models. Time series models have a different nature: statistical, based on neural networks, fuzzy, etc. The use of models is complicated by data characteristics. Data may require preprocessing, normalization, may have a high degree of uncertainty.

The nature of the fuzzy time series is caused by using expert evaluations. Expert evaluations have a high level of uncertainty and fuzziness. Fuzziness, unlike stochastic uncertainty, makes it difficult or impossible to use statistical methods and models. However, fuzziness can be used to make subject-oriented decisions based on approximate reasoning of an expert. The formalization of intellectual operations that simulate fuzzy human statements about the state and behavior of a complex system nowadays forms an independent area of scientific and applied research called "fuzzy modeling" [3].

First-order sets (type 1) are usually used to build a process model. Type 1 fuzzy sets are used to represent or to create a model of domain uncertainties [4]. L. A. Zade introduced in 1975 second-order fuzzy sets (type 2) and higher-order fuzzy sets (type n) to eliminate the shortcomings of type 1 fuzzy sets [5]. The main disadvantage is the mapping of the membership function to exact real numbers. The solution to this problem can be the use of type 2 fuzzy sets, in which the boundaries of the membership regions are themselves fuzzy [4]. For each value of the x variable from the universe X the value itself is a function, not a value at a point. This function represents a type-2 fuzzy set, which is three-dimensional, and the third dimension itself adds a new degree of freedom for handling uncertainties.

The union, intersection, and complement operations can be over type 2 fuzzy sets. Type 2 fuzzy sets have an extended representation of uncertainties, which creates additional computational complexity.

The designation of the fuzzy type 2 membership function can graphically shown as a region called the footprint of uncertainty (see fig. 1). In contrast to using the membership function with crisp boundaries, the values of the type 2 membership function are themselves fuzzy functions.

This approach has given an advantage in bringing the fuzzy model closer to the model in linguistic form. People may have different evaluations of the same uncertainty. There was a need to exclude an unambiguous comparison of the obtained value of the degree of mem-

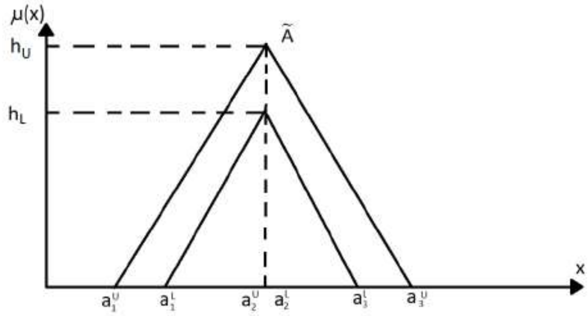


Figure 1. Example of a type 2 fuzzy sets.

bership function. Thus, the risk of error accumulation is reduced when an expert sets the membership degrees due to not including points located near the boundaries of the function and being in uncertainty.

Blurring boundaries is the first step in moving from type 1 fuzzy sets to type 2 fuzzy sets. It is required to choose the type of membership function at the second step as is done for type 1 fuzzy sets (for example, triangles).

The uncertainty present in the tasks of managing the activities of any enterprise and characterized by the statements of experts containing incomplete information, with fuzzy and unclear information about the main parameters and conditions of the analyzed problem. Thus, the solution to the control problem becomes complicated and is generated by multiple factors. The combination of these factors in practice creates a wide range of different types of uncertainty. Therefore, it becomes necessary to use methods that allow the use of blurry values of indicators.

Type 2 fuzzy sets $\tilde{A}$ in the universum U can be defined using type 2 membership function. Type 2 fuzzy sets can be represented as:

$$\tilde{A} = ((x,u), \mu_{\tilde{A}}(x,u)) | \forall x \in U, \forall u \in J_x \subseteq [0,1]$$

where $x \in U$ and $u \in J_x \subseteq [0,1]$ in which $0 \leq \mu_{\tilde{A}}(x,u) \leq 1$. The main membership function is in the range from 0 to 1, so the appearance of the fuzzy set is expressed as:

$$\tilde{A} = \int_{x \in U} \int_{u \in J_x} \mu_{\tilde{A}}(x,u)/(x,u) J_x \subseteq [0,1]$$

where the operator $\int\int$ denotes the union over all incoming x and u.

Time series modeling needs to define interval fuzzy sets and their shape. The figure 1 shows the appearance of the sets.

Triangular fuzzy sets are defined as follows:

$$\begin{gathered}\tilde{A}_i = (\tilde{A}_i^U, \tilde{A}_i^L) = \\ = ((a_{i1}^u, a_{i2}^u, a_{i3}^u, h(\tilde{A}_i^U)), \\ (a_{i1}^l, a_{i2}^l, a_{i3}^l, h(\tilde{A}_i^l))).\end{gathered}$$

where $\tilde{A}_i^U$ and $\tilde{A}_i^L$ is a triangular type 1 fuzzy sets, $a_{i1}^u, a_{i2}^u, a_{i3}^u, a_{i1}^l, a_{i2}^l, a_{i3}^l$, is reference points of type 2 interval fuzzy set $\tilde{A}_i$, h is the value of the membership function of the element a_i (for the upper and lower membership functions, respectively).

An operation of combining fuzzy sets of type 2 is required when working with a rule base based on the values of a time series. The combining operation defined as follows:

$$\begin{gathered}\tilde{A}_1 \oplus \tilde{A}_2 = (\tilde{A}_1^U, \tilde{A}_1^L) \oplus (\tilde{A}_2^U, \tilde{A}_2^L) = \\ = ((a_{11}^u + a_{21}^u, a_{12}^u + a_{22}^u, a_{13}^u + a_{23}^u; \\ min(h_1(\tilde{A}_1^U), h_1(\tilde{A}_2^U)\tilde{A}_1^U)), min(h_2(\tilde{A}_1^U), h_2(\tilde{A}_2^U)),); \\ (a_{11}^l + a_{21}^l, a_{12}^l + a_{22}^l, a_{13}^l + a_{23}^l; \\ min(h_1(\tilde{A}_1^L), h_1(\tilde{A}_2^L)), min(h_2(\tilde{A}_1^L), h_2(\tilde{A}_2^L)));\end{gathered}$$

B. Algorithm for smoothing and forecasting of time series

The main principle of the proposed algorithm is closely related to the nature of the time series. Type 2 fuzzy sets are used for modeling in the process of smoothing and forecasting of time series because the time series has the interval nature. [4].

The proposed algorithm can be represented as a sequence of the following steps:

1) Determination of the universe of observations. $U = [U_{min}, U_{max}]$, where U_{min} and U_{max} are minimal and maximal values of a time series respectively.
2) Definition of membership functions for a time series $M = \{\mu_1, ..., \mu_l\}, l << n$, where l is the number of membership functions of fuzzy sets, n is the length of a time series. The number of membership functions and, accordingly, the number of fuzzy sets is chosen relatively small. The motivation for this solution is the multi-level approach to modeling a time series. It is advantageous to reduce the number of fuzzy sets at each level to decrease the dimension of the set of relations. Obliviously, this approach will decrease the quality of approximation of a time series. However, creating the set of membership functions at the second and higher levels will increase the approximation accuracy with an increase in the number of levels.
3) Definition of fuzzy sets for a series. In that case, the superscript defines the type of fuzzy sets. $A^1 = \{A_1^1, ..., A_l^1\}, A^2 = \{A_1^2, ..., A_m^2\}$, where l is the number of type 1 fuzzy sets, m is the number of type 2 fuzzy sets.

4) Fuzzification of a time series by type 1 sets. $\forall x_i$ $\tilde{y}_i = Fuzzy(x_i)$
5) Fuzzification a time series by type 2 sets.
6) Creation of relations. The rules for the creation of relations are represented in the form of pairs of fuzzy sets in terms of antecedents and consequents, for example: $A_1^1 A_1^2 ... \longrightarrow A_2^1 A^2 1$.
7) Doing forecasting for the first and second levels based on a set of rules. The forecast is calculated by the centroid method, first on type 1 fuzzy sets $A^1 = \{A_1^1, \ldots, A_l^1\}$, then on type 2 fuzzy sets.
8) Evaluation of forecasting errors.

C. Knowledge discovery

Modern organizations have a large amount of accumulated knowledge presented in the form of various corporate knowledge bases. Thus, the development of a technological platform (TP) is necessary to solve the following tasks:

- the TP should not require additional skills and knowledge from the user;
- the TP should provide the programmer the familiar and easy to use data access mechanism;
- the TP should provide an inference mechanism to implement knowledge discovery (KD) functionality;
- the TP should provide functions to the automation of the process of obtaining essential knowledge about problem area (PrA) in the internal knowledge base (KB).

At the moment, a lot of researchers use the ontological approach for the organization of the knowledge bases of expert and intelligent systems: F. Bobillo, U. Straccia [6], [7], M. Gao, C. Liu [8], D. Bianchini [9], N. Guarino [10], G. Guizzardi [11], R.A. Falbo [12], G. Stumme [13], T.R. Gruber [14], A. Medche [15].

The inference is the process of reasoning from the premises to the conclusion. Reasoners [16] are used to implementing the function of inference and form logical consequences from many statements, facts, and axioms. Currently, the Semantic Web Rule Language (SWRL) is used to record logical rules for reasoners [17].

D. The model of the fuzzy domain knowledge base content with contexts support

Contexts of the KB represent the parts of ontology in space and time. Each space context is associated with a value from 0 to 1 defining the expert level of expertise in the part of the problem area (PrA). Time contexts allow using versioning of the PrA ontology and give an opportunity to monitor the dynamics of the ontology development. The fuzzy nature of the KB appears in the process of integration of contexts of the domain ontology.

The problem of developing the model of fuzzy domain KB with support of logical rules for inference is coming up. One of the KB main objectives is providing the mechanism for adapting the TP [18], [19] to the concrete PrA with the use of methods of ontological analysis and data engineering.

Let the following definition represents the model of the KB content:

$$O = \langle T, C^{T_i}, I^{T_i}, P^{T_i}, S^{T_i}, F^{T_i}, R^{T_i} \rangle, i = \overline{1,t}, \qquad (1)$$

where t is a number of the KB contexts,
$T = \{T_1, T_2, \ldots, T_t\}$ is a set of KB contexts,
$C^{T_i} = \{C_1^{T_i}, C_2^{T_i}, \ldots, C_n^{T_i}\}$ is a set of KB classes within the i-th context,
$I^{T_i} = \{I_1^{T_i}, I_2^{T_i}, \ldots, I_n^{T_i}\}$ is a set of KB objects within the i-th context,
$P^{T_i} = \{P_1^{T_i}, P_2^{T_i}, \ldots, P_n^{T_i}\}$ is a set of KB classes properties within the i-th context,
$S^{T_i} = \{S_1^{T_i}, S_2^{T_i}, \ldots, S_n^{T_i}\}$ is a set of KB objects states within the i-th context,
$F^{T_i} = \{F_1^{T_i}, F_2^{T_i}, \ldots, F_n^{T_i}\}$ is a set of the logical rules fixed in the KB within the i-th context, logical rules are used to implement the functions of inference by the content of KB,
R^{T_i} is a set of KB relations within the i-th context defined as:

$$R^{T_i} = \{R_C^{T_i}, R_I^{T_i}, R_P^{T_i}, R_S^{T_i}, R_F^{T_i}\},$$

where $R_C^{T_i}$ is a set of relations defining hierarchy of KB classes within the i-th context,
$R_I^{T_i}$ is a set of relations defining the "class-object" KB tie within the i-th context,
$R_P^{T_i}$ is a set of relations defining the "class-class property" KB tie within the i-th context,
$R_S^{T_i}$ is a set of relations defining the "object-object state" KB tie within the i-th context,
$R_F^{T_i}$ is a set of relations generated on the basis of logical KB rules in the context of the i-th context.

Figure 2 shows an example of the translation of the OWL representation of the ontology of family relations into the entities of the KB.

As seen in figure 2:

- OWL class "Person" was translated into the KB class with the same name;
- OWL individuals "Alex", "Helen", "Kate" and "17" was translated into the KB objects with same names;
- KB objects "Alex", "Helen" and "Kate" are objects of KB class "Person";
- KB object "17" is the object of built-in KB class "Integer";
- OWL data property "hasAge" was translated into the KB property with the same name;
- OWL object properties "hasFather" and "hasSister" was translated into the KB properties with the same names;
- OWL data property assertion "Helen hasAge 17" was translated into the KB state with the same name,

the range of this state is "Helen", the domain of this state is "17";

- OWL object property assertion "Helen hasFather Alex" was translated into the KB state with the same name, the range of this state is "Helen", the domain of this state is "Alex";
- OWL object property assertion "Helen hasSister Kate" was translated into the KB state with the same name, the range of this state is "Helen", the domain of this state is "Kate".

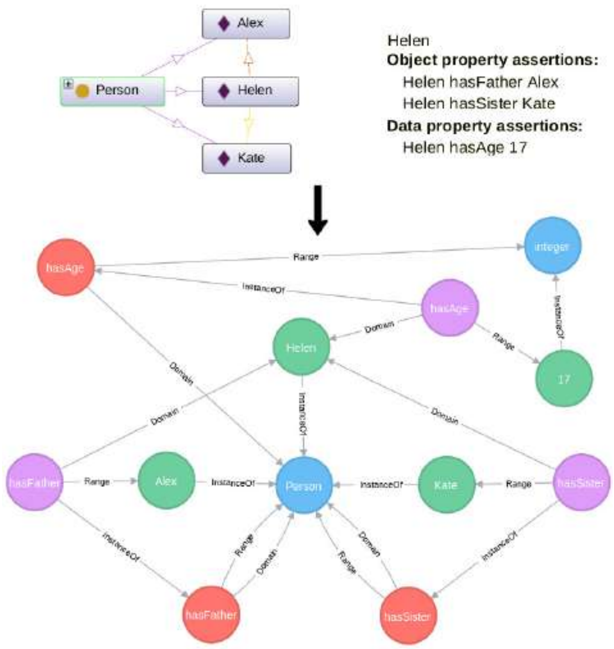


Figure 2. Example of the translation of the OWL representation of ontology of family relations into the content of the KB.

E. Dynamical user interface

The dynamic graphical user interface (GUI) mechanism is used to simplify the work with KB of untrained users and control of user input [20], [21].

Need to map the KB entities to the GUI elements to build a GUI based on the contents of the KB. Formally, the GUI model can be represented as follows:

$$UI = \langle L, C, I, P, S \rangle, \tag{2}$$

where $L = \{L_1, L_2, \ldots, L_n\}$ is a set of graphical GUI components (for example, ListBox, TextBox, ComboBox, etc.),
$C = \{C_1, C_2, \ldots, C_n\}$ is a set of KB classes,
$I = \{I_1, I_2, \ldots, I_n\}$ is a set of KB objects,
$P = \{P_1, P_2, \ldots, P_n\}$ is a set of KB properties,
$S = \{S_1, S_2, \ldots, S_n\}$ is a set of KB states.

The following function is used to build a GUI based on the content of KB:

$$\phi(O) : \{C^O, I^O, P^O, S^O, F^O, R^O\}^{T_i} \rightarrow \\ \rightarrow \{L^{UI}, C^{UI}, I^{UI}, P^{UI}, S^{UI}\},$$

where $\{C^O, I^O, P^O, S^O, F^O, R^O\}^{T_i}$ is a set of KB entities represented by definition 1 within the i-th context; $\{L^{UI}, C^{UI}, I^{UI}, P^{UI}, S^{UI}\}$ is a set of GUI entities represented by the definition 2.

Thus, the contents of the KB are mapped to a set of GUI components. This mapping makes it easier to work with KB for a user who does not have skills in ontological analysis and knowledge engineering. It also allows you to monitor the logical integrity of the user input, which leads to a reduction in the number of potential input errors.

We are confident that the creation of tools for the formation of an ontology for "simple" users will simplify and speed up the process of generating new knowledge [22].

If in the OSTIS technology will be a tool for working with ontology contents in the form of catalogs, this will speed up the process of solving the problem of integrating semantic data.

IV. CONCLUSION

Thus, OSTIS technology aims to unify and standardize models, methods, and tools for developing knowledge-based intelligent systems.

OSTIS technology uses approaches that have proved their in various implemented projects aimed at solving problems of a wide range of prolem areas [23]–[25].

As for recommendations for the further development of OSTIS technology, we can distinguish:

1) It is necessary to develop tools for translating ontologies in OWL and RDF formats into the knowledge format of OSTIS technology.
2) It is necessary to develop tools to translate SWRL rules into fragments of scp-programs.
3) It is necessary to develop the centralized repository to store up-to-date documentation on OSTIS technology, covering all aspects of installing, configuring, launching, using and developing OSTIS technology.
4) It is necessary to build a container image to allow the use of the latest version of OSTIS technology.

ACKNOWLEDGMENT

The reported study was funded by RFBR and the government of Ulyanovsk region according to the research projects: 18-47-732016, 18-47-730022, 18-47-730019, and 19-47-730005.

REFERENCES

[1] V. Golenkov, N. Guliakina, I. Davydenko, "Methods and tools for ensuring compatibility of computer systems," in Open semantic technologies for intelligent systems, vol. 3, 2019 pp. 25-53.

[2] V. Golenkov, D. Shunkevich, I. Davydenko, "Principles of organization and automation of the semantic computer systems development," in Open semantic technologies for intelligent systems, vol. 3, 2019 pp.53-91.

[3] I. Perfilieva, N. Yarushkina, T.Afanasieva, A. Romanov, "Time series analysis using soft computing methods," in International Journal of General Systems, vol. 42, 2013, pp. 687-705.

[4] J.M. Mendel, R.I.B. John, "Type-2 Fuzzy Sets Made Simple," in IEEE Transactions on Fuzzy Systems, vol. 10, 2002, pp. 117-127.

[5] L.A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," in Information Sciences, vol. 8, 1975, pp. 199-249.

[6] F. Bobillo, U. Straccia, "FuzzyDL: an expressive fuzzy description logic reasoner," in Proceedings of the 17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2008), 2008, pp. 923-930.

[7] F. Bobillo, U. Straccia, "Representing fuzzy ontologies in OWL 2," in Proceedings of the 19th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010), 2010, pp. 2695-2700.

[8] M. Gao, C. Liu, "Extending OWL by fuzzy description logic," in Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2005), 2005, pp. 562-567.

[9] D. Bianchini, V. De Antonellis, B. Pernici, P. Plebani, "Ontology-based methodology for e-service discovery," in Information Systems, vol. 31, 2005, pp. 361–380.

[10] N. Guarino, M.A. Musen, "Ten years of Applied Ontology," in Applied Ontology, vol., 2015, 10, pp. 169–170.

[11] G. Guizzardi, N. Guarino, J.P.A. Almeida, "Ontological Considerations About the Representation of Events and Endurants in Business Models," in International Conference on Business Process Management, 2016, pp. 20-36.

[12] R.A. Falbo, G.K. Quirino, J.C. Nardi, M.P. Barcellos, G. Guizzardi, N. Guarino, "An ontology pattern language for service modeling," in Proceedings of the 31st Annual ACM Symposium on Applied Computing, 2016, pp. 321-326.

[13] A. Hotho, S. Staab, G. Stumme, "Ontologies improve text document clustering Data Mining," in ICDM 2003, 2003, pp. 541-544.

[14] T. Gruber, "Ontology," available at: http://tomgruber.org/writing/ontology-in-encyclopedia-of-dbs.pdf (accessed 2019, Dec)

[15] A. Maedche, S. Staab, "Ontology learning for the Semantic Web," available at: https://www.csee.umbc.edu/courses/771/papers/ieeeIntelligentSystems/ontologyLearning.pdf (accessed 2019, Dec)

[16] K. Dentler, R. Cornet, A. ten Teije, N. de Keizer, "Comparison of reasoners for large ontologies in the OWL 2 EL profile," in Semantic web, vol. 2, 2011, pp. 71–87.

[17] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Available at: https://www.w3.org/Submission/SWRL (accessed 2019, Dec)

[18] N. Yarushkina, A. Filippov, V. Moshkin, "Development of the Unified Technological Platform for Constructing the Domain Knowledge Base Through the Context Analysis," in Creativity in Intelligent Technologies and Data Science (CIT&DS-2017), 2017, pp. 62-72.

[19] N.G. Yarushkina, A.A. Filippov, V.S. Moshkin, L.I. Filippova, "Application of the fuzzy knowledge base in the construction of expert systems," in Information Technology in Industry, vol. 6, 2018, pp. 32-37.

[20] S. Hattori, Y. Takama "Recommender System Employing Personal-Vallue-Based User Model," in J Adv. Comput. Intell. Intell. Inform. (JACIII), vol. 18, 2014, pp. 157–165.

[21] F.B. Ruy, C.C. Reginato, V.A. Santos, R.A. Falbo, G. Guizzardi, "Ontology Engineering by Combining Ontology Patterns," in 34th International Conference on Conceptual Modeling (ER2015), 2015, pp. 173-186.

[22] A.S. Boriskin, D.N. Koronchik, M.E. Sadowski, "Ontology-based design of user interfaces," in Open semantic technologies for intelligent systems, vol. 1, 2017 pp.95-107.

[23] V. Taberko, D. Ivaniuk, V. Kasyanik, "Design principles of integrated information services for batch manufacturing enterprise employees," in Open semantic technologies for intelligent systems, vol. 3, 2019, pp. 215-225.

[24] V. Zahariev, T. Lyahor, N. Hubarevich, "Semantic analysis of voice messages based on a formalized context," in Open semantic technologies for intelligent systems, vol. 3, 2019, pp. 103-113.

[25] N. Iskra, V. Iskra, M. Lukashevich, "Neural network based image understanding with ontological approach," in Open semantic technologies for intelligent systems, vol. 3, 2019, pp. 113-123.

Обзор технологии OSTIS в контексте разработки нечетких систем

Романов А.А., Филиппов А.А., Гуськов Г., Васин А.О., Скалкин А.М., Строева Ю.В.

В данной статье представлен обзор технологии OSTIS. Рассмотрены перспективы использования технологии OSTIS в области анализа данных и разработки нечетких систем.

Описан подход к моделированию временных рядов на основе нечетких множеств типа 2. Представлен подход к построению нечеткой базы знаний.

Received 29.12.2019

The approach to "Big Data" keeping with effective access in multi-tier storage

Globa L.S., Filimonov N.
Information and Telecommunication Department
Institute of Telecommunication Systems
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv, Ukraine
lgloba@its.kpi.ua

***Abstract*—Big Data has too much cumulative amount of weakly linked data availability of a large amount of structured and unstructured data, their substantial diversity. Such data is handled by means of horizontally scalable software tools and alternative traditional database management systems as well as Business Intelligence class systems. One of the challenges of this Big Data approach is the need to effectively control and process data flows.**

The paper discusses the approach of distributed data storage in multi-tier repositories, taking into account the interdependence of IT infrastructure and location of business objects, as well as the importance and reliability of data over time. The task of placing blocks of data in multi-tier repositories is formulated so that the average access time to them is minimal. A method for solving this problem using a genetic algorithm, the effectiveness of which is confirmed by the results of mathematical modeling is proposed.

***Keywords*—distributed Big Data storing, data storage system, database management system, datacenter.**

I. Introduction

Every year, as a result of human and business activities, more and more data are generated that need to be stored, processed, and analyzed. The vast accumulation of loosely coupled data has been called Big Data, which means that there is a significant amount of structured and unstructured data and a significant variety of them. This data can be processed by means of horizontally scalable software tools used recently and by alternative traditional database management systems, as well as Business Intelligence class systems [1] [2].

Big Data characteristics define at least "three V": volume – physical volume of data, velocity – both speed of increasing volume of data, and necessity of high-speed processing with obtaining results, variety – possibility of simultaneous processing of various types of structured and semi-structured data.

The emergence of large volumes of data storages, Cloud and Frog technologies, the development of the concept of central data storage in certain subject areas of knowledge and business with the simultaneous possibility of virtualization of such a "logical" centralized storage, while physically it is separated by unknown geographically distributed computing datacenters, has exacerbated data management issues. There are problems with reliability – veracity ("fourth V" proposed by IBM), viability – ("fifth V"), value in terms of business – value ("sixth V"), as well as variability – the ability of data to change constantly and visualization – the ability of quick perception the meaning of the results obtained by the user ("seventh and eighth V").

To overcome Big Data challenges, tools of mass-parallel processing with indefinite structure by means of NoSQL database management systems, MapReduce algorithms, and Hadoop [3] are used. Various analytical software components and technologies, approaches to structuring, clustering, and data grouping, that would provide more or less similar characteristics of processing of large data sets, began to refer to Big Data technologies.

It is obvious that the process of constant accumulation of data complicates the work with them. That is why the task of accessing such logically centralized, and physically geographically dispersed, data storages with an uncontrolled end-user data architecture is a daunting problem related to the control and management of information flows by the end-user of data precisely in terms of the requirements of certain processing procedures [4]. The implementation of this approach to Big Data processing covers the following major areas of data processing improvement: databases (both storage structures and processing approaches), mathematical foundations for analytics, the visualization of significant volumes in clear and fuzzy concepts, architecture and flow processing.

The classic approach to storing and processing data no longer fulfills its functions. One of the most urgent tasks is the development of new and modification of old methods that adapt the DSSs (data storage systems) to more stringent conditions [5,6].

The data storage system is a complex software and hardware solution for the organization of reliable storage of information resources and providing guaranteed access to them [7].

In the process of working with data in the DSS the following common properties can be determined:

- accessibility
- capacity
- scalability
- security
- manageability

However, the individual stored information has its own static and dynamic properties. These include value, intensity of usage, and others depending on the tasks. Values and dynamics of change of values define a life cycle of the data manage of which determines efficiency of use of the stored information [8].

The paper is structured as follows: Section 2 contains a state-of-the-art approach to distributed Big Data storing in tiered data storages. Section 3 introduces the proposed approach to distributed Big Data storing in tiered data storages in the requirements of processing procedures terms and places. Section 4 presents the results of the proposed approach to designing distributed Big Data storing in tiered data storages based on data access intensity. Section 5 includes a summary and outlook on future work.

II. State of the art and background

In order to securely store information resources and provide them with guaranteed access when using remote, decentralized, modern data storage systems, lifecycle management processes called ILM (Information Life-cycle Management process) are used, and systems that support data lifecycle management are ILM systems.

The ILM process is a process of data transformation, from the moment you create the data to the deletion, which includes all stages of processing throughout the entire period of work with specific data. ILM systems allow you to organize processes and create data management technologies that allow you to efficiently process and apply specific data throughout their lifecycle. Such systems consider the interdependence of IT infrastructure and the location of business objects, as well as the importance and reliability of data over time. Data that is important to business today may be almost unnecessary tomorrow, called the data aging process. In the case of aging data, the latter can be automatically moved to cheaper and less time-efficient access to the data storage and to use low-performance technologies, depending on their relevance, to save money and improve the use of available data storage resources [9,10].

The implementation of the ILM process includes the following set of actions. First of all, it is the classification of the data itself and the programs for processing them, then – defining a strategy for creating and deleting data, depending on their class, volumes and other parameters of influence on the importance of certain datasets.

A separate aspect of the ILM process is the management of the processes of storing input streams, uploading them to data storages and extracting them. As a logical conclusion, the concept of ILM describes the multi-tier storage of data on different types of media, which, in turn, divided by the parameters of price, speed of access, reliability, etc. [11].

To integrate the ILM process into application systems that support a particular business environment, perform the following actions:

- establish a data storage system
- classify data (tiers, cycles, policies)
- moving between tiers manually
- automate basic business processes
- fully integrate the individual steps of the ILM process into all software components of the application system.

Moving data manually requires the complex work of analysts on each individual implementation of multi-tier DSS, so there is a transition to a DSS architecture with fully automated FAST (Fully Automated Storage Tiering) storage.

For automated multi-tier storage architectures, the DSS hierarchy has the following tiers:

- Tier 1 – designed for business-critical data (mission-critical, business-critical), characterized by minimal access time and high availability. Currently typically implemented on SSD storage systems.
- Tier 2 – online access reference data, sometimes archive systems, document storage systems for which speed of data access and system availability are important, but not critical for business, traditionally the main storage array of the organization. It is implemented on the whole variety of arrays and storage systems, based on hard drives (FC, SAS, SATA, SCSI)
- Tier 3 – mainly archive systems and backup systems. In addition to disk arrays can be implemented including on tape libraries.

The moving between tiers approach is called Tiering. Each of the tiers is characterized by certain properties (price, speed, capacity, security and others). Typically, the tiering process is called a mechanism for moving data between disks of different tiers, or between disks and a magnetic stripe, or ROM storage. Often, it is used to organize the ILM data storage in accordance with the status of the data and the QoS level.

Discs used today in DSSs differ in performance, capacity, cost. And at the same price, productivity and capacity are usually inversely proportional (larger capacity – lower productivity (SATA), lower capacity – greater productivity (SAS), even greater productivity – even smaller capacity (SSD). DSSs are generally characterized by an IOPS/cost ratio where IOPS is the number of I / O operations performed by DSSs per second (input / output operations per second).

In general, a small block of data requires most of the total data processing time, so it is usually determined by the speed of the system. In contrast, large blocks of data refer to so-called "cold" data, which have almost no effect on performance. A data processing center (storage center) can consist of many tiers that use different types of media.

It is well known that when creating all business systems, they require cost reduction without loss of quality. Another way to save, and therefore optimize, is to de-energize individual repositories (racks, disks) with unnecessary data at specific times. For example, night-time shutdown of devices that store “cold” data or data that will not be needed exactly during off-hours. This approach takes into account the additional requirements when the analysis and further tuning of the system must occur in such a way that shutdowns for the sake of economy do not interfere with business processes, and the tuning of the process of de-energization occurred without fatal consequences, because the processing datacenter, apart from the carriers themselves, has configuration servers and other infrastructure.

Based on the foregoing, there is a complex set of problems associated with the organization and control of large dynamic data streams, the collection, storage, processing of large amounts of data, operational support of users for adequate decision making.

For large systems, tier partitioning is quite a challenge because most of the application software is located in the processing datacenter, and the distributed application components are located in several processing datacenters at once. Such an application software infrastructure may use more than one DSS, but several, making it difficult to break down the entire set of DSSs at the tier.

Each datacenter has its own DSSs, application servers, and other infrastructure that allows you to quickly access application software components to information that is situated directly within the DSSs of a particular datacenter. But if the system is distributed, then the application software needs to receive information hosted by different datacenters, and the speed of data access is highly dependent on the distance to other datacenters, speed, response and other network settings.

III. Problem definition for Big Data distributed storing in tiered data storages based on data access intensity

With regard to optimizing the placement of data in distributed systems, it is obvious that the faster the application software receives information from the DSS, the faster it processes the data. To improve the performance of the system, it is advisable to place the data on media that can be accessed as quickly as possible, taking into account the features of the location and processing of the specific application software component that uses them. There is a problem of optimal placement of data in distributed DSSs to improve performance of a particular application component. And, unlike a system consisting of one DSS and one application server, there is a situation where it is more advantageous to keep information on less fast media but less time accessing the application server. In such a system, it is also important to consider the intensity of data usage by individual servers in order to properly calculate the optimal location [12].

For application servers located in a single datacenter, it is possible to break up the DSS, which is physically hosted at different datacenters, at the following criteria:

- Speed of access to storage. The repository acts as a unit.
- Media performance. The disk acts as a unit.
- Media access speed. The disk acts as a unit.

By unit we mean an inseparable part of the tier.

For application servers located in different datacenters:

- The sum of the output of the speed of access to the repository and the intensity of access to the application server with which the exchange is being performed. Storage is a unit.
- The sum of the output of the speed of access to the media and the intensity of calls of the application server with which the exchange is being performed. Storage is a unit.

This is the idea underlying the tiering process. If a particular type of data is active, it is automatically transferred to more modern and faster, performance-enhancing media and less active data – to less productive and outdated disk types, resulting in an IOPS / cost ratio.

IV. Mathematical model for the proposed approach

Consider the case where the functions of application software components are accessed by several blocks, and each block can be accessed by several functions, and also, each data block must be on only one data carrier (see Fig. 1).

We introduce the following notation:

F_k – function calls, $k = 1, \dots, l$;

μ_k – the average intensity of function calls F_k;

D_i – data blocks accessed by the function, $i = 1, \dots, n$;

d_i – volume of data block D_i;

λ_{ki} – the intensity with which the function F_k accesses the data block D_i;

S_j – physical data carriers on which data blocks are stored, $j = 1, \dots, m$;

t_j – the time of receiving data from the S_j repository;

t_{ki} – time to access the F_k function;

s_j – volume of storage S_j;

x_{ij} – a Boolean variable that determines whether the data block D_i is located in the S_i repository;

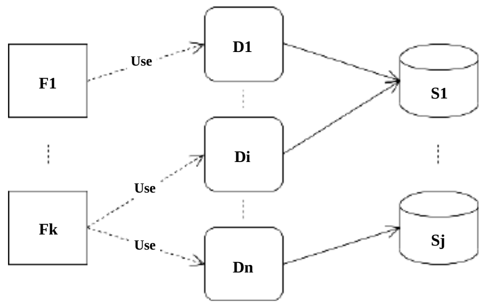


Figure 1. Scheme of organization of function calls of application software to the data of physically distributed DSS

Since each block of data must be on only one data media, the condition must be:

$$\sum_{j=1}^{m} x_{ij} = 1 \qquad (1)$$

where $i = 1, \ldots, n$

Formulation of the problem. In a situation where resources in storage systems are limited and there are no strict requirements for the response time of functions, the task arises to optimally place blocks of data on data carriers so that the average system performance reaches the maximum value and the average access time to the data is minimal.

The average intensity with which the function F_k accesses the data block D_i can be represented as follows:

$$\mu_{D_i} = \sum_{k=1}^{l} \mu_k \lambda_{ki} \qquad (2)$$

and the criteria for minimizing the average access time can be written:

$$T_{min} = min \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{D_i} \cdot x_{ij} \cdot (t_j + t_{ki}) \qquad (3)$$

Then the problem of minimizing the average access time can be formulated: minimizing the average time of access to the data (3), provided the execution of the constraint (1).

V. The problem solution based on genetic algorithm

The above problem is a linear Boolean programming problem and, with a small number of independent variables, can be effectively solved by using exact methods. When a distributed application infrastructure uses more than one physically distributed DSS that changes dynamically over time, is itself territorially distributed, the dimension of the problem is high, then it is advisable to use a genetic algorithm (GA) [13].

Since each block of data must be on only one data carrier, in order to encode the genes according to the algorithm, it is necessary to move from the Boolean matrix n x m of variables x_{ij} to the vector of discrete variables $y_i \in [1, m]$, in which each element contains a data carrier number S_j, $(j = 1, \ldots, m)$, at which the corresponding block of data is located at the data instant.

Example of gene coding:

$$x_{ij} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \rightarrow y_i = \begin{bmatrix} 4 \\ 2 \\ 3 \\ 1 \\ 3 \\ 4 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

It is recommended to use a controlled genetic algorithm to solve problems of this class [14].

The block diagram of the algorithm is shown in Fig. 2.

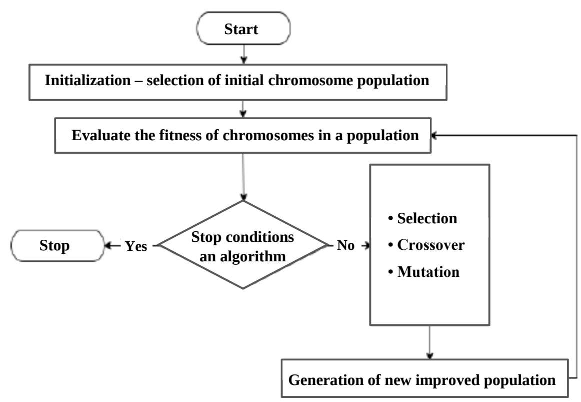


Figure 2. Flowchart of a managed genetic algorithm

Consider the basic steps of executing a managed genetic algorithm:

Step 1: Initialization – selection of the initial chromosome population;

Step 2: Evaluate the fitness of chromosomes in a population;

Step 3: If the end condition is positive then stop, and the choose of the "best" chromosome in current population;

Step 4: Create a new population with repeating next steps:

- Selection – Extract a subset of genes from a population, according to any definition of quality;
- Crossover – crossing parents genes;
- Mutation – forms a new population of offspring;

Step 5: Go to step 2.

VI. The results of proposed solution

The effectiveness of the proposed approach to storing “big data” in multi-tier “logically” centralized repositories in accordance with the tasks of the business environment, which function as physically distributed DSSs, was tested using mathematical modeling tools. Below are graphs that compare the quality of data access (storage speed, media access speed, media performance) and the data volume change graph.

During the experiment, an architecture with automated multi-tier data storage was applied, for which the media was broken down by media types and the average amount of data in the database of a specific fragment of the DSS. The (Fig. 3) depicts the dependence without applying the proposed approach, where the Y axis represents the μk – average intensity of function calls to the DSS, and the X axis shows the average amount of data in the database of a specific node of the DSS.

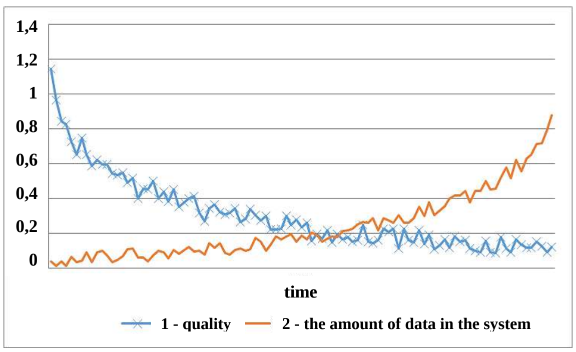


Figure 3. Quality of access to information, depending on the amount of data stored in the system according to their life cycle, without applying the proposed approach

The graph in Figure 3 shows that the quality of access to information in the process of increasing it is not constant and at different points in time changes, leading to low rates of data access speed, which in turn does not meet the requirements of the application systems.

As a result of applying the proposed approach, after structuring and grouping the data, the partitioning of the media at the tier by media types and the average amount of data in the database of a specific node of the DSS, we obtain the following result (Fig. 4).

As a result of dynamic restructuring we get: with the increased amount of information, the quality of access to data at any address remains almost at the same tier, which allows to provide stable access to all information resources while maintaining maximum access efficiency.

VII. Summary and outlook

The paper deals with the pressing issue of efficient placement of data in multi-tier storage systems, which arises when dealing with a large amount of data, limited resources and stringent requirements for storage performance.

The application of dynamic models of optimal data placement, depending on their volume in the system,

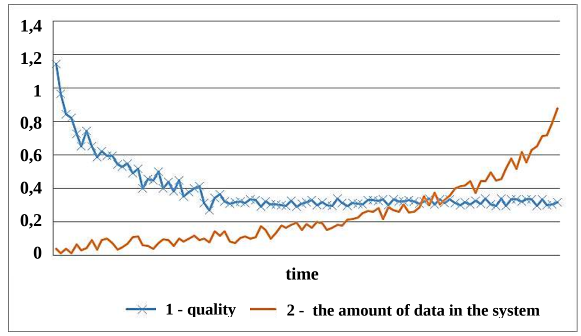


Figure 4. Quality of access to information, depending on the amount of data stored in the system according to their life cycle, using the proposed approach

taking into account their life cycle, allows to increase the performance indicators of the distributed data storage system, which was proved by the results of the experiment. The greatest increase in the efficiency of the functioning of the distributed storage system resulted in the movement between the carriers of different productivity.

It is recommended to apply the proposed model when designing system software for multi-tier data storage systems using partitioning of databases and setting up information lifecycle management (ILM) processes.

Further, there is a need to improve the proposed approach in terms of increasing the diversity of media types, allowing data to be accessed so that access to important and frequently used data is performed as quickly as possible, and outdated data is moved to less rapid media or archived at all.

References

[1] Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung. Big Data. Related Technologies, Challenges, and Future Prospects. – Spinger, 2014. – 100 p. – ISBN 978-3-319-06244-0. – DOI:10.1007/978-3-319-06245-7.
[2] James Manyika et al. Big data: The next frontier for innovation, competition, and productivity (engl.) (PDF). McKinsey Global Institute, June, 2011. McKinsey (9 August 2011).
[3] Preimesberger, Chris Hadoop, Yahoo, 'Big Data' Brighten BI Future (engl.). EWeek (15 August 2011).
[4] "Cisco Global Cloud Index: Forecast and Methodology 2015-2020" Cisco Public.
[5] Mayer-Schönberger V. Big Data: A Revolution That Will Transform How We Live, Work, and Think / V. Mayer-Schönberger, K. Cukier. – Houghton Mifflin Harcourt, 2013. – 240 p.
[6] Marz N. Big Data: Principles and Best Practices of Scalable Real-time Data Systems / N. Marz, J. Warren. – Manning Publications Company, 2015. – 328 p.
[7] Schulz G. Cloud and Virtual Data Storage Networking / G. Schulz. – CRC Press Taylor & Francis Group, 2011. – 370 p.
[8] Horizontal, vertical, and functional data partitioning [Online Resource] // Microsoft – Resource Access Mode: https://azure.microsoft.com/en-us/documentation/articles/best-practices-data-partitioning/ – 2018
[9] Information Lifecycle Management. – 2005. – Resource Access Mode: http://i.bnet.com/whitepapers/sun00.pdf
[10] Colarelli D. Massive arrays of idle disks for storage archives / D. Colarelli, D. Grunwald. // Proceedings of the 2002 ACM / IEEE conference on Supercomputing. – 2002. – P. 1-11.

[11] Information Lifecycle Management for Business Data // Oracle Help Centr. – Resource access mode: http://www.oracle.com/us/026964.pdf. – 2007
[12] Storage systems [Electronic resource]. – Resource Access Mode: http://www.compdoc.ru/peripherals/drives/categorization-storage-systems – 2016
[13] Goldberg D.E. Genetic algorithms in search, optimization, and machine learning / D.E. Goldberg. – Addison-Wesley, 1989. – 412 p.
[14] Telenik SF / S.F. Telenik, OI Roller, M.M. Bukasov, S.A. Androsov // Automation. Automation. Electrotechnical systems and systems. – 2010. – № 1 (25). – P. 106-120.

Подход к хранению «больших данных» с эффективным доступом в многоуровневом хранилище

Глоба Л.С., Филимонов Н.

Большие данные имеют слишком большой совокупный объем слабосвязанных данных, большой объем структурированных и неструктурированных данных, их существенное разнообразие. Такие данные обрабатываются с помощью горизонтально масштабируемых программных средств и альтернативных традиционных систем управления базами данных, а также систем класса Business Intelligence. Однако, обработка больших данных на основе такого подхода требует эффективных как контроля, так и доступа к хранилищам данных, контроля процессов обработки потоков данных.

В статье рассматривается подход распределенного хранения данных в многоуровневых хранилищах, учитывающий взаимозависимость ИТ-инфраструктуры и местоположения бизнес-объектов, а также важность и достоверность данных во времени. Задача размещения блоков данных в многоуровневых хранилищах сформулирована таким образом, чтобы среднее время доступа к ним было минимальным. Предложен метод решения этой задачи с использованием генетического алгоритма, эффективность которого подтверждается результатами математического моделирования.

Received 13.12.2019

Analysis of clustering algorithms for use in the universal data processing system

Buhaienko Y., Globa L.S., Liashenko A., Grebinechenko M.
Information and Telecommunication Department
Institute of Telecommunication Systems
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv, Ukraine
yura.buhaenko@gmail.com
lgloba@its.kpi.ua
andrey.lyashenko44@gmail.com

***Abstract*—Today, humanity is moving to work in the digital space. While processing large amounts of information in digital space, data processing systems can analyze data, create logical inference systems and building some conditions in which data must be homogeneous, similar, and grouped into clusters. Such systems can analyze the data and create logical output systems according to the rules. Some conditions are required to build such systems: the data must be homogeneous, similar to each other and grouped into clusters. One way to help achieve this is through clustered data analysis. There are a large number of cluster analysis methods available today, but not all of them provide positive results. The article offers criteria that allow you to select algorithms that are appropriate for application in a specific subject area, namely the algorithmic complexity of the algorithm, the accuracy of the object's belonging to the cluster, the attribution of the object to one or another cluster at the boundary of two clusters, the ability to build from clusters fuzzy logic rules. Also described is the architecture of the created system, which clusters the data with different algorithms to use from the desired area. The algorithms were tested to solve the problem of data clustering to predict the degree of server energy efficiency at certain values involved in the processor frequency and number of cores calculations. The use of algorithms allowed us to analyze and look at the features of each, to determine the fuzzy C-average as the most appropriate for this task.**

***Keywords*—clustering algorithms, clustering, fuzzy logic**

I. Introduction

Data volumes that require some analysis and processing are growing steadily today. According to a source [1], the average annual amount of data in the world that needs processing is increasing exponentially. As a result, information management takes more time and resources than ever before. In order to minimize the computational complexity of these processes, a number of methods are being actively used and explored today to improve efficiency and minimize the cost of processing large amounts of data.

According to a research made by We Are Social and Hootsuite [2], the number of people using the Internet has increased over the past year, with more than 1 million daily users appearing from 2018.

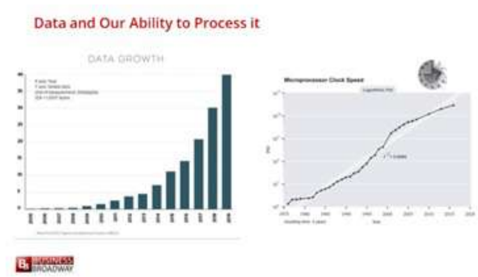


Figure 1. Increasing of average amount of data per year that needs processing

- There are now 5.1 billion unique mobile users in the world, an increase of 100 million over the last year.
- The number of internet users reaches almost 4.4 billion, an increase of 366 million over the last year.
- The number of active users of social media reaches 3.4 billion, an increase of 288 million over the last year.
- The number of mobile users on social media reaches 3.2 billion, an increase of 297 million.

Looking at this data, it is clear that the amount of content and data that users produce will also increase steadily. According to a source [3], the amount of data generated every second is more than 30,000 Gigabytes. You can use clustering to analyze this data, which can group similar computing objects into clusters and show how many and how busy the servers are during computing. Each of these groups should be processed independently, which allows to manage the direct process of processing, to determine the type of analysis and work on information, to effectively allocate resources. It is better for analysts to select groups of similar objects and study their features separately than to study the entire sample of data. When designing and developing real smart systems, there is a problem - the formation of a complex mathematical model that will include various technical features of the equipment and the physical characteristics of the environment (complex calculations may require additional computational resources). In the course of further development, it may be necessary to take into account new conditions, which may lead to a complete revision of the existing model. In this case, different conditions can be defined as some rules (IF ... THAN) using fuzzy logic. To get these rules, you can use experts who set the terms themselves. And you can teach the system to build these rules on its own, using different clustering algorithms. Non-hierarchical algorithms according to [4] are not suitable for large volumes of data. Since the K-Means and Fuzzy C-Means algorithms are universal, hierarchical clustering algorithms are proposed in [5]. In order to have an idea of how to choose the optimal algorithm for a particular subject area, we need to solve the problem of which algorithm is more appropriate to choose. Each algorithm fits to its subject area, according to certain criteria. In [6] it is proposed to determine the algorithmic complexity of the algorithms, and in [7] it is proposed to determine which of the algorithms is best suited for the construction of fuzzy knowledge bases.

Consider solving the problem of predicting the degree of energy efficiency of servers at certain values involved in the process of calculating the processor frequency and number of cores.

Clustering plays an important role in dealing with large amounts of data, both in the modeling technique and as a pre-processing step in many implementations of the data mining process.

II. Algorithms

- K-means algorithm Advantages of the method: ease of use; speed of use; clarity and transparency of the algorithm. Disadvantages of the method: the algorithm is too sensitive to emissions that can distort the mean; slow work on large databases; the number of clusters must be set. The algorithm of clear clustering has certain disadvantages: in the process of work it is very difficult to determine the degree of fuzzy (blurring) of cluster boundaries, and the actual number of clusters cannot be calculated mathematically, but is given by an expert at the beginning of processing.
- Fuzzy C-means fuzzy clustering algorithm. Advantages: The fuzzy definition of an object in a cluster allows you to identify objects on the border into clusters. Disadvantages: Computational complexity, cluster number assignment. [8]

K-means Clustering K-means Clustering is one of the machine learning algorithms. This algorithm is a non-hierarchical, iterative method of clustering, it has become very popular due to its simplicity, clarity of implementation and high quality of work. The basic idea of the k-means algorithm is that the data should be broken down into clusters, after which the center of mass for each cluster obtained in the previous step is iteratively recalculated, then the vectors are again broken down into clusters with new centers of mass. The purpose of the algorithm is to divide n observations into k clusters so that each observation belongs to only one cluster. [9]

Fuzzy C-Means Clustering Fuzzy C-means algorithm is similar to the human thinking style, which in the future will allow to create rules of fuzzy inference of type Mamdani or Sugeno, allows to control degree of blur of borders of clusters and to take into account fuzzy belonging of certain data to a certain group, provides more flexibility, than clear methods, allowing each point in space belongs to several clusters. This algorithm allows each data sample object to be located in each cluster with a different degree of belonging and is less sensitive to emissions. This method should solve the problem of defining the identity of an object in a cluster if it is on the boundary of several clusters. The fuzzy C-mean algorithm is suitable if the clustering objects have the following requirements:

- High dimensionality of data space - objects are described by a large number of attributes, therefore, the algorithm should be adapted to work in high dimensional data spaces.
- Large amount of data.
- Clustering is done to obtain fuzzy rules based on built clusters. [1]

III. Comparison of clustering algorithms

Algorithmic complexity and performance To use clustering algorithms, you need to understand their algorithmic complexity, because typically large datasets need to be clustered. The larger the input sequence enters the clustering input, the longer the algorithm will run. According to [6], the time complexity of the K-means $O(ncdi)$ algorithm and the FCM of the $O(ncd^2i)$ algorithm, where n is the number of points, the input sequence, d is the amount of space dimension, c is the number of clusters, and the number of iterations. That is, as the number of clusters increases, the performance of the FCM algorithm will decrease. The time spent calculating the K-means algorithm will be 2 times greater than that of the FCM algorithm. We can conclude that K-means is the best algorithm for performance.

IV. Building fuzzy knowledge bases

You also need to understand the need for fuzzy systems. If clustering is performed to construct fuzzy rules, then it is better to choose the FCM algorithm. This algorithm makes it possible to see the affiliation of each object to the cluster and to determine which cluster it belongs to, especially relevant for data that are at the boundary of two clusters, that is, there is some uncertainty about the choice of a point. [7]

Figure 2. The data set

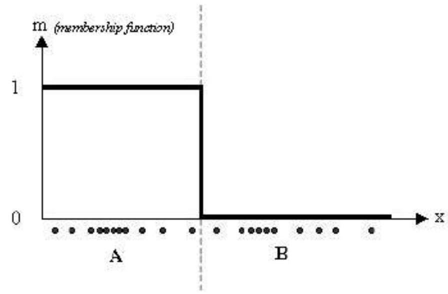


Figure 3. Clustering with clear allocation of clusters

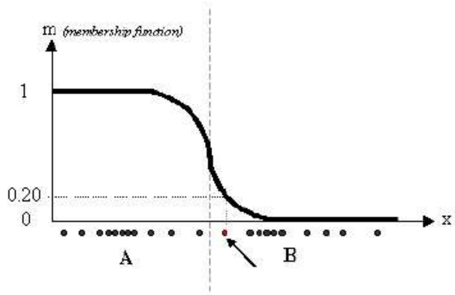


Figure 4. Fuzzy clustering. Each point has its own degree of belonging

In Fuzzy C Means clustering, more accuracy can be achieved to determine the degree of belonging to each point. For example, by adjusting the constant ε, this can be achieved to minimize the criterion J. From the article [7] we can conclude that by reducing ε, in order to minimize the objective function, better data distribution can be achieved. But for such improvement, it should be noted that the execution time is significantly increased, as the number of iterations increases. For K-clustering, it is also suitable for fuzzy systems. However, looking at Figure 3, it can be seen that this clustering “roughly” separates points into clusters that are not accurate enough for points at the cluster boundary.

Practical implementation of the clustering algorithm by example The algorithm was written in C # programming language using the ASP.NET Core platform. In terms of architecture, the application is completely client-server. There is an explicit client level in the application, such as a browser. And an application layer that uses an algorithm to execute a clustering algorithm. The application also uses REST technology. The project uses an ASP.NET Core MVC Web-based API.

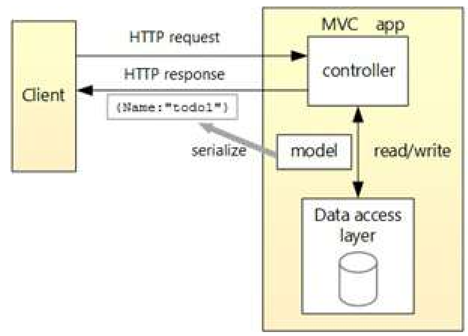


Figure 5. Architecture of the created application

The server side receives the data transmitted in the file through the interface and performs fuzzy clustering with the initial parameters specified by the expert (number of clusters, precision, minimization). The client side of the web application includes a graphical interface for data input for processing, and a page for outputting data results. The clustered data is displayed on the page as a graph written using the open-source Javascript library Plotly.js. Written on the popular libraries d3.js and slack.gl, it is a high-level declarative graph library. For a clearer and more convenient output, we have used the division into three clusters, each responsible for red, green or blue. Each point on the graph represents one dimension in the dataset: its color is represented in RGB (red, green, blue) format, where each parameter is a cluster belonging. With this display method, you can show to which cluster this dimension is most relevant, which data has not been clustered accurately (these points are marked in indeterminate gray). To obtain the result of the experiment, a clustering of server energy efficiency data (number of threads, data processing frequency, energy consumed by the machine) was collected at the Dresden University of Technology. In addition, processing was performed with different clarity of cluster boundaries, which made it possible to correct processing uncertainty.

As a result, an application was developed in which the expert can process his data on different algorithms and also to perform analytic on the basis of visualization, adjusting the parameters - to set up clustering for the conditions of his task.

V. Conclusions

This article provides a comparative analysis of K-means Clustering algorithms and Fuzzy C-means fuzzy clustering algorithm and tested each for data taken from Article [10] on the performance of computing servers. Comparative analysis is carried out on the criterion of algorithmic complexity of algorithms and on the criterion of building fuzzy knowledge bases. And showed that the K-means algorithm has less algorithmic complexity,

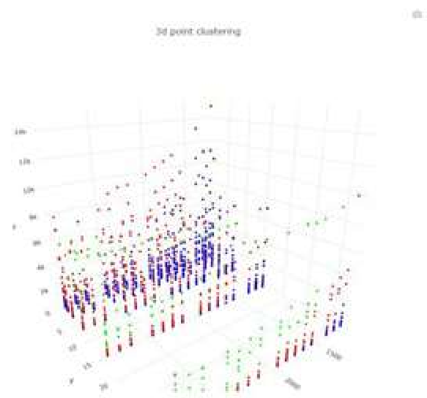


Figure 6. Visualization of clustered data with minimization of criterion J for $\varepsilon < 0.05$

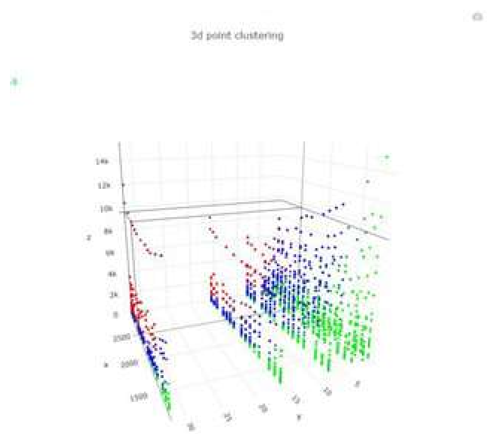


Figure 7. Visualization of clustered data with minimization of criterion J for $\varepsilon < 0.0001$

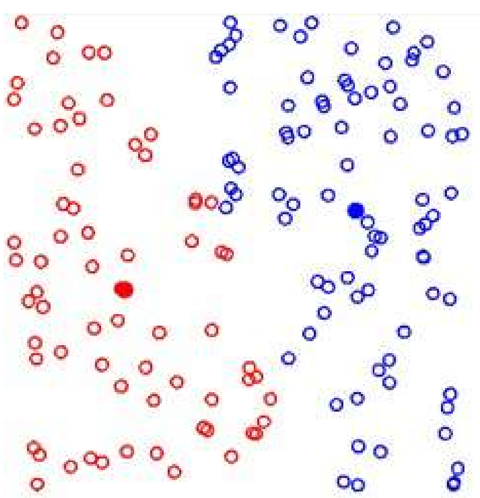


Figure 8. Visualization of the K-means algorithm for 2 clusters in 2-dimensional space

which will allow to spend less resources to execute clustering algorithms. But fuzzy C-means is better suited to building fuzzy knowledge bases, which allows you to more gently divide objects into clusters and show more precisely which cluster the object belongs to. This will give a better opportunity to build fuzzy knowledge bases. In the future, it is planned to use already given clustering algorithms - to build fuzzy knowledge bases, using different types of fuzzy inference rules.

References

[1] Liashenko A., Buhaienko Y. "A clustering method for processing large volumes of data". MODERN CHALLENGES IN TELECOMMUNICATIONS NTUU "KPI", Kiev 2019.

[2] DIGITAL 2019: GLOBAL INTERNET USE ACCELERATES [Electronic resource]. – Resource Access Mode: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates – 2019

[3] Nathaz Marz, James Warren. Big data. Principles and best practices of scalable real-time data systems. p.[185-192] – 2015 – ISBN-13: 978-1617290343.

[4] Methods of cluster analysis. Iterative methods. [Electronic resource]. – Resource Access Mode: https://www.intuit.ru/studies/courses/6/6/lecture/184?page=4 – 2016 – ISBN: 978-5-9556-0064-2

[5] Pegat A. Fuzzy modeling and control. p. [520-530]. – Moscow, 2009.

[6] Soumi Ghosh, Sanjay Kumar Dubey «Comparative Analysis of K-Means and Fuzzy CMeans Algorithms». International Journal of Advanced Computer Science and Applications(IJACSA), Volume 4 Issue 4. – 2013.

[7] A Tutorial on Clustering Algorithms. Fuzzy C-Means Clustering. [Electronic resource]. – Resource Access Mode: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html – 2014

[8] Magerramov Z., Abdullayev V., Magerramova A. «BIG DATA: PROBLEMS, ANALYSIS METHODS, ALGORITHMS» – 2017

[9] K-means algorithm. [Electronic resource]. – Resource Access Mode: http://algowiki-project.org/en/K-means_clustering – 2018

[10] Dmytro Pukhkaiev, Sebastian Götz «BRISE: Energy-efficient Benchmark Reduction». The 6th International Workshop. Dresden – 2018

Анализ алгоритмов кластеризации для использования в универсальной системе обработки данных

Бугаенко Ю., Глоба Л.С.,
Ляшенко А., Гребинеченко М.

Аннотация – В статье предложены критерии, которые позволяют выбрать алгоритмы, целесообразны для применения в конкретной предметной области, а именно алгоритмический сложность алгоритма, точность принадлежности объекта к кластеру, отнесение объекта к одному или другому кластера на границе двух кластеров, возможность построить из кластеров нечеткие логические правила. Также описана архитектура созданной системы, кластеризирует данные различными алгоритмами в зависимости от потребностей предметной области.

Received 15.12.2019

Workflows Generation based on the Domain Ontology

Larysa Globa, Maryna Popova, Andrii Bakai
Information and Telecommunication Department
Institute of Telecommunication Systems
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv, Ukraine
lgloba@its.kpi.ua, pma1701@gmail.com, andruha.bakay1997@gmail.com

***Abstract*—The digital economy, which is a prerequisite for the 4th Industrial Revolution, requires a rapid increase in services provided in the digital global space. However, the level of computer-aided workflows underlying the various services is imperfect due to the lack of meta-models (ontologies) that would describe the conditions of use of the services, take into account the meaningful connections between them, and a mathematical apparatus that would make it easy to execute certain interactions of elements of such models.**

The paper describes the approach to automation of the design of workflows and their components (functional services, communications and rules of interaction) by computer-aided generation of both a set of services that are components of workflows and the sequence of their execution, using ontologies – meta-models of workflows, services, domain, logical rules that establish relationships between functional services.

The application of the proposed approach allowed to generate workflows based on the parameters defined in the meta-model (ontology) of the register of services and limitations imposed by the domain ontology, taking into account both functional and non-functional requirements, realizing the transformation of the workflow in its program execution models regardless of the calculation.

For the computer-aided workflows generation, a software environment has been developed, the performance of which is tested on the example of workflow generation with the use of IT TODOS tools for constructing and modifying ontologies and generating a workflow scheme using BPMN.

***Keywords*—ontology, workflow, functional services, computer-aided business process design.**

I. Introduction

Nowadays is characterized by the rapid development of information technology, which transforms all spheres of society. The emergence of the digital economy in the industries and business has been called the 4th Industrial Revolution – Industry 4.0 [1], which has created a virtual (digital) business environment, which is constantly increasing the volume of services.

Under such conditions, technological innovation and reconfiguration of complex software solutions to adapt to the requirements of the global business environment are an integral part of ensuring business competitiveness. The process of providing a variety of services in the global business environment requires a high level of computer-aided workflows that underlie the various services, due to the constant change of requirements for them. Besides, workflows that perform an ordered sequence of computations in a virtual digital space during providing services to the end user are formed based on algorithms for executing an ordered sequence of domain actions.

The sequence of actions (tasks), ordered in a logical and orderly way with cause and effect relationships between them, the purpose of which is to obtain a result (product, service, information, decision, etc.) is called a *business process*.

Workflow is a software tool (complex software component) used to execute a business process of a domain in a virtual digital space, which is both a graphical and a code representation of the execution sequence of software components (services) in a computing environment, including specific calculations, information dependencies and sequences of decisions and computational works. Thus, it is a set of technologies and tools that allow you to appropriately follow a certain sequence of work of business processes of the domain, to implement certain computational procedures to achieve the goals.

Computer-aided workflows with the help of modern information technologies is carried out, as a rule, to reduce the number of failures, errors, downtime and losses, increase the reliability and accuracy of procedures. With the implementation of workflow modeling and computer-aided design, it is possible to deeply analyze the situation and develop a plan of action to reduce costs and increase productivity overall.

Workflow Management System is a software system that serves to prepare, execute and monitor a defined workflow order and tasks to improve productivity and reduce the cost required to complete the workflow.

The following elements are required to develop a workflow scenario:

Method: A series of steps for submitting and modeling domain information.

Meta-Model: An information model that describes the properties of the individual components of the domain that are important in terms of content, as well as their meaningful relationships that determine the conditions of using the individual components.

Notation: Symbols and rules for presenting information.

Tools: Computing tools (software) for information processing.

To represent both sequential and state-based workflows, it is suggested to use ontology as their metamodel. In terms of modern software workflow is a sequence of instances of a class (or subclass) in the ontology, representing the individual stages (states) of the computational process.

An important aspect of the need to use ontology for workflows generation is the distribution, heterogeneity, and multiagency of components of the global virtual (digital) business environment. For example, ontologies were developed for the automotive industry, describing the entire logistics chain of production (Odette [2], RosettaNet [3]).

Due to the transition to digital space and the constant modification of business processes in all industries under Industry 4.0, as well as the implementation of artificial intelligence systems, the need to develop specific ontologies is expected to increase. Accordingly, there is a growing need for workflow automation methods based on the computer-aided generation of their steps from ontological models of both the domain and the global virtual business environment. According to [4], research in this area is at an early stage.

Taking into account that the ontology of workflows allows you to formally describe the process, structure and visualize parts of it, combining functional services with links in the form of domain constraints that reflect the interaction between them, and the ontology of functional services will retain their formal meta descriptions, an intellectual environment based on which it is possible to computer-aided workflows generation depending on the data state of the domain.

An advantage of the ontological approach is that the ontological model presents the relationships between computational components and the rules for their application in the form of graphs, allowing the use of graph theory mathematical apparatus to generate workflows in real time, based on the description of the input and output parameters of each functional service.

This paper is organized as follows. After the Introduction section 2 contains a state of the art approach to computer-aided workflow design, methods and tools for their analysis. Section 3 is dedicated to the formalized description of ontology-based meta-models. Section 4 provides an example of software implementation and further research perspectives. Section 5 includes a summary and outlook on future work.

II. State of the Art and Background

Many studies have been devoted to the computer-aided workflow generation [2-18], by which computational independent workflows are developed using graphical standards, allowing them to be formalized to reflect possible flows and transitions in a schematic form. The analysis showed that in practice, computational independent workflows are usually developed using graphical notation by tools such as BPMN 2.0 [5], UML AD, USLD [6] and tools such as CA ERwin Process Modeler [7], Enterprise Architect [8] and MS Visio [9].

A brief overview of workflow analysis methods and tools showed that there are two types of analysis that take into account the performance of the computing process:

1.Analysis of design time (modeling and verification). Monte Carlo modeling tools and Petri analysis can be used. For example, there are analysis tools for BPMN, UML AD [12], EPC [13], BPEL [14] using the Petri net mathematical apparatus and their subsequent analysis. USLD diagrams can be analyzed using ontological analysis of services [15].

2.Analysis of execution time (e.g., process of determining execution time based on execution logs) [16].

Software such as Pegasus, Cactus, ASKALON, GLUE, etc. is used for such areas of analysis [17]. All mentioned and analyzed current opportunities for this stage of the task are very limited. The disadvantages of methods and tools of workflow analysis are clearly described in [18]. The central disadvantage is that the requirements analysis stage is mainly applied manually.

Thus, it can be concluded that the strategies of adapting systems to constantly changing user requirements through computer-aided design and re-engineering of end-user services (i.e. designing and modifying existing business processes) are poorly formalized and verified, requiring several iterations with the participation of analysts, system architects, significant investment of time and resources.

There are quite a few tools for describing and building workflows, but all of them are poorly automated and require human involvement. There are currently no solutions that would allow you to reengineer workflows due to changing meta descriptions of functional services.

III. Formalized Description of Workflow Generation from Ontology

In general, a workflow is a scheme of organizing a sequence of execution tasks within a single service that is provided to end users, which may include related computational subprocesses, information dependencies, decision making sequences, and some computations.

A flowchart or graph consisting of operations (functional services or microservices), symbols of logic, connections is used to represent the workflow. The sequence of calculations is indicated by arrows.

Typically, the workflow can be described using formal or informal flowchart methods that show directional flows between processing steps. Individual processing steps or workflow components can be basically defined by three parameters:

- Input description: information, data, and resources required for the workflow.
- Conversion rules: algorithms that can be executed by a person, program, or both.
- Output description: information, data and resources that are processed by the workflow and ready to be forwarded.

Components can be combined only if the output information of one component corresponds to the input information of the component, which performs further processing, and satisfies the conditions of the domain. Thus, a semantic description must include a description of the input, output and meta description of the functional services performed in this workflow. If there are several methods (functional services) for processing the same data, then you need to add a meta description of the algorithms that perform the given function, and it is desirable to specify the performance characteristics of the given functions, such as accuracy, speed and others.

This is due to the fact that the information that the workflow is weakly cohesive, multi-structured and can be processed by different methods depending on the characteristics of the input flow. But from a data processing point of view, information needs to be structured and the system trained on specific data in order to be able to use different methods more effectively.

A. Description of the Service Presentation Model of the Workflow Generation:

In many cases, the same problem can be solved by different methods. The nature of data processing is determined by the probabilistic properties of observations made on some data, so it is proposed to use ontological models to describe all the rules for establishing connections based on input data in order to determine the most effective method of solving the same problem.

Consider a computer-aided workflow generation approach based on the domain ontologies, computing environment, namely workflows and all the components (functional services, relationships, and rules) that form the workflow.

Functional Service (S) is a service, an element (object) that accepts a dataset, performs operations on it, and transmits new data to the output. Functional services are saved using a file system, and service meta descriptions are in the database. All meta descriptions of the functional services of each domain are stored in separate database collections. Functional services can be simple and complex.

Consider a simple service model:

$$S = \{Code, M, P\},$$

where:

$Code$ - the functional code of the functional service;

M - meta description of the functional service;

P- the physical path to a functional service.

Functional service meta description (M) is a description of a function, its input and output parameters, to be further displayed in the user interface and to take into account constraints when building relationships in the workflow.

A meta description of a simple service can be represented as:

$M = \{ID, F, \{I_1, \ \dots \ I_n\}, \{O_1, \ \dots, \ O_n\}, T\}$,

where:

$M = \{ID, F, \{I_1, \ \dots \ I_n\}, \{O_1, \ \dots, \ O_n\}, T\}$ – a meta description that includes the following parameters:

ID – a meta description identifier, for example: ObjectID («5cdd4fd0598ec622d023e048»)

$F = \{N, D, R, P\}$ – a set of parameters for describing a functional service;

$I_i = \{N, D, R, T\}$ – the set of parameters to describe the input parameters;

$O_i = \{N, D, R, T\}$ – a set of parameters to describe the output parameters, where:

N – the name of the object (functional service, input / output parameters), such as *"Sustainability calculation"*

D – the description of the object, the characteristic of the computing operations of the functional service, for example – «The service checks whether this card exists at the destination bank»;

R – the name of the object that is convenient for a person to understand, for example – «Checking the presence of a payee's bank account»

P – the path to a functional service, for example – «./bank/checkBankAccount.js»;

T – type of service, for example – «function», «process»; variable type «number», «string» ...

$Proc = \{CPU, R, t_{max}, \dots\}$ – a description of non-functional service parameters, where:

CPU – the number of computing kernels required to perform the service;

R – the amount of RAM;

t_{max} – maximum time for service execution;

$\dots$ – other parameters.

The meta descriptions of the functional service are stored in the database as follows:

```
{
 _id
 function: {name, description, readable_neme, path},
 input: [{name, description, readable_neme, type}],
 output: [{name, description, readable_neme, type}],
 proc: {cpu, ram, tmax, ...}
 type
}
```

In algebraic form, a complex service can be written:
$S_{complex} = \{I, \{R_1,\ \ldots,\ R_n\}, \{S_1,\ \ldots,\ S_n\}, O, Proc\}$,
where:
$S_{complex} = \{I, \{R_1,\ \ldots,\ R_n\}, \{M_1,\ \ldots,\ M_n\}, O, \}$ – a set of parameters that describe a complex service, namely:

I – service input;

R_i – the rule of choice of method;

$S_i = \{I \bigcap R_i\}$ – a data processing method or simple service that can be selected when applying the R_i rule to the input parameters;

O – the output of the service.

$Proc = \{CPU, R, t_{max}\}$ – a set of non-functional parameters that acquire the value of the selected processing method.

B. Formalized Workflows Description

The ontology of workflows consists of functional services and relationships between them, which is built taking into account the limitations of the domain.

A domain is a set of all objects, classes, and attributes linked together by logical connections.

Workflow – a process that consists of one or more functional services interconnected by the domain and logical connections.

The workflow is represented as follows:

$$W_f = \{I, \{F_1,\ \ldots,\ F_n\}, O, Proc\},$$

where:

$W_f = \{I, \{F_1,\ \ldots,\ F_n\}, O, Proc\}$ – a workflow that includes the following parameters:

I – description of the input parameters;

F_i – used functional services;

O – description of the output parameters;

$Proc = \{CPU_{max}, R_{max}, t_{max}\}$ – non-functional workflow parameters where:

CPU_{max} – the maximum number of computing kernels required;

R_{max} – the maximum amount of RAM required;

t_{max} – the maximum permissible cumulative runtime that consists of the sum of t_{max} of all functional services.

Relationships between services (L) – logical connections establish the order of data processing, the order of transferring parameters from one service to another, can be established only with the condition that the two services are logically connected and at least one of the output parameters of the first service corresponds to at least one input parameter of another service.

C. Formalized Description of Workflow Support Infrastructure

In computer-aided workflow systems, their workflows are accomplished through workflow management systems, which are described using data flow diagrams.

Formally, the workflow management system can be represented by:

$$WfMS = \{S_r, S_d, S_c, S_a, S_{ai}\},$$

where:

S_r – routing system, routing the flow of information or objects, transmits information from one work item to the next;

S_d – a distribution system that detects exceptional circumstances and transmits information to designated system components;

S_c – coordination system, coordinates simultaneous activities, preventing resource conflicts (allocation of CPU time and memory) or conflicts of priorities;

S_a – agent system, automatically starts the service, controls its execution and receives data from it;

S_{ai} – a helper system that regulates the choices of the following methods and provides suggestions which of the following methods can be used.

The components of a Data Flow Diagram (flowchart):

$$DFD = \{D_p, D_{df}, D_w, D_t\},$$

where:

D_p – the process, part of a system that converts inputs into outputs;

D_{df} – data flow, shows the transfer of information from one system to another;

D_w – a data storage used to store data for later use

D_t – access terminal, interface of the system interaction with the external environment.

D. Ontology-based Computer-aided Workflows Generation

To generate a workflow, you must perform an intersection operation with each one for all services, provided that there is a logical connection, as well as the data connection between them. In this case, the input data will determine the service from which the calculation begins, provided the data coincides with their meta descriptions (M) in the service. The meta descriptions (M) of the services (S) are a set of:

$$M = \{M_{use}, M_{domain}, M_{all}\},$$

Formally, the workflow can be represented as:

$$W_f = \{\forall M_{use,i} \bigcap \forall M_{use,j} | \exists L_{ij}\},$$

where:

$W_f = \{\forall M_{use,i} \bigcap \forall M_{use,j} | \exists L_{ij}\}$ – a generated workflow consisting of:

$M_{use} = \{M_{def} \bigcap M_{domain}\}$ – a selection of services used in the workflow where:

M_{use} – functional services used in the workflow;

M_{def} – user-defined functional services;

M_{domain} – functional services of this domain;

$L_{ij} = \{M_{use,i}.O = M_{use,j}.I\}$ – define workflow relationships based on compatible input and output rules, where:

L_{ij} – workflow connections;

$M_{use,i}.O$ – the output parameter of the i-th functional service;

$M_{use,j}.I$ – the input parameter of the j-th functional service;

$M_{domain} = \{M_{all} \in SD\}$ – selects all domain-specific services where:

M_{domain} – functional services of a domain;

M_{all} – all functional services available in the repository (register);

SD – Specified domain.

The steps of workflow generation using an ontological register of services based on a formal algebraic system are described below.

Step 1 – Determine the parameter/parameters (meta descriptions M) of the service / microservice (S) that needed to solve a specific problem.

Step 2 – Selection the service / microservice parameter values given in the ontology as input. In this case, the parameter values can be *static* (predefined by the expert) and *dynamic* (the user sets the parameter values in the expert's range «from» – «to»).

Step 3 – Definition of microservices and their clustering into services on the basis of simple operations: *elementary* (addition, multiplication), *logical* (conjunction, disjunction, equivalence, negation, implication), *over sets* (union, cross-section, belonging, etc.), according to the specified rules.

Step 4 – Compare the values of service / microservice parameters based on the intersection operation with each for all services / microservices. In this case, based on the static values of the parameters specified in the ontology, a predefined service / microservice is selected from the register, and on the basis of dynamic values, a number of services / microservices are offered to the user by which the problem can be solved in the best way (faster, easier).

Step 5 – Combining multiple services / microservices into a workflow.

The order of inclusion of services in the workflow is determined by the presence of parameter values in stages 2 and 4. For example, the $n+1$ service will be executed after the n service in the workflow if the input for its execution is data obtained from the n service execution.

Changing the order of the steps will result the wrong answer, which makes it impossible to use the inversion operation and the logical combination operation for the selected service.

IV. Example of Software Implementation Of Domain Ontology-based Workflows Generation

The domain «Temporary norms for calculation of strength of power elements of ITER magnetic systems» is chosen as an example. Services are engineering calculations of superconducting electromagnetic systems – choice of basic parameters and calibration calculation, and microservices – separate calculations of selected parameters.

The service ontology was generated by TODOS information technology [19]. It consists of 2 classes, 10 subclasses and 53 terminal nodes (56 nodes in total), linked by «IS-A» type connections and has the format *.xml and contains information about each service / microservice as well as their input parameters. The service selected for the example, «Stability calculation» is described by metadata, the values of which are calculated during the computer-aided workflow. Tools for viewing ontologies in the TODOS system will allow to present the register of services in the form of a frame, hierarchically ordered applications, graph and a table. Searching for services in the register to run an engineering workflow is performed by IT TODOS based on an algorithm for solving a choice task by ranking alternatives by a set of metadata values [20].

Using the XML file parsing / generation modules, the original file was converted to JSON format. Further processing of the ontology file allowed us to generate a model describing the workflow in BPMN format.

The bpmn-js module visualized the workflow, data flow and variables in the web user interface (Fig. 1), which allowed the user to edit the input at each stage of the process.

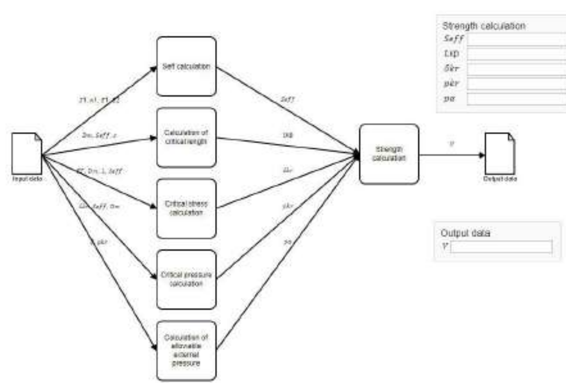


Figure 1. BPMN-visualization of workflow

Each service has its own GUID (Globally Unique Identifier) – unique key, by which each service workflow can be activated from the service ontology.

Software implementation of the computer-aided workflow algorithm involves the formation of a plan for its execution, that is, the sequence of execution of its services by BPEL-PM. Existing modern software tools and environments, in conjunction with the TODOS Ontology Model System, allow the creation and application of ontology models for various domains, form the single distributed intelligent business environment support platform for Industry 4.0.

V. Summary and Outlook on Future Work

The paper proposes an approach to computer-aided workflows design of and their components (functional services, communications and rules of interaction) by computer-aided generation of both a set of services that are components of workflows and sequence of their execution, using ontologies – meta-models of work process, services, domain, logical rules that establish relationships between functional services.

A formalized description of ontological domain models, functional services and workflows, as well as operations of computer-aided workflow generation using relationships and rules of application established between ontological models, are offered.

The proposed approach makes it possible the computer-aided workflows generation from functional services described in the domain ontology, and management the computation process depending on the input data.

The use of the proposed approach to the workflow generation makes it possible to computer-aided the choice of processing method (functional service) of input data, which is a very important factor in real-time systems. Depending on the data, the most efficient method for faster processing can be selected.

Using the domain ontology as a register of meta descriptions of functional services will help computer-aided workflow generation and define a functional service from many similar services that most closely match the conditions of use determined by the input data flow.

Further studies will be devoted to a more detailed consideration of the process of computer-aided workflows generation from sets of functional services, in particular, the computer-aided generation of program code for the execution of workflows.

References

[1] Industry 4.0: Building the Digital Enterprise. Available at: https://www.pwc.com/gx/en/industries/industries-4.0/landingpage/industry-4.0-building-your-digital-enterprise-april-2016.pdf

[2] Odette. Available at: https://www.odette.org/

[3] W. F. Boh, C. Soh, S. Yeo. Standards development and diffusion: a case study of RosettaNet. *Communications of the ACM*, 2007, vol. 50 (12), pp. 57–62.

[4] Flensburg P. Generation of ontologies for workflows. IN FLENSBURG P. & IHLSTRÖM C. (Eds.) *Learn IT, Know IT, Move IT – IRIS27*. Falkenberg, 2004. [Online]. Available: http://www.perflensburg.se/texts/iris27-1118.pdf

[5] Business Process Model And Notation. Available at: http://www.omg.org/spec/BPMN/2.0/.

[6] I. Grobelna, M. Grobelny, M. Adamski. Model Checking of UML Activity Diagrams in Logic Controllers Design. *Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, Advances in Intelligent Systems and Computing*, 2014, Vol. 286, pp. 233-242.

[7] ERwin Process Modeler. Available at: https://erwin.com/products/erwin-business-process/ http://erwin.com/products/erwin_process_modeler/.

[8] SPARXsystems. Available at: http://www.sparxsystems.com/products/index.html.

[9] Visio. Available at: http://office.microsoft.com/en-us/visio/.

[10] E. Börger. Approaches to modeling business processes: A critical analysis of BPMN, workflow patterns and YAWL. *Software & Systems Modeling*, 2012, vol. 11. pp. 1-14.

[11] L.Ivancic (2019) Business Process Modeling Vs. Process Mining: A Comparative Study of Methods and Software Tools. *10th Scientific Conference of Croatian Society for Quality*, 2019. pp. 41-41.

[12] Y. Imtiaz Khan, N. Guelfi. Slicing High-level Petri nets. *International Workshop on Petri Nets and Software Engineering (PNSE′14)*, 2014, vol. 1160, pp. 201-220.

[13] F. Kozisek, I. Vrana. Business Process Modeling Languages. *Agris on-line Papers in Economics and Informatics*, 2017, vol. 9, pp. 39-49.

[14] E. Stachtiari, A. Mentis, P. Katsaros. Rigorous Analysis of Service Composability by Embedding WS-BPEL into the BIP Component Framework. *Proceedings – 2012 IEEE 19th International Conference on Web Services*, ICWS 2012, 2012, pp. 319-326.

[15] T. Erçelebi Ayyıldız, A. Koçyiğit. Size and Effort Estimation Based on Problem Domain Measures for Object Oriented Software. *International Journal of Software Engineering and Knowledge Engineering*, 2018, vol. 28, No. 02, pp. 219-238.

[16] L. Silva, R. Braga, F. Campos. Scientific Workflow Composition in E-Science. 25th Brazilian Symposium on Software Engineering, 2011. pp. 273-282.

[17] T. Allard, P. Alvino, L. Shing, A. Wollaber, J. Yuen.. *A dataset to promote automated workflow analysis*. Article (PDF Available) in PLoS ONE 14(2):e0211486, 2019.

[18] T. Kot, A. Reverchuk, L. Globa, A. Schill A novel approach to increasing the efficiency of OSS / BSS workflow planning and design. *BIS 2012 Conference*. Springer, 2012.

[19] V. Yu. Velychko, M. A. Popova, V. V. Prykhodniuk. TODOS – IT platform for the formation of Transdisciplinary Information Environments. *Systems of Arms and Military Equipment*, 2017. Vol. 1, pp. 10-19.

[20] O.Ye. Stryzhak, V.V. Gorborukov, O.V. Franchuk, M.A. Popova. Ontology of the problem of choice and application in the analysis of limnological systems. *Environmental Safety and Natural Resources*, 2014, vol. 15, pp. 172-183.

Генерация рабочих процессов из онтологии предметной области

Глоба Л. С., Попова М. А., Бакай А. Н.

В работе освещён подход к автоматизации проектирования рабочих процессов и их компонентов (функциональных сервисов, связей и правил взаимодействия) путем автоматизированной генерации как набора сервисов, которые являются составляющими рабочих процессов, так и последовательности их выполнения, используя онтологии – метамодели рабочего процесса, сервисов, предметной области, логических правил, устанавливающих связи между функциональными сервисами.

Применение предложенного подхода позволило генерировать рабочие процессы на основе параметров, определенных в мета-моделях (онтологиях) реестра сервисов, и ограничений, наложенных онтологией предметной области, учитывая как функциональные, так и нефункциональные требования, реализуя преобразования рабочего процесса независимо от вычислений в его программной модели выполнения.

Для автоматизированной генерации рабочих процессов разработана программная среда, работоспособность которой проверена на примере генерации рабочего процесса с применением инструментальных средств построения и модификации онтологий ИТ ТОДОС и генерации схемы рабочего процесса средствами языка BPMN.

Received 12.12.2019

Use of Similarity of Wiki Pages as an Instrument of Domain Representation for Semantic Retrieval

Julia Rogushina
Institute of Software Systems
of National Academy of Sciences of Ukraine
Kyiv, Ukraine
Email: ladamandraka2010@gmail.com

***Abstract*—We propose an ontology-based approach to the Web semantic search that uses thesaurus representation of user task. Domain ontologies are considered as a source of semantic markup of the Wiki resources pertinent to retrieval domain. We use practical aspects of ontological approach to organization of Wiki-based information resources. Ontological model of Wiki resource formalizes the knowledge base structure and explicitly represents its main features. Domain ontologies, Wiki resources and task thesauri are generated independently by different applications but are used in general technological chain of user-oriented semantic retrieval. Open information environment is considered as an external data base with great volumes of heterogeneous and semi-structured information.**

Wiki ontologies are considered as the basis for establishing a semantic similarity between domain concepts pertinent to user task. Such Wiki-ontology elements as classes, property values of class instances and relations between them are used as parameters for the quantitative assessment of semantic similarity.

***Keywords*—semantic search, domain ontology, task thesaurus, semantic Wiki**

I. Introduction

The development of information technologies shows the tendencies to transition from traditional means of data processing to semantic computer systems oriented on work into the open environment. Now is the time for integration of the traditional information technologies with achievements of artificial intelligence. For example, theoretical principles and practical means of semantic computer systems development are designed by the Open Semantic Technology for Intelligent Systems Design (OSTIS Technology) [1]. But other important problem deals with interoperability of knowledge created and used by intelligent systems of different developers.

The growth in the use of the Web brings with it an increase in the number of interconnections among information systems and resources supporting the various aspects of human activities. Such interconnections have to be carefully prescribed to ensure interoperability. Standards of the Semantic Web [2] provides the universal ontology-based means of knowledge representation. Increasingly, the kinds of information structures being standardized today are much more complex than they were even a decade ago. However every practical task needs in specific methods of their use. In this work we consider an important component of open intelligent systems that deals with information retrieval on semantic level.

Open intelligent systems need in information retrieval tools and methods that search user-oriented information into the Web open sources. Such retrieval requires a formalized model of the search domain and description of user needs and interests in this process.

Ontologies are widely used to describe domains, but this causes a number of problems.

- Creating ontologies is a complex process that requires the involvement of a knowledge engineer and domain expert.
- The domain ontology is usually quite complex and contains a lot of unnecessary information for the specific task pertinent to user query.
- Processing an ontology and its matching with other information resources (such as unstructured natural-language texts) is a long and complicated process that requires the use of other background knowledge (e.g. linguistic knowledge bases). Therefore, it is advisable to use simpler information structures to formalize domain knowledge in information retrieval tasks.

We consider in this work the use of task thesauri as special cases of ontologies. Task thesaurus T is based on domain ontology O and consists of the ontological concepts (classes and individuals) joined by the semantic similarity to the user task in domain described by this ontology O.

II. Problem definition

An analysis of research works in the sphere of distributed information technologies shows that many intelligent tasks need in external sources of background knowledge from the Web, corporative and local networks, data warehouses etc. However, the problem of extracting such knowledge in the general case is extremely complex, and one of its components is semantic search that applies knowledge about user and user current task for selection of pertinent information resources.

To ensure the effective use of ontologies for semantic search by the various intelligent applications and to simplify the knowkedge processing process we propose to generate simplified ontology-based such information structures as thesauri. Every task thesaurus contains only such part of the domain knowledge that is needed to search for information that is pertinent to the user's current task. Thesaurus is a representation of semantically similar (in the local sense of current task) domain concepts related to this task.

This approach requires to justify the ways of ontology knowledge representation by means of thesaurus, to develop an algorithm for generating of such thesaurus based on the domain ontology and the description of the task. In addition, we need in development of methods for processing of this task thesaurus in applied retrieval systems and justification of their effectiveness of the received results for various types of such systems. It is also important to determine what information resources are used for creation of domain ontologies and what restrictions are imposed on the ontologies created in this way. In particular, ontologies that are generated by semantically marked Wiki resources often contain enough knowledge to carry out semantic search, but their processing is much simpler due to restrictions on their structure.

III. Task thesauri and their features

Wikipedia defines a thesaurus as a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order [3]. It is important that thesaurus as opposed to dictionary does not contain all word synonyms and their definitions.

Such definition does not use ontological approach but reflects the main characteristic of thesauri deal with it orientation on some particular task.

In the context of information retrieval, thesaurus is a form of controlled vocabulary that seeks to dictate semantic manifestations of metadata in the indexing of content objects. Its use is aimed to minimize semantic ambiguity by ensuring uniformity and consistency in the storage and retrieval. In this meaning thesaurus has to contain at least three elements: - list of words that correspond to domain terms, – the hierarchical relations between these words (e.g. parent/broader term; synonym, etc.), - a set of rules for thesaurus usage.

Thesaurus can be used for domain representation. If thesaurus represents ontological concepts as terms and uses ontological relations to link these concepts then we can consider such thesaurus as a special case of domain ontology oriented on analyses of natural language texts. Thesaurus contains only ontological terms (classes and instances) but does not describe all semantics of relations between them.

Some methods of thesauri generation use ontologies as a source of domain knowledge and integrate it with the current task description. Task thesaurus is a thesaurus that is generated automatically on base of the domain ontology selected by user and the NL description of particular task that is interesting for this user [4]. A simple task thesaurus is a special case of task thesaurus based on the terms of a single domain ontology. A composite task thesaurus is a task thesaurus that is based on the terms of two or more domain ontologies by operations on simple task thesauri of these ontologies.

Generation of the simple task thesaurus uses as input data two parameters:

- domain ontology selected by user;
- task description – the natural language text defines the current problem.

The text of task description contains elements related to the ontology concepts.

The process of simple thesaurus constructing contains two main steps:

- Step 1. Automated generation of the subset of ontology concepts correlates with fragments of task description:
  Substep 1.1 User explicitly and manually selects task-pertinent terms from the automatically generated list of classes and instance X. In the simplest cases, the construction of the thesaurus may be completed in this step, but it requires a lot of efforts from the user.
  Substep 1.2 Thesaurus is expanded with the help of various methods for processing of natural-language applied to task description (linguistic analysis, statistical processing, semantic markup analysis) that allow to detect NL fragments related to terms from O.
- Step 2. Expansion of the simple thesaurus by other ontology concepts according to the set of conditions that can use all elements of the O ontology.

Linguistic knowledge bases (KB) can be used for thesaurus construction. We can apply specific domain-oriented linguistic KBs that accumulate a large amount of lexical information. Such information is not universal and depends either from domain and natural language used in task definition. Therefore we cannot use Text Mining systems oriented on processing English texts. We apply direct updating of the domain lexical ontology by users and export linguistic knowledge from relevant vocabularies and knowledge bases, as well as from semantically marked Ukrainian texts.

In many cases, information about properties of ontological classes and individuals, their allowed values and their relations with other terms is appropriate in thesaurus constructing. Such information can be processed for refining the initially formed thesaurus in accordance with explicitly formulated user conditions. These conditions are defined by the specific nature of the task, but are not derived from its description and can be considered as meta-rules of retrieved information.

Complex task thesauri are generated from the built earlier task thesauri (simple or complex) with the help of set theory operations such as sum of sets, intersection of sets etc.

IV. Existing approaches to similarity measures

Similarity is a fundamental and widely used concept. Many researchers analyze the principles and measures of semantic similarity of domain concepts. In the cognitive domain, similarity is treated as a property characterized by human perception but use of this property in information systems requires the quantitative evaluations.

Now many similarity measures are used in various applications, such as information content, mutual information [5], Dice coefficient [6], distance-based measurements [7] etc. In [8] similarity defined in terms of information theory is applicable if domain has a probabilistic model. The similarity measure is derived from a set of assumptions about similarity and is not defined directly by some formula.

The similarity of concepts is also related to their content. One of the key factors in the similarity of the two concepts is the degree of information sharing in the taxonomy. The edge-counting method takes this into account indirectly. The information content of concept can be quantified by the logarithmic function of probability of concept use. Thus, the level of concept abstraction (i.e., its place in taxonomy) causes the less informational content. If there is a unique upper concept in taxonomy then its information content is 0. This quantitative characterization of information provides a new way of measuring semantic similarity based on the extension of concepts.

The more information is shared by two concepts, the more similar they are, and the information co-shared by the two concepts is determined by the information content of the concepts included in them into taxonomy.

Some measures of similarity [9] take into account only the depth of the nodes of the terms. Although the similarity is calculated taking into account all the upper bounds for the two concepts, the information measure allows to identify the minimum upper bound, but no class is less informative than its superclasses.

Measures to determine the semantic similar concepts (SSC) on the basis of ontologies use various semantic features of these concepts – their properties (attributes and relations with other concepts), the relative position in ontological hierarchies. The SSC set is a fuzzy set of concepts with the semantic distance less than the selected threshold.

Similarity is an important and fundamental concept in AI and many other fields. Various proposals for similarity measures are heuristic in nature and tied to a particular domain or form of knowledge representation.

The most general definitions of similarity are based on three intuitive assumptions:

- the similarity between A and B depends directly on their commonality;
- the similarity between A and B depends inversely on the differences between them;
- the maximum similarity between A and B is reached if A and B are identical, no matter how much commonality they share.

The similarity of two objects is related to their commonality depends directly on number of their common features and depends inversely on number of their differences. Concept similarity can be defines by similarity of strings and words. Feature vectors are widely used for knowledge representation, especially in case- based reasoning and machine learning. They can be applied for representation of words. Weights of features is used to account the importance of various features for word similarity. Some special features are applicative for natural language words and non-applicative for arbitrary strings of characters.

The similarity measures suppose that words derived from the same root as some initial word have the better similarity rankings. Other similarity measures are based on the number of different trigrams in the matching strings and on proposed by user definition of similarity under the assumption that the probability of a trigram occurring in a word is independent of other trigrams in the word. Similarity measures between words correspond to their distribution in a text corpus.

Semantic similarity can be based on similarity between concepts in domain taxonomy (such as the WordNet or CYC). The semantic similarity between two classes characterize not the set of their individuals or subclasses classes. Instead, generic individuals of these classes are compared.

A problem with similarity measures is that each of them is tied to a particular application or assumes a particular domain model. For example, distance-based measures of concept similarity assume that the domain is represented in a network. Another problem with the similarity measures is that their underlying assumptions are often not explicitly stated. Without knowing those assumptions, it is impossible to make theoretical arguments for or against any particular measure.

Methods aimed at SSC finding in different ontologies can be used to analyze the semantic similarity between the domain concepts. The assessment of similarity of concepts may be based on their positions in the hierarchy of classes with defined similarity: if the subclasses and superclass of these concepts are similar, then the same concepts are also similar.

The following parameters (features) can be considered into quantified similarity assessment of the two ontological classes:

- similarity assessing of their direct superclasses;
- similarity assessing of all their superclasses;
- similarity assessing of subclasses of concepts;
- similarity assessing of instances of classes.

Semantic similarity is a special case of semantic affinity. For the individual case of ontology, where the only relation between concepts is applied - the hierarchical relation of type IS-A, - taxonomy - the similarity of the two terms can be estimated by the distance between the concepts into the taxonomy.

The semantic distance between the concepts depends on the length of the shortest path between the nodes and the overall specificity of the two nodes. The shorter the path from one node to another, the more similar they are. If there are several paths between elements then the length of the shortest path is used [10]. This length is determined by the number of nodes (or edges) in the shortest path between two corresponding nodes of the taxonomy [11], taking into account the depth of the taxonomic hierarchy.

However, this approach is compounded by the notion that all taxonomy edges correspond to homogeneous distances. Unfortunately, in practice the homogeneity of distance in taxonomy is not supported.

In real taxonomies, there is great variability of distances covered by a single taxonomic relation, especially if some subsets of taxonomies (such as biological categories) are much denser than other ones. For example, WordNet [12] contains direct links between either fairly similar concepts or relatively distant ones. Therefore, it is advisable to take into account the semantics of relations between concepts for different taxonomic relationships and to consider the number of instances in subclasses.

In [13] a measure of semantic similarity is based on domain taxonomy that take advantage of taxonomic similarity in resolving syntactic and semantic ambiguity.

Semantic similarity represents a special case of semantic relation between concepts. In [11] the assessment of similarity in semantic networks is defines with the help of taxonomic links. Although other types of links such as “part-of” can also be used for assessment of similarity [14].

All these researches use only some aspects of ontological representation of knowledge limited by:

- hierarchical relations – taxonomic relations between classes and relations between classes and their individuals;
- other types of relations which semantics influence on their weight for the similarity but does not used in logical inference;
- properties of class individuals and their properties that matched in process of similarity estimation but do not analyzed at the level of expressive possibilities. However all these ontological features can be represented by semantic Wiki resources. Therefore we propose to use such Wiki resources as a source of semantic similar concepts for other intelligent systems.

Now a lot of software supports Wiki technologies. One of the most widely used is MediaWiki. The basic functionality allows to create pages connected by hyperlinks, set their categories and publish their content with some structure elements etc. Semantic MediaWiki (SMW) extends semantically this Wiki engine by use of semantic properties of pages [15]. SMW definitely displays content with these annotations in the formal description using the OWL DL ontology language [16].

V. SEMANTIC SIMILARITY OF CONCEPTS INTO THE WIKI RESOURCES

We approve the proposed above approach in development of semantic search and navigation means implemented into e-VUE – the portal version of the Great Ukrainian Encyclopedia (vue.gov.ua). This resource is based on ontological representation of knowledge base. To use a semantic Wiki resource as a distributed knowledge base we develop knowledge model of this resource represented by Wiki ontology [17]. This model provides semantic markup of typical information objects (IOs) by domain concepts [18].

Application of semantic similarity estimation for this IR provides the functional extension of Encyclopedia by new ways of content access and analysis on the semantic level.

One of the significant advantages of e-VUE as a semantic portal is the ability to find SSCs. Criteria of e-VUE concept similarity is based on the following assumptions:

- concepts that correspond to Wiki pages of the same set of categories are semantically closer than other e-VUE concepts;
- concepts corresponded to Wiki pages with the same or similar meanings of semantic properties are semantically closer than concepts corresponded

to Wiki pages with different values of semantic properties or those ones with not defined values;
- concepts defined as semantically similar by the both preceding criteria are more semantically similar than concepts similar by one of criteria.

e-VUE users can apply SSC search if they are unable to select correctly the concept category or if they enter concept name with errors. Similar concepts help to find the desired Wiki page. We propose to user retrieval of globally similar (by the full set of categories and values of semantic properties) and locally similar (by some subset of these features) IOs.

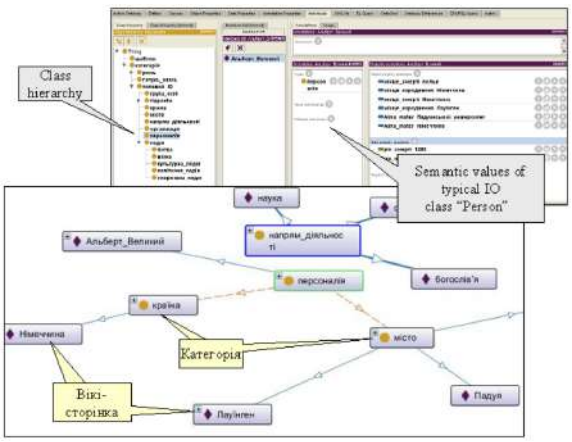


Figure 1. Wiki ontology defined the structure of e-VUE typical information objects (fragment).

Wiki ontology is a basis for research of similarity between concepts of Wiki resource because the estimates of similarity are based on such elements of semantic markup as categories and properties of these concepts. We can process in these estimates only those characteristics of IOs that are explicitly represented by ontology elements (Fig. 1).

Therefore the development of Wiki ontology defines the expressiveness of search procedure on base of Wiki resource marked by this ontology. Similarity can be defined by any subset of ontological classes and values of their properties but all other content of Wiki pages is not available for this analysis (these characteristics can be received by statistic analyses of from NL processing systems but they are over the consideration of this work).

According to the specifics of encyclopedic IR, it is impractical to search for pages that match all available parameters because some parameter groups are unique (for example, last name and year of birth) and some other ones dependent functionally on other parameters (although they have independent importance e.g. the name in the original language).

Therefore we realize the following examples of local SSPs retrived by:

- the fixed subset of categories of current page (Fig. 2);
- the values of the fixed subset of semantic properties of current page;
- the combination of categories and values of semantic properties of current page.

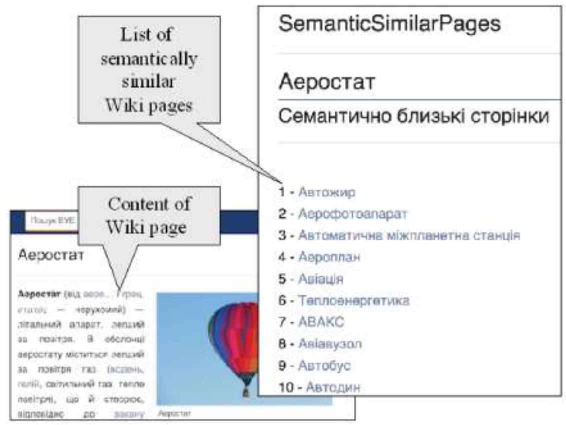


Figure 2. Semantic similar concepts of e-VUE for concept "Aerostat".

It should be noted that built-in tools of Semantic MediaWiki don‘t support search for SSCs (local and global) and all of these requests are realized by special API queries that analyze code of the Wiki pages.

VI. Use of SSCs for information retrieval

The set of SSCs can be considered as a thesaurus of a user's task for intelligent retrieval systems that support personified search of information pertinent to user needs. An example of such system is semantic retrieval system MAIPS based on ontological representation of background knowledge [19].

This system is oriented on users with stable informational interests into the Web. Ontologies and thesauri provie formalized conceptualization of subject domain pertinent to user tasks. The search procedure in MAIPS is personified by indexes of natural language text readability.

MAIPS uses OWL language for representation of domain ontologies and thesauri, it supports automated thesauri generation by natural language documents and set-theoretic operations on them. Task thesaurus in MAIPS is constructed directly by the user in order to display the specifics of the task which causes these information needs.

We propose the possibility to import this information from external Wiki resources where the set of thesaurus terms is generated as a group of semantically similar concepts. The most pertinent results user receives in situation if Wiki resource is matched semantically by terms of pertinent ontology.

User can improve this thesaurus on base the selected domain ontology by weights of concept importance for

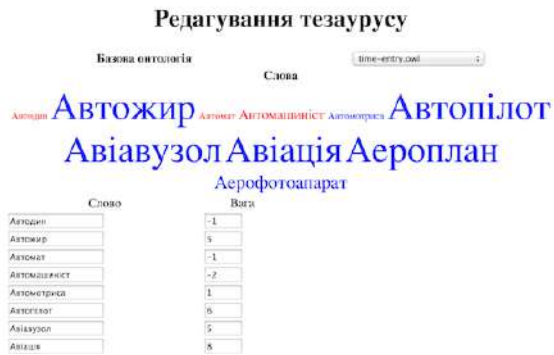


Figure 3. Visualization of domain thesaurus in MAIPS.

task. MAIPS visualise thesaurus by the cloud of tags (Fig. 3). Users can also manually edit any previously created thesaurus by adding or deleting some terms. In addition, MAIPS realizes such set-theoretic operations on thesauri as union, intersection and complement.

VII. Conclusion

The main idea for the study is to ensure the integration of various intelligent systems that use domain knowledge represented by anthologies. In order to simplify the processing of such knowledge we propose to pass from ontologies to their special case – thesauri. Actuality of this problem is caused by development of intelligent applications based on the Semantic Web technologies [20]. Thesaurus of task contains only limited subset of domain concepts and their relations.

Such knowledge structures are more understandable for users, their creation and processing take less time and qualification. We demonstrate some methods of automatic generation of thesauri by appropriate ontologies and Wikis, and on example of MAIPS we show the usage of such thesauri as a source of domain knowledge for intelligent information retrieval.

References

[1] Golenkov V., Shunkevich D., Davydenko I., Grakova N. Principles of organization and automation of the semantic computer systems development, 2019.

[2] Ray S.R. Interoperability standards in the semantic web, J. Comput. Inf. Sci. Eng., 2002, Vol.2(1), pp.65-69.

[3] "Thesaurus," Available at: https://en.wikipedia.org/wiki/Thesaurus (accessed 11.11.2019)

[4] Gladun A., Rogushina J. Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources, International Journal of Intelligent Systems and Applications, 2012, Vol.1, pp.11-20.

[5] Hindle D. Noun classification from predicate-argument structures, ACL-90, 1990, pp. 268-275.

[6] Frakes W. B. , Baeza- Yates R. Information Retrieval, Data Structure and Algorithms, 1992.

[7] Lee J. H., Kim M. H., Lee Y. J. Information Retrieval Based on Conceptual Distance in is-a Hierarchies, 1989, Vol.49(2), pp.188-207.

[8] Lin D. An information-theoretic definition of similarity, ICML, 1998, Vol. 98, pp. 296-304.

[9] Wu Z., Palmer M. Verb semantics and lexical selection, 32nd Annual Meeting of the Association for Comput. Linguistics. Las Cruces, 1994, pp. 133-138.

[10] Rada R. , Bicknell E. Ranking documents with a thesaurus, JASIS, 1989, Vol.10(5), P.304-310.

[11] Rada R., Mili H., Bicknel E., Blettner M. Development and application of a metric on semantic nets, IEEE Transaction on Systems, Man, and Cybernetics, 1989, V.19(1), pp.17-30.

[12] Fellbaum C. WordNet, Theory and applications of ontology: computer applications, 2010, pp. 231-243.

[13] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, Journal of Artificial Intelligence Research, 1999, Vol. 11, pp.95-130.

[14] Richardson R., Smeaton A.F., Murphy J. Using WordNet as a knowledge base for measuring semantic similarity between words, Working paper CA-1294, Dublin City University, School of Computer AppUcations, 1994.

[15] Rogushina J. Processing of Wiki Resource Semantics on Base of Ontological Analysis, Open semantic technologies for intelligent systems, 2018, pp.159-162.

[16] McGuinness D. L., Van Harmelen F. OWL web ontology language overview, W3C recommendation, 2004, 10(10).

[17] Rogushina J. Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies, International Journal of Mathematical Sciences and Computing (IJMSC), 2017, Vol.3, No.3, pp.50-58.

[18] Rogushina J. The Use of Ontological Knowledge for Semantic Search of Complex Information Objects, Open semantic technologies for intelligent systems, 2017, pp.127-132.

[19] Rogushina J. Use of the Ontological Model for Personification of the Semantic Search, International Journal of Mathematical Sciences and Computing (IJMSC), 2016, Vol. 2, No. 1, pp.1-15.

[20] Semantic Web, Available at: https://www.w3.org/standards/semanticweb/ (accessed 20.10.2019)

Использование подобия страниц Вики-ресурсов как инструмента представления предметной области для семантического поиска

Рогушина Ю.В.

Мы предлагаем основанный на онтологиях подход к семантическому поиску в Веб, использующий тезаурусное представление задачи пользователя. Онтологии домена рассматриваются как источник семантической разметки Вики-ресурсов домена.

Мы используем онтологии для формализации структуры базы знаний Вики-ресурсов, которая явно представляет ее основные функции. Онтологии, Вики-ресурсы и тезаурусы задач создаются независимо различными приложениями, но используются в общей технологической цепочке семантического поиска, ориентированного на пользователя. Открытая информационная среда рассматривается как внешняя база данных, содержащия большие объемы гетерогенной и частично структурированной информации.

Вики-онтологии рассматриваются как основа для установления семантического подобия между понятиями предметной области, которые относятся к задаче пользователя. В качестве параметров для количественной оценки семантического подобия используются элементы Вики-онтологии (классы, значения свойств экземпляров классов и отношения между ними).

Received 12.12.2019

The problem of convergence of intelligent systems and their submergence in information systems with a cognitive decision-making component

1st Kobrinskii B.A.
Federal Research Center
"Computer Science and Control"
of the Russian Academy of Sciences,
Institute of Artificial Intelligence Problems
Pirogov Russian National Research Medical University
Moscow, Russia
kba_05@mail.ru

2nd Yankovskaya A.E.
National Research Tomsk State University
Institute of Applied Mathematics and Computer Science
Tomsk, Russia
ayyankov@gmail.com

***Abstract*—In this paper the problems of convergence or synergy of knowledge in intelligent systems of one subject area, implemented on different approaches, are considered. The basis is the addition and synonymy of signs (terms) in the knowledge metabases. A version of the work of an information system user with various intelligent decision-making support systems using synonymous concepts is presented. It is proposed to use 2, 3-simplexes to make and justify a decision for converged intelligent systems, and in the dynamics of changes -– 2-simplex prism.**

***Keywords*—convergence of intelligent systems, computer-assisted intelligence design, knowledge metabase, synonym base, semantic interoperability, cognitive component**

«Harmony, or rather, “complementarity”, manifests itself only in general...»

— Henri Bergson [1]

«Assimilate the method of considering all things in their constant transition into each other»

— Mark Aurelius [2]

I. Introduction

The structure of an intelligent system includes a knowledge base, a decision inference mechanism, and an intelligent interface [3]. Intelligent systems (IntS) are divided into a number of types and are based on various principles of extraction (acquisition) and analysis of knowledge.

The first in time of occurrence was the class of expert systems (ES), presented on the basis of knowledge and assessments of experts.

Another class of intelligent systems, which can be called traditional, is based on revealing knowledge from literature data, documents or (in medicine) from arrays of health records. The knowledge revealed in IntS, as well as in ES, is presented in the form of a rule base. In some cases, the knowledge of highly qualified experts is also included in the knowledge base of the intelligent system [4].

A special class is made up of JSM-IntS [5], the solver of which is implemented on the basis of intelligent qualitative analysis of data, which is arguments and counterarguments for hypotheses accepting, followed by the presentation of selected signs as elements for explanation. The JSM-method for automated generating of hypotheses presents formalized heuristics for establishing the causes (presence or absence) of the effects studied, presented in open (updated) databases of structured facts.

Case-based reasoning [6] contains situations or precedents, and the solution comes down to abduction from particular to particular (to search by analogy).

L. Zade proposed to call hybrid those systems in which fuzzy logic is combined with neurocomputational, genetic algoritms and probabilistic computational [7]. In the future, other combinations began to be used.

The totality of concepts in intelligent systems of various types can correspond to judgments and hypotheses implemented on various approaches. In the future, it is necessary to search for a single solution to represent knowledge at various stages of intelligent systems building, which will allow the user to receive a number of hypotheses from different IntS in one session with a single software product for entry. However, given that the full integration of various intelligent systems seems theoretically impossible, we consider the concept of convergence, which is accepted in foreign literature to be considered as sinergy.

II. Integration or convergence of intelligent systems?

We introduce the concept of convergence (from Lat. Convergere – to approach, converge) as a way of convergence on the basis of common approaches to building the same type of modules with an agreed standardization of certain data, according to the option proposed for information systems (IS) [8].

The movement in the direction of convergence can be based on the principle of “weak integration” according to the principle of corporate information systems, which implies a certain degree of redundancy with a simultaneous margin of system flexibility [9]. In a broad sense, convergence refers to the interpenetration of several sciences and scientific fields.

On the convergence of scientific trends, some intelligent recognition systems are based. Among them, one can single out the theory of discrete automata, test pattern recognition, logical-combinatorial, logical-probabilistic algorithms, soft computing, cognitive graphics tools (CGT), dividing systems, optimization of the selection of fault-tolerant diagnostic tests [4], [10], [11].

We assume that the greater the interpenetration, the higher the degree of convergence.

In studies at the intersection of sciences and scientific fields, the use of convergence allows, through interpenetration, to obtain a synergistic effect in solving problems of analyzing data and knowledge, identifying various patterns, making and substantiating decisions, including using cognitive means.

The convergence of knowledge in IntS, based on various methods of artificial intelligence, should be considered at the stages of its creation: extracting knowledge from various sources and presenting it in a single structured form, forming knowledge bases (rule bases) and inference mechanisms, explaining proposed hypotheses, implementing interfaces, as well as fact bases. Let's consider some aspects of the example of medical applications. In this regard, it should be noted that the signs observed in patients are data, but for the doctor, relevant indicators correspond to knowledge about diseases and serve as arguments for putting forward diagnostic or prognostic hypotheses.

The ontology of taxonomic terms may be a structurally formalized representation in the model Resource Description Framework (Resource Description Framework – RDF), which is represented in the form of a triplet "subject -– predicate -– object" (for example, in the clinical version of the medical description "skin – has color – pale"). Many RDF-statements form a directed graph in which the vertices are subjects and objects, and the edges represent relationships. RDF is an abstract data model, that is, it describes the structure, methods of processing and interpreting data [12]. Based on RDF, for example, the Metagenome and Microbes Environmental Ontology was built [13], [14].

The complex of knowledge sub-bases of intelligent systems of the same subject area complementing each other, connected by certain relationships, will be called the metabase of knowledge. In such converging systems controlled by a body of knowledge, the exchange of information is possible through a common "memory" based on a unified representation of the information being processed through common terms. A complex of interrelated sub-bases of knowledge can carry out real interaction only under the condition of semantic interoperability.

III. The problem of semantic interoperability

The concept of compatibility of software systems should serve as the basis for fundamental communication, integration and effective sharing of knowledge and data by systems, organizations and users. In particular, semantic compatibility is a way of a common understanding of context and meaning, which ensures the correct interpretation of the message received, in particular the disease of a particular person, which is based on unambiguously understood factual data. In this regard, we can talk about a critical approach to the concept of semantic compatibility for communication and exchange of medical data.

Particular emphasis is needed in relation to clinical terminology [15]. An effective solution to this issue is possible with a coordinated exchange of information between intelligent systems on the basis of common classifiers and formats of knowledge about various indicators characterizing the manifestations of diseases at different stages of their progression. But this requires appropriate development of the composition of semantic information, its coding and rules for the exchange of knowledge.

The practical implementation of semantic compatibility requires vocabularies, taxonomies, ontologies and a description of relationships. That is, a semantic model of a subject domain can be implemented on ontologies integrated into a semantic network of a subject domain with their properties and relations between objects. Ontologies can perform an integrating function, providing a common semantic basis and a single platform [16]. On this basis, it is possible to represent knowledge, the subsequent combination of individual modules of various intelligent systems.

To each concept characterized by the presence of synonyms, ontologies of synonymous series must be added. Using a block of synonyms, primary input terms will be replaced with synonymous concepts presented in knowledge bases and case studies of various systems of the subject area. Thus, the semantic core of the subject domain metasystem should ensure the interoperability of intelligent systems built on various principles using classifiers, including standards for coding and exchange of characteristic features with the semantic "core" of the subject area (Fig. 1).

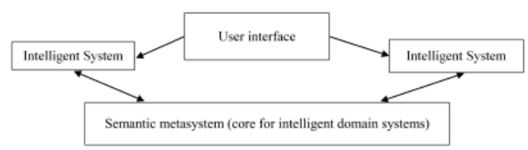


Figure 1. The semantic exchange scheme for the unification of terms

In medicine, these are clinical health indicators, for example, the Systematized Nomenclature of Medicine – Clinical Terms [17] or the search for hierarchical relationships between multiple terms that support multiple parents in a single term, for which semantic networks and trees were used [18].

The knowledge base should provide the possibility of "understanding" of synonyms as a result of appropriate labeling of synonymous concepts and the exchange of knowledge with the base or module of synonyms in the process of replenishment under the supervision of experts when a new concept is discovered (Fig. 2). New concepts may be included in the metabase or rejected. Meta-descriptions of domain objects in the framework of the ontological model make it possible to take into account their semantic meaning [19]. In healthcare, syntax for adapting knowledge has been developed for specific tasks. Arden Syntax for Medical Logic Module was originally developed to share health knowledge [20].

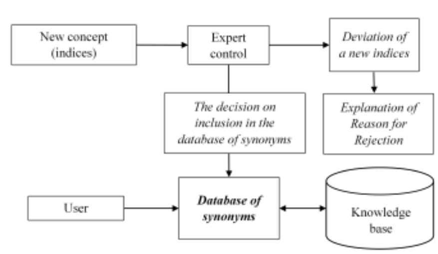


Figure 2. Refillable base synonymous concepts

For users of systems, unlike experts, it is necessary to provide access to the database of synonyms as read.

IV. Convergence at the stages of building intelligent systems

Convergence is impossible without the use of homogeneous classifiers, including, if necessary, synonymous concepts. These classifiers should be applied at every stage of building an intelligent system for a specific subject area (possibly for a separate problem area) – in the process of knowledge extraction, structuring of knowledge (including visual structuring), development of an explanation module and user interface.

Consider the option of building a converged IntS user interface. It is possible to enter data terminologically familiar by name for a particular user, followed by checking for the presence of synonyms and replacing it with a name that is adopted by agreement of experts in the subject domain. For this, there must be a base (module, block) of synonymous concepts of the subject domain, shown in Fig. 2.

When forming an explanation of hypotheses put forward by the system, it is also necessary to use a base of synonymous concepts.

The convergence of intelligent systems, implemented on a semantic network that combines the knowledge bases of individual intelligent systems through key concepts of subject domains, should generally correspond to the proposed concept of an OSTIS system [21], that is, all stored information should be structured on the basis of a thesaurus (multi-level classifier), which should serve as the basis for the component construction of system interfaces and should ensure the simplification and acceleration of their correction, if necessary.

V. Embedding of intelligent support in information systems

In the process of a user working with an information system, a need arises to support various solutions, for which more than one intelligent system can be used. The problem arises of knowledge-driven IS [22] while ensuring the semantic interoperability discussed above. Consider this with healthcare. In medicine, the idea of “embedding” decision-making support systems in electronic medical records arose in the last century. Its implementation can be represented by the example of the versatile American system Siegfried (Fig. 3) [23]. This system allows you to take into account 1700 guidelines for a variety of nosologies and give reasonable recommendations for the treatment of patients. In the 4th generation of improved databases of medical information systems (MIS), computer-assisted software design [24], although not intelligent, was constructed. Accordingly, at this stage there were still no solutions necessary to ensure user dialogue with the embedded system.

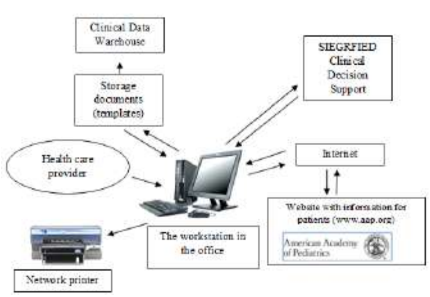


Figure 3. Siegfried: Medical system for Interactive Electronic Guidelines

At the same time, the task is to implement the necessary functional properties and characteristics that meet the requirements of doctors in the decision-making support modules. With this in mind, within the framework of the INTERIS information system [25], a specialized module for assessing the state of central hemodynamics in cancer patients was created. Based on six hemodynamic parameters, 12 syndromes were identified that reflect the state of blood circulation and the severity of hemodynamic disturbances in operated patients with oncological pathology. The results of the measured parameters are recorded in the MIS database, which provides a link between it and the intelligent decision-making support system. A resuscitator uses the information obtained when making decisions in the process of forming medical prescriptions. The DOCA+ clinical information system [26] provides an analysis of the fundamentally important functions of the body by signaling violations of the norm limits set for various indicators -— heart rate, blood pressure, etc., which allows you to issue warnings about potential risk for patients. From modern perspectives, the most efficient use of such functions in the treatment process is possible when moving to intelligent systems and using the method of situational management [27], which is especially necessary for dynamic intelligent systems. Similar situational management solutions can be applied in various fields.

The medical informatics manual [28] emphasizes the need to make greater use of active computer systems that generate various clinical signals or treatment recommendations. It is also proposed to use artificial intelligence as a cognitive assistant to a doctor, for example, to search for specific patterns in clinical data that indicate important changes in the patient’s condition.

VI. Interactive data exchange

The concept of “MIS-Mentor” was introduced in the five-level MIS paradigm, which implies an interactive dialogue between the intelligent module and the user to request signs that are missing for decision-making. The refusal to obtain additional data in the dialogue, as some authors believe [29], should be recognized as incorrect, since it is impossible to imagine that the doctor will introduce all the data necessary for the formation of hypotheses. Accordingly, the decision on an incomplete data set will lead to a decrease in the efficiency of the hypotheses proposed by the intelligent system. Previously, it was shown that IntS can be more confident in the diagnosis when receiving information about the patient’s manifestations of the disease in accordance with the recommendations on additional studies important for decision-making [30].

In the information system may be incorporated ("immersed") series of intelligent systems to provide the solution of various problems. In this case, the problem of their compatibility with each other arises. It is necessary to solve the problem of data and message exchange structure. The structure of the terminological representation of domain knowledge should be unified and replenished, otherwise no convergence is possible. O.I. Larichev et al. [31] drew attention to the need at any time to return to the stage of structuring and supplement the set of criteria.

As for the convergence of MIS [8], it seems appropriate to switch to a modular structuring of knowledge in artificial intelligence systems, which will simplify and accelerate re-structuring when new or adjusted knowledge is included in the thesaurus. Previously, the hierarchy of the structure of concepts was considered in the data field using the so-called multidimensional tubes, and the decomposition of the structure was presented as a meaningful analysis by its specialist [32]. This work was written to solve problems by computational methods of pattern recognition, but at a critical stage in the creation of knowledge-based systems, and included the evaluation of data by specialists in the subject area. However, the modularity of knowledge at the stage of its extraction and structuring has not yet been adequately reflected in the construction of intelligent systems. And without solving this issue, it is impossible for the user to switch in the process of working from one intelligent decision-making support system to another. The transition to the modular structure of classifiers will be the basis for the ability to interact with their individual sections-modules. At the same time, full user support should include various aspects of process control for which separate knowledge bases can be used.

It should be emphasized once again that when switching to the inclusion ("immersion") of intelligent systems in information systems, the main problem is on-line data exchange in those cases when there is not enough data available in the information system to make a decision and they must be additionally requested or clarified from user. Off-line mode is not acceptable for many situations.

VII. The Cognitive Component of Decision-Making

A large number of publications by the authors of this article are devoted to the cognitive component of decision-making. The Cognitive Graphics Tool (CGT) visually mapping a complex object, phenomenon or process on a computer screen, allowing users to formulate a new solution, idea or hypothesis based on visual elements. Cognitive graphic tools were developed as one of the areas of Artificial Intelligence in the 80s of the 20th century in the works of D. A. Pospelov and A. A. Zenkin. For the first time, A. E. Yankovskaya proposed the CGT n-simplex ($n = 2$ is an equilateral triangle, $n = 3$ is a regular tetrahedron) and developed in publications [4], [33], [34]. Initially, 2, 3 – simplices were also used to create dynamic intelligent systems, but in 2015 A. E. Yankovskaya and A. V. Yamshanov proposed to use a 2-simplex prism for various problem areas to create dynamic intelligent systems [34].

A 2-simplex prism is a triangular prism with identical equilateral triangles (2-simplexes) at its bases. The height of the 2-simplex prism [34] in dynamic intelligent systems corresponds to the considered time interval of the dynamic process. It is divided into several time intervals. The number of time intervals corresponds to the number of diagnostic or prognostic decisions. The distance between two adjacent 2-simplexes is proportional to the time interval between them (Fig. 4).

Under the direction of A. E. Yankovskaya user-oriented graphical, including cognitive, tools were developed, designed to visualize information structures, revealed regularities, make and substantiate decision-making both for specific problem areas (having a mapping in ordinary reality (naturalistic CGT) [4], [33], and for invariant to problem areas, that is, not mapping in ordinary reality [4], [33], [34].

Unfortunately, the scope of this article does not allow to consider even a part of the created cognitive graphics tools (CGT) to be stated. Here are just some of them devoted to the cognitive component of decision-making and the substantiation for decision-making [4], [33], [34]. The publication [4] presents the cognitive tool n-simplex ($n = 2$, $n = 3$), which allows you to save the sum of the distances to the faces and the relationship between the distances. This cognitive graphics tools of decision-making and substantiation of decision-making does not have a mapping in ordinary reality and, since it is invariant to problem areas, it is oriented to any problem areas, which is especially important for converging intelligent systems. The publication [33] presents CGT for specific problem areas that has a mapping in ordinary reality (naturalistic CGT). The cognitive way of mapping in ordinary reality is based on the use of images of real objects. In the publication [34], for synergistic IntS, a cognitive tool 2-simplex prism is presented, which is invariant to problem areas and implemented for medical and other problem areas that take into account changes in time (in the dynamics of the pathological process), and for geoinformation systems that take into account changes in space (for example, the distance on the map).

For the differential diagnosis of various diseases, the 3-simplex cognitive tool was used [4], [10], [11], which visualizes the degree of disease manifestation [33] (the presence of one or more of 4 diseases with an indication of the accuracy of each decision).

The cognitive tool "2-simplex prism" can be considered as the cognitive component of the convertible dynamic IntS (online dynamics options are not considered, the dynamics is considered as changes over time).

Distance from the basis of the 2-simplex prism to the i-th 2-simplex h'_i, representing the object under study at a given point in time, is calculated by the following formula:

$$h'_i = H' \cdot \frac{T_i - T_{min}}{T_{max} - T_{min}}$$

where H' is the height of the 2-simplex prism preassigned by a user and corresponded to the study duration, T_i is the moment of the i-th fixation of the parameters of the object under study, T_{min} is the moment of the first fixation of the parameters of object under study, T_{max} – the moment of the last fixation of the parameters of the object under study.

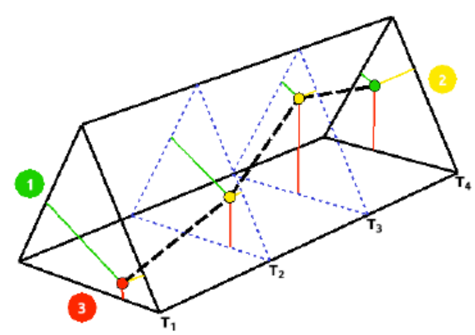


Figure 4. Example of 2-simplex prism usage for visualization of dynamical processes

Since the 2-simplex prism is based on the use of 2-simplex, all the properties of representing objects on a 2-simplex are also valid for a 2-simplex prism.

2-simplex prism allows you to model and map dynamic processes.

Examples of the use of a 2-simplex prism in synergistic dynamic IntS are presented in medicine for the differential diagnosis of various diseases [33], [34] and in geoinformation systems [34].

VIII. Conclusion

In the process of development of convergence (synergy) of intelligent systems and their integration with information systems, new opportunities are opened up for users in terms of working simultaneously in IS while providing broad support for decision-making in an interactive process with systems based on domain knowledge. But for this need good solutions for hooking together software components from various developers, network communication technologies provide standards that can be leveraged for low-level hookup and semantic integration [35]. The problems of interoperability at the semantic level are discussed above with respect to the construction of interfaces, knowledge bases, and explanation blocks. The cognitive graphics tool proposed for use in converged intelligent systems are based on two approaches: those that have and do not have a mapping in ordinary reality. Cognitive graphics tool are applicable for modeling various processes in problem and interdisciplinary fields, such as genetics, psychology, sociology, ecology, geology, construction, radio electronics, economics and others. For the decision-making and substantiation converged intelligent systems cognitive graphics tools 2, 3-simplexes have been developed, and for dynamic – 2-simplex prisms. It is shown that the use of cognitive graphics tool allows you to make and justify decisions in IntS both in statics and in dynamics.

The integration of convertible intelligent ostis-systems for decision-making support into information systems and their implementation in the format of cloud technologies will significantly expand the possibilities of using large systems that are not available for financial reasons to individual institutions/firms.

REFERENCES

[1] H. Bergson, “Creative evolution". Mineola, New York: Dover Publ., 2014. (in Russian).

[2] M. Aurelius, “Razmyshleniya” [Reflections]. Magnitogorsk: “AMRITA-URAL”, 1994, (in Russian).

[3] A.N. Averkin, M.G. Gaaze-Rapoport, and D.A. Pospelov, “Tolkovyj slovar’ po iskusstvennomu intellektu” [Explanatory Dictionary of Artificial Intelligence]. Moscow: Radio i svyaz, 1992, (in Russian).

[4] A. E. Yankovskaya, “Logicheskie testy i sredstva kognitivnoj grafiki” [Logical Tests and Means of Cognitive Graphics]. LAP LAMBERT Acad. Publ., 2011, (in Russian).

[5] V.K. Finn, “Epistemological foundations of the JSM method for automatic hypothesis generation”, Automatic Documentation and Mathematical Linguistics, vol. 48, iss. 2, pp.96-148, 2014.

[6] D. Leake, “Problem Solving and Reasoning: Case-Based”, International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Elsevier Ltd, 2015.

[7] L.A. Zadeh, “Rol’ myagkih vychislenij i nechetkoj logiki v ponimanii, konstruirovanii i razvitii informacionnyh / intellektual’nyh sistemi“ [The role of soft computing and fuzzy logic in the understanding, design and development of information / intelligent systems], Iskusstvennyj intellekt [Artificial Intelligent], no.2-3. pp.7-11, 2001, (in Russian).

[8] B.A. Kobrinskii, “Konvergenciya federal’nyh i territorial’nyh informacionnyh sistem v edinom informacionnom prostranstve monitoringa sostoyaniya zdorov’ya naseleniya” [Convergence of federal and regional information systems in a single information space for monitoring public health], Vrach i informacionnye tekhnologii [Physician and information technology], no.3, pp.46-49, 2008, (in Russian).

[9] V. Ananyin, “Granicy primeneniya servisa. ITIL i IT-arhitektury” [The boundaries of the service. ITIL and IT architecture], Intelligent Enterprise, no.2, pp.41-45, 2008, (in Russian).

[10] A.E. Yankovskaya, “Analiz dannyh i znanij na osnove konvergencii neskol’kih nauk i nauchnyh napravlenij” [Analysis of data and knowledge on the basis on the sinergy of several sciences and scientific directions], Intellektualizaciya obrabotki informacii: 8-ya mezhdunarodnaya konferenciya. Respublika Kipr, g. Pafos, 17-24 oktyabrya 2010g.: Sbornik dokladov [Intellectualization of information processing: 8th international conference. Republic of Cyprus, Paphos, October 17-24, 2010: Collection of reports]. M.: MAKS Press, 2010, pp.196-199, (in Russian).

[11] A.E. Yankovskaya, A.A. Shelupanov, and V.G. Mironova, “Construction of Hybrid Intelligent System of Express-Diagnostics of Information Security Attackers Based on the Synergy of Several Sciences and Scientific Directions”, Pattern Recognition and Image Analysis, vol. 26, no. 3, pp.524–532, 2016.

[12] G. Klyne, and J. Carroll, eds., “Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, 2004”, Available: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[13] J.C. Wooley, D. Field, and F.O. Glockner, “Extending standards for genomics and metagenomics data: a research coordination network for the Genomic Standards Consortium (RCN4GSC)”, Standards in Genomic Sciences, vol.1, pp.87–90, 2009.

[14] (2017, Sept.) “Metagenome and Microbes Environmental Ontology” [Online]. Available: https://bioportal.bioontology.org/ontologies/MEO

[15] T.N. Arvanitis, “Semantic interoperability in healthcare”, Studies in Health Technology and Informatics, vol.202, pp.5-8, 2014.

[16] A.A. Filippov, V.S. Moshkin, D.O. Shalaev, and N.G. Yarushkina, “Edinaya ontologicheskaya platforma intellektual’nogo analiza dannyh” [Unified ontological platform for intellectual data analysis], in Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nyh system: Materialy VI mezhdunarodnoj nauchno-tekhnicheskoj konferencii (Minsk, 18-20 fevr. 2016 goda) [Open Semantic Technologies for Intelligent Systems (OSTIS-2016): Proceedings of the VI international scientific and technical conference (Minsk, 18-20 febr. 2016)]. Minsk: BSUIR, 2016, pp.77-82, (in Russian).

[17] C. Price, and K. Spackman, “SNOMED Clinical Terms”, British Journal Healthcare Computing & Information Management, vol.17, no.3, pp.27-31, 2000.

[18] Y. Tao, E.A. Mendonça, and Y.A. Lussier, “A “Systematics” Tool for Medical Terminologies”. AMIA Annual Symposium Proceedings. vol.2003, pp.1028-1030, 2003.

[19] L.S. Globa, “Ispol’zovanie ontologij dlya sozdaniya scenariev obrabotki dannyh” [Using ontologies to create data processing scripts], in Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nyh system: Materialy VI mezhdunarodnoj nauchno-tekhnicheskoj konferencii (Minsk, 18-20 fevr. 2016 goda) [Open Semantic Technologies for Intelligent Systems (OSTIS-2016): Proceedings of the VI international scientific and technical conference (Minsk, 18-20 febr. 2016)]. Minsk: BSUIR, 2016, pp.51-56, (in Russian).

[20] J. Choi, Y.A. Lussier, and E.A. Mendoca, “Adapting current Arden Syntax knowledge for an object oriented event monitor”, AMIA Annual Symposium Proceedings, vol.2003, pp.814-816, 2003.

[21] V.V. Golenkov, and N.A. Gulyakina, “Semanticheskaya tekhnologiya komponentnogo proektirovaniya sistem, upravlyaemyh znaniyami” [Semantic technology of component design of knowledge-driven systems], Materialy V mezhdunarodnoj nauchno-tekhnicheskoj konferencii “Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nyh system” (Minsk, 19-21 fevr. 2015 goda) [Proceedings of the V international scientific and technical conference of Open Semantic Technologies for Intelligent Systems (Minsk, 19-21 febr. 2015). Minsk: BSUIR, 2015, pp.57-78, (in Russian).

[22] B.A. Kobrinskii, “Formirovanie medicinskih informacionnyh sistem, upravlyaemyh znaniyami” [Forming of medical information systems, knowledge management], in Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nyh system: Materialy VI mezhdunarodnoj nauchno-tekhnicheskoj konferencii (Minsk, 18-20 fevr. 2016 goda) [Open Semantic Technologies for Intelligent Systems (OSTIS-2016): Proceedings of the VI international scientific and technical conference (Minsk, 18-20 febr. 2016)]. Minsk: BSUIR, 2016, pp.89-92, (in Russian).

[23] D.F. Lobach, C.S. Gadd, J.W. Hales, and P. Baskaran “Siegfried: System for Interactive Electronic Guidelines with Feedback and Resources for Instructional and Educational Development”, Studies in Health Technology and Informatics, vol. 52, pt. 2, pp.827–831, 1998.

[24] A.G. Dean, “Computerizing public health surveillance systems”, in Principles and Practice of Public Health Surveillance, S.M.

Teutsch and S.M. Churchill, eds., 2nd ed. Oxford Univ. Press, 2000. P.229-252.

[25] E.G. Belyaev, M.V. Petrova, S.L. Shvyrev, and T.V. Zarubina, “Ispol’zovanie informacionnoj sistemy «INTERIS» dlya sindromal’noj ocenki narushenij central’noj gemodinamiki u onkologicheskih bol’nyh v rannem posleoperacionnom periode” [Using the INTERIS information system for the syndromic assessment of central hemodynamic disturbances in cancer patients in the early postoperative period], Informacionno-izmeritel’nye i upravlyayushchie sistemy. Informacionnye tekhnologii v medicine [Information-measuring and control systems], vol.9, no.12, pp.63-67, 2011, (in Russian).

[26] E. SHul’man, “Aksioma proaktivnosti medicinskih informacionnyh system”, [Axiom of proactivity of medical information systems], PC WEEK DOCTOR, no.2, pp.15-16, 2008, (in Russian).

[27] D.A. Pospelov, “Situacionnoe upravlenie: teoriya i praktika”, [Situational management: theory and practice]. M.: Nauka, 1968, (in Russian).

[28] E. Coiera, “The Guide to Health Informatics”, 3rd ed. Boca Raton, London, New York: CRC Press, 2015.

[29] V.L. Malyh, S.V. Rudeckij, and M.I. Hatkevich, “Aktivnaya MIS” [Active MIS], Vrach i informacionnye tekhnologii [Physician and information technology], no.6, pp.16-24, 2016, (in Russian).

[30] B.A. Kobrinsky, B.V. Marjanchik, P.A. Temin, and A.Yu. Ermakov, “Expert system for infantile spasm’s diagnostics”, in Epilepsia: Abstract, 22nd International Epilepsy Congress, vol.38, suppl.3, p.117. Dublin, Ireland, 1997.

[31] O.I. Larichev, I.V. Ashihmin, G.V. Rojzenzon, and E.M. Furems, “Podderzhka vybora luchshego ob"ekta na osnove nezavisimosti kriteriev po predpochteniyam i tranzitivnosti” [Support for choosing the best object based on the independence of criteria for preferences and transitivity], Novosti iskusstvennogo intellekta [News of Artificial Intelligence], no.4, pp.12-19, 2003, (in Russian).

[32] V.E. Levit, and V.S. Pereverzev-Orlov “Struktura i pole dannyh pri raspoznavanii obrazov” [Structure and data field in pattern recognition]. M.: Nauka, 1984, (in Russian).

[33] A.E. Yankovskaya, D.V. Galkin, and D.E. Chernogoryuk, “Computer visualization and cognitive graphics tools for applied intelligent systems”, in Proceedings of the IASTED International Conferences on Automation, Control and Information Technology, vol.1, pp. 249-253, 2010.

[34] A. Yankovskaya, “2-Simplex Prism as a Cognitive Graphics Tool for Decision-Making”, in Encyclopedia of Computer Graphics and Games, N. Lee, ed. Springer Nature Switzerland AG, 2019.

[35] K.R. Thórisson, “Integratea A.I. systems”, Minds and Machines, vol.17, no.1, pp.11-25 (2007).

Проблема конвергенции интеллектуальных систем и погружения их в информационные системы с когнитивной компонентой принятия решения

Кобринский Б.А., Янковская А.Е.

В настоящей работе рассматриваются проблемы конвергенции или синергии знаний в интеллектуальных системах одной предметной области, реализованных на различных подходах. В основу положены дополнение и синонимия признаков (терминов) в метабазах знаний. Представлен вариант работы пользователя информационной системы с различными интеллектуальными системами поддержки решений в рамках использования синонимичных понятий. Для принятия и обоснования принятия решения для конвергируемых интеллектуальных систем предлагается использовать 2, 3-симплексы, а в динамике изменений — 2-симплекс призму (2-simplex prism).

Received 12.01.2020

Modeling of Reasoning a Cognitive Agent with Significant Time Restrictions

1st Igor Fominykh
National Research University "MPEI"
111250 Moscow, Russia
igborfomin@mail.ru

2nd Alexander Eremeev
National Research University "MPEI"
111250 Moscow, Russia
eremeev@appmat.ru

3rd Nikolay Alekseev
National Research University "MPEI"
111250 Moscow, Russia
alekseevnp@list.ru

4th Natalia Gulyakina
Belarusian State University of Informatics and Radioelectronics
220013 Minsk, Belarus
guliakina@bsuir.by

***Abstract*—The problems of developing logical system to model the reasoning of cognitive agent are faced. Such an agent should be able to make conclusions based on its knowledge and observations in solving problems in the case of hard real-time. The hard real-time is characterized by setting a critical time threshold that considerably influences on agent's problem-solving.Exceeding this time threshold can lead to serious, sometimes catastrophic consequences and it is unacceptable for the agent. The formal basis of the modeling system is a logical system that integrates the concepts of active temporal logic and logical programming. Among the original approaches and methods proposed by the authors the following one should be mentioned. An integrated logical system that combines the concepts of active logic and logical programming has been built.. An approach to constructing a paraconsistent declarative semantics based on the concept of active logic has been introduced. The method of representing agent's temporal non-monotonous reasoning by active temporal logic has been proposed. The temporal granulation technique in logical system to formalize meta-reasoning has been suggested. Taking into account that the agent has to make decisions under the lack of time. the problem of the decision quality arises. In this context, it is useful to take branching time logics that allows us to infer various consequences of agent's decision. A subclass of such logics oriented to real-time systems applications has been considered. In general, the proposed methods and algorithms provide the conceptual and algorithmic bases for developing new generation intelligent systems able to function in the case of hard real-time.**

***Keywords*—cognitive agent, hard time constraints, active logic, step theory, temporal reasoning, logical programming**

I. Introduction

Various versions of active logic have been proposed for modeling reasoning in hard real time [1]- [4], which make it possible to observe the agent's reasoning process during its implementation. The creators of active logic emphasize its fundamental difference from traditional non-monotonic systems, such as default logic, auto-epistemic logic, etc. At present, there are dozens of different temporal logics, the purpose of which is to formalize the reasoning about time. The process of reasoning thus occurs as if out of time: the world as if stops while the system "thinks". For hard real-time systems, when solving problems, it is important to be able to estimate the amount of time available to them "to think" until it is too late to think. To do this, it is necessary to be able to correlate the steps and results of the conducted reasoning with the events occurring in the external environment. This type of reasoning is called reasoning situated in time. The General concept of active logic is described in [1]. As a model of deduction, active logic is characterized by language, many deductive rules, and many "observations". Reasoning situated in time is characterized by performing cycles of deduction called steps. Since the active logic is based on a discrete model of time, these steps play the role of a time standard – time is measured in steps. Agent knowledge is associated with the index of the step at which it was first obtained. The principal difference between active logic and other temporal epistemic logics is that temporal arguments are introduced into the language of agents own theories. A common drawback of most systems of Active Logic is the interpretation of time, in a sense, as the internal essence of these systems, the course of which is determined by the structure of the rules of inference used to obtain new formulas from existing ones. In all cases, the measure of time (standard) implicitly refers to the duration of the deductive cycle (=output step). Each execution of the deductive cycle corresponds to one "tick" of the virtual internal clock. It is also implicitly assumed that the duration of execution does not change from cycle to cycle, or that the changes are so small that they can be ignored. In reality, the duration of the deductive cycle is influenced by changes in the composition and structure of knowledge as a result of ongoing reasoning and

This work was supported by the Russian Foundation for Fundamental Research (RFFR) and Belarusian Republican Foundation for Fundamental Research (BRFR) (projects 18-51-0007 Bel_a, 20-57-00015 Bel_a)

observations of the external environment. In addition, the duration of deductive cycles can be influenced by random factors, such as power failures, in the operation of other technical systems, etc. in fact, the assumption of a constant duration of deductive cycles is akin to logical omniscience [5] and, like the latter, it often conflicts with reality. The report presents an approach in which time is treated as an external entity that is not related to the structure of knowledge and the speed of deductive cycles. Moreover, we propose a logical system (extended step theory [6], [7]) that integrates the concepts of active logic and logical programming, which allows us to optimize the relationship between the expressive capabilities of a cognitive agent and the complexity of calculations.

II. Main Results

Further, the main results in the field of modeling reasoning of a cognitive agent in the "hard" real-time mode, obtained by the authors to date, are considered.

An analytical review of studies has been carried out, including studies on the capabilities of existing logical systems for formalizing reasoning (meta-reasoning) of a cognitive agent in hard real time; by methods of granulation of time, methods of solving the problem of logical omniscience.

In the field of formalizing reasoning with limited time resources, there are several different directions, each of which is to one degree or another connected with solving the problem of logical omniscience, without which the conduct of reasoning (meta-reasoning), strictly limited in time, is not correct. Moreover, within the framework of existing epistemic logics, various restrictions on the ability of cognitive agents to logical inference (rational behavior) were proposed. This allowed us to solve the problem of logical omniscience, but at the same time there was a significant decrease in the capabilities of the cognitive agent, whose behavior was modeled. Today, only a few approaches to the creation of logical systems are known in which the problem of logical omniscience is solved without a serious limitation of the “mental abilities” of agents. One such example is the epistemic logic proposed by D. Ho [8]. It is based on the idea of introducing into the logical language special modal operators interpreted as “mental efforts” necessary to obtain any knowledge expressed in the form of logical formulas. This system overcomes the problem of logical omniscience. but it does not allow modeling the reasoning of a cognitive agent when it is necessary to determine whether it is able to solve a problem without going beyond certain time boundaries. Another example is active logic created by a team of specialists from Active Logic Grupp, and the like) [1]- [4]. It is a fairly general concept, which meets the logic presented in the report, a system based on the interpretation of reasoning as a process that proceeds in time. For this purpose, a temporal parameter is introduced into the metalanguage of logical systems that meet this concept. However, today there are a number of problems associated with active logic and other similar systems that hinder its actual practical application. Among the most important, it is necessary to highlight the absence of logical systems that meet the concept of active logic and have paraconsistent semantics, which makes it difficult to use active logic systems if there are contradictions in the information available; the lack of estimates of the computational complexity of reasoning and meta-reasoning, which are formalized by systems that meet the concept of active logic; lack of research results regarding the completeness and semantic consistency of systems that meet the concept of active logic. The logical system proposed in this report is largely free from these shortcomings due to the integration of formalisms of active logic and logical programming implemented in it.

The concept of time granulation as a special case of information granulation is developed, and like information granules, this representation of time in the form of granules - indistinguishable objects. Formally, granules can be represented as a neighborhood of points, intervals, fuzzy sets, etc. The concept of “granulation of time” was first introduced in the formalism of TLC (Temporal Logic with Clocks) [9]. In the Active Logic formalism, time granulation is introduced by analogy with TLC and reduces to the fact that the duration of deductive cycles, assumed constant in classical active logic, is not performed for hard real-time systems.

In reality, the duration of the deductive cycle is influenced by changes in the composition and structure of the agent’s knowledge as a result of his reasoning and observations of the external environment. In addition, random factors, such as power outages, other technical systems, etc., may affect the duration of deductive cycles. Also, “thinking abilities”, in this case, the duration of the computational cycles of different agents, ceteris paribus, can be different. To simulate the possibility of changing the duration, we propose a modification of the classical active logic - Step Logic (a lot of rules with a binary preference relation specified on it), which provides these capabilities by assigning the so-called. hours of model run, simulating the behavior of the system in various conditions (runs). A model run clock is a finite or infinite strictly increasing subsequence of a global clock whose members are interpreted as time instants (on a global clock) of the completion of deductive cycles, for example, <3, 5, 7, 10, ...>. By changing the model’s running hours, it is possible to simulate various operating conditions of the system and better reflect, for example, features such as an increase in the duration of deductive agent cycles as the amount of information known to him increases or in connection with the failure of part of his computing resources.

The concepts of metacognition, counting the time spent on conducting reasoning, paraconsistency of agent metareasoning based on active logic formalisms are developed.

The term metacognition was proposed in the works of J. Flavell [10] and is defined by him as the individual's awareness of his cognitive processes and related strategies, or, in other words, as reasoning about reasoning, meaning "cognition second order". The difference between cognitive and metacognitive strategies should be noted. The former helps the individual to achieve a specific cognitive goal (for example, to understand the text), and the latter are used to monitor the achievement of this goal (for example, self-inquiry for understanding this text). Metacognitive components are usually activated when cognition fails (in this case, it may be a misunderstanding of the text from the first reading). Such failure activates metacognitive processes that allow the individual to correct the situation. Thus, metacognition is responsible for the active control and sequential regulation of cognitive processes. The concept of "metacognitive cycle" was proposed by M. Anderson [11] in the context of using the principles of metacognition to improve resistance to anomalies of a rational agent with a limited time resource. It is defined as the cyclical implementation of the following three stages: self-observation (monitoring); self-esteem (analysis of the revealed anomaly), self-improvement (regulation of the cognitive process). At the self-observation stage, meta-reasoning comes down to checking for the presence in the argument of an agent that solves a certain problem, formal signs of the presence of anomalies. These formal features are often direct contradictions in the agent's knowledge (the presence of a counter pair of formulas expressing the agent's current knowledge). At the self-assessment stage, the degree of threat to the quality of the agent's functioning that the identified anomaly bears is established, and at the self-improvement stage, if the threat is real, a new strategy for solving the problem faced by the agent is selected.

The countdown is achieved using the special predicate now (.) Introduced in the rules. Moreover, now (t) takes the value "true" if and only if t is the time moment of completion of the last of the deductive cycles already completed, that is, in other words, when t is the current time. At the same time, the time counting principle used in this project is free from the unrealistic assumption of a constant duration of deductive cycles inherent in other existing approaches to solving the problem of modeling metaraguments. As you know, logic is called paraconsistent if it can be the basis of conflicting, but not trivial theories. In turn, a contradictory theory is a logical theory in which a certain proposition and its negation are simultaneously provable, and a trivial theory is a contradictory logical theory in which any proposition is provable or formally: for any formulas A and B,$\{A, \neg A \vdash B\}$. As a result of the analysis, it was found that the paraconsistency of classical active logic has not yet received a theoretical justification. At the same time, it was shown that the proposed step theory, based on the integration of the concepts of active logic and logical programming, is paraconsistent in the sense that the existence of contradictions in step theories does not lead to their destruction, as is the case in standard logical systems.

A method has been developed for evaluating the time resource available for a cognitive agent based on the proposed logical system.

It seems obvious that for agents with a strictly limited time resource, it is impossible to control this resource without correlating the results obtained in the course of the cognitive process (the process of solving the problem) with the times when these results were obtained. In accordance with the concept of step logic, this process, which proceeds in time, is characterized by the execution of deduction cycles (output steps).

As noted earlier, the time is counted using the special single predicate now (.). The following inference rule applies to this predicate:

$$\frac{t : now(t)}{t+1 : now(t+1)}, \tag{1}$$

moreover, the $now(t)$ formula is not inherited at the time instant (at the output step) t + 1, as is the case with "ordinary" formulas due to the inference rule

$$\frac{t : A}{t+1 : A}, \tag{2}$$

Also, a formula of the form $resource(t)$ is not inherited, obtained using the following inference rule, which allows you to evaluate a temporary resource at any time:

$$\frac{t : resource(t_1)}{t+1 : resource(t_1 - 1)}, \tag{3}$$

where t_1 is the time resource available to the agent at time t.

An important feature of step logic systems is the introduction of temporal parameters into the metalanguage, which determines their operational semantics, and the output steps play the role of a temporary reference. Agent knowledge is associated with the index of the step at which it was first acquired. This illustrates the inference rule, which is the "active" analogue of the modus ponens rule:

$$\frac{t : A, A \vdash B}{t+1 : B}, \tag{4}$$

This rule "says" that if at an instant t, an agent derived from reasoning or obtained from observing the external environment formulas A and $A \rightarrow B$, then at time t +1, formula B will be derived.

Moreover, the assumption that the duration of deductive cycles is always the same is unfair. The time moments of completion of the output steps form a sequence ("clock"), which is a subsequence of a sequence of natural numbers, for example, clock = <1, 3, 5, 7, 10, 14, ...>.

Methods have been developed to control the intermediate results and the time of their receipt on the basis of the proposed logical system. In the conditions of severe time constraints, it is extremely important to control the course of the reasoning process, primarily identifying the anomalies that arise. To do this, it is necessary for the agent to be able to realize not only what he knows at a given moment in time, but also what he does not know at this moment. In accordance with the concept of step logic, such an ability (which can be called self-cognition) is achieved thanks to two rules of inference:

$$\frac{t:C}{next(t):K(t,C)}, \frac{t:C, sub(A,C), [A]}{next(t):\vdash K(t,A)}, \tag{5}$$

where A is any formula that is not known to the agent at time t, but is a subformula of some known formula C, i.e. recognized by the agent, $sub(.,.)$ is a double meta-predicate expressing the relation "to be a subformula", $[A]$ is a notation meaning that the formula A is absent in the agent's current knowledge at time t. $K(.,.)$ Is a double meta-predicate (and not a modal operator!), Expressing the fact that the agent knows some formula at some point in time.

The above rules are used in order to be able to compare the current state of knowledge of the agent with those expectations that he had regarding the reasoning process carried out by the agent. If these expectations conflict with reality, this indicates a possible anomaly and the need to take measures to eliminate its consequences.

The syntax and declarative semantics of the language of the logical system are developed, combining the concepts of active logic and logical programming (LP) - the Extended Stepping Theory. The theory got its name by analogy with the extended logical programs introduced by A. Lifshitz and M. Gelfond [12] as applied to the logic programming paradigm. The language of this logical system includes two types of negation. One of them corresponds to the usual ("strong") logical negation, while the second, called "subjective", in a sense is similar to the default negation (negation as failure) in the LP, but has the following important difference. While in LP the meaning of negation by default lies in the fact that the negated formula (in the LP is always a literal) could not be deduced using the given logical program, the subjective negation in the considered logical system means that the negated literal could not be deduced by the current moment in time. Thus, in the system under consideration, the principle of self-knowledge is implemented, which consists in the fact that an agent whose behavior is modeled by a logical system is able to recognize and express explicitly not only what he knows, but also what he does not know at the moment. Note that such an opportunity is especially in demand when managing the cognitive process in the conditions of severe time constraints. This allows you to make managing the process of solving the problem more efficient compared to using other existing meta-reasoning formalisms in which this principle is not implemented.

Extended step theories are pairs of the form $T = (R, Ck)$, where R is a finite set of named rules, Ck is a clock of step theory, which is a finite subsequence of a sequence of natural numbers. The members of this subsequence characterize the duration of sequentially performed deductive cycles that determine the process of reasoning in all systems of active logic. At the same time, in this project, in contrast to classical active logic systems, the original principle of time granulation is used, which is implemented using the concept of clocks of step theories. The latter allow one to take into account the difference in the duration of deductive cycles and increase the temporal sensitivity of the step theory (i.e., the dependence of the results of the argument on how quickly the available time resource is spent). At the same time, it is interpreted as an external entity, independent of the internal structure of the set of rules of step theory, while in early versions of active logic systems, time is rigidly tied to the internal structure of the knowledge base used, which is a drawback.

The properties of completeness and correctness of the declarative semantics of a logical system for formalizing meta-reasoning in relation to various existing semantics of logical programs, including the semantics of stable models, the semantics of response sets, and others are investigated. The conditions and restrictions imposed on the language of the logical system for formalizing meta -reasoning, in which its semantics is correct and / or complete in relation to the semantics of logical programs listed above. The concepts of correctness and completeness of two semantics are refined using the relationship of logical sequence. It has been established that the declarative semantics of the logical system are correct and incomplete with respect to all the semantics listed above, except when the extended logical program is stratified. It was also established that the declarative semantics of a logical system are not only correct, but also complete with respect to the semantics of the set of answers of stratified extended logical programs.

The paraconsistency of the declarative semantics of the logical system for formalizing meta-reasoning is proved. The consistency of semantics informally means that the presence of contradictions in the theory does not lead to its destruction (i.e., it turns out to be trivial in a sense). The proof of the paraconsistency theorem for the declarative semantics of step theories with two types of negations is constructed similarly to the previously proved theorem on the consistency of semantics of step theories with only strong negation.

Studies have been carried out to determine the relationship of step theories of the logical system to formalize meta-reasoning and advanced logical programs. Formalisms of step theories with two types of negation and advanced logical programs were created for various purposes. The main difference between these formalisms is that in step theories, reasoning is treated as a process that develops over time, while from the point of view of the semantics of extended logical programs, reasoning is static. A comparison of the formalisms under consideration was reduced to a comparison of the logical sequence relations that they define and was carried out using some translation that transforms the rules of an extended logical program into the rules of step theory.

The temporal logic of branching time (branching temporal logic) can be used to solve the problems of training, forecasting and modeling in intelligent systems, when it is necessary to consider time branching into the future. The application of this logic allows us to simulate, as noted earlier, the possible consequences of the solution (or solutions) found by the agent under rather tight time constraints. In work [13], various temporal logics were considered in terms of their application in intelligent real-time systems (IRS). As a basis for use in the IRS, the recommended BPTL (Branching-Time Propositional Temporal logic), proposed in [14] and is an extension of propositional temporal logic (PTL). PTL is a modal temporal logic built on the basis of classical logic with added modal operators for discrete linear time.

The PTL syntax is defined as follows. The L_p PTL language is a countable set of propositional symbols $p, q, r, s, \ldots$ Formulas are constructed using the following symbols:

- A variety of propositional characters of L_p;
- Classic ligaments: $T, F, \neg, \wedge, \vee, \Rightarrow, \Leftrightarrow$;
- Temporal operators of the future tense: unary - $\boldsymbol{O}, \Box, \Diamond$; binary - U, W;
- Temporal operators of the past: unary - $\otimes, \bullet, \blacklozenge, \blacksquare$; binary J, Z;

A variety of well-formed formulas (wffs) PTL are:

- All propositional characters of L_p are wffs;
- If A and B are wffs, then
  $T, F; \neg A, A \wedge B, A \vee B, A \Rightarrow B, A \Leftrightarrow B$;
  $\boldsymbol{O}A, \Box A, \Diamond A, AUB, AWB$;
  $\otimes A, \bullet A, \blacklozenge A, \blacksquare A, AJB, AZB$ are wffs too;

The intuitive (informal) meaning of modal operators is as follows. Unary: $\boldsymbol{O}$ - next, $\bullet$ - last, $\otimes$ - last), $\Box$ - always in the future, $\blacksquare$ - always in the past, $\Diamond$ - sometime in the future, $\blacksquare$ - sometime in the past; binary: U - until, W - unless, J - since, Z - zince. If A and B are propositional formulas, then the intuitive meaning of modal formulas is defined as follows: "wff $\boldsymbol{O}A$ is true at the moment (in a given state) if wff A is true at the next moment; wff $\Box A$ ("always" A) is true at the moment, if and only if A is true at all future moments (states, including the current one); wff $\Diamond A$ ("eventually - finally" A) is true at the present moment if and only if A is true at some future moment.The strict wff until AUB is true at the moment if and only if the wff B is finally true, i.e. at the moment $s > n$, where n is the current moment, and the wff A is true for all moments t such that $n \leq t < s$. The operator W is a weak version of the operator U when it is not guaranteed that wff B is true at some future moment. Temporal operators of the past tense are defined as a strict version of the past tense for the corresponding operators of the future ("future" twins), i.e. the past does not include the present.

Semantics of PTL. To define the semantics of PTL, the semantics of the possible worlds of Kripke are used. A possible world is considered as a set of states in time, connected by temporal relations from a set of permissible relations R. Formally, the world is defined by a pair $A = (S, R)$, where S is a nonempty set of possible states, R is a binary relation, $R \subseteq S \times S$.

Considering L_p as a set of atomic statements, the model of the world can be defined as $M = (R, S, V)$, where V is the valuation function defining the map $V : S \times L_p \to \{T, F\}$, that is, calculating a propositional value for each state $s \in S$. By introducing various constraints on the relation R, various model structures are obtained. For example, if we introduce the restriction of antireflexivity (<), then we obtain a discrete model. For discrete linear models, the set S can be considered as a sequence of states, R - as the relation of following or successor. The interpretation is given by the pair $\langle M, i \rangle$. where M is the model and i is an integer indexing the states $s_i \in S$ in the model.

The semantics for temporal wff are defined using the relationship $\models$ between interpretation and wff. Thus, the statement $\langle M, i \rangle \models A$ means that the wff A is interpreted in the model M as a state with index i. The axiom system for linear PTL is consistent and complete. In a linear discrete PTL, the time model is an ordered sequence of natural numbers, i.e. each state has one and only one successor.

In the branching BPTL logic, a single successor is not necessary for each state and there can be many possible paths from any given state and, therefore, several different "future" ones are possible. The time model is an infinite tree, each vertex of which has a finite number of successors. The top of the tree is regarded as a possible state, and a branch or path is considered as the history of a possible world. The semantics of BPTL are defined in terms of the model structure $M = (S, R, V)$, where S, R and V are defined similarly to PTL. The concept of branching time requires the introduction of the linearity condition into the past and the transitivity of R. BPTL wffs are state formulas, and path formulas are auxiliary objects introduced in order to facilitate the determination of the semantics of state formulas. A system of axioms for BPTL is defined and BPTL is proved to be complete with respect to all branching time structures. Inference algorithms for BPTLs with a focus on IRS were proposed in [13].

On the whole, the obtained results create the necessary conceptual basis for constructing promising systems for modeling the reasoning of a cognitive agent that functions under strict time constraints by combining the concepts of active logic and logical programming.

Conclusion

The principal differences of active logic from traditional nonmonotonic logics such as default logic, auto-epistemic logic, etc. are formulated. (rejection of logical omniscience, the presence of temporal sensitivity and self-knowledge). The advantages of the step theory are formulated in comparison with other Active Logic systems (improved characteristics of computational complexity, paraconsistency, implementation of the principle of time granulation). The consistency of the step logic allows one to avoid the destruction of the entire system of reasoning, despite the presence of conflicting information. To further improve the management of the process of reasoning, formalisms of extended step theories are used, which differ from standard step theories by the introduction, along with a strong negation of subjective negation, which allows the cognitive agent to recognize and express explicitly not only what he knows, but also what he does not knows at the moment. This improves the expressive capabilities of the theory and, in particular, the property of temporal sensitivity.

The use of temporal logic of branching time is proposed, which allows modeling (deriving) various consequences of a solution found by an agent. A subclass of such logic, oriented to application in real-time systems, is considered.

The results can be used in the design of complex dynamic systems of hard real-time, including the design of control systems for vehicles (ships, aircraft), power systems, power plant units and their simulators.

References

[1] Elgot-Drapkin J. Step Logic: Reasoning situated in time. PhD thesis. Department of computer science, University of Maryland, Colledge-Park, Maryland, 1988

[2] Purang K., Purushothaman D., Traum D., Andersen C., Perlis D. Practical Reasoning and Plan Executing with Active Logic. Proceedings of the IJCAI'99 Workshop on Practical Reasoning and Rationality, 1999

[3] Perlis D., Purang K., Purushothaman D., Andersen C., Traum D. Modeling time and metareasoning in dialog via active logic. Working Notes of AAAI Fall Symposium on Psychological Models of Communication, 2005

[4] Hovold J. On a semantics for active logic. MA Thesis, Department of Computer Science, Lund University, 2011

[5] Vinkov M.M., Fominykh I.B. Discussions about knowledge and the problem of logical omniscience. Part 1. Modal approach. Artificial Intelligence and Decision Making. vol. 4, 2011, pp. 4-12 (In Russian)

[6] Vinkov M.M., Fominykh I.B. Stepping Theories of Active Logic with Two Kinds of Negation. Advances in electrical and electronic engineering. 2017, vol. 15, no. 1,pp. 84-92.

[7] Vinkov M.M., Fominykh I.B. Extended Stepping Theories of Active Logic: Paraconsistent Semantics. Proceedings of the 16th Russian Conference RCAI 2018 Moscow, Russia, September 24-27, 2018. Communications in Computer and Information Science v.934 Springer, pp.70-78, 2018

[8] D. N. Ho, Logical omniscience vs. logical ignorance. On a dilemma of epistemic logic. Progress in Artificial Intelligence. Proceedings of EPIA'95, LNAI, vol. 990, Springer Verlag, 1995, pp. 237-248

[9] C.Liu, M.A. Orgun. Verification of reactive systems using temporal logic with clocks. Theoretical Computer Science 220 (1999) 377—408

[10] Flavell, J. H. Metacognition and cognitive monitoring: A new era in cognitive- developmental inquiry. American Psychologist, 34(10), 1987. - pp. 906-911

[11] Michael L. Anderson and Tim Oates and Waiyian Chong and Don Perlis. The metacognitive loop I: Enhancing reinforcement learning with metacognitive monitoring and control for improved perturbation tolerance. Journal of Experimental and Theoretical Artificial Intelligence. 2006. 18. 3. 387-411

[12] Gelfond, M., Lifschitz, V. Classical Negation in Logic Programs and Disjunctive Databases. New Generation Computing 9, 365–385 (1991)

[13] Eremeev A.P., Kurilenko I.E. An Implementation of Inference in Temporal Branching Time Models. Journal of Computer and Systems Sciences International, 2017, Vol. 56, No. 1, pp. 105–124

[14] I.S. Torsun, Foundations of Intelligent Knowledge-Based Systems (Academic, London, 1998).

Моделирование рассуждений когнитивного агента при существенных временных ограничениях

Фоминых И. Б., Еремеев А. П.,
Алексеев Н. П., Гулякина Н. А.

Рассматриваются вопросы проектирования системы моделирования рассуждений когнитивного агента, способного на основе своих знаний и наблюдений за внешней средой делать умозаключения, решая задачи в режиме «жёсткого» реального времени. Для работы в таком режиме характерно существование критического временного порога, установленного для решения стоящей перед агентом задачи. Превышение порога чревато тяжёлыми, подчас катастрофическими последствиями и для агента является неприемлемым. Формальной основой системы моделирования является логическая система, объединяющая концепции активной темпоральной логики и логического программирования. Среди предлагаемых авторами в работе оригинальных методов следует отметить подход к объединению концепций активной логики и логического программирования в одной логической системе; подход к построению паранепротиворечивой декларативной семантики логической системы, имеющей в основе концепцию активной логики; метод формализации темпоральных, немонотонных рассуждений агента средствами активной темпоральной логики; метод грануляции времени в логической системе для формализации мета-рассуждений. Учитывая, что в системах жесткого реального времени агенту часто приходится принимать решения в условиях недостатка времени, то возникает вопрос о качестве найденного решения. В этом плане полезно использование темпоральной логики ветвящегося времени, позволяющей смоделировать (вывести) различные последствия найденного агентом решения. Рассматривается подкласс такой логики, ориентированный на применение в системах реального времени. В целом, предлагаемые методы создают концептуальные и алгоритмические основы для построения перспективных интеллектуальных систем жёсткого реального времени нового поколения.

Received 17.01.2020

On the Approach to Intelligent Data Analysis in the Social Sciences

Maria A. Mikheyenkova
Federal Research Center "Computer Science and Control", Russian Academy of Sciences,
Russian State University for the Humanities,
Moscow, Russia
Email: mikheyenkova@yandex.ru

***Abstract*—The paper considers the intelligent data analysis in subject areas with open sets of empirical data, where formal apparatus is absent and the procedures for the theories' formation are heuristic. Approaches to solving these problems by means of the JSM Method of automated support for research are described. Implementing JSM Method intelligent systems are partner human-machine systems that automatically reproduce an imitation of the set of natural (rational) intelligence' abilities. In the interactive mode, it is possible to imitate such abilities as adaptation and correction of knowledge and the choice of reasoning strategy. JSM Socio intelligent system has been created to solve various problems of sociological data analysis. The generation of hypotheses on behavior prediction and its determinants, empirical regularities revealing in expanding empirical data give reason to consider the JSM Method as a tool for knowledge discovery in the social sciences.**

***Keywords*—knowledge discovery, JSM Method, formalized qualitative analysis, social data analysis**

I. Introduction

Artificial Intelligence researches can be divided into two parts – an epistemological part and a heuristic part [1]. The first one means formalization and automatization (in computer systems) the cognition process as such. The heuristic problem of how to create constructive tools for knowledge acquisition has a variety of approaches. The most important result of practical realization of theoretical principles and procedures in intelligent systems (AI-systems) is the possibility of new knowledge generation as a result of intelligent analysis of empirical data.

It is important to distinguish data analysis as pattern extraction from data (data mining) from intelligent data analysis (IDA) as new knowledge generation (knowledge discovery). The final product of the empirical data analysis should be a new knowledge, which is provided by the full process of Knowledge Discovery [2]. Data Mining represents one of the stages of this process – application of specific algorithms for extracting models (patterns) from data. The most important principle of intelligent data analysis is the tools' adequacy to the subject domain and the nature of the problem in hand [3] – in contrast to the dominant role of the tools in data analysis.

Intelligent data analysis (knowledge discovery) is an inherent part of empirical research in areas where there are no developed formal tools and, accordingly, the procedures for theories formation are heuristic in nature. Approaches that formalize the relevant research heuristics with implementation in computer systems are considered to be fruitful for the purpose of discovering new (with respect to the existing fact bases, BF, and knowledge bases, KB) knowledge. Such systems using artificial intelligence (AI) methods can be a tool for automated research support in sociology.

II. Problems of Epistemology and Heuristics of Social Sciences

The mass nature of many social phenomena and the obvious difficulties of taking into account the many factors influencing them led to the dominance of quantitative (statistical) methods of studying social reality. These methods are quite effective in the analysis of global and mass phenomena, but are of little use at the microsociological level, where mechanisms, motivation, and incentives of social behavior – individual and group – are considered. Insufficiency of Quantitative approach in AI design for social sciences is the main idea in contemporary research [4].

Studying the actions of human individuals (notably in their social relations with each other) is based on qualitative methods that transform subjective personal experience into typical models by informal means [5], [6]. Qualitative Comparative Analysis (QCA) [7], which is rather popular in sociological practice, can be considered as the certain approximation to the formalization of the typical qualitative methodology of case-study. The method is based on minimization of Boolean functions describing the dependence of the studied effect (presence or absence of phenomena, processes, structures) on presence or absence of some independent variables (possible causal conditions) and their combinations. Of course, formalism of this level seems obviously insufficient to study complex social phenomena [8].

The need for objectification of the results of qualitative analysis is to some extent satisfied due to the development of CAQDAS (computer aided/assisted qualitative data analysis systems) [9]. The use of these tools provides systematic nature, efficiency and reliability of standard procedures for processing qualitative data, thereby increasing the validity of conclusions.

However, this is far from solving the actual problem of formal imitation and computer realization [10] of the inductive strategy of qualitative analysis. In general, research heuristics of sociologists aimed at building theories based on revealing dependencies from an analysis of empirical facts can be represented by a universal cognitive cycle "data analysis – prediction – explanation". The formalization of this process – IDA – provides a transition from phenomenology to a knowledge system in the sciences with a poorly developed formal apparatus.

The irreducibility of human reasoning to axiom-based proof schemes has determined both the "cognitive turn" of modern logic – a departure from "antipsychologism" [11], and the emergence of ideas of "humanization" of AI systems among AI founders [12], [13]. At the current stage of AI development, the main goal of research in this area is constructive imitation (only to some extent) and the strengthening of human cognitive abilities in intelligent systems (IS) [14](pp. 256–277).

III. JSM Method in the Sociological Research

Examples of such systems are intelligent JSM-IS systems that implement the JSM Method of automated support for research [15]. The method reproduces research heuristics of mentioned type, using plausible reasoning with ampliative conclusions in the open world in the form of synthesis of non-elementary cognitive procedures: empirical induction (analysis), structural analogy (prediction), abductive acceptance of hypotheses. A formal language with a descriptive and argumentative functions [16] has been created for this purpose. Descriptive function provides initial data and knowledge structurization (with possibility of similarity determination) and formation of relation system. Argumentative function enables to formalize reasoning – analytic and prognostic procedures as well as procedures of explanation, falsification and possible verification of the results obtained [14](pp. 170–231). The induction is represented by formal elaborations and extensions of J.S. Mill's inductive methods [17], abductive acceptance of hypotheses is based on the initial data explanation [18]. This kind of abduction solves the problem of forming a criterion for sufficient reason for inductive hypotheses acceptance and has a fundamental importance for open areas with poorly developed (or completely absent) formal apparatus.

The realization of JSM-procedures and their combinations – strategies – in IS-JSM for the analysis of sociological data JSM Socio [19] is aimed at building a theory based on empirical facts which correlates with the methodological approach of qualitative analysis (using informal inductive inference). JSM Socio is considered to be a tool for formalized qualitative analysis of sociological data (FQASD). Expert-sociologist requirements reflected in the principles of formation of the information environment (fact base BF and the knowledge base KB) and the features of the user interface. Knowledge base includes both a priori (conventional) and obtained new knowledge as a result of the application of procedures. According to the microsociological paradigm, the social interaction of individuals is forced by internal motivation and possible external influences, which necessarily requires a multiparametric description [20]. This circumstance, coupled with the discrete nature of qualitative variables and the need to form a relational system that displays the semantics of the subject area, is taken into account by the descriptive function of the JSM-language intended for the FQASD.

The procedural semantics of the JSM Method can be formulated for various data structures that preserve the algebraic expressibility of similarity. The basic representation is the Boolean data structure. Accordingly, finite sets $\mathbf{U}^{(i)}$ and Boolean algebras defined on them $B_i = \langle 2^{\mathbf{U}^{(i)}}, \varnothing, \mathbf{U}^{(i)}, \text{—}, \cap, \cup \rangle, i = 1, 2, 3,$ are considered. Thus, subjects of behavior are characterized by a set of differential indicators $\mathbf{U}^{(1)} = \{d_1, \dots d_{r_1}$ that include elements of a social character (including value-normative attitudes), individual personality characteristics and biographical data. $\mathbf{U}^{(2)} = \{a_1, \dots a_{r_1}\}$ is a set of behavioral effects (actions and attitudes), $\mathbf{U}^{(3)} = \{s_1, \dots s_{r_1}\}$ is a set of situational parameters.

The individual variables $X, Z, V, \dots$ of the 1-st kind (perhaps with sub-indices) and constants $C, C_1, C_2, \dots$, being the values of the variables for objects and subobjects X, Z, V, etc., are introduced to represent persons (subject of behaviour) in the language, $X \in 2^{\mathbf{U}^{(1)}}$ The objects properties (for example, subjects' behavioural effects) are represented with the individual variables of the 2-nd kind $Y, U, W, \dots$ (perhaps with lower indices) and constants $Q, Q_1, Q_2, \dots$, $Y \in 2^{\mathbf{U}^{(2)}}$. The variables $S, S_1, \dots, S_n$ and the constants $\bar{S}, \bar{S}_1, \dots,$ $\bar{S}_n, S \in 2^{\mathbf{U}^{(3)}}$ of the 3-rd kind are introduced for the context (situational) parameters.

Social phenomena reflect the interaction of motivated, purposefully acting individuals taking into account important factors for them. Accordingly, the most important component of the JSM-language for FQASD is the representation of the opinion φ – the individual's personal perception of various aspects of social reality. Opinion is formed on the basis of the respondent's evaluation of the statements $p_1 \dots p_n$, characterizing the situation of interaction and argue the attitude towards it [21]. Statement $J_\nu p_i$ is the answer to the question "what is the value v of the statement p?" $(i = 1, \dots, n)$; $J_\nu p_i = t$ if $v[p_i] = \nu$; otherwise, $J_\nu p_i = f$; t and f are truth values of two-valued logic "true" and "false", respectively. In the general case of an m-valued poll (if there are m variants of sociologically interpreted estimates of statements $p_1 \dots p_n$) the evaluation function $v[p_i]$, $(i = 1, \dots, n)$ takes values $\nu \in \left\{0, \frac{1}{m-1}, \frac{m-2}{m-1}, 1\right\}$, $v[p_i] = \nu$. The j-th individual's opinion is the maximal conjunction $\varphi_j = J_{\nu_1^{(j)}} p_1 \& \dots \& J_{\nu_n^{(j)}} p_n$, where $\nu_i^{(j)}$ is corresponding evaluation of statements p_i $(i = 1, \dots, n)$, $\nu_i^{(j)} \in \left\{0, \frac{1}{m-1}, \frac{m-2}{m-1}, 1\right\}$, $j = 1, \dots, m^n$. Let's $[\varphi_j] = \left\{J_{\nu_1^{(j)}} p_1, \dots, J_{\nu_n^{(j)}} p_n\right\}$ be the set of corresponding conjunction's atoms.

Thus, the subject of social interaction is defined by the term $\bar{X}$ (complete object), $\bar{X} = \langle X, S, [\varphi] \rangle$. The complex multi-parameter structure of social systems and the various mechanisms of social interactions require an epistemologically adequate language for data representation (in particular, their parametrization), the choice of effective analysis procedures and strategies, and the conscious formation and enlargement of empirical facts set. In general case initial data are represented by (+)-facts $FB^+ = \left\{\langle \bar{X}, Y \rangle \mid J_{\langle 1,0 \rangle}(\bar{X} \Rightarrow_1 Y)\right\}$ ("object (person, for example) $\bar{X}$ possesses the set of properties (effect of behavior) Y"), (−)-facts $FB^- = \left\{\langle \bar{X}, Y \rangle \mid J_{\langle -1,0 \rangle}(\bar{X} \Rightarrow_1 Y)\right\}$ and facts that describe objects with previously undefined properties, $FB^\tau = \left\{\langle \bar{X}, Y \rangle \mid J_{\langle \tau,0 \rangle}(\bar{X} \Rightarrow_1 Y)\right\}$, $FB = FB^+ \cup FB^- \cup FB^\tau$. This allows us to vary the relational structure depending on the sociological model [19], [22]. Types of truth values in JSM Method are $1, -1, 0, \tau$ (factual truth, factual falsity, factual contradiction ("conflict"), uncertainty) correspond to the semantics of the four-valued logic of argumentation [14](pp. 312–338).

The JSM-research strategies are formed taking into account the empirical situation of the study. The key procedures for inductive generation of causal hypotheses are formalization of Mill's inductive methods, as well as their extensions and elaborations [15]. The minimal predicates representing the inductive similarity method are the predicates $M^+_{a,n}(V, W)$ and $M^-_{a,n}(V, W)$ for generating possible hypotheses on the causes of (+)- and (−)- facts, respectively (parameter n shows the number of applications of the rules of plausible inference to the FB, a – agreement – is the "name" of the Mill's inductive similarity method). The premises of the inductive inference rules $(I)^\sigma_n$, $\sigma \in +, -, 0, \tau$ include the corresponding Boolean combinations of the predicates $M^+_{a,n}(V, W)$ and $M^-_{a,n}(V, W)$. Thus, induction in the JSM Method includes an argumentation condition that ensures mutual falsifiability of conclusions and constructiveness of their truth values generating.

The similarity predicates can be strengthened by additional conditions, including those allowing formalizing other Mill's inductive Methods. Let I^+ be the set of $M^+_{a,n}(V, W)$ strengthening (indices), I^- be the set of $M^-_{a,n}(V, W)$ strengthening. Then the JSM strategies $Str_{x,y}$ will be the sets of rules

$(I)_n^\sigma$, $\sigma \in \{+,-,0,\tau\}$ such that they are formed by possible combinations of $M^+_{x,n}(V,W)$ and $M^-_{y,n}(V,W)$ predicates. The partial order relations based on the relation of logical deducibility are generated on the sets of predicates $M^+_{x,n}(V,W)$ and $M^-_{y,n}(V,W)$. The partially ordered sets of predicates $M^+_{x,n}(V,W)$ and $M^-_{y,n}(V,W)$, as well as the rules of plausible inference including them, form distributive lattices, and the direct products of these lattices form possible strategies $Str_{x,y}$ of JSM reasoning [23]. Thus, the strategies of the JSM Method have an algebraically definable structure, and the difference in the plausibility degrees of the hypotheses generated as a result of application of various strategies is given constructively. The use of various strategies characterizes the mechanism of causal forcing of the studied effects, which means the realization of the idea of syntax adequacy to the semantics of the subject area and the method's adaptability to the class of problems being solved.

A characteristic feature of empirical sociological research is the incompleteness (openness) of knowledge about the world, facts available to the researcher and describing their data. Developed logical means of the method provide the research possibility: empirical regularities (ER) (nomological statements) discovery. ER are inductive operationally definable (non-statistical) generalizations of the results of formalized JSM heuristics when enlarging (changing) data. ER are defined as regularities in sequences of embedded $FB(p)$, $p = 1,\ldots,s$ using various JSM strategies from the set $\overline{Str} = \{Str_{x,y} \,|\, x \in I^+, y \in I^-\}$ [15]. Semantically, this means recognition of the conservation of the cause – effect relation, i.e. the constancy of the truth values type in inductive hypotheses about $\pm$-causes and hypotheses-predictions obtained using causal hypotheses in the conclusion by analogy. Acceptance of the results of the JSM study on the basis of a non-singular assessment of the quality of reasoning and hypotheses allows correction of open (quasi-axiomatic) empirical theories. In combination with falsification tools built into the JSM procedures, this forms a enhancement of the K.R. Popper demarcation criterion [16], which separates the completed scientific research from the pre-research and provides sufficient reason for grounded decision-making.

IV. Application Experience

The most complete analysis of the social behavior of individuals is realized when considering the relational structure $\bar{X} \Rightarrow_1 Y$. The representation of the initial fact base by the predicates $\langle X, S, [\varphi]\rangle \Rightarrow_1 Y$ was used in the analysis of the constructive social activity, performed in collaboration with the Institute of Sociology, RAS. The chosen representation is related to the complex and multiple influence of the society characteristics on social activity. The focus of the study was the problem of society typology, based on the generation of determinants of political or civil forms of social activity.

A concept and model for the study of the determinants of social behavior (political/civic participation/non-participation) was formed, parameterization of the initial data with the inclusion of situational parameters – a set of socio-economic and functional characteristics of the respondent's area of residence (administrative status of the city, population income, cultural status, etc.), representing the territorial context of actions was proposed. The set of potential determinants included individual characteristics of the respondents' status; opinions, assessments that characterize the civil position. Different levels of determinations (situational, value, normative) were taken into account. Political activists – participants in political actions, (+)-examples in the JSM Method language – opposed (was considered as (–)-examples) civil activists (members of public organizations, do not participate in political activities), as well as helping individuals and nowhere involved passive citizens.

Visualization of the results of a computer experiment in the form of a "hypothesis tree" provides the sociologist with the opportunity to interpret the results and build a typology based on the revealed determinants. The basis of typologization – the "core" – is formed by the maximal intersections of respondents' descriptions. Similarities of subsets of respondents included in the maximum intersection allow to identify additional "peripheral" features. Peripheral features in different combinations form subtypes, which makes it possible to characterize the nuances of the position of subjects belonging to the same type of behavior, i.e. to suggest typology clarification.

As a result, the characteristic features of social types that implement various forms of social activity – political, civil, individual and passive – were described, and the features of interaction between these types were revealed. A non-trivial meaningful conclusion concerns the self-reference of "political activists" and "active citizens". Political activists consistently attribute the status of the subject of social action to themselves, denying this status to others. In other words, a feature of all political activists (who are representatives of systemic political parties in Russia regions) is the "leader" ideology, which is transmitted from political authorities to political activists. This seems to be due to the fact that the recognition of civic partners as social actors is uncomfortable for political activists, since it implicitly calls into question their social role. However, the rejection of partnerships with active citizens, attributing them the role of followers destroys the civil dialogue. On the contrary, active citizens ascribe to citizens the status of subjects of social action and declare their own readiness to participate in solving social problems and to unite.

A significant contribution of the regional context to the difference in forms and degrees of civic engagement was identified. This allows us to talk about the influence of the social system on the formation of individual behavioral strategy and the danger of transferring contextual features to the individual level.

The results aroused interest in the study of various forms of non-political (civic) activity. The material was a study of helping (prosocial) behavior, including semantic opposition "private helping behavior – volunteering". The first is situational, sporadic, i.e. is an act of social action. The second is a reflective, prolonged social activity, value collective behavior that creates the phenomena of practical solidarity. The work is also carried out in collaboration with researchers from the Institute of Sociology on the empirical data provided by them.

Based on the proposed conceptual model of the research object, a structure of empirical data (a set of variables and indicators) was formed on the basis of a sociological study in different organizations and different regions of the country. Respondents are described by socio-demographic and status characteristics, situational characteristics represent the development of the locality and other parameters of the regional context, opinions and assessments characterize the value aspects of relations between people, their attitude to volunteering and to the organization of this activity where respondents work.

The similarity of socio-demographic characteristics and the similarity of basic signs of social behavior of corporate, independent and combining both types of civil activity of volunteers, in particular, a high level of interpersonal trust, was found. The phenomenon of "open social borders" between "systemic" and "non-systemic" civil activists, discovered by means of the JSM Method, turned out to be interesting. A rather intense mobility between these groups was revealed, which indicates that there is no value confrontation among representatives of different socially active communities. This effect is difficult to detect by statistical methods. Further work is aimed at identifying the motivation of various forms of volunteer movement, determining complexes of value, socio-demographic and ideological characteristics that distinguish corporate volunteers from individuals, and those and others from those who refuse to help behavior

V. Conclusion

IDA (knowledge discovery) is performed by computer systems that implement an intellectual process represented by the interaction of the mental process and the cognitive process controlled by it [24]. The formal representation of the universal cognitive process "data analysis – prediction – explanation" provides imitation of natural (rational) intelligence abilities (reasoning, argumentation, learning, explanation of the results obtained) and allows reproduction in intelligent computer systems in automatic mode. However, a poorly formalized mental process, including attitudes, imperatives, goal-setting, the formation of open empirical theories and their adaptation in the context of data and knowledge correction, requires human participation and can be implemented in such systems only in an interactive mode.

It obviously follows that the IDA effective implementation is possible only by means of partner human-machine intelligent systems (see, for example, [25], p. 64). Even the successful implementation of the descriptive function of a formal language depends (to a large extent) on the meaningful interpretation of the cognizing subject (expert). Interactive pre-processing of open empirical data, control of the use of formalized heuristics, expert evaluation of generated empirical regularities ensure meaningfulness of the results obtained and determine the effectiveness in a specific study of the argumentative function of the language. The interpretability and explainability of the results generated by the IS tools play a fundamental role in the acceptance of IAD results, since the responsibility for final decisions is the human prerogative. This is confirmed by the attention to research in Explainable Artificial Intelligence [26], which arose in the context of spectacular successes of AI methods, the results of which are not interpreted.

Intelligent systems that implement the JSM Method are a technological means of exact epistemology and are partner human-machine systems. They effectively implement the generation of new knowledge, but at the same time do not replace, but support and strengthen the meaningful work of the researcher in various subject areas, including social sciences.

References

[1] McCarthy, J. Epistemological Problems of Artificial Intelligence //*Fifth International Joint Conference on Artificial Intelligence IJCAI-77 (August 22 – 25, 1977, Cambridge, Massachusetts, USA)*, Proceedings, vol. 2, pp. 1038 – 1044.

[2] Fayyad U.M., Piatetsky-Shapiro G., and Smyth P. From data mining to knowledge discovery in databases // *AI Magazine*, 1996, vol. 17, No 3, pp. 37–54.

[3] Sorokin P. *Fads and Foibles in Modern Sociology and Related Sciences.* Westport, Connecticut: Greenwood Press Publishers, 1956. 357 p. Pp. 102–130.

[4] Sloane M., Moss E. AI's social sciences deficit // *Nat. Mach. Intell.*, 2019, Vol. 1, August, pp. 330–331.

[5] Hammersley M. *What is qualitative research?* London: Bloomsbury Academic, 2013. 136 p.

[6] Glaser B., and Strauss A. *Basics of qualitative research. Techniques and procedures for developing grounded theory*, 4th ed. Thousand Oaks, California: SAGE Publications, 2015. 456 p.

[7] Marx A., Rihoux B., Ragin C. C. The Origins, Development, and Application of Qualitative Comparative Analysis: The First 25 Years // *Eur. Political Sci. Rev.*, 2014, vol. 6, No 1, pp. 115–142.

[8] Mikheyenkova M.A. On the logical tools of intelligent sociological data analysis // *Sci. Tech. Inf. Process*, 2010, vol. 37, Issue 6, pp. 386–397.

[9] Silver C., and Lewins A. *Using Software in Qualitative Research: A Step by Step Guide*, 2nd ed. London: SAGE, 2014. 304 p.

[10] Fielding N. G. Qualitative research and our digital futures // *Qualitative Inquiry*, 2014, vol. 20, No 9, pp. 1064–1073.

[11] Benthem J. van. Logic and Reasoning: Do the Facts Matter? // *Studia Logica*, 2008, vol. 88, No 1, pp. 67–84.

[12] McCarthy J. From here to human-level AI // *Artif. Intell.*, 2007, vol.171, No 18, pp. 1174–1182.

[13] McCarthy J. The well-designed child // *Artif. Intell.*, 2008, vol. 172, No 18, pp. 2003–2014.

[14] Finn V.K. *Iskusstvennyj Intellekt: metodologija, primenenija, filosofija [Arificial Intelligence: Methodology, Applications, Philosophia].* Moscow: URSS, 2018. 448 p.

[15] Finn V.K. On the Heuristics of JSM Research (Adds to Articles) // *Autom. Doc. Math. Ling.*, 2019, vol. 53, No. 5, pp. 250–282.

[16] Popper K. Evolutionary Epistemology // In: *Evolutionary Theory: Paths into the Future / Ed. by J. W. Pollard*. Chichester and New York: John Wiley &. Sons., 1984, 294 p. Pp. 239–255.

[17] Mill J.S. *A System of Logic Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. London, England: Parker, Son and Bowin; 1843. Available at (accessed 2019, Dec): http://www.gutenberg.org/files/27942/27942-h/27942-h.html

[18] Josephson J.R. Smart inductive generalizations are abductions // In: *Abduction and Induction. Essays on their Relation and Integration / Flach P.A., Kakas A. (Eds.)*. Dordrecht, Boston, London: Kluwer Academic Publishers, 2010. 336 p. Pp. 31–44.

[19] Mikheyenkova M.A., Klimova S.G. Knowledge Discovery in Social Research // *Autom. Doc. Math. Ling.*, 2018, vol. 52, No. 6, pp. 318–329.

[20] Luhmann N. *Social Systems*. Stanford: SUP, 1995. 627 pp.

[21] Finn V.K., Mikheyenkova M.A. Plausible Reasoning for the Problems of Cognitive Sociology // *Log. Log. Philos.*, 2011, vol. 20, pp. 111–137.

[22] Gusakova, S.M., Mikheyenkova, M.A. Knowledge Discovery as a Tool for the Formation of the Structure of Society // *Autom. Doc. Math. Ling.*, 2016, vol. 50, No. 5, pp. 179–188.

[23] Finn V.K. On the Class of JSM Reasoning That Uses the Isomorphism of Inductive Inference Rules // *Sci. Tech. Inf. Process*, 2017, vol.44, No. 6, pp. 387–396.

[24] Mikheyenkova M.A., Finn V.K. Myslitel'nye I poznavatel'nye procedury v kognitivnoj sociologii [Mental and Cognitive Procedures in Cognitive Sociology] // *VIII mezdunarodnaja konferencija po kognitivnoj nauke: tezisy dokladov [VIIIth International Conference on Cognitive Science: Abstracts*] – Moscow: Izd-vo "Institut Psihologii RAN", 2018. 1368 p. Pp. 1133–1136.

[25] The AI Index 2018 Annual Report // *AI Index Steering Committee, Human-Centered AI Initiative*. Stanford University, Stanford, CA, December 2018. 94 p.

[26] IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI), Proceedings. Available at (accessed 2019, Dec): https://drive.google.com/file/d/1ma5wilaj31A0d5KC4I2fYaTC_Lqm_d9X/view

Об одном подходе к интеллектуальному анализу данных в социальных системах

Михеенкова М.А.

Рассматриваются проблемы интеллектуального анализа данных в областях с открытыми множествами эмпирических данных, где отсутствует формальный аппарат и формирование теорий носит эвристический характер. Описаны подходы к решению этих проблем средствами ДСМ-метода автоматизированной поддержки исследований. Реализующие ДСМ-метод интеллектуальные системы являются человеко-машинными системами, воспроизводящими в автоматическом режиме имитацию ряда способностей естественного (рационального) интеллекта. В интерактивном режиме возможна имитация таких способностей, как адаптация и коррекция знаний и выбор стратегии рассуждений. Для решения задач анализа социологических данных создана интеллектуальная система JSM Socio. Порождение гипотез о детерминантах поведения и его прогнозе, обнаружение эмпирических закономерностей в расширяющихся базах эмпирических данных дают основание считать ДСМ-метод инструментом интеллектуального анализа данных в науках о человеке и обществе.

Received 15.12.2019

Neural Networks in Semantic Analysis

1st Alexey Averkin
Dorodnicyn Computing Centre
FRC CSC RAS
Moscow, Russia
averkin2003@inbox.ru

2nd Sergey Yarushev
Department of Informatics
Plekhanov Russian University of Economics
Moscow, Russia
Yarushev.SA@rea.ru

***Abstract*—This paper presents research of the possibilities of application deep neural networks in semantic analysis. This paper presents the current situation in this area and the prospects for application an artificial intelligence in se-mantic analysis and trend and tendencies of this science area. For better un-derstanding future tendencies of researches in semantical area we present detailed review of the studies in semantic analysis with using artificial intelligence, studies about a human brain.**

***Keywords*—Semantic Analysis, Deep Neural Networks, Forecasting, Processing of Nat-ural Language.**

I. Introduction

Machine learning is a central technology of artificial intelligence. In recent years, in this area all the attention was focused on the deep learning technology as a method for automatically extracting characteristic values. Huge amounts of time series data collected from the devices, especially in the era of Internet of Things. Applying deep learning that data and classifying them with a high degree of accuracy, it is possible to carry out further analysis with a view to creating new products and solutions and open up new areas of business. Deep learning [1] technology “Fig. 1”, which is seen as a breakthrough in the development of artificial intelligence, delivers the highest accuracy of image recognition and speech, but it is still applicable only to limited types of data. In particular, it has so far been difficult to classify accurately automatically variable time series data coming from the devices connected to the Internet of things. The theory and practice of machine learning are experiencing this "deep revolution" caused by the successful application of methods Deep Learning (deep learning), representing the third generation of neural networks. In contrast to the classical (second generation) neural networks 80-90-ies of the last century, a new learning paradigm allowed to get rid of some of the problems that hindered the dissemination and successful application of traditional neural networks. Networks trained by deep learning algorithms, not just beat the best in terms of accuracy, alternative approaches, but also in a number of tasks showed the beginnings of understanding the meaning of information supplied (for example, image recognition, analysis of textual information, and so on). Most successful modern industrial methods of computer vision and speech recognition built on the use of deep networks, and IT-industry giants such as Apple, Google, Facebook, buying groups of researchers involved in deep neural networks. Deep Learning - is part of a broader family of machine learning methods - learning concepts, which feature vectors are located directly on a variety of levels. These features are automatically determined and associated with each other, forming output data. At each level, abstract presented features based on attributes of the previous level. Thus, the deeper we move forward, the higher level of abstraction. The neural network is a plurality of layers of a plurality of levels with feature vectors that generate output data.

The reported study was funded by RFBR, project number 20-010-00828.

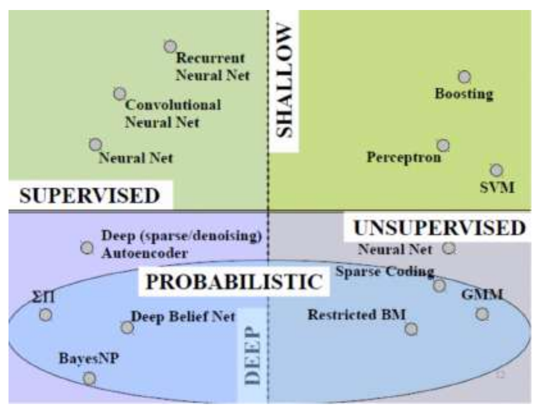


Figure 1. Classification of deep architectures.

II. Histrory of Deep Learning

The team of graduate students from the University of Toronto (Canada), led by Professor Jeffrey Hinton [2] (Geoffrey E. Hinton), won the competition held by the pharmaceutical company Merck. With access to a limited set of data describing the chemical structure of the molecules 15, a group of Hinton was able to create and use a special software system, which is to determine which of these molecules will be more effective than other work as drugs. The peculiarity of this operation lies in the fact that system designers have used artificial neural network based on the so-called "deep learning"

(Deep Learning). As a result, their system was able to conduct the necessary studies and calculations on the basis of very small set of input data: usually for training neural networks before using the required load in a vast array of information. Achieving Hinton team looks particularly impressive when you consider that the application form has been submitted at the last moment. Moreover, the system of "deep learning" was created in the absence of concrete data on the interaction of molecules with the proposed target. The successful application of the technique of "deep learning" was another achievement in the development of artificial intelligence, which is very rich in the second half of 2012. So, Jeff Dean this summer (Jeff Dean) and Andrew Ng (Andrew Y. Ng) from Google revealed a new system for image recognition to determine the accuracy of the cat in the picture 15.8%, where the cluster system for training of 16 thousand units to use the network ImageNet, contains 14 million images of 20,000 different objects. Last year, the Swiss scientists have demonstrated a system that is better recognized a person signs in the photos (the accuracy was 99.46% on a set of 50,000 images, with people maximum accuracy was 99.22%, and the average accuracy for a group of 32 people was "only" 98.84%). Finally, in October of this year, Richard Rashid (Richard F. Rashid) [3], coordinator of the Microsoft academic programs shown at a conference in Tianjin (China) technology live translation from English into Mandarin Chinese, maintaining the original voice. All these technologies, demonstrating a breakthrough in the field of artificial intelligence, in one way or another based on the technique of "deep learning". The main contribution to the theory of in-depth training is now making just Professor Hinton, who, incidentally, is the great-grandson of George Boole, an English scientist, the inventor of Boolean algebra underlying modern computers. The theory of deep learning complements conventional machine learning technology with special algorithms for the analysis of the input data at several levels of presentation. The peculiarity of the new approach is that the "deep learning" is studying the subject, until it finds enough informative presentation levels to take into account all the factors that could affect the characteristics of the studied subject. Technology Artificial Intelligence (AI) in general, and in-depth training in particular, are now widely used in different systems, including voice assistant Apple Siri-based Nuance Communications technologies and recognition of addresses on Google Street View. However, scientists carefully evaluate progress in this area since the history of AI is replete with big promises and no less loud crunches. Thus, in the 1960s, scientists have believed that before the establishment of full-fledged AI there are only about 10 years. Then, in the 1980s, there was a whole wave of young companies that offer a "ready-AI", then in this area has come true "Ice Age" - right up to the present time, when the enormous computing power available in the cloud services, opened a new level implementation of powerful neural networks using new theoretical and algorithmic basis. New learning paradigm realizes the idea of training in two phases. In the first stage of a large array of un-partitioned data using avtoassotsiatorov (stratified by their learning without a teacher) extracted information about the internal structure of the input data. Then, using this information in a multi-layer neural network, trained with her teacher (tagged data) by known methods. The number of untagged data desirable to be as large as possible. Markup data may be much less. In our case, this is not very important. Deep learning the most active and the most powerful tool in artificial intelligence area now. Based on the well-reviewed seen on many popular technologies of deep learning. Major corporations sent huge amounts of money in research in this area. The most popular field of application of deep learning today are the systems of computer vision, facial recognition, text, search engines and ontological system. To date, the use is gaining popularity depth training in forecasting time series. Based on this, in our further work we will carry out experiments in this direction. Today, deep learning technology based on the principles of service, which makes their use more convenient and faster in most areas. Processing information in natural language analysis of the relationship between the collection of documents and terms presented in this document, understanding and determination of the direction and theme of the text - all tasks of semantic analysis. Latent semantic analysis are the search giant to find the text of one subject. A lot of work is carried out in the field of construction of semantic models for processing quality of the text, understanding the logical relationships, optimization of knowledge bases, as well as the widest range of applications. How do children learn language? In particular, as they associate it with a sentence structure meaning? This issue is certainly associated with issue that is more global - how the brain connects the sequence of characters for building symbolic and sub-symbolic representations? In this paper, we consider a number in the field of semantic analysis using artificial neural networks applications.

III. Review of text classification approaches

Chinese scientists [4] in their study presented a new model of text classification using a neural network and its learning by back propagation, as well as with the modified method. Use an effective method to reduce the characteristics of selecting the sample dimension, that of a bear increasing system performance. Standard learning algorithm backpropagation trained quite slow, so the researchers modified the algorithm to increase the speed of learning. Traditional word-matching based on the classification of text using vector space model. However, in this approach takes no account semantic relationships between terms, which can lead to deterioration of the classification accuracy. Latent semantic analysis can overcome the problems caused by the use of statistically derived conceptual indices and not just individual words. It creates kontsetualnye vector spaces, in which each term or a document is submitted as a vector in space. This not only significantly reduces the dimension, but also can detect important associative relations between terms. The researchers tested their model on a set of 20 news data, experimental results showed that the model with the modified method of backpropagation proposed in this paper, have surpassed the traditional teaching method. As well as the use of latent semantic analysis for this system allows you to dramatically reduce the dimension that can achieve good results classification. Indian scientists compared the performance of common neuronal back propagation network with a combination of the neural network with the method of latent semantic indexing of text classification problem. In traditional neural networks with backpropagation error process of adjusting the scales locked in a local minimum, as well as the speed of learning network of this type is rather low, which entails a decrease in performance. In view of these shortcomings, scientists have decided to make a combination of latent semantic indexing and this neural network. Latent semantic representation of the structure of data in the low-dimensional space in which the documents, the terms and words sequences were also compared. The one-dimensional decomposition technique used in latent semantic analysis, multi-dimensional matrix in which the terms are broken down into a set of K orthogonal factors, which the original text data is changed to a smaller semantic space. The new vector documents can be found in the K-dimensional space. Also, the new coordinates are queries. Performance tested a combination of these methods on the basis of the classification methodology 20 newsgroups from different categories, such as sports, medicine, business, politics, and others. As a result, these methods can significantly reduce the dimension and get the best results of the classification of the text. One of the key figures in the study of human brain is Mauntkasl [5]. In this paper, he summarized his many years of research, he argues that, despite the diversity of their functions, all sections of the cerebral cortex are arranged, in principle, the same. This means that learning and pattern recognition takes place uniformly in the cortex, and the variety of its functions is a consequence of the diversity of signals processed by different parts of the cortex. According Mauntkaslu bark has a two-dimensional honeycomb structure. The basic functional unit of the cortex is a mini-column of about 30 microns in diameter, consisting of about 100 neurons. Such mini-columns interconnected by positive and negative lateral connections. Moreover, the latter shall be sharp, but with a certain delay relative to the first. As a result, at the same time excited by a pool adjacent mini-column, involuntarily forcing the recall T. Kohonen self-organizing map [6]. As a result, throughout the cortex, we are seeing signs of self-organizing maps: detectors similar signals are located next to each other. Experiments show that the elementary detectors in the area of these maps the order of 0.1 mm 2, ie, they contain a mini-columns 102 or 104 neurons. Such functional units Mauntkasl calls the macro columns. It is they who determine the "resolution" of the cortex, and limit the number of signs that can remember people (only a few million). But the reliability of the memory is guaranteed by a large number of neurons that make up the macro column. So we keep his memory throughout life, even with the death of a substantial part of the neurons. Thus, Kohonen maps are evidently the most appropriate tool for simulating the operation of the cortex. You just need to teach them how to work with dynamic patterns, with which only the brain works, because its main task - foresight. As people acquire language, as well as two or more different languages with a nervous system remains an open question. To solve this problem, the French scientists, led by Peter blast furnace [7] proposed to build a model that will be able to learn any language from the very beginning. In this work, they offer the neural network approach, which handles proposals for like, word for word, without prior knowledge of the semantics of words. The proposed model does not "pre-bound" structure, but only random and trained compound model is based on Reservoir Computing technology. Scientists for robotic platforms through which users can teach the robot the basics of the English language, to later give him a different job, developed earlier model. In this paper, was added ability to handle rare words in order to be able to keep the dictionary size is quite small in the processing of natural language. Moreover, this approach was extended to the French language and shows that the neural network can learn two languages at the same time. Even with a small body language model is capable to learn and generalize knowledge in a monolingual or bilingual. This approach may be a more practical alternative for small shells of different languages than other training methods based on large data sets. Many studies performed in the language processing by neural networks [8], and in recent years with the use of so-called Echo State Networks [9]. How the human brain processes the proposals that a person reads or hears? The task of understanding how the brain does it, occupies a central place in the research of scientists from the field. proposals processed in real time. Previous words in a sentence may affect the processing time in hundreds of milliseconds. Recent neurophysiological studies suggest that it is the frontal part of the brain plays a crucial role in this process. Hinaut [10] conducted a study that gives some insight into how certain aspects of the treatment of the proposals in real time occur, based on the dynamics of periodic cortical networks and plasticity in cortico-striatal system. They model the prefrontal area BA47 using recurrent neural network to obtain on-line input word categories in the processing of proposals. The system is trained on pairs of sentences in which the meaning is encoded as a function of the activation of the corresponding role of verbs and nouns in sentences. The model examines the expanded set of grammatical structures and demonstrates the ability to generate new designs. This shows how much early in the sentence a parallel set of predicates makes sense. The model shows how the online responses to the speech are influenced by the preceding words in a sentence and the previous proposals in the discourse, providing a new perspective on the neurophysiology of the brain cortex to recognize grammatical structure. The study found that recurrent neural network can decode the grammatical structure of the sentences in real-time in order to get an idea of the meaning of sentences. Neural processing of natural language. The focus in cognitive science today is focused on the study of how the neural networks in the brain are used to read and understand the text. This issue explores a huge number of scientists in neuroscience along with recent studies that are designed to examine the brain processes involved in learning non-linguistic sequences or artificial grammar learning. Peter Ford Dominey [11] in their study attempted to combine data from several neurophysiological proposals processing models, through the specification of neural network model, the architecture of which is based on the well-known cortico-striatal-thalamo-cortical (KSTK) neuroanatomy system of human language. The challenge is to develop simulation models that take into account the limitations and neuroanatomical connections, and functional image data. In the proposed model, the structural cues encoded in a recurrent neural network in Kortikova BA47, activate circuit (KSTK) for modulating the flow lekskicheskoy semantic information in an integrated view on the meaning of the sentence level. The simulation results are shown in Caplan [12]. Modeling Language Authority spent S.A. Shumsky. In the work [13], the author puts forward three hypotheses: The first hypothesis is that the processing of the time series in the crust is done like modules, recognizing the typical temporal patterns, each with its own input. For example, the bark of the site responsible for the morphological analysis of words, recognize about 105 words and their constituent morphemes and syllables. Another cortical portion defining sentence structure, operates in the same way, but with a different primary alphabet, each character is no longer encodes a letter and a word unit. This area stores the characteristic patterns of the combination of words in a grammatically correct sentence. According to the second hypothesis, the input for the next cortical module, responsible for the analysis of the temporal structures of higher order thalamus serves as a compressed output signal from the previous module. According to the third hypothesis, in the "body language" there are two interchannel "deep learning": the grammatical and semantic. Similarly, the dorsal (scene analysis) and ventral (object recognition) channels the analysis of visual information.

References

[1] Ranzato, M. A., & Szummer, M.: Semi-supervised learning of compact document representations with deep networks. In: Proceedings of the 25th international conference on Machine learning. pp. 792-799, ACM, 2008.

[2] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye: "A fast learning algorithm for deep belief nets." Neural computation 18.7: pp. 1527-1554, 2012.

[3] Deep Learning: Microsoft's Richard Rashid demos deep learning for speech recognition in China, http://deeplearning.net/2012/12/13/microsofts-richard-rashid-demos-deep-learning-for-speech-recognition-in-china/, last accessed 2020/01/17

[4] Yu B., Xu Z., Li C.: Latent semantic analysis for text categorization using neural network. //Knowledge-Based Systems, . 21, №. 8, pp. 900-904, 2008.

[5] Mountcastle V.: The columnar organization of neocortex // Brain, vol. 120, P. 701–722, 1997.

[6] Kohonen T.: Self-Organizing Maps. Springer Verlag, 2001.

[7] Hinaut X. et al.: A Recurrent Neural Network for Multiple Language Acquisition: Starting with English and French //CoCo@ NIPS, 2015.

[8] Miikkulainen, R.: Subsymbolic case-role analysis of sentences with embedded clauses. Cognitive Sci 20: pp. 47–73, 1996.

[9] Hinaut X., Dominey P. F.: Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing //PloS one, .8, №. 2, p. 52946, 2013.

[10] Frank, S. L.: (2006). Strong systematicity in sentence processing by an Echo State Network. In Proc. of ICANN, pp. 505–514, 2006.

[11] Peter Ford Dominey, Toshio Inui, Michel Hoen: Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing // Brain and Language. doi:10.1016/j.bandl.2008.08.002, 2008.

[12] Caplan, D., Baker, C., Dehaut, F.: Syntactic determinants of sentence comprehension in aphasia. Cognition, 21, 117–175, 1985.

[13] Shumski, S.: Brain and language: Hypotheses about structure of a natural language, 2017.

Нейронные сети в семантическом анализе

Аверкин А.Н. Ярушев С.А.

В статье описываются глубинные архитектуры искусственных нейронных сетей и возможности их применения в семантическом анализе. Рассматривается история разработки глубинных нейронных сетей. Исследуются современные тенденции в задачах семантического анализа, а также представляется кратких обзор исследований в данной области.

Received 21.12.19

Hybridization Method of Intelligent Models

Vadim V. Borisov
Department of Computer Engineering,
The Branch of National Research University "Moscow Power Engineering Institute" in Smolensk,
Energeticheskiy proezd, 1, Smolensk 214013, Russia
vbor67@mail.ru

***Abstract*—The development trends of intelligent models are generalized. The approach to systematization and classification of hybrid models is described. Examples of classification of known hybrid intelligent models related to each considered approach to their hybridization are given. The proposed method of hybridization of intelligent models is considered.**

***Keywords*—hybridization of intelligent models, fuzzy model, neural network model, evolutionary modelling**

I. Introduction

At present the researches in the field of creation of intelligent models are actively conducted. The main tendencies of their development are the following:

- these intelligent models are actively used for the solution of poorly structured and unstructured problems;
- these intelligent models based on representation and processing of heterogeneous data and knowledge;
- principles of hybridization and adaptability are used more and more widely;
- problems can be solved in conditions of various types of uncertainty (stochastic, statistical, not stochastic, linguistic).

At present intensive researches are carried out in the field of the following intelligent modelling [1]: fuzzy models based on theories of fuzzy computing, fuzzy sets and relations, fuzzy measure, fuzzy logic and inference; neural networks models based on artificial neural network theory; methods and technologies of evolutional modelling and genetic algorithms.

For each intelligent model is possible to define the development trends.

The development trends of fuzzy models are the following:

- development of fuzzy mathematics (of fuzzy computing, fuzzy sets and relations, fuzzy measure, fuzzy logic etc.);
- designing of fuzziness into traditional mathematical models;
- designing of fuzzy models which correspond to features of tasks;
- hybridization of fuzzy models of various types;
- use of fuzzy models, methods and technologies for problem solving in various subject areas.

The development trends of neural networks models are the following:

- designing of neuron-like elements of artificial neural networks;
- designing of neural networks which correspond to the features of tasks;
- designing of algorithms of artificial neural networks training;
- combination of neural networks models of various types;
- use of neural networks models, methods and technologies for solving problems in various subject areas.

The development trends of evolutionary modelling methods and technologies are the following:

- designing of genetic operators, criteria and search strategy;
- designing of genetic algorithms providing the greater accuracy and rapid convergence;
- developing of approaches to adaptation of genetic algorithms for problem solving in various subject areas.

At present the intensive researches are conducted in the field of creation of hybrid intelligent models, methods and technologies. Hybridization of these intelligent models allows them to complement each other.

Hybridization compensates the restrictions of each of these models and it also provides tolerance to uncertainty and partial validity of the data used to achieve flexibility of decisions [2].

However, up to now, there is no unified approach to hybrid models systematization. The purpose of this paper is the following: to systematize of existing hybrid models; to develop a mechanism for hybridization of intelligent models.

II. HYBRIDIZATION METHOD OF INTELLIGENT MODELS

Hybridization can be realized in different ways: either between models related to different intelligent technologies, or within the same intellectual technology [3]. In both cases hybridization of intelligent models is possible on the basis of one of following approaches:

- hybridization "with functional replacement" – one of the model is picked out as a dominating model, and its separate components are replaced with components of other model;
- hybridization "with interaction" – several models are used relatively independently, and thus they solve various particular tasks in achieving the final goal, exchanging data.

Let's briefly consider these approaches of hybridization, firstly, by the example of various intelligent technologies and, secondly, within the limits of a separate technology. Let's demonstrate realization of the hybridization approach "with functional replacement" on the example of fuzzy models' and neural networks. We will consider fuzzy rule-based models as the dominant, and we will design each separate component of these models on the basis of neural networks. Possible ways of such hybridization allow to distinguish a separate class of fuzzy rule-based neural models (the first sign mentioned – fuzzy – is dominant) [3-5].

In case the neural networks are used as dominating models, then realization of separate components of these models can be executed on the basis of fuzzy models' components. Possible ways of such hybridization allow distinguish a separate class of neural fuzzy models (the first mentioned sign – neural – is dominating) [6].

Hybridization "with functional replacement" in the context of separate technology is most peculiar to neural networks technology (i.e. hybridization neural networks models only).

Let's demonstrate realization of the hybridization approach "with interaction" by the example of fuzzy models and neural networks technologies (to be more precise, fuzzy rule-based models and neural networks as above). According to this method the following models can be offered:

- fuzzy rule-based models with use of neural networks for space partition of input variables and formation multidimensional membership functions of antecedents for them;
- joint using of fuzzy rule-based models and neural networks for computing and "interchange" by parameter values [7].

The approach of hybridization of intelligent models "with interaction" is the most important at performance of complex problems. Complex problems are the problems that require the use of various models for performance of separate stages or particular tasks to achieve general purpose.

Let's consider the hybridization method "with interaction" within the limits of one technology by the example of fuzzy models at realization of complex problems [3]. This method consists of the following stages.

Stage 1. Decomposition of a complex problem into set of particular tasks.

$$Problem = \{Task_1, Task_2, \dots, Task_{\mathrm{n}}\}.$$

Here *Problem* – complex problem; $Task_i$ – particular task, $i = 1, \dots, n$.

As an example, let's consider the decomposition of a complex problem – *situational analysis* – on following particular tasks:

- $Task_1$ – identification of a situation;
- $Task_2$ – estimation of alternative decisions;
- $Task_3$ – prognosis of the resulting situation.

Stage 2. Specifying the requirements to fuzzy models for each of particular tasks.

The *following classification of requirements to intelligent models for particular tasks is offered:*

- *R_1 – subject of knowledge: {$R_{1,1}$ – about object of management; $R_{1,2}$ – about a management method; $R_{1,3}$ – about interrelation of events; $R_{1,4}$ – about the decision-maker "relation" to object of management; $R_{1,5}$ – about the decision-maker "relation" to a management method};*
- *R_2 – type of time analysis: {$R_{2,1}$ – statics; $R_{2,2}$ – short-term dynamics; $R_{2,3}$ – long-term/complete dynamics};*
- *R_3 – uncertainty type: {$R_{3,1}$ – the stochastic; $R_{3,2}$ – not stochastic; $R_{3,3}$ – stochastic and not stochastic};*
- *R_4 – method of knowledge acquisition: {$R_{4,1}$ – primary; $R_{4,2}$ – secondary}.*

For each particular task $Task_i$(i=1, …, n) one or several possible groups of these requirements can be determined:

$$Task_i, i = 1, \dots, n : \{G_{i,1}, G_{i,2}, \dots, G_{i,mi}\}$$

Here m_i – number of possible groups of requirements for a task $Task_i$.

The example of requirements to fuzzy models for particular tasks ($Task_1$, $Task_2$ and $Task_3$) is presented in Table 1.

Stage 3. Classification of intelligent models and concretization of their possibilities to meet the requirements for particular tasks.

The following classification of fuzzy models depending on their purpose is offered:

- universal fuzzy models (rule-based, relational, functional);
- task-oriented fuzzy models:
  - functional and relational estimation models;
  - models of events (linguistic lotteries, event trees, failure trees, Bayesian networks, game models);
  - state models and management models ("state-action" models, situational networks with the direct/inverse method of model design, cognitive maps, Markovian and semi-Markovian

Table I
REQUIREMENTS TO FUZZY MODELS FOR PARTICULAR TASKS

Particular tasks	**Group of requirements**	**Requirements**			
		R_1	R_2	R_3	R_4
$Task_1$	$G_{1,1}$	$R_{1,2}$	$R_{2,1}$	$R_{3,3}$	$R_{4,1}$
	$G_{1,2}$	$R_{1,2}$	$R_{2,2}$	$R_{3,3}$	$R_{4,1}$
$Task_2$	$G_{2,1}$	$R_{1,4}$	$R_{2,1}$ OR $R_{2,2}$ OR $R_{2,3}$	$R_{3,2}$	$R_{4,1}$
	$G_{2,2}$	$R_{1,2}$	$R_{2,2}$	$R_{3,3}$	$R_{4,1}$
	$G_{2,3}$	$R_{1,2}$	$R_{2,3}$	$R_{3,3}$	$R_{4,2}$
$Task_3$	$G_{3,1}$	$R_{1,1}$	$R_{2,2}$ OR $R_{2,3}$	$R_{3,3}$	$R_{4,1}$
	$G_{3,2}$	$R_{1,2}$	$R_{2,2}$ OR $R_{2,3}$	$R_{3,3}$	$R_{4,1}$

models, Petri nets, automatons, decision trees) [3].

Stage 4. Specification of set of the intelligent models that meets requirements for particular tasks.

"The covering tree" of complex problem is formed by a set of corresponding fuzzy models on the basis of Stages 2 and 3. The example of "the covering tree" for a considered problem is presented in Fig. 1.

Stage 5. Selection of a subset of the intelligent models providing minimal "covering" of a complex problem.

Minimal "covering" of a complex problem is such a subset of fuzzy models where even removal of one of the models leads to impossibility of the problem solution.

The following subset of fuzzy models for a considered example is selected:

- fuzzy situational network with the direct method of model design (basic model);
- fuzzy model "state-action" (supplementary model);
- linguistic lotteries (supplementary model).

III. EXAMPLE OF THE COMPOSITIONAL FUZZY MODEL FOR THE EFFICIENCY ESTIMATION OF ENERGY- AND RESOURCE SAVING

Figure 2 shows the structure of the compositional fuzzy model for the efficiency estimation of energy- and resource saving, based on the proposed method [8].

The compositional fuzzy model consists of the sets of the following models:

- fuzzy cognitive models for estimating the effects of actions on the subsystems' indicators (power supply, heat and water supply);
- fuzzy rule-based models for efficiency estimating of energy- and resource saving of subsystems: power supply, heat and water supply;
- generalized fuzzy rule-based model for efficiency estimating of energy- and resource saving of system as a whole.

The fuzzy cognitive models are affected by events from groups of measures $A_{s_j} = \{a_{k_j}^{(s_j)}\}, k_j = 1, \ldots, K_j$, corresponding to subsystems $\forall s_j \in S$. As a result, the values of the output variables $p_1(s_j)$ and $p_2(s_j)$ of these models, which (together with indicators $p_3(s_j)$ and $p_4(s_j)$, do not depend on the impact of the events) are fed to the inputs of fuzzy rule-based models for efficiency estimating $e(s_j), j = 1, \ldots, J$ of energy- and resource saving of subsystems. Then, these estimates are aggregated into a generalized indicator $E(S)$ of energy and resource efficiency of the entire system S using generalized fuzzy rule-based model.

IV. CONCLUSIONS

The development trends of intelligent models are generalized. The approach to systematization and classification of hybrid models is offered. Examples of classification of known hybrid intelligent models related to each considered approach to their hybridization are given.

The proposed method of hybridization of intelligent models is considered. This method consists of the following stages:

(i) decomposition of a complex problem into set of particular tasks;
(ii) specifying the requirements to models for each of particular tasks;
(iii) classification of intelligent models and concretization of their possibilities to meet the requirements for particular tasks;
(iv) specification of set of the intelligent models that meets requirements for particular tasks;
(v) selection of a subset of the intelligent models providing minimal "covering" of a complex problem.

ACKNOWLEDGMENT

Research is supported by the grant of the Russian Foundation for Basic Research, project no. 18-07-00928_a.

REFERENCES

[1] S. Rajasekaran, G.A. V. Pai, "Neural Network, Fuzzy Logic, and Genetic Algorithms – Synthesis and Applications," Prentice Hall, 2005.

[2] V. K. Lamba, "Neuro Fuzzy Systems," Laxmi Publications PVT Limited, 2008.

[3] V. V. Borisov, "Hybridization of Intellectual Technologies for Analytical Tasks of Decision-Making Support," Journal of Computer Engineering and Informatics. 2014, Vol. 2, Iss. 1. PP. 148–156.

[4] C. H. Chen, "Efficient Self-Adaptive Learning Algorithm for TSK-Type Compensatory Neural Fuzzy Networks," International Journal of Fuzzy Systems, vol. 14, no. 4, pp. 510-518, 2012.

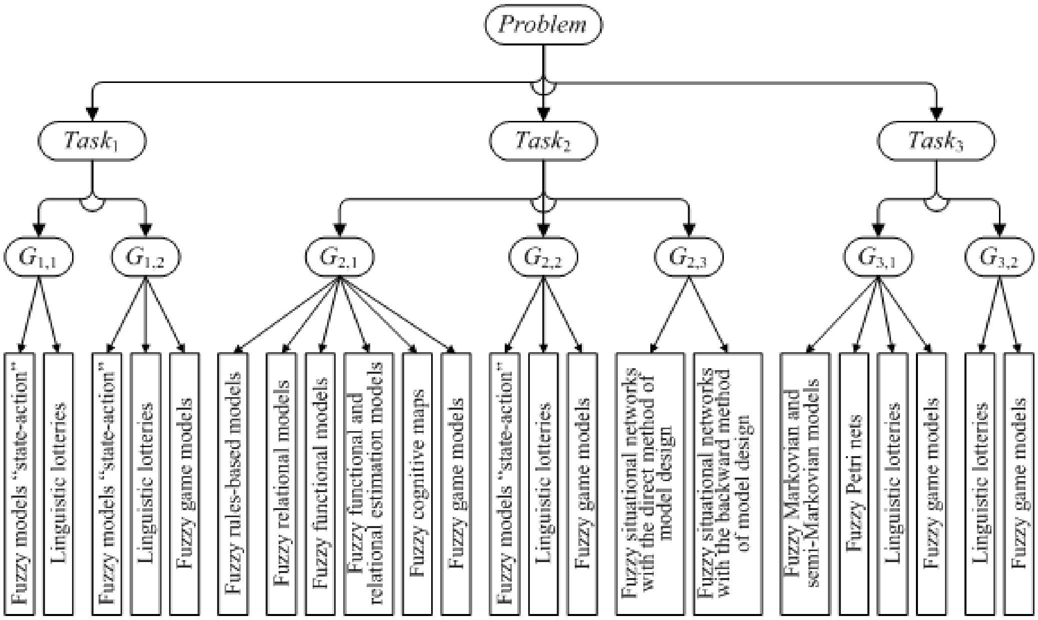


Figure 1. "The covering tree" of complex problem – situational analysis – by a set of fuzzy models

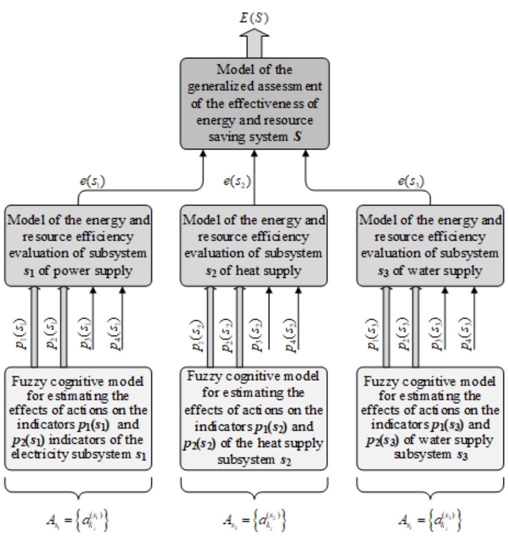


Figure 2. Structure of the compositional fuzzy model

[5] B. Pizzileo, Kang Li, G. W. Irwin, and Zh. Wanqing, "Improved Structure Optimization for Fuzzy-Neural Networks," IEEE Transactions on Fuzzy Systems, vol. 20, no.6, pp. 1076-1089, 2012.
[6] D. Golmohammadi, "Neural network application for fuzzy multi-criteria decision making problems," Int. J. Production Economics, vol 131, pp. 490-504, 2011.
[7] V. V. Borisov, V. V. Kruglov, and A. S. Fedulov, "Fuzzy Models and Networks," 2nd ed., Moscow, Publishing House "Hot Line-Telecom", 2018. (in Russian)
[8] A. V. Bobryakov, V. V. Borisov, A. I. Gavrilov, E. A. Tikhonova Compositional Fuzzy Modeling of Energy- and Resource Saving in Socio-Technical Systems // EAI Endorsed Transactions on Energy Web and Information Technologies. 09.2018–01.2019. Vol. 6. Iss. 21, e9. DOI: 10.4108/eai.12-9-2018.155863.

Метод гибридизации интеллектуальных моделей

Борисов В.В.

Обобщены тенденции развития интеллектуальных моделей. Описан подход к систематизации и классификации гибридных моделей. Приведены примеры классификации известных гибридных интеллектуальных моделей, в соответствии с рассмотренным подходом к их гибридизации. Рассмотрен предложенный метод гибридизации интеллектуальных моделей.

Received 21.12.19

Neuroinformatics Presumptions of Semantic Representations Evolution

Alexander Kharlamov
Neuroontogenesis lab. Institute of Higher Nervous Activity and Neurophysiology of RAS
Moscow State Linguistic University
Higher School of Economics
Moscow
kharlamov@analyst.ru

***Abstract*—To date, traditional information technology and artificial intelligence technology have evolved independently of each other. Now is the time to fundamentally rethink the experience of using and evolution of traditional information technology and its integration with artificial intelligence technology. Currently, the key problem in the development of information technology in general and artificial intelligence technology in particular is the problem of ensuring the information compatibility of computer systems, including intelligent systems. One of the advantages of combining knowledge bases — the core of modern information systems — is the use of a more or less standard way (in the form of semantic networks) to represent knowledge. However, the sole use of semantic networks to represent knowledge does not solve the problem of standardization. For such a representation to become generally accepted, it must be based on the fundamental principles of describing semantics. The most fundamental concept in this sense is the representation of knowledge in human consciousness. Semantic representations in the modern sense – that is, thesauri, semantic networks, ontologies – are a very rough description of the highest level of reality where we live, which can only conditionally be called a model of the world. In order to understand the architecture of the semantic representations that adequately describe the human world (as the modeled object), it is necessary to reproduce the architecture of the human world model as it is formed in the human consciousness. To understand this architecture, one need to look at the human brain (a natural neural network that forms semantic representations), at its (brain) organs that form this architecture, at the cognitive networks that form in it in response to particular situations at its input and output in the process of solving specific problems, at the informatics of individual sensory and effector modalities, the levels of hierarchical representations of which contain images of events of the external and interoceptive world of human of varying degrees of complexity, at the combinations of these hierarchies in multimodal representations, including those in the description of entire situations, as well as the sequence of situations, how they are used in the process of the unconscious and purposeful behaviour – its planning and implementation control. Obviously, the brain is both very large and very heterogeneous neural network that is complex in architecture. Such representations are not only well supported by a comparison with the architecture and informatics of the brain, but are also effectively modelled in applications. TextAnalyst, a software technology for automatic semantic analysis of unstructured texts (which is based on an artificial neural network based on neurons with time summation of signals), effectively implements the functions of forming a homogeneous semantic network, automatic abstracting of texts, comparing texts by their meaning, as well as classifying and clustering texts. It can be assumed that this technology will also effectively analyze code sequences obtained in the analysis of video sequences, if this analysis is sufficiently bionic. A single approach to the processing of textual and visual information will enable constructing effective multimodal systems for processing and presenting information, which is the only accurate approach to the modelling of human intellectual functions.**

***Keywords*—integration of information technology, semantic networks as the basis for standardizing the representation of knowledge, semantic networks in the human consciousness, artificial neural networks, neurons with temporary summation of signals, semantic analysis of texts, integration of modalities**

I. Introduction

To date, traditional information technology and artificial intelligence technology have evolved independently of each other. Now is the time to fundamentally rethink the experience of using and evolution of traditional information technology and its integration with artificial intelligence technology. This is necessary to eliminate a number of shortcomings of modern information technology [1].

Currently, the key problem in the development of information technology in general and artificial intelligence technology in particular is the problem of ensuring the information compatibility of computer systems, including intelligent systems. One of the advantages of combining knowledge bases — the core of modern information systems — is the use of a more or less standard way (in the form of semantic networks) to represent knowledge. However, the sole use of semantic networks to represent knowledge does not solve the problem of standardization. For such a representation to become generally accepted, it must be based on the fundamental principles of describing semantics. The most fundamental concept in this sense is the representation of knowledge in human consciousness, which is the objective of this paper.

Semantic representations in the modern sense – that is, thesauri, semantic networks, ontologies – are a very rough description of the highest level of reality where we live, which can only conditionally be called a model of the world. In order to understand the architecture of the semantic representations that adequately describe the human world (as the modeled object), it is necessary to reproduce the architecture of the human world model as it is formed in the human consciousness. To understand this architecture, one need to look at the human brain (a natural neural network that forms semantic representations), at its (brain) organs that form this architecture, at the cognitive networks that form in it in response to particular situations at its input and output in the process of solving specific problems, at the informatics of individual sensory and effector modalities, the levels of hierarchical representations of which contain images of events of the external and interoceptive world of human of varying degrees of complexity, at the combinations of these hierarchies in multimodal representations, including those in the description of entire situations, as well as the sequence of situations, how they are used in the process of the unconscious and purposeful behaviour – its planning and implementation control. Obviously, the brain is both very large and very heterogeneous neural network that is complex in architecture.

Understanding the architecture of the brain enables correct modelling of its individual elements, their combinations and the architecture as a whole, which is the only tool to implement truly intelligent systems (including the highest level of intelligence, an integrated robot) that will solve intellectual problems no worse than humans, and maybe better if one can implement a model of this architecture compactly enough.

Understanding how the architecture of the human brain is structured will make it possible to approach the solution to the tasks of intelligence cloning, as well as the tasks of the natural informational combination of separate natural and artificial intelligence into some hybrid system (bypassing the sensory channels of exchange, a bottleneck in the exchange of information, but necessary due to significant differences in internal representations of models of the world of various people), which will enable the transition from a virtual repository of social knowledge (currently available) to some real unified storage of knowledge of mankind.

To understand all this, one has to turn to our knowledge of the human brain. The more accurate our understanding of its structure and functions is, the better the artificial intelligent systems that we are trying to build will be. Therefore, let us look at the human brain and its informatics in relation to the semantic representations that model the world of human.

II. Brain informatics

The transition from the signal level of information processing to the symbolic one occurs in the periphery of sensory organs, which differs for different modalities, but brings the streams of input sensory information of various modalities to a single form that corresponds to its (sensory information) representation in the cortical parts of the analyzers. The human brain includes three main organs that handle specific information at the symbolic level. This is the cortex of the hemispheres, cerebrum, hippocampus and thalamus. The cerebral cortex forms a model of the human world as a hierarchy of images of events of varying complexity of various modalities. The hippocampus forms representations of the images of situations and brings these representations into line with real situations encountered by human in the process of his/her activity. Finally, the thalamus is the energy commutator of the brain, which implements the attention mechanism, either focusing it (attention) on one process, or defocusing it to provide completely parallel solving of many tasks simultaneously. Let us consider the first two organs in more detail, since it is they that perform the basic processes of forming a model of the human world, manipulating specific sensory and effector information. One important point to bear in mind: this paper deals exclusively with information processing; intentional moments are left out of the scope of the paper.

A. Cerebral cortex

In the columns of the cerebral cortex, which are mainly composed of groups of pyramidal neurons of the third layer, from the flows of sensory information previously pre-processed in the periphery of sensory organs, hierarchies of images of human world events of various modalities are formed. Each modality at the periphery is processed in its own way, but at the input to the cortex, all modalities are represented by one-type sequences (either lineups or matrixes of sequences, depending on the type of information being analyzed; visual information is more complex than auditory information) of event codes.

These sequences in the columns of the cortex form sequences of pyramidal neurons that respond to them (each pyramidal neuron is a filter for its own fragment of the input sequence, including (fragment) n symbols of this sequence), which is a sequence of passed vertices of the n-dimensional signal space (trajectory), where the coordinates of the vertices correspond to the addresses of the filtering neurons in the corresponding column modelling a fragment of the n-dimensional space R^n. Such a mapping of a symbolic sequence into a multidimensional space leads to the structural processing of input information, that is, to the formation of a hierarchy of dictionaries of event images of various complexity for a particular modality [2], as well as to the formation of syntactic sequences including relations of words from the dictionary of the previous level in the input sequence received at the input of the next processing level.

Hierarchy of representations for a separate modality

The frequency of occurrence of various events of varying complexity at the input of sensory organs is different. So, if we consider textual representations, such events will be letters (or speech sounds), morphemes, words, syntactic groups, steadily repeating pairs of notions. In a natural language, this is called level-forming elements, and the corresponding language levels are called, respectively: graphematic (acoustic-phonetic, for spoken speech), morphemic, lexical, syntactic levels, and semantic level of a separate sentence. Such a representation is characteristic not only for the textual modality, but also for other modalities as well, since the sensory sequences of these modalities are encoded quasi-texts observed by human – the sequences that are meaningful and structured at several levels of complexity of their constituent elements (for example, a video sequence).

The hierarchy of such representations in the form of dictionaries of event images is formed in the columns of the cortical parts of the analyzers; each modality (in each analyzer) has its own one. A feature of the formation of such a hierarchy is the presence of a previous formed representation level as a filter needed for the formation of the next level: when forming a dictionary of the next level, the input information sequence interacting with the formed dictionary of the previous level, generates a so-called syntactic sequence (a sequence of word relations from the dictionary of the previous level in the input sequence, which is used at the input of the next level to form the dictionary of the following level). Thus, a hierarchy of dictionaries is formed in which the words of lower-level dictionaries are inserted by association into the words of dictionaries of the following levels (the words of the higher-level dictionaries are grammars for the words of the lower-level dictionaries), and this entire hierarchy can be considered a world model of the human (in terms of a particular modality), as well as a set of interconnected elements – images of events at various levels – a static model of the world in terms of this particular modality.

Combining hierarchies – a multimodal model of the world

The hierarchies in the dictionary of event images of various modalities formed in this way are virtually combined into a single representation, being combined by associative level-by-level relations between elements describing the same phenomena in terms of different modalities. The same hierarchies are combined informationally (at the higher levels of representations): syntactic sequences lose some amount of information (it remains in dictionaries) as they rise from level to level, therefore, these informational sequences that become increasingly sparse are combined, complementing each other in places of lacunae turning into a single multimodal information sequence. This combination takes place in the parietal cortex that collects information of all modalities into a single representation, a multimodal static model of the world.

Anterior and posterior cortex

Everything written above concerned the sensory (posterior) cortex. There is also the anterior cortex, which is essentially a motor cortex. It is also a hierarchy of representations, the lower level of which – the motor cortex proper – controls the human motor activity. What are the upper levels of the anterior cortex?

In processing information by the posterior and anterior cortex, a hippocampus is involved at their border. First, let us talk about what happens in the process of processing sensory information in the hippocampus, and how it participates in the formation of representations in the anterior cortex.

B. Hippocampus

The main function of the hippocampus is the formation of sensory representations of the situation, integral in space and time, that include images of events represented in the cortex in a way they are included in particular situations of the external and interoceptive world of human. In individual hippocampal lamellae (CA_3 field represents the Hopfield associative memory [3]), images of situations are formed as spatio-temporal groups of event images represented in the corresponding columns of the cortex.

Another function of the hippocampus is to filter the input flow of situations through the images of situations presented in the lamellae, and to identify the degree of their similarity to the images of situations (averaged in some sense) stored in them. After identifying the degree of similarity, the representation of the old and new images of situations is averaged.

Hippocampus in the processing of sensory information

The situations represented in the lamellae of the hippocampus are used by the higher levels of the motor (front) cortex in the process of formation and control of purposeful behaviour.

This representation can be illustrated by the extended predicate structure of the sentence (which actually describes the situation), which is a graph that includes the subject, predicate, main and secondary objects and attributes. The graph describing the situation and represented in the lamella of the hippocampus is similar to the graph of the extended predicate structure, and includes multimodal representations of the same images: subject, main and secondary objects, attributes. This representation includes both linguistic and multimodal representations, both sensory and motor ones.

Representations of situations in the hippocampus due to the influence of representations of current implementations of situations at the input to the cortex are averaged in free time of the human. The sensory image of the situation projected onto the columns of the cortex by association selects those columns where there is a representation of the images of events included in the input situation. This image is projected (also associatively) onto the hippocampus, where those lamellae respond, the image of the situation in which most closely matches the input situation. During the iterative procedure, the found lamellae affect the images of the initial events of the cortical columns (the so-called long-term potentiation [4]) providing further training for the images of events stored in the columns to an average state (average between the state stored in the lamella and the state received from the input). In the process of this procedure, further training of images of situations stored in the hippocampus also takes place.

Hippocampus in the processing of motor information

The situations presented in the lamellae of the hippocampus are used by the higher levels of the motor (anterior) cortex in the process of formation and control of purposeful behaviour.

In the images of situations represented in the lamellae of the hippocampus, motor information is also recorded in addition to sensory information. Thus, along with sensory images, their effector analogues are stored there. It is they that form at the lower level of the anterior cortex (the motor cortex proper) the representations that control the effector organs of the human.

C. Anterior (motor) cortex

The world represented in the effector model of the world (the anterior cortex) is, in some sense, mirrored by its representation in the sensory cortex. A hierarchy of representations is also formed in the effector part of the world model, but these representations work in the opposite direction: while in the sensory cortex information flows come from the bottom up (analysis), they are directed from top to bottom (synthesis) in the motor cortex. But before this control flow from top to bottom occurs, such images of effector (motor) images of events of various levels should be formed.

Inner speech as an example of forming a hierarchy of representations of the anterior cortex

Let us show how motor representations are formed by the example of the formation of internal speech, as A.R. Luria [5] described it (managing internal speech also belongs to motor skills). Inner speech is initially formed with the participation of the teacher. Mother tells her son: "Do this, do that." And he does. In the son's hippocampus lamella, a situation is remembered when he hears, repeats what he hears and does what he is told. In the first level of the anterior cortex (the motor cortex proper) following the last level of the posterior cortex, the least variable information is recorded – repeating the phrase ("Do this!").

Over time, a fair amount of such events accumulate in the motor cortex (lower-level dictionaries are formed). However, life moves forward: in addition to simple actions, the son, under the control of his mother, performs more complex (increasingly complex) actions, in which those simple events are components. They form representations of the following levels. These representations of the anterior cortex are related to the hippocampus by association, as well as the representations of the sensory cortex: the corresponding hippocampal lamella by association responds to images of the anterior cortex that are involved in the situation description presented in this lamella ("Do this first, then that!"). But the same lamella also includes links to images of the sensory cortex, which describe the mentioned situation.

At higher and higher levels of the anterior cortex, increasingly complex effector events are presented – sequences of lower-level events of the motor cortex with links to the lamellae of the hippocampus, that is, to entire spatio-temporal images of situations. Thus, the higher levels of the anterior cortex manipulate sequences of images of situations.

D. Purposeful behaviour

Actually, the formation of such sequences of situations (from the current situation to the target one), and control over the implementation of these sequences, that is, correcting the discrepancies in the situation representations in the hippocampus (plus in the columns of the cortex, on which the representation in the hippocampus is based) and in reality is a purposeful behaviour. Inner speech is also purposeful behaviour.

III. Integration of modalities as a perspective for the development of semantic representations

The representation described above is, to some degree, implemented in modern intelligent applications that model the processing of information in the human brain. However, currently such models are somewhat one-sided. If speech recognition is modeled, the task of processing speech information only is to be solved. However, sometimes processing of visual information about articulatory movements is also added (the so-called lip-reading). If the processing of visual information is modelled, the task of processing visual information only is to be solved.

The human brain is not so one-sided. To process information of any modality, any suitable information is involved. If speech is recognized, then any necessary accessible context is involved, including information not only related to the language model of the world, but also extralinguistic information of modalities other than textual. The same thing happens when processing information of other modalities. It is the involvement of the widest (necessary) context that allows achieving such accuracy in speech recognition, even in noisy conditions. But one has to pay for it with the resources needed to support the provision of the appropriate context (including other modalities).

A. *World models in the dominant and subdominant hemispheres*

The human model of the world, due to the features of the information representation in the dominant and subdominant hemispheres, is divided into three independent parts: two of them are in the dominant hemisphere and the third in the subdominant one [6]. Since the cortical fields responsible for the perception and articulation of speech are represented in the dominant hemisphere of the human brain, it is here where the linguistic model of the world is formed, including the language model, and the world model described in terms of language. This model is formed in parallel with another model of the dominant hemisphere (multimodal/ extralinguistic), which is formed under the influence of society through the language model. The multimodal model of the dominant hemisphere is multi-level (knowledge of society is deep and wide), but schematic, since socialized knowledge is usually represented by few examples. In the subdominate hemisphere, the situation is fundamentally different. There is no direct influence of society (there is no speech perception and synthesis), therefore the model of the world is multimodal and very individual: it contains only the information that the person handled in the process of his/her development. But this model has only two levels of representation (part-whole) due to the very great variability of the events presented there.

These three parts of the model of the world are connected level-by-level by associative relations, and represent a single whole in the process of manipulation.

B. *Integral robotics*

The only modern type of intelligent systems where the use of a wide context is justified from this point of view, and is even necessary regarding the task assignment, includes integrated robots, which by their nature include sensors and effectors of many modalities in their architecture. But this is an advantage of tomorrow, and today resources are being saved there, and the sensory (and motor) skills only work as much as it is strictly necessary for solving the task. If this is a speech control, then a command recognition system with speaker-dependent tuning and restrictions in the dictionary (and other restrictions) is used.

IV. Mechanisms for the Formation of Semantic Representations

At the moment there are few approaches to the representation of semantics. Traditionally, these are: (1) logical languages; (2) production rules systems; (3) frame representations; and (4) semantic networks. Recently, they were completed by (5) ontologies [7]. However, ontologies are just another (non-graphical) form of either frame representations or semantic networks. Therefore, we will not consider them separately.

Of the above, production rules systems (in expert systems), and semantic networks are still successfully used. The effectiveness of using these representations depends solely on the availability of mechanisms for automatic generation and manipulation of the semantic representation itself. In this sense, one can only talk about the presence of mechanisms for the automatic formation of semantic networks. All other types of semantic representations are formed exclusively manually. And even the deep learning mechanisms do not leave any hope for the automation of these processes [8].

It is easiest to automatically form homogeneous (so-called associative) semantic networks. They work almost exclusively on statistics of the processed text (for now, only mechanisms for analyzing natural language texts have been implemented). Using linguistic information, it is possible to automate the formation of heterogeneous semantic networks as well.

A. *Homogeneous semantic networks*

Homogeneous semantic networks are graphs, where vertices are connected by arcs that represent only one type of relations, that is, “be together”. Therefore, they are called associative. They are easily generated automatically, since identification of a dictionary of words (recall that so far only analysis of texts of natural language modality has been implemented) in a text and identification of a dictionary of pairwise occurrence of words (these dictionaries are necessary for constructing an associative network) is not difficult. And the calculation of the weights of the network vertices is a completely manageable procedure [9]. It is also easy to use them in applications [10].

B. *Heterogeneous semantic networks*

Inhomogeneous semantic networks require for their construction information on the types of relations between the notions of the network that characterize arcs. There are two ways to automatically construct heterogeneous semantic networks. One can first construct a homogeneous semantic network, and then mark up the relations between the vertices by examining the sentences of the text (corpus of texts), on the basis of which a homogeneous semantic network is constructed. One can also first examine the sentences of the text, on the basis of which a heterogeneous semantic network is constructed, and then calculate the weights of the network vertices in the same way as in calculating the weights of the vertices of a homogeneous semantic network.

In the first case, we obtain a somewhat "roughened" network, but it is resistant to distortions generated by a linguistic interpretation of link analysis. Since the procedure is statistical in nature, it is almost independent of the language. In the second case, the construction of the network is complicated by the fact that no more than 85% of the sentences of the Russian-language text are analyzed correctly when identifying the types of relations between their constituent words [11]. In addition, the mark-up of sentences according to the types of relations depends on the language, and the resulting network with very blurry vertices (each type of relation for the same pair of notions generates additional branches in the network) is poorly ranked by vertex weights.

V. Neural network interpretation of the world model architecture

We still have to say a few words about the possibility of a neural network implementation of the procedure for automatic formation of the world model. As mentioned above, only two types of semantic representations are implemented automatically. They will be discussed below.

A homogeneous semantic network can be interpreted as a model of the world (linguistic - textual), for example. The same mechanisms can be used to form an extralinguistic model of the world (so far this has not been implemented). To do this, it is necessary to replace natural language text sequences at the input of the semantic network formation process with quasi-text sequences (for example, a video sequence encoded accordingly).

A. Neural network based on neurons with temporal summation of signals – cortical column model

Constructing a homogeneous semantic network is an ordinary procedure, if artificial neural networks based on neural-like elements with time summation of signals are used for this [2]. Such artificial neural networks are naturally designed to identify dictionaries of words and word pairs in natural language texts, as described above. They model the processing of information in columns of the cortex.

B. Hopfield associative memory - hippocampal lamella model

Applying a procedure similar to the Hopfield associative memory mechanism [12], it is just as easy to calculate the weights of the network vertices [10].

C. Formation of a heterogeneous semantic network

Probably, one can look for evidence that the mark-up of relations between network vertices can also be implemented using artificial neural networks [13]. Well, then one can rely on a completely neural network mechanism for constructing a heterogeneous semantic network.

D. Analysis of texts and quasi-texts

All that was said about the analysis of texts can be said about the analysis of quasi-texts, which are understood as meaningful code sequences relating to the interpretation of the human world formed by sensors other than textual. Since quasi-texts are similar to natural-language texts, that is, they contain level-forming units of varying degrees of complexity, then, at least, theoretically, nothing prohibits them from being processed in ways similar to those described above, when forming both a homogeneous and a heterogeneous semantic network.

VI. Possible applications

The processes described above are quite easily algorithmized at the information level. And software solutions are different from similar ones relying on simpler and less natural bases, and work for the benefit of users. So, on the basis of a simplified language model of the world formed in the columns of the cortex, which includes only some levels of language representation supplemented by an iterative procedure for knowledge reordering, as it is implemented in the Hopfield associative memory, which models the CA_3 fields of hippocampal lamellae, a program system was created for semantic analysis of texts that uses a network representation of semantics and has proved quite effective [14].

A. Semantic analysis of texts

TextAnalyst, a software system for the semantic analysis of texts, was developed on the principles presented above. It includes a limited language model where there is no syntactic level (it is replaced by a dictionary of deleted words), and where morphological analysis is implemented by the simplest means, but an artificial neural network based on neural-like elements with time summation of signals is used to represent the lexical level and semantics of a separate sentence. A simplified representation is also taken from the informatics of the hippocampal lamella, including iterative reordering of the weights for the vertices of a homogeneous semantic network obtained at the semantic processing level [2].

Such a network modeling of text semantics makes it possible to take into account the word relations in the analyzed text by n steps in the semantic network. It is convenient for analyzing the text semantics, since it does not require a huge training sample of texts as in n-gram language modelling. The ranking of the text notions obtained as a result of the iterative procedure allows the extraction of a topic tree of texts, as well as abstracting and comparison of texts by meaning and classification of texts.

B. Analysis of a video sequence as a quasi-text

An effective solution to the problem of semantic analysis of natural language texts leaves hope for the effective use of the presented approach for the analysis of video sequences (and even individual two-dimensional images), as opposed to quite successful convolutional networks.

However, it requires implementation of a very specific image preprocessing, which is very different from the usual sweep from left to right and from top to bottom. If we model image processing by the eye of a person who sees images as large contoured surfaces filled with some texture (for example, color) and scans these images from the point of greatest informativity [15] to the point of greatest informativity (as a rule, these are contour turning points) according to a given rule (rules differ depending on the task assignment), a code sequence is obtained that resembles text. It repeats elements of varying complexity (they differ from those in natural language texts), that is, it becomes possible to structurally analyze them, as well as ordinary text.

But one can save the whole information processing mechanism, similar to that of the TextAnalsyt technology. But scanning itself differs in one aspect from traditional methods of image processing: we can scan a poorly represented part of the image in the model for as long as it takes to present it in detail.

C. Combining modalities

And finally, there is another very important advantage of this way of representing semantics. A single way of presenting and processing information of textual and visual modalities makes it possible to combine these representations into a single model of the world [16], in which fragments of the visual modality are named with the corresponding names of the textual modality. Such a two-modal representation is much more resistant to recognition problems than a single-modal visual one. In a two-modal representation, one of them complements the gaps (lacunes) of the other. But, most importantly, this representation makes it possible to interpret visual images and situations in terms of a natural language, greatly simplifying the communication of the system with the user.

VII. Conclusion

The tasks to be solved when creating artificial intelligent systems are effectively achievable provided that architectures similar to the human brain architecture used to solve a similar problem are used when solving them. Including the use of representations about the model of the world implemented on the basis of any approaches based on semantic representations. A prerequisite for the formation of a model of the task being solved in this case is the automatism of the formation of the mentioned semantic representations. At the moment, only semantic networks can be efficiently generated automatically, including involvement of artificial neural networks.

Such representations are not only well supported by a comparison with the architecture and informatics of the brain, but are also effectively modelled in applications. TextAnalyst, a software technology for automatic semantic analysis of unstructured texts (which is based on an artificial neural network based on neurons with time summation of signals), effectively implements the functions of forming a homogeneous semantic network, automatic abstracting of texts, comparing texts by their meaning, as well as classifying and clustering texts. It can be assumed that this technology will also effectively analyze code sequences obtained in the analysis of video sequences, if this analysis is sufficiently bionic. A single approach to the processing of textual and visual information will enable constructing effective multimodal systems for processing and presenting information, which is the only accurate approach to the modelling of human intellectual functions.

The standardization of knowledge representation on the basis of semantic networks, as they are presented in the human consciousness, is an appropriate basis for standardizing the ways of representing knowledge in information systems.

References

[1] V. Golenkov, N. Guliakina, and I.Davydenko, “Methods and tools for ensuring compatibility of computer systems,” [Open semantic technologies for intelligent systems, 2019, iss. 3, Pp. 25-53, 2019].

[2] A.A. Kharlamov, Assotsiativnaya pamyat’ – sreda dlya formirovaniya prostranstva znanij. Ot biologii k prilozheniyam. [Associative memory as an environment for the formation of a space of knowledge. From biology to applications]. Germany: Palmarium Academic Publishing, 2017. (in Russian)

[3] E.T. Rolls, “Theoretical and Neurophysiological Analysis of the Functions of the Primate Hippocampus in Memory,” [Cold Spring Harbor Symposia on Quantitative Biology, Vol. LV, Cold Spring Harbor Laboratory Press. Pp. 995 – 1006, 1990].

[4] O.S. Vinogradova, Gippokamp i pamyat’. [Hippocampus and memory]. Moscow, USSR: Nauka, 1975.(in Russian)

[5] A.R. Luria, Yazyk i soznanie. [Language and consciousness]. Rostov-on-Don: Feniks, 1998.(in Russian)

[6] V.L. Bianki, Morfofunktsionalnaia organizatsiia mezhpolusharnykh sviazei. [Morphofunctional organization of interhemispheric connections]. Leningrad, USSR: Nauka, 1989.(in Russian)

[7] I.V. Efimenko, and V. F. Khoroshevskii, Ontologicheskoe modelirovanie: podkhody, modeli, metody, sredstva, resheniia [Ontological modelling: approaches, models, methods, tools, solutions]. Moscow: HSE Publ. House, 2011. (in Russian)

[8] N.V. Lukashevich, and B.V. Dobrov, “Combining ontological and statistical methods in the tasks of automatic processing of large text data [Kombinirovanie ontologicheskikh i statisticheskikh metodov v zadachakh avtomaticheskoi obrabotki bolshikh tekstovykh dannykh],” [Coll. of papers of the Conference “Artificial intelligence: issues and solutions”. Pp. 32-37, 2018].

[9] A.A. Kharlamov, “Formirovanie n-grammnoj tematicheskoj modeli teksta. [Formation of an n-gram topic model of the text],” J Speech technology 1-2: 15–23, 2016. (in Russian)

[10] A.A. Kharlamov, Nejrosetevaya tekhnologiya predstavleniya i obrabotki informatsii (estestvennoe predstavlenie znanij). [Neural network technology of information representation and processing (natural representation of knowledge]. Moscow, Russia: Radiotekhnika, 2006. (in Russian)
[11] G.S. Osipov, I.A. Tikhomirov, and I.V. Smirnov, "Exactus – an intelligent meta-search system on the Internet [Exactus – sistema intellektualnogo metapoiska v seti Internet]," [the 10th National Conference on Artificial Intelligence with KII-2006 participation. Moscow: Fizmatlit. Vol. 3, pp. 859-866, 2006]. (in Russian)
[12] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Natl. Acad. Sci. 79, Pp. 2554 – 2558, 1982.
[13] L. He, K. Lee, O. Levy, L. Zettlemoyer, and P.G. Allen, "Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling," [Annual Meeting of the Association for Computational Linguistics (Short Papers) , Melbourne, Australia, July 15 - 20, Pp. 364–369, 2018].
[14] D. Sullivan, Document Warehousing and Textmining. NY: Wiley publishing house, 2001.
[15] N.V. Zavalishin, and I.V. Muchnik, Model zritelnogo vospriyatiia i algoritmy analiza izobrazhenij [Visual perception model and image analysis algorithms]. Moscow: Nauka, 1974. (in Russian)
[16] A.A. Kharlamov, R.M. Zharkoy, and V.V. Arzumanov, "Kollektivnaya model mira mobilnogo robota. Yazykovoy analiz stseny i upravlenie peshekhodnym perekhodom [Collective model of the world of a mobile robot. Language scene analysis and pedestrian crossing management]," J Speech technology 1. Pp. 99-108, 2019. (in Russian)

Перспективы развития семантических представлений, основывающиеся на тенденциях нейроинформатики

Харламов А.А.

До настоящего времени традиционные информационные технологии и технологии искусственного интеллекта развивались независимо друг от друга. Сейчас настало время фундаментального переосмысления опыта использования и эволюции традиционных информационных технологий и их интеграции с технологиями искусственного интеллекта. Ключевой на текущий момент проблемой развития информационных технологий в целом и технологий искусственного интеллекта в частности является проблема обеспечения информационной совместимости компьютерных систем и в том числе интеллектуальных систем. Одной из возможностей информационного совмещения баз знаний – ядра современных информационных систем – является использование для представления знаний более или менее стандартного способа в виде семантических сетей. Однако само использование семантических сетей для представления знаний не решает проблемы стандартизации. Чтобы такое представление стало общепринятым, оно должно базироваться на фундаментальных принципах описания семантики. Наиболее фундаментальным в этом смысле представлением является представление знаний в сознании человека. Семантические представления в современном понимании – тезаурусы, семантические сети, онтологии – это очень грубое описание самого верхнего уровня действительности, в которой мы живем, которое лишь условно можно назвать моделью мира. Для понимания релевантной моделируемому предмету – реальному миру человека – архитектуры семантических представлений, адекватно описывающих этот мир, необходимо воспроизвести архитектуру модели мира человека, как она формируется в его сознании. Для понимания этой архитектуры необходимо посмотреть на мозг человека (естественную нейронную сеть, формирующую семантические представления), на его (мозга) органы, которые формируют эту архитектуру, на когнитивные сети, возникающие в нем в ответ на появление конкретных ситуаций на его входе и выходе в процессе решения конкретных задач, на информатику отдельных сенсорных и эффекторных модальностей, уровни иерархических представлений которых содержат образы событий внешнего и интероцептивного мира человека разной степени сложности, объединение этих иерархий в многомодальных представлениях, в том числе, в рамках описания целых ситуаций, а также последовательностей ситуаций, как они используются в процессе неосознанного, а также целенаправленного поведения – его планирования и контроля исполнения. Очевидно, что мозг – это и очень большая, и очень неоднородная, а потому, сложная по архитектуре нейронная сеть. Такие представления не только хорошо подтверждаются сравнением с архитектурой и информатикой мозга, но и эффективно моделируются в приложениях. Программная технология для автоматического смыслового анализа неструктурированных текстов TextAnalyst (в основе которой лежит искусственная нейронная сеть на основе нейронов с временной суммацией сигналов) эффективно реализует функции формирования однородной семантической сети, автоматического реферирования текстов, сравнения текстов по смыслу, классификации и кластеризации текстов. Можно предполагать, что также эффективно подобная технология будет анализировать кодовые последовательности, полученные при анализе видеорядов, если этот анализ будет достаточно бионичен. Единый подход к обработке текстовой и зрительной информации позволит говорить о построении эффективных многомодальных систем обработки и представления информации, что является единственно верным подходом в развитии моделирования интеллектуальных функций человека.

Received 09.01.2020

Open Integrated Semantic Component Design Technology

Georgy Evgenev
Computer manufacturing automation systems
Bauman Moscow State Technical University
Moscow, Russia
g.evgenev@mail.ru

***Abstract*—The Digital revolution in industry is supposed to cover all stages of the product life cycle, including product design and planning of manufacturing processes. At these stages, goods and processes are not accomplished as real things but formed as models in the virtual world. Therefore, the Internet of Things concept, the basis of the "Industry 4.0" project, is not sufficient to conduct a full-scale digital revolution.**

The key methodology to examine this problem is the methodology of artificial intelligence. It provides for comprehensive consideration of the problems that arise at all stages of the life cycle of engineering products.

The Internet of Knowledge has an ontological basis and includes meta-ontology, which comprises the ontology of objects, the ontology of tasks and the ontology of optimization. The Digital Revolution should give the knowledge carriers without programming skills an opportunity to enter pieces of information into the computer without intermediaries. The materials of the paper are of practical value for the creation of integrated automation systems of engineering products design and production.

***Keywords*—digital manufacturing, intelligent systems, computer-aided process planning, computer-aided manufacturing, manufacturing execution system**

I. Introduction

Before the digital revolution, written sources were the bearers of knowledge. As a result of the digital revolution, software has become knowledge carriers. Software was originally built on an algorithmic basis using algorithmic languages. A non-programming knowledge carrier could not enter them into a computer. The digital revolution should radically change this scheme and enable a non-programming knowledge carrier to enter them into a computer without intermediaries. This became possible using the expert programming methodology [1, 2]. In this methodology, knowledge is described in the language of business prose, as close as possible to the literary language, but formalized so that it is possible to automatically generate software tools that correspond to the source texts. Below is an example of the use of expert programming.

II. Wikipedia and Expertopedia

The most common computer encyclopedia currently is Wikipedia (Fig. 1). Wikipedia is a publicly accessible multilingual universal Internet encyclopedia with free content [3]. Its conceptual principles are multilingualism and the ability of users to replenish and adjust the content. Encyclopedia is a review of all branches of human knowledge or a circle of disciplines brought into the system that together constitute a separate branch of knowledge. In this case, we are interested in the meta category "Technique". Only information on various devices, mechanisms and devices that do not exist in nature and are made by humans should be directly placed in this category.

Traditional wikis have a number of drawbacks, which include, in particular, lack of consistency in content. Due to the frequent duplication of data on wikis, the same information may be contained on several different pages. When changing this information on one wiki page, users should ensure that the data is also updated on all other pages.

Another drawback is the difficulty of accessing the knowledge available on wikis. Large wiki sites contain thousands of pages. Performing complex search queries and comparing information obtained from different pages is a task that is quite time-consuming in traditional wiki systems. The program can only show the text of a Wikipedia article in a certain context and cannot take additional steps related to the object.

Traditional wikis use flat classification systems (tags), or classifiers organized in taxonomy. The inability to use typed properties generates a huge number of tags or categories. In this regard, semantic wikis appeared (Fig. 1). Semantic Wiki is a web application that uses machine-processed data with strictly defined semantics in order to expand the functionality of a wiki system [4].

Regular wikis are populated with structured text and untyped hyperlinks. Semantic wikis allow you to specify the type of links between articles, the type of data within the articles, and page information (metadata).

A variety of such systems based on the technology of expert programming described below for creating knowledge bases can be called "Expertopedia" (Fig. 1).

The semantics of Expertopedia is determined by the

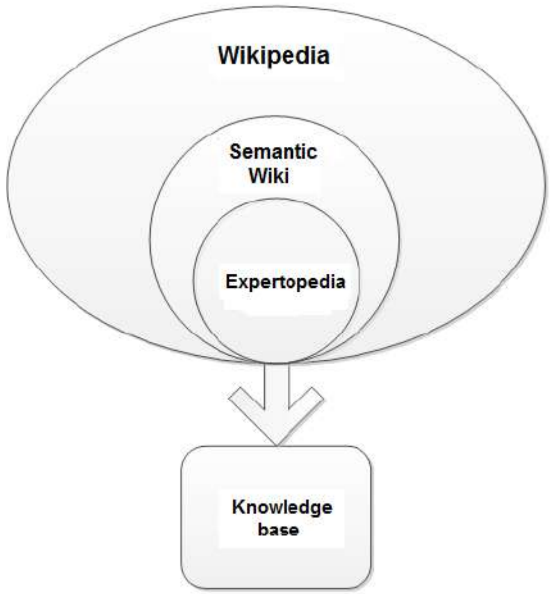


Figure 1: Wikipedia and Expertopedia

metaontology used in all cases. In this case, the root meta object is the knowledge base, which has its own name, identifier name, and version. The methods connected to the knowledge base allow you to sort and search the knowledge modules of which it consists. A subset of knowledge modules integrated into a semantic network is a method that can be exported to other knowledge bases, as well as imported and merged into a single semantic network.

From a wiki point of view, the knowledge base is a page. Moreover, the knowledge base can be considered as a module as part of the upstream knowledge base, which in this case consists of a set of pages. The structural components of a knowledge base are a knowledge base dictionary and many modules. The dictionary has its own identifier name and consists of words. It is possible to sort and search for words, as well as import words from text documents. Words also have an identifier name and a meaningful name in one of the languages of the world, as well as a type of word from a set of integers and real numbers and a set of characters. There are operations for adding, deleting and determining the occurrence of a word.

Words of an enumerated type can have associative lists containing valid word names. An associative list has an identifier name, a name, and a type of values from among the above. For lists, you can add, delete, sort values, and search for a list. The list consists of many elements with their values and the ability to add and remove values. The ability to replace the knowledge base dictionary allows for the fulfillment of the conceptual requirements of wiki systems: the use of any world language.

The second structural component of a knowledge base is a multitude of knowledge modules (KM). KMs perform the main functional purpose of the knowledge base - transforming the current state of the data in order to obtain new objects that satisfy design goals. Each module has a name, as well as identifier names of the module itself, preconditions for its execution and version.

The knowledge base has the ability to add modules, select an analog module, translate the knowledge module into one of the programming languages, test the result, as well as determine whether the module is in the knowledge base and delete the selected module. These actions can be performed by any operator, which allows you to implement the second conceptual basis of wiki systems: providing the possibility of independent replenishment and adjustment of the content of the system.

Each knowledge module has its own dictionary, which is a subset of the knowledge base dictionary. Due to the fact that modules are an object-function, the dictionary has two subsets: input and output variables of the module.

Knowledge modules can have preconditions that determine the possibilities of its implementation. Preconditions are logical expressions made up of words from a dictionary of a knowledge base, restrictions on the value of variables, and logical connectives.

The functions of the data conversion module are performed by mechanisms, which can be formulas, tables, working with databases, generating 3D models of objects, as well as mathematical models. When working with databases, it is necessary to select a database from among the available ones and the tables in it. Next, the access conditions are formed and the fields and properties used in the database and when creating the knowledge base are matched. The operation itself is performed by the corresponding DBMS.

It is possible to create and modify geometric 3D models based on the calculation results using the capabilities of various CAD systems. In this case, it is necessary to form a model and parameterize it.

Another variety of mathematical knowledge needed to perform calculations is models of continuous systems based on differential-algebraic systems of equations. To implement these capabilities, there must be a tool with the ability to generate the mentioned models. Such a tool should provide: support for object-oriented modeling technology compatible with the UML language; convenient and adequate description of the model in a generally accepted mathematical language without writing any program code; automatic construction of a computer model corresponding to a given mathematical model, with the possibility of autonomous use of this computer model.

To the greatest extent, these requirements are met by the Model Vision Studium (MVS) package [5]. The core element of the MVS language is the Active Dynamic

Object (ADO). An active dynamic MVS object provides modeling not only of continuous behavior, but also of a discrete or hybrid one. As noted above, semantic wikis allow you to specify the type of links between articles, the type of data within the articles, and information about pages (metadata). In Expertopedia, articles are knowledge bases that can go into one another in the form of knowledge modules.

III. Knowledge Base Examples

We will consider the formation of a knowledge base by the example of designing caps for bearings. The knowledge base consists of knowledge modules (MZ). Each module has input and output properties, as well as a mechanism for converting the first to the second. In this example, formulas and tables are used as mechanisms. Using knowledge tools, knowledge modules are converted into software tools, with the help of which in this example a 3D model of an object with calculated dimensions is generated (Fig. 2).

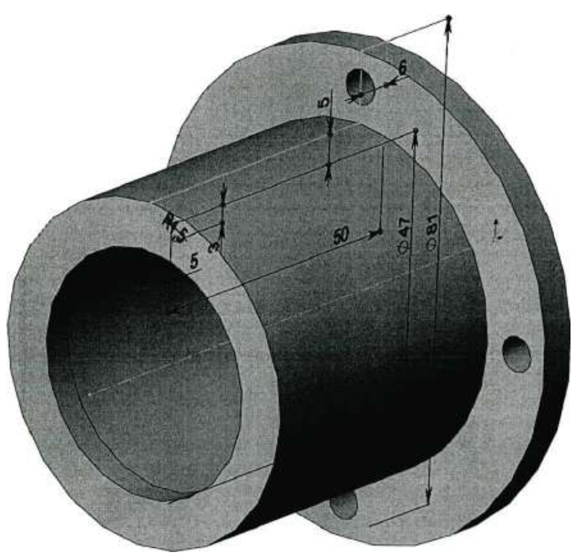


Figure 2: 3D model of the designed screw cap

Determining the design of the covers is the diameter D of the hole in the bearing housing 204 GOST 8338-75. Table I gives recommendations for choosing the wall thickness del, diameter d and number z of screws that secure the cover to the housing, depending on the D bearing:

Table I

D, mm	0...50	50...62	63...95	100...145	150...220
del, mm	5	6	8	10	12
d, mm	6	6	8	10	12
z	4	4	4	6	6

Dimensions of other structural elements of the cover: del2 = 1.2 * del; del1 del Dph = D + (4 ... 4, 4) d; C d.

MZ: "FORMULA" — Calculation of the size of the bearing cap

Input properties

Name	Description	Type	Value
D	Bearing diameter	REAL	47
dh	Hole diameter	REAL	6
r_	Bearing radius	REAL	1.5
del	Side wall thickness	REAL	5

Mechanism — Formula

dell=del
del2=1.2*del
Da=D+2*del
Df=Da+4*dh
C_=dh
t_=2*r_

Output Properties

Name	Description	Type	Value
C_	The distance from the side wall to the centers of the holes	REAL	6
t_	Thrust Shoulder Height	REAL	3
Df	Flange diameter	REAL	81
del1	End wall thickness	REAL	5
del2	Flange thickness	REAL	6
Da	Outside Diameter	REAL	57

MZ: "tabledel" — Selection of bearing cap sizes

Input Properties

Name	Description	Type	Value
D	Bearing diameter	REAL	47

Mechanism — Table

Table property configuration

D
Del
Dh
Z
r

Table

(0,50)	(50, 62]	[63, 95]	[100,130]	(130,145]	[150,220]
5	6	8	10	10	12
6	6	8	10	10	12
4	4	4	6	6	6
1.5	1.5	2	1.5	3	3.5

Output Properties

Name	Description	Type	Value
Z	Number of holes	INTEGER	4
r	Bearing radius	REAL	1.5
del	Side wall thickness	REAL	5
dh	Hole diameter	REAL	6

IV. Results

In order to integrate the Industrial and Digital revolutions that are taking place at present, it is necessary to consider two worlds together: the virtual world realized by the Internet of Knowledge (IoK), and the real world realized by the Internet of Things (IoT).

The Internet of Knowledge has an ontological basis and includes meta-ontology, which comprises the ontology of objects, the ontology of tasks and the ontology of optimization.

The Digital revolution should enable the non-programming knowledge carriers to enter knowledge into the computer without intermediaries. That can be done by way of expert programming methodology, in which knowledge is described in the language of business prose, which is very close to the literary language, but formalized so that it becomes possible to automatically generate software matching the source texts. Business prose can be formed in any languages, and software can be generated in different programming languages.

Artificial intelligence methods allow creating semi-automated systems for products' 3D model generation using knowledge bases integrated with CAD systems. It should be possible to integrate them with various CAD systems.

Knowledge bases are generated on the basis of knowledge modules representing a condition-action rule, which has an identifier and name, a precondition, input and output properties, and a mechanism for converting the first to the second. Modules are automatically translated into subprograms in the programming language selected by the user. Thus, the user can choose both the input language of the knowledge representation and the resulting language of the software generation.

References

[1] Evgenev G. B. Sprut ExPro is a tool for generating multi-agent design systems in mechanical engineering. Part 1. // News of higher educational institutions. Engineering., No. 6, 2017, p.66-77

[2] Evgenev G. B. Sprut ExPro is a tool for generating multi-agent design systems in mechanical engineering. Part 2. // News of higher educational institutions. Engineering., No. 7, 2017, pp. 60-71

[3] Todd R. Weiss. Wikipedia simplifies IT infrastructure by moving to one Linux vendor, Computerworld.com (October 9, 2008 (Computerworld))

[4] Sikos, Leslie F. (2017). Description Logics in Multimedia Reasoning. Cham: Springer International Publishing. doi:10.1007/978-3-319-54066-5. ISBN 978-3-319-54066-5

[5] Kolesov Yu.B., Senichenkov Yu.B. Modeling systems. Object oriented approach. Bhv. 2016, p. 192

Открытая интегрированная технология компонентного проектирования

Евгенев Г. Б.

Предложена методология создания семантических вики-систем для представления знаний с использованием технологий искусственного интеллекта. Методология основана на многоагентных методах создания баз знаний и пригодна для разработки систем проектирования и управления для цифровых интеллектуальных производств. При формировании внешнего представления баз знаний могут использоваться национальные языки. Проанализированы схемы ввода знаний в компьютер. На основе технологии экспертного программирования для создания баз знаний предложена семантическая вики-система, которая может быть названа «Экспертопедией». Возможность замены словаря базы знаний позволяет обеспечить выполнение концептуального требования викисистем: использование любых языков. В базе знаний имеются возможности добавления модулей, выбора модуля-аналога, трансляции модуля знаний на один из языков программирования, тестирования полученного результата, а также определения входимости модуля в базы знаний и удаления выбранного модуля. Эти действия могут выполняться любым носителем знаний, что позволяет реализовать вторую концептуальную основу вики-систем: обеспечение возможности независимого пополнения и корректирования содержания системы. Разработана метаонтология Экспертопедии. Подробно рассмотрены методы создания многоагентных баз знаний для проектирования и управления в машиностроении. Приведена архитектура агента САПР. Даны примеры применения многоагентных систем для создания интеллектуальных систем полуавтоматического проектирования изделий машиностроения. Описана технология экспертного программирования.

Received 05.12.2019

The mechanism of optimal control of the evolution of continuous technological systems

1st Boris V. Paliukh
dept. Information System
Tver State Technical University
Tver, Russia
pboris@tstu.tver.ru

2nd Irina A. Egereva
dept. Information System
Tver State Technical University
Tver, Russia
irina.egereva@gmail.com

***Abstract*—Due to competition in the conditions of continuous development of technology, enterprises have to constantly optimize production processes. There is a problem of choosing the most suitable solution to a particular situation managing the evolution of a technological system because the activities of enterprises must take into account a large number of external factors, take into account various kinds of limitations (resource, material and others).**

This paper presents the mechanisms of optimal control of continuous technological systems based on the application of the theory of active systems, the theory of fuzzy systems, artificial intelligence methodologies, multi-agent systems. A method is proposed for forming a minimum set of solutions to problems with given constraints by combining a number of control mechanisms for the functioning of the technological system.

***Keywords*—optimal control, evolution of continuous technological systems, semantic systems, theory of fuzzy systems, artificial intelligence methodologies, multi-agent systems**

I. Introduction

Achieving target production efficiency level is made possible by developing technological processes through continuous control of main production metrics and generation of management decisions for technological systems evolution. This work presents a management model for fuzzy multi-stage technological system evolution, based on algorithmic support to managing fuzzy multi-stage technological system evolution. A method is proposed for forming a minimum set of solutions to problems with given constraints by combining a number of control mechanisms for the functioning of the technological system.

II. The task of managing the technological system evolution

Managing the system functioning can be represented as a multi-stage process. Figure 1 shows a formal model of the technological system evolution in the management process.

This work was supported by the Russian Foundation for Fundamental Research (RFFR) (projects 20-07-00199).

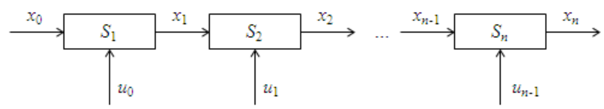


Figure 1. Example of a figure caption.

Let X be the set of states of the stages of the evolution of the system, U be the set of corresponding controls. Let $x_0 \in X$ be the state of the technological system at the input of the first stage of the system evolution process. As a result of using the control $u_0 \in U$ at the output of the first stage, a state is formed $x_1 \in X$ that is not known in advance. It is only known in advance that the variables x_0, u_0, x_1 are interconnected by a fuzzy relation S_1 with the membership function $\mu_{S_1}(x_0, u_0, x_1)$. Moreover, at the end of the functioning of the first stage, the actual state of the process x_1 is available for observation.

Similarly, if $x_{n-1} \in X$ is the state of the process at the input of stage n, $n = 1, \ldots, N$, where N is the number of evolution stages, then as a result of using the control $u_{n-1} \in U$ at the output of stage n, a state is formed $x_n \in X$. Variables x_{n-1}, u_{n-1}, x_n are interconnected by a fuzzy relation S_n with the membership function $\mu_{S_n}(x_{n-1}, u_{n-1}, x_n)$.

The task of the evolution technological system control is to find the control mechanisms $u_0, u_1, \ldots, u_{N-1}$ of the set U that maximize the goal achievement G, provided that the initial state x_0 is given.

The substantiation of this problem is considered in detail in [1].

III. Formation of the optimal control set of the technological system evolution

To manage the evolution technological system, it is necessary to process a huge amount of information to select the best option of solving various problems.

Let a finite set of solution options be the intersection of several sets, including the options obtained in the following ways:

- set $M1$ - expert information (external consulting services);

- set $M2$ - search for solutions in a distributed computing environment;
- set $M3$ - information generated in the process of applying motivational management;
- set $M4$ - the construction of a set of preferred states.

Briefly consider the methods of forming each sets.

A. Expert information (external consulting services)

Today consulting services of third-party organizations are quite common. The demand for services for the development of management systems, the introduction of new financial technologies, business valuation, and marketing research has increased. However, the results of the analysis, for example, in [2], showed a number of problems in the consulting services market, such as the lack of a clear pricing policy in consulting services; uniformity of consulting programs associated with an underdeveloped scientific and methodological base for the provision of services; the presence of a large number of non-adapted translation programs; the high cost of consulting services for small and medium enterprises. Of course, the enterprise management should recur to the use of external consulting services, however, this method cannot be used as the only one for solving complex problems of managing the technological system.

B. Search for solutions in a distributed computing

Distributed environments are focused on supporting research in domains united by thematic areas. To provide users with access to services and packages contained in distributed environments thematic information catalogs have been developed. In order to avoid restricting access to the full amount of information contained in the catalogs of scientific and computing services of distributed environments, it is advisable to submit a request to the system in a qualitative form (see [3]).

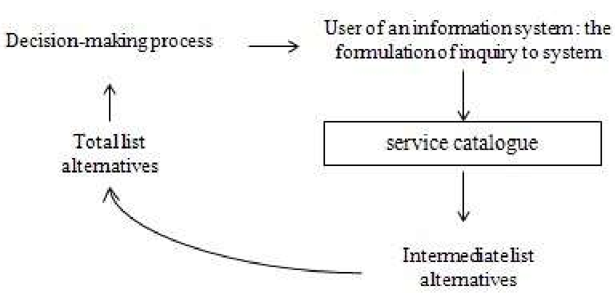


Figure 2. Multistep decision making, general view.

Briefly give a mechanism for the formation of many decisions in managing the technological system evolution (figure.2).

Let the set $X = \{x\}$ be the service catalog containing the services $\{x_1, x_2, \ldots, x_n\}$. The author, who includes information in the catalog, registers the service using a catalog card. Each $x_n \in X$ corresponds to a set of characterizing attributes:

$$x_j \left(a_1(x_j), a_2(x_j), \ldots, a_k(x_j)\right),$$

where x - services, $j = 1 \ldots n$, a - attributes, $k = 1 \ldots n$.

Each corresponds to the semantic service description, characterized by semantic units:

$$x_j \left((s_1(x_j), s_2(x_j), \ldots, s_k(x_j)\right),$$

where x - services, $j = 1 \ldots n$, s - semantic units, $k = 1 \ldots n$.

A user request to the system formulated in an informal form can be represented as follows:

$$L = L_1 \cup L_2 \cup L_3 \cup L_4$$

where L_1 is the set of terms from the search string, L_2 is the additional set of synonyms that are similar in meaning and words that are close in meaning; L_3 - an additional set of associative words and phrases; L_4 - a lot of translated keywords, depending on the settings for connecting dictionaries (dictionaries of automatic word processing are used when forming sets).

Denote the set of semantic descriptions generated on the basis of a user inquiry as

$$L = \{l_1(s_j), l_s(s_j), l_q(s_j), j = 1 \ldots m\}$$

When choosing the most suitable service for solving the problem, it is necessary to use a given set of parameters $p_j \in P$ evaluating the content of services and their functionality.

User sets priority service requirements

$$P = \{p_1(a_j), p_2(a_j), \ldots, p_r(a_j), j = 1 \ldots k\}$$

where p - parameters, a - attributes.

For simplicity, we denote the intersection of the fuzzy sets A (set of attributes) and P (set of parameters) as $Y = A \cap P$ the intersection of the sets S (set of semantic units) and L (set of users semantic units) as $H = S \cap L$.

To construct the membership function μ_Y of the set Y, we associate each $y_i(x_i)$ with a number $\mu_{y_i}(x_j)$ where $i = 1, \ldots k, j = 1, \ldots n$, $0 \leq \mu_{y_i}(x_j) \leq 1$

Similarly, to construct the membership function μ_X of the set X, we associate each $h_i(x_i)$ with a number $\mu_{h_i}(x_j)$ where $i = 1, ..., l, j = 1, \ldots, n$, $0 \leq \mu_{h_i}(x_j) \leq 1$

For $\mu_{y_i}(x_j)$ and $\mu_{h_i}(x_j) = 0$ - according to this description, the service does not fit exactly, with $\mu_{y_i}(x_j) = 1$ and $\mu_{h_i}(x_j) = 1$ - the service fits exactly, with $0 < \mu_{y_i}(x_j) < 1$ and $0 < \mu_{h_i}(x_j) < 1$ are intermediate variants.

Let us assume that a solution is a set of D services formed during the operation of a multi-step alternative search system that meets the conditions set in the request as much as possible. Let

$$\mu_y(x_n) = min\{\mu_{y_k}(x_n), \ldots, \mu_{y_k}(x_k)\}$$

and

$$\mu_h(x_n) = min\{\mu_{h_k}(x_n), \ldots, \mu_{h_k}(x_l)\}.$$

Following the Bellman-Zade formula [4], we will present the solution as a fusion of goals and limitations. Then the result of solving the problem is the set goal $D = Y \cap H$ coincides with the solution. Thus, to find a suitable scientific computing service from the set D, one selects the service for which the membership function will be the largest

$$\mu_D(X) = min\{\mu_Y(x), \mu_H(x)\}$$

Then the optimal solution comes down to finding

$$x^* = \arg\max_x \mu_D(x).$$

The software implementation of the proposed mechanism for the formation of many options for solving problems is presented in [5]

C. Information generated in the process of applying motivational management

Using the terminology of the theory of active systems [6], the procedure for the interaction of a center (enterprise manager) with agents (employees) as follows: first, the center learns from agents only an approximate description of their technological sets and finds a solution to the problem; the center asks the agents for evaluating the plan, clarifies their interests in the vicinity of the decision in exchange for incentives for awareness; Having received new information, the center carries out a review of the solution and poses new questions to the agents until an exact or close solution is received.

D. The construction of a set of preferred states

The analysis by the Z center of the efficiency of production functioning shows that the volume and quality of finished products directly depends not only on the level of modernization of technological equipment, the quality of raw materials, etc., but also on the degree of coordination of interests of the managing and controlled systems, the agent's target functions f and the center F. Competitive production requires high qualifications and training of employees, as well as motivation, interest and intellectualizing [7]. Specialists involved in projects to improve existing technological processes need to be aware of the significance and importance of the tasks assigned to them.

Since the specific form of the objective function of agent f and the composition of the sets P and X are not completely known, it is advisable to solve the problem of constructing preferred states by the center Z using "reducible" algorithms, that is, to solve the problem on the basis of solved locally optimal problems by agents a_k and obtaining from them for more information.

The formation of counter information with this approach is to implement a set of sequential procedures designed to search for intermediate solutions, on the basis of which the agent clarifies its capabilities and forms the final decision. The agent's full cycle of generating information about its capabilities consists of the following steps:

- step 1 - Agent a_k at the $r-th$ step receives from the center a version of the plan $v(r)$ and a control action $u_k(r)$. Based on this, the agent forms many P_k of its potential capabilities and x_k marginal technological capabilities. Here is the formation of a point

  $$y^{\omega*} = y(x), x^t(x_k(p)), p \in P_k$$

  determining the point of assessing the reliability of the result $\omega(y^{\omega*})$.
- step 2 - At this stage, the task of finding a potentially preferred set of actions is being solved

  $$x^* = x(p)$$

  Wherein $x^* \in C(p)$. If there is no such solution, you should try to find a compromise solution corresponding to the initial sets P_k and X_k and go to step 3; otherwise, go to step 5.
- step 3 - Here we analyze the directions of the possible expansion of the set P. By studying the properties of the situation and organizing the search for new information (knowledge) for

  $$P_k(r+1) \supseteq P_k(r)$$
- step 4 - If the extension of the set $P_k(r)$ is possible and there exists $P_k(r+1)$ such that inclusion (1) holds, then we go to step 1.
- step 5 - Based on the procedure for constructing the sets P_k+1 and X_k+1, the procedure for searching for the minimum-preferred point in the space of evaluating the value of the situation of a purposeful state is carried out.
- step 6 - The possibility of expanding the sets P and X is determined, the procedure is based on the search for additional information. In the case of a positive result, the go to step 1.

The implementation of the proposed mechanism for controlling the evolution of the system was considered in [8].

As a result of the formation of sets $M1$, $M2$, $M3$, $M4$, a single set M is formed, containing possible solutions to the problem when managing the evolution of the technological system (figure 3).

Then, by analogy with the formation of the set $M2$, we take into account the existing limitations and goals of the decision maker. As a result of the intersection of the set M, the set of goals and the set of restrictions, we get

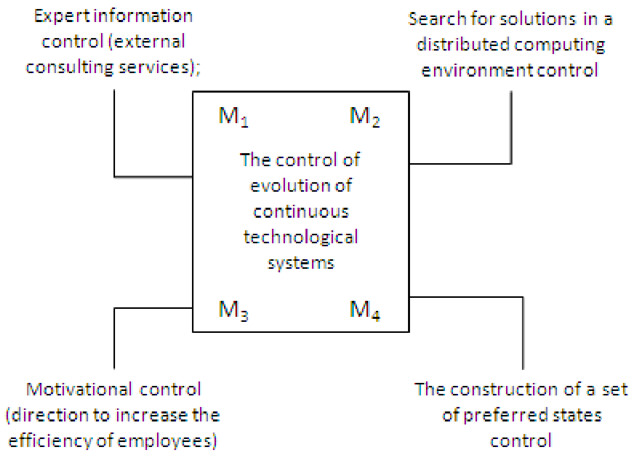


Figure 3. Optimal control set of the technological system evolution

the minimum possible set of solutions to the problems. This work was supported by the Russian Federal Property Fund (project 20-07-00199)

Conclusions

The paper proposes and considers the main ways of forming a multitude of options for solving problems arising in the process of functioning of a technological system. As a result of the application of the proposed mechanisms, management is carried out at various levels of the system, the proposed solutions of both enterprise managers and external consultants, employees of the enterprise, as well as information obtained in specialized thematic distributed systems and catalogs are taken into account. Most of the proposed mechanisms are implemented.

References

[1] A.N.Sotnikov, S.M. Dzyuba, I.A. Egereva, et al., "Control of The Final State of Fuzzy Dynamical Systems", Lobachevskii J Math (2019) 40: 599. https://doi.org/10.1134/S1995080219050202

[2] A.M. Magomedov, V.Yu. Kolyvanov. "Problems and features of the development of the Russian market of consulting services" // Regional problems of economic transformation. - 2009. - No. 4 (21).

[3] V.K. Ivanov "Features of organizing a data warehouse based on the intellectualization of a search agent and an evolutionary model for selecting targeted information". // Bulletin of the Tver State Technical University. Series: Engineering. 2019.No 1 (1). S. 75-84.

[4] R. Bellman, L. Zade "Decision making in vague conditions". Mir, Moscow: 1976, 46 p.

[5] B.V. Palyukh, I.A. Egereva. "Multistep system for searching alternatives in the information catalog" // Software products and systems. 2013. No3. S.291-295.

[6] D.A. Novikov D.A. "Theory of management of organizational systems". 3rd ed. - M .: Publishing house of physical and mathematical literature, 2012.

[7] G.P. Vinogradov "Individual decision making: behavior of a motivated agent". Scientific monograph. Tver: 2011.

[8] B.V. Palyukh, G.P. Vinogradov, I.A. Egereva. "Management of the evolution of a chemical-technological system" // Theoretical Foundations of Chemical Technology, 2014. - T. 48. - No. 3. - S. 349-355. [Paliukh B.V., Vinogradov G.P., Egereva I.A. Managing the Evolution of Chemical Engineering System // Theor. Found. Chem. Eng. 2014. Vol. 48, No. 3. - pp. 325-331].

[9] N.G. Andronikovs, S.V. Leont'ev, D.A. Novikov. "Mechanisms of Active Fuzzy Expertise" // Automation and Remote Control. 2002. Vol. 63, No. 8. P. 1322-1328.

[10] J. Ramik, R. Perzina "A method for solving fuzzy multicriteria decision problems with dependent criteria" // Fuzzy Optimization and Decision Making, Springer-Verlag, Vol.9, No. 2 (2010), P. 123-141.

[11] S. Ramezanzadeh, A. Heydari "Optimal control with fuzzy chance constraint" // Iranian Journal of Fuzzy Systems, 2011, vol. 8, pp. 35-43.

[12] M. Najariyan, H. Farahi "Optimal control of fuzzy linear controlled systems with fuzzy initial conditions" // Iranian Journal of Fuzzy Systems, 2013, vol. 10, pp. 21-35.

Механизм оптимального управления эволюцией непрерывных технологических систем

Палюх Б.В., Егерева И.А.

В связи с постоянно растущей конкуренцией в условиях непрерывного развития технологий промышленные предприятия вынуждены постоянно оптимизировать технологические процессы на производстве. Так как деятельность предприятий должна учитывать большое количество внешних факторов, учитывать различного рода ограничения (ресурсные, материальные и другие) при управлении эволюцией технологической системы возникает проблема выбора наиболее подходящего для конкретной ситуации варианта решения задачи. В данной работе приводятся механизмы оптимального управления непрерывными технологическими системами на основе применения теории активных систем, теории нечетких системы, методологий искусственного интеллекта, многоагентных систем. Предлагается способ формирования минимального множества вариантов решений задач с заданными ограничениями путем объединения ряда механизмов управления функционированием технологической системы.

Множества решений формируются на основе информации, получаемой путем привлечения на предприятие внешних консультантов, разносторонне исследующих функционирование системы и предлагающих различные варианты решений по оптимизации организационного управления. Другое множество представляет собой знания, получаемые в процессе взаимодействия руководителей предприятия непосредственно с сотрудниками. Следующее множество формируется на основе применения информации, полученной из тематических распределенных систем путем поиска и выбора оптимального решения задачи с последующей адаптацией.

Приведенные механизмы управления эволюцией непрерывных технологических систем более подробно рассмотрены в [1], [5], [8]. Предложенные алгоритмы использованы при разработке экспертных систем и информационных систем управления предприятием.

Received 15.01.20

The intelligent control concept of the transportation process on the Belarusian Railway

Alexander Erofeev
Belarusian State University of Transport
Gomel, Belarus

***Abstract*—directions for the development of transportation control systems have been established. The basic provisions of the concept of creating an intelligent system for managing the transportation process on the Belarusian Railway are described. The generalized objective function of the effectiveness of intelligent control is formulated. An end-to-end control loop is described. The main functions of the intelligent control system and stages of system implementation are determined.**

***Keywords*—transportation process, transportation organization, operational task, end-to-end management cycle, transportation resources, intelligent management**

Improving the efficiency of the functioning of the transport system is embedded in many strategic documents of the Republic of Belarus. For example, [1] provides for «Acceleration of scientific and technological progress and the scale of the digitalization of the economy». «. . . In the context of the exhaustion of factors of extensive production growth, the main accelerator of economic development is its digitalization. Digital platforms and artificial intelligence technologies gradually go beyond the industry, covering all spheres of life».

In accordance with [1], at the first stage, as a priority area of innovative development of transport, an increasing of the transport processes informatization level through the intelligent control introduction and monitoring systems is defined. The second stage 2026-2035 provides the accelerated development of the intellectual transport system, which will allow to reach the world level of development of the transport complex of the Republic of Belarus and increase its competitiveness.

The intelligent transport systems implementation is defined in more detail in [2], which provides the «creation of an intelligent transport system of the state integrated with the transport systems of the EU and the EAEU».

To implement the principles of intelligent management in railway transport, scientists at the Belarusian State University of Transport have developed a concept for introducing an Intelligent Transportation Control System (ITCS).

The goal of creating the ITCS is to form a unified transportation management business process that will increase the efficiency of railway transportation activity and the level of transportation safety. These goals are achieved by the using of information and communication and intelligent technologies in the operational management system.

Using ITCS allows you to use the accumulated array of experience in the formation of effective management decisions (MD), manage production processes in real time, plan work, simulate and predict the development of operational situations at the entire control range.

Intelligent transportation management aims to:

- implementation of integrated operational management and the use of a single information model by all participants in this activity. The information model includes a database, knowledge base, experience base and describes end-to-end production processes of all departments and all levels of management;
- formation of operational information services and technological interaction of participants in the transportation process. Technological services include daily and ongoing planning, implementation of agreed and approved plans;
- implementation of end-to-end automated control of the execution modes of technological processes for managing operational work;
- operational and process economic evaluation of the developed plans and technological processes of operational work;
- operational forecasting and cost estimation of unproductive losses of technological processes.

For all levels of management (network, road, linear) and operational tasks, unified principles are used to describe the problem environment and search for rational MD. These principles include:

1) The use of adaptive planning methods and scenario approaches, which in real time, taking into account the current situation, should ensure the development and adjustment of plans (both voluminous and detailed) for all business processes implemented at the training ground.
   Adaptive planning methods should be used for operational management, while scenario methods are used for annual and monthly planning. Technical norms (volume plans) formed using scenario methods should be the source data and target indicators for adaptive planning (detailed plans). The results of annual and monthly planning are stored in IAS PUR GP and can be used as target documents for other systems.
2) Automation of end-to-end business processes formed on the basis of a comprehensive digital model of the transportation process, which should be the basis for the design, implementation and operation of the ITCS. Using a digital model will allow us to close existing information gaps in management processes.
   To build a comprehensive digital model within the ITCS creation process specific measures must be provided to ensure the development and implementation of end-to-end management processes that take into account the specifics of the activities of various subdivisions of the warhead with the simultaneous alignment, regulation and proper information support of the processes of interaction

of the transportation process participants. An integrated digital model of the transportation process should be formed in the IAS PUR GP and used by all the functional subsystems of the ITCS in the development and implementation of MD.

3) The use of a single platform for the development of ITCS subsystems to achieve maximum synchronization and information security of automated business processes.

   At the same time, a single platform should provide:

   - Creation of a unified information environment for automation of MD development processes for the planning, implementation and control of end-to-end production processes, taking into account reliability, safety and risk assessment indicators;
   - integration and ensuring the interaction of ITCS and automated systems operating on warheads in a single information space;
   - the using of complete, consistent regulatory and operational reliable information in the ITCS.

   The implementation of a single platform should include an ontological description of the subject area, the construction of a conceptual model and a digital model of the transportation process, distributed architecture, intelligent planning methods and a number of other functions [3].

   The unified platform and functional subsystems of the ITCS form the planned and regulatory MDs for subsequent automatic implementation or for consideration by dispatch personnel.

   The technological subsystems of the ITCS ensure the implementation of SD at specific control facilities.

4) Technical and economic assessment. All the results of the functional subsystems of the ITCS should be accompanied by an economic assessment and analysis of the achieved technological effects in the implementation of the developed plans. In ITCS, a comprehensive assessment of MD should be implemented, including modules for the intellectual formation of assessment criteria and weighting criteria for their significance for various production situations [4].

5) The ITCS software should have an effective system of settings for further replication. Implementation at a new landfill should be carried out by adjusting the system using the appropriate settings and adding reference information, and not by developing subsystems again for each landfill.

The use of unified approaches to the description of the problem environment will provide the following advantages compared to “traditional” methods of solution:

- reduction of uncertainty in the source data through the use of harmonized source data;
- reduction of entropy in solving operational problems by using the results of solving some problems as source data for solving others;
- the ability to search for global extrema in assessing the effectiveness of the transportation process, and not local for each individual task;
- providing an objective assessment of the magnitude of the impact of the results of solving one operational problem on the development of control solutions to another problem.

The main priorities for the creation of the ITCS are:

- improving the quality of adopted MD and solving new operational problems on the basis of flexibly formed optimality criteria;
- optimization of the total costs of all participants in the transportation process by improving the quality of planning and monitoring the implementation of operational plans;
- improving the safety of the transportation process as an essential condition for the effective functioning of the ITCS;
- creation of ITCS as a modular adaptive system, which subsequently should be effectively integrated into the digital information space of the warhead, the transport system of the Republic of Belarus, transit transport corridors, and the EAEU.

The function of the functioning efficiency of the ITCS can be described in general terms by the expression

$$F_i(\overline{a_i}, \overline{b_i}, \overline{c_i}) = F_{1i}(\overline{a_{1i}}, \overline{b_{1i}}, \overline{c_{1i}}), \ldots, \\ F_{Wi}(\overline{a_{Wi}}, \overline{b_{Wi}}, \overline{c_{Wi}}), \ldots, \\ F_{Yi}(\overline{a_{Yi}}, \overline{b_{Yi}}, \overline{c_{Yi}}) \rightarrow max_{a_i \in \Omega},$$

where F_i – vector of optimality criteria characterizing the quality of adoption of MD at the i-th stage of decision-making; $i = 1, \ldots, j$ – number of decision stages; $(a_i) = (\overline{a_{1i}}, \overline{a_{ti}}, \overline{a_{\psi i}})$ – vector of technical parameters at the i-th stage of decision making; $(b_i) = (\overline{b_{1i}}, \overline{b_{ti}}, \overline{b_{\psi i}})$ – vector of technological parameters at the i-th stage of decision making; $(c_i) = (\overline{c_{1i}}, \overline{c_{ti}}, \overline{c_{\psi i}})$ – vector of unmanaged parameters at the i-th stage of decision making; $\psi = 1, \ldots, t$ – the number of subsystems of the ITCS that are involved in the development and implementation of MD; $y_i = 1, \ldots, w_i$ – number of optimality criteria at the i-th stage of MD development.

Criteria for the transportation process optimality include, for example, the car’s turnover, route speed, the need for transportation resources, the capacity of infrastructure elements, fulfillment of contractual obligations, labor productivity, the need for locomotives and locomotive crews, and the cost of fuel and energy resources. The list of criteria, depending on the operational task to be solved, can vary over a wide range, while the number of optimality criteria and their structure do not affect the efficiency of adoption of MD.

The set of options $\overline{a_i}, \overline{b_i}, \overline{c_i}$ can be represented as a combination of technical, technological parametric groups and uncontrolled environmental parameters.

The group of technical parameters include: for train work: the number of tracks on stages and intermediate stations, the number of train locomotives with details on series and types of work, the number of locomotive teams, etc. for station work: the number of technical and commercial maintenance teams in the subsystems of stations, the number of tracks in the parks of technical stations, the capacity of cargo fronts, etc.

The group of technological parameters include:

- for train work: train mass norms, train length norms; the presence of technological "windows" and their duration; type of train schedule. train formation plan, local cargo distribution schemes, etc.;
- for station work: sorting slide technology (with one, two or three locomotives); specialization of tracks in the sorting fleet, specialization of shunting locomotives, number of rides at local work stations, permissible deviations in length and mass of formed trains, etc.

To uncontrollable parameters: the technical condition of the car, weather conditions, etc.

Depending on the formulation of the control problem, individual groups of parameters can be considered both technological and uncontrollable. For example, the dimensions of the movement when developing a train schedule are technological, and when developing a station work plan, they are unmanageable. In the future, we will assume that the parameter refers to technical or technological (controllable), if for at least one ITCS subsystem it is. A parameter refers to uncontrollable parameters if no subsystem of the ITCS can form a control unit providing a change in the parameter.

Managed and unmanaged parameters can be described quantitatively (train mass rate), in the form of logical relationships (local cargo distribution scheme) or have a fuzzy description (flea runner, difficult weather conditions, etc.). In ITCS there must be mechanisms for describing and then comparing the parameters of various forms of presentation.

In ITCS a lot of controlled parameters $\overline{a_i}, \overline{b_i}$ are both initial data for the formation of rational MD, and the results of the functioning of the ITCS.

The set of uncontrollable parameters are subject to identification as describing environmental conditions, i.e.

$$F_i(\overline{a_i}, \overline{b_i}, \overline{c_i}) = F_{i+1}(\overline{a_{i+1}}, \overline{b_{i+1}}).$$

That is, at what environmental parameters the MD was obtained and how effective it was. The set $\overline{c_i}$ one of the components of the system knowledge base.

Thus, when modeling technological processes in the ITCS, among the admissible set Ω, it is necessary to find such control parameters and form such MD that will ensure the maximization of the target function F.

An end-to-end cycle of the transportation process management should be implemented at ITCS, including all levels (road, department, linear) and periods (long-term, medium-term, operational, current) of planning. The end-to-end cycle of the transportation process control includes modules (figure 1):

- infrastructure management;
- on-line digital model of the events of the transportation process;
- on-line modeling of processes;
- a module for the formation of control decisions (planners and problem solvers);
- module for assessing financial results and the formation of an array of experience;
- strategic planning.

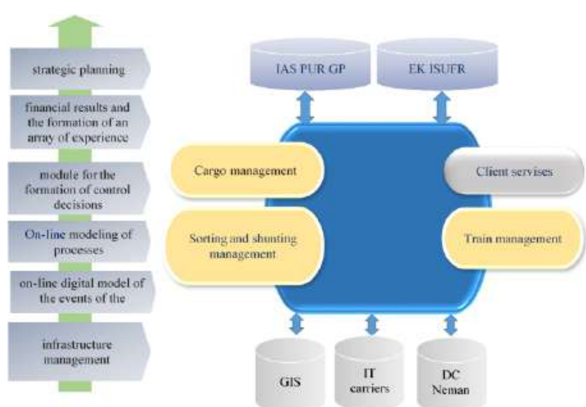


Figure 1. Diagram of the functional interaction of ITCS with external systems.

The functions of the ITCS are implemented through technological and functional subsystems.

Technological subsystems of control objects

A.1 Railway stations;

A.2 Dispatch sites (circles).

Functional Subsystems

B. Annual and monthly planning:
  - B.1 Applications from customers for the transport of goods and the development of a plan for the transport of goods;
  - B.2 Applications from carriers for the use of infrastructure services;
  - B.3 Approval of applications;
  - B.4 Development of a plan for the formation of trains;
  - B.5 Development of a regulatory train schedule;
  - B.6 Development and selection of a variant plan for the formation of trains;
  - B.7 Development and selection of a variant train schedule;
  - B.8 Technical regulation of operational work;
  - B.9 Calculation of the required operating fleet of wagons and locomotives.

C. Operational planning:
  - C.1 Operational planning of train and freight work:
  - C.2 Ongoing planning of train and freight work;
  - C.3 Determination of the current need for an operated fleet of wagons and locomotives;
  - C.4 Linking rolling stock to applications;
  - C.5 Linking compounding with a predicted train schedule. Garter locomotives and locomotive crews to trains.

D. Operational management and regulation:
  - D.1 Dispatching and traffic regulation;
  - D.2 Development of a forecast train schedule, the implementation of the "Auto dispatcher" function;
  - D.3 Implementation of the function "Auto";
  - D.4 Operations associated with the movement of loaded and empty wagons;
  - D.5 Shunting work at stations;
  - D.6 Loading, unloading and other initial and final operations;
  - D.7 Technical inspection and maintenance operations;
  - D.8 Commercial service operations;
  - D.9 Registration of electronic shipping documents;
  - D.10 Other operations with objects of transportation.

E. Control, accounting and analysis:
  - E.1 Formation of accounting and reporting data;
  - E.2 Analysis of the performed transportation.

The creation of the ITCS is a time-distributed process in which functional subsystems are developed in series and parallel. When designing subsystems, the possibility of their autonomous functioning is provided. In this regard, it is necessary to determine such conceptual approaches to the creation of the ITCS elements, which will allow to obtain a synergistic effect during aggregation and achieve the goals of creating the system.

The creation of the ITCS involves three main stages:

1) Creation of information and mathematical models of the transportation process on the basis of a unified road data transmission network, development and implementation of information and analytical systems. Integration of microprocessor systems, diagnostic and monitoring devices for infrastructure and rolling stock with information management systems. The introduction of GIS technology. Completion of work on the creation of automated planning systems for train and freight work. The goal of the first stage is to create a digital model of the railway.
2) Development and implementation of intelligent planning information systems focused on the operational

control center of the freight control center and line level facilities. Development of forecasting modules for the transportation process of the situation for 3, 4, 6, 12 hours, taking into account external factors for the transportation process. Development of technical and economic assessment modules for management decisions (plans) and their integration into existing MD. Target: implementation of management decision support systems and technical and economic assessment modules in existing MD.

3) Transition to intelligent forecasting, planning, management and decision support systems (automation of management, decision making automation to prevent difficulties in train work, optimization of regulatory measures), development of promising measures, development of the necessary technical, technological and regulatory documentation. The formation of an array of experience and training subsystems ITCS.

Target: the introduction of intelligent systems for the automatic control of the transportation process.

References

[1] *Kontseptsiya Natsional'noi strategii ustoichivogo razvitiya Respubliki Belarus' na period do 2035 goda* [Concept of the National Strategy for Sustainable Development of the Republic of Belarus for the period until 2035]. Available at: http://www.economy.gov.by/uploads/files/ObsugdaemNPA/Kontseptsija-na-sajt.pdf. (accessed 15.12.2019)

[2] *Strategii innovatsionnogo razvitiya transportnogo kompleksa Respubliki Belarus' do 2030 goda* [Strategies for the innovative development of the transport complex of the Republic of Belarus until 2030]. Available at: https://www.rw.by/corporate/press_center/reportings_interview_article/2015/03/strategija_innovacionnogo_razv/. (accessed 15.12.2019)

[3] Erofeev A.A. Semioticheskaya model' perevozochnogo protsessa i ee ispol'zovanie pri proektirovanii intellektual'nykh sistem [Semiotic model of the transportation process and its use in the design of intelligent systems]. *Intellektual'nye sistemy upravleniya na zheleznodorozhnom transporte. Komp'yuternoe i matematicheskoe modelirovanie [Intelligent railway management systems. Computer and mathematical modeling]*. Moscow, 2017, pp. 24-26. (In Russian)

[4] Erofeev A.A. Multi-criteria evaluation of management decisions in the intellectual system of transportation management. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2019, Iss. 3, pp. 205-208.

[5] Erofeev A.A. Sistema intellektual'nogo upravleniya perevozochnym protsessom: modeli povedeniya [Intelligent Transportation Control System: Behavior Models]. *Sbornik trudov Mezhdunarodnoi nauchno-prakticheskoi konferentsii «Transportnye sistemy: tendentsii razvitiya», 26–27 sentyabrya 2016 goda*. B.A. Levina Ed. M.: MIIT, 2016, pp. 380-383 (In Russian)

[6] Erofeev A.A. Predposylki sozdaniya intellektual'noi sistemy upravleniya perevozochnym protsessom [Prerequisites for the creation of an intelligent transportation process control system]. *Vestnik Belorusskogo gosudarstvennogo universiteta transporta: Nauka i transport*, 2017, Iss. 1 (34), pp. 42-45. (In Russian)

[7] Golenkov V.V. Ontology-based Design of Intelligent Systems. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2017, Iss. 1, pp. 37-56

[8] Shunkevich D.V Ontology-based Design of Knowledge Processing Machines. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2017, Iss. 1, pp. 73-94

[9] Vargunin V.I., Moskvichev O.V. *Informatsionnye tekhnologii i avtomatizirovannye sistemy upravleniya na zheleznodorozhnom transporte* [Information Technology and Automated Control Systems in Rail Transport] : uchebnoe posobie dlya vuzov zh.-d. transporta. Samara: Sam GAPS, 2007. 234 p. (In Russian)

[10] Erofeev A.A., Erofeeva H. Intelligent Management of the Railway Transportation Process: Object Model. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2017, Iss. 1, pp. 281-284

[11] Gapanovich V.A., Rozenberg I.N. Osnovnye napravleniya razvitiya intellektual'nogo zheleznodorozhnogo transporta [The main directions of development of intelligent railway transport]. *Zheleznodorozhnyi transport*, 2011, Iss. 4, pp. 5-11. (In Russian)

[12] Kozlov P.A., Osokin O.V. Postroenie avtomatizirovannykh analiticheskikh sistem na zheleznodorozhnom transporte [Construction of automated analytical systems in railway transport]. *Upravlenie bol'shimi sistemami: sb. trudov*, 2006, Iss. 12-13, pp. 78–89. (In Russian)

[13] Misharin A. S. *Effektivnoe funktsionirovanie zheleznodorozhnogo transporta na osnove informatsionnykh tekhnologii [Effective functioning of railway transport based on information technology]*, M.: VINITI RAN, 2007, 300 p. (In Russian)

[14] Gruntova P.S., eds.*Avtomatizirovannye dispetcherskie tsentry upravleniya ekspluatatsionnoi rabotoi zheleznykh dorog [Automated Dispatch Centers for Railway Operational Management]*, M.: Transport, 1990, 288 p. (In Russian)

[15] Borodin A.F., Kharitonov A.V., Prilepin E.V. «Poligon» – novaya avtomatizirovannaya sistema tekushchego planirovaniya [«Polygon» – a new automated system for current planning]. *Zheleznodorozhnyi transport* 2002, Iss. 4, pp. 10–16. (In Russian)

Система интеллектуального управления перевозочным процессом на Белорусской железной дороге

Ерофеев А. А.

Установлены направления развития систем управления перевозочным процессом. Описаны основные положения концепции создания интеллектуальной системы управления перевозочным процессом на Белорусской железной дороге. Сформулирована обобщенная целевая функция эффективности интеллектуального управления. Описан сквозной цикл управления. Определены основные функции системы интеллектуального управления. Определены этапы внедрения системы.

Received 12.01.2020

ADAPTIVE CONTROL OF ROBOTIC PRODUCTION SYSTEMS

Viktor Smorodin
Department of Mathematical Problems
of Control and Informatics
Francisk Skorina Gomel State University
Gomel, Belarus
Email: smorodin@gsu.by

Vladislav Prokhorenko
Department of Mathematical Problems
of Control and Informatics
Francisk Skorina Gomel State University
Gomel, Belarus
Email: snysc@mail.ru

***Abstract*—The purpose of the work, that is presented in this paper, is to develop a method for adaptive control of a technological production cycle based on a software and hardware system that includes indicators of the hardware units states, parameters of the technological production cycle operation, simulation model of the probabilistic technological process and a built-in decision-making system. Operational interaction of the software and hardware system components and construction of the feedback control connections is implemented through the control parameters and variables of the simulation model based on the output of the neuroregulator model. To address the described problem, tasks related to implementation of the neural network technologies when constructing architecture and mathematical model of the neuroregulator were solved. The mathematical model of the neuroregulator is based on parameters of operation of the physical prototype, construction of the feedback connections for the real-time control (adaptive control) is based on the procedure of training of a recurrent neural network that includes LSTM-cells. Considering the testing results is was found out that recurrent neural networks with LSTM-cells can be successfully used as an approximator of the Q-function for the given problem in the conditions when the observable region of the system states has complex structure.**

***Index Terms*—technological production process, parameters of operation, probabilistic network chart, state indicators, methods of adaptive control, neural network, LSTM, Q-learning**

I. INTRODUCTION

The modern analysis of the research in the field of controlled production systems demonstrates that the problem of determining the parameters of real-time operation for such objects of study arises primarily in the cases when they involve production of complex technical items, which require precision of production methods and high labor productivity.

During that process a multi-criteria control optimization problem is considered. It sets strict requirements for quality and algorithmic execution of the production process, minimization of the human factor effects on the implementation quality of the technological production cycle, prevention of the occurrence of technogenic emergencies. Such situation is specific for the robotic production systems, that operate under control of software and hardware controllers, which administer the operation of the technological cycle control system in accordance with the implemented programs.

However, the occurrence of emergency situations due to hardware failures, random external control influences, including the human factor, leads to deviation of the operation parameters of the production system from the intended values. This leads to the necessity of adjustment of the control parameters in real-time, based on the neuroregulator models that operate within the means of software and hardware pairing with the technological production cycle.

The formalization of the control structure for technological production cycle is based on the research results in the field of the analysis of operation of probabilistic technological systems, that include a technological production process as a controlled object. The technological production process is characterized by a relatively slow speed of performance of technological operations, which are interconnected within the cycle, and it's structure is defined by a simulation model of the probabilistic network chart[1][2].

Special AI models such as artificial neural networks (ANNs) have unique qualities, can be used as universal approximators[3], and were able to produce in the recent years a variety of impressive results in different kinds of control tasks[4][5][6], including complex decision-making tasks[7][8], and being able to reach human level performance in some of them[9][10].

While the idea of applying ANNs to the control problems is not new[11][12][13][14], some of the recent interesting developments in the field are related to implementation of deep reinforcement learning methods[5]. The possibilities of application of such methods to the considered problem of constructing an effective controller model for the technological production process are explored in this paper.

II. PARAMETERS, STATISTICS AND RESPONSES OF THE CONTROL OF TECHNOLOGICAL PRODUCTION PROCESS

Operational interaction of the system components and the technological production cycle that operates in real-time is carried out based on constant monitoring of the hardware unit states and the control parameters through the registers-indicators and special technical pairing means.

The system of software and hardware pairing for the operational control for the technological production cycle consists of a simulation model of the technological production process, which has a structure defined by a probabilistic network chart; a specialized decision making system, which performs analysis and control of the current developments in the operating environment of the technological production cycle; a control block, that is used as a mediator between the decision making system and the neuroregulator model. The scheme of the system is shown in Fig.1.

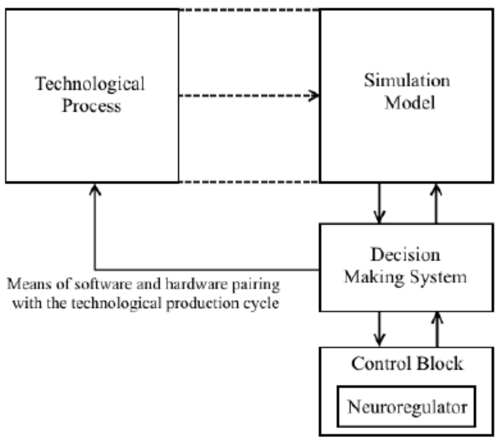


Figure 1. Scheme of the software and hardware system for the operational control in the described formalization.

Operational interaction of the system components is implemented through the parameters and variables of the simulation model of the control system: reliability characteristics for the hardware units operation G^*_{rh}; indicators of the current hardware units states ind^*_{rh}, that accumulate each unit's time before failure; indicators of the emergency occurrences due to hardware failure π^*_{abrh}, that have effect on the configuration of the technological cycle and the control system; current values U^*_{fh} of the variables for the control of the technological process; values of the control variables adjustments ΔU^*_{fh} that are based on the output of the neuroregulator model; parameters of the operational conditional of the technological process Z^*_{fh}, that depend on the values of the control variables U^*_{fh}. The decision making system sends three types of signals to the hardware and the control registers of the technological process: corrective adjustments ΔU^*_{fh} for the control variables of the technological production process; signals α^*_{rh} that allow to switch to the backup hardware or signals to initiate maintenance or repairs for hardware units.

III. Features of the technological process control structure based on a neuroregulator model

In this section the technological and operational control of the simulation model is described, as well as the features of the control structure. For the implementation of the simulation model in the described formalization, the following parameters are used:

- number of hardware units N;
- list of the K operation stages of the technological process, that consists of pairs $(t_k, \{R(M_{ik}\})$, where t_k is duration of the stage, $R(M_{ik})$ - acceptable mode of operation for the i-th unit at this stage;
- maximum number of the possible simultaneous maintenance jobs for the hardware units S;
- reliability characteristics for the hardware units operation G^*_{rh}, which include:
  1) distributions for the duration of operation till failure for the i-th hardware unit $F_{ir}(t_{wf})$ in the operation mode $R(M_i)$;
  2) distributions for the duration of the required maintenance (repairs) of the i-th hardware unit in the event of failure $F_{if}(t_r)$;
  3) distributions for the duration of the liquidation of the emergency due to failure of the i-th hardware unit $F_{ife}(t_{re})$;
  4) probabilities of the emergency occurrence due to failure of the i-th hardware unit P_{ie}.

The simulation model operates during the given time interval, restarting the production cycle and carrying on before each start the maintenance actions if necessary. In order to make a decision if such actions are necessary, a neuroregulator model is used, which operates based on the current environment state data.

IV. Procedure of neuroregulator generation

When solving this task a group of reinforcement learning algorithms, and particularly Q-learning, is interesting to consider as a basis for the method to determine the optimal control strategy that has feedback connections and takes into account the complex structure of the space of available actions.

Reinforcement learning algorithms assume implementation of an agent capable of observing the environment and manipulating it with actions, as well as the mechanism of numerical evaluation of agent's performance - the reward function[15]. When training a controller for the agent during the reinforcement learning process, an exploration process is involved in order to determine the agent's actions that lead to the highest rewards. Such approaches have a potential to train a controller capable of implementing the optimal strategy.

In this paper reinforcement learning (Q-learning) is used to train a neuroregulator capable of forming a set of maintenance recommendations for the technological production cycle at each point of time.

During the model training the values of the reward function are calculated according to the following rules during each "maintenance-cycle" procedure:

- $r = 0$
- $r = r - REW\_REPAIR\_COST$ – penalty for each maintenance (repairs), initiated due to the agent's decision during the maintenance stage of the given iteration;
- $r = r - REW\_FAILURE\_COST$ – penalty for each hardware unit failure that happened during the cycle stage of the given iteration;

In the case of emergency that happened as a result of a hardware unit failure:

- $r = r - REW\_EMERGENCY\_COST$ – penalty for the emergency that as caused by one of the hardware units failure during the cycle stage of the given iteration.

It is presumed that the penalty value REW_REPAIR_COST is significantly less than the values for $REW\_FAILURE\_COST$ and $REW\_EMERGENCY\_COST$, so that the action of the hardware unit maintenance is more preferable than the situation when the unit fails during the cycle, causing the cycle to stop and possibly causing an emergency. On the other hand, this value is non-zero because the maintenance procedure consumes resources and time and should be used only when necessary.

The choice of the specific numerical values for the parameters of the reward function with regard to the practical requirements for the reliability and cost reduction will have effect on the action selection policy of the agent learned during training. For example shaping the reward function with usage of the calculated cost values may result in agent learning to prefer lower maintenance costs at the expense of higher failure rates.

The idea of Q-learning is to train the agent's controller to approximate Q^* - the function that estimates reward values for the next environment state as a result of the chosen control actions[16][17]. A neural network can be used as an approximator for that function. In this case the training task is to find such values of the trainable parameters of the network, so that the approximate function Q is close enough to the optimal function Q^* and therefore the determined policy of action selection π is close enough to being optimal[9]:

$$Q(s, a) \approx Q^*(s, a) = \max_{\pi} E\left[R_t | s_t = s, a_t = a\right] \quad (1)$$

And Bellman equation is true for Q^*:

$$Q^*(s, a) = E(r + \gamma \max_{a'} Q^*(s', a') | s, a) \quad (2)$$

When solving complex real-world problems, the agent often has the whole information about the environment unavailable. In this case the agent that operates based upon a Q approximator calculating values depending only on the current observable environment state may be inefficient when the environment structure has high complexity or is of temporal nature and involves processes, both dependent and independent of the agent's actions. In the described situation it is necessary to use some mechanism of memory for the agent. A recurrent neural network with a hidden state preserved across multiple iterations can be used for such purpose[7][18]. LSTM-cells as a structural component of the neural network architecture allow the network to approximate temporal dependencies stretched over long periods of time[19].

Therefore the neural network model choice for the problem in the described formalization:

DQRN1 – agent that uses a recurrent neural network based on the multi-layer perceptron architecture with LSTM-cells. The current environment state is feed to the network input.

Neural network structure:

1) Dense x16 ReLU;
2) Dense x16 ReLU;
3) Dense x32 ReLU;
4) LSTM x32 ReLU;
5) Dense x6 no activation.

The input vector of the network is formed at each point of time based on values of the indicators ind^*_{rh} of the current state of the hardware units.

The output vector of the network has dimension $N + 1$. The values of the elements of the output vector define a set of maintenance recommendations for the N hardware units. In case value of an element exceeds some given threshold level, it is considered that the model recommends to perform maintenance of the corresponding hardware unit. In case the last element's value exceeds the threshold, it is considered that the model recommends not to perform any maintenance at this time.

Based on the values of the elements of the output vector of the model, the signals for performing the maintenance (repairs) for the corresponding hardware units α^*_{rh} are determined.

To implement model training based on Q-learning algorithm, a custom environment was developed that includes a simulation model of a technological process in the given formalization. The environment was written in Python language, using simpy library for simulations, tensorflow and keras libraries for the neurocontroller model.

In this paper experience replay was used during training[9] - the biologically inspired mechanism to select randomly and demonstrate during training the previous experience of agent's interactions with the environment. The updates of the Q values were performed after each run of a simulation based on all the saved experiences. The neural network is trained on sequences of 16 timesteps.

Training procedure:

1) Agent acts during a single simulation run (multiple runs of the production cycle, determined by a set of K stages during the given period of model time). The agent receives the current observable state of the environment and acts according to the chosen exploration strategy that determines how the random actions are selected. In this paper the softmax exploration strategy was used[15]:
$$p_t(\alpha^*) = \frac{exp(Q_t(\alpha^*)/\tau)}{\sum_i exp(Q_t(\alpha_i)/\tau)} \quad (3)$$
2) Experiences of agent's interactions with the environment are saved as sequences $\{(s_t; a_t; r_t; s_{t+1})\}$ to memory.
3) The experiences to be used for the next update of the trainable parameters of the neural network are sampled from the memory. According to the selected training parameters, these experiences are sampled party from the whole memory randomly, and partly from the latest sequences.
4) The updated Q values are calculated based on the following observable states and rewards from the experiences:
$$Q(s, a^*) = r + \gamma \max_{a} Q(s, a) \quad (4)$$

5) The neural network is trained on a set of size 32 for 1 epoch using the RMSprop algorithm[20]. The trainable parameters are updated after presenting each set of 4 sequences.
6) Go back to 1) and restart the simulation.

The training continues until the simulation number limit is reached.

V. Parameters of the simulation model used in the experiments

One cycle run under normal condition lasts for 48 ticks of simulation time. One simulation run lasts for 64*48 = 3072 ticks of simulation time.

Number of hardware units:

$N = 5$

Modes of operation $R(M_i)$:

operating/not operating – for all hardware units.

Number of maintenance jobs that can be performed simultaneously:

$S = 3$

$K = 2$ cycle stages:

$\{(4, (1, 0, 0, 0, 0)), (44, (1, 1, 1, 1, 1))\}$

First stage - operation of only the first hardware unit during 4 ticks of simulation time. The second stage - all of the units operate for 44 ticks of simulation time.

Costs:

- $CYCLE\_NON\_OPERATING\_COST = 20$ – non-operation of a production process for one simulation tick;
- $FAILURE\_DURING\_CYCLE\_COST = 150$ – hardware unit failure during the operation of the cycle;
- $EMERGENCY\_DURING\_CYCLE\_COST = 1000$ – emergency during the operation of the cycle caused by a hardware unit failure.

Distributions and probabilities:

- distributions for the operation time before failure of the i-th hardware unit$F_{ir}(t_{wf})$:
  1) uniform $U(150; 200)$;
  2) normal $N(\mu = 1000, \sigma^2 = 250)$;
  3) normal $N(\mu = 350, \sigma^2 = 50)$;
  4) uniform $U(300; 500)$;
  5) normal $N(\mu = 800, \sigma^2 = 300)$.
- distributions for the duration of the required maintenance (repairs) of the i-th hardware unit in the event of failure $F_{if}(t_r)$:
  1) uniform $U(3; 5)$;
  2) uniform $U(2; 4)$;
  3) uniform $U(3; 6)$;
  4) uniform $U(5; 10)$;
  5) uniform $U(2; 3)$.
- probabilities of the emergency occurrence due to failure of the i-th hardware unit P_{ie}: $\{0.05; 0.25; 0.1; 0.01; 0.2\}$
- distributions for the duration of the liquidation of the emergency due to failure of the i-th hardware unit $F_{ife}(t_{re})$:
  1) uniform $U(10; 15)$;
  2) uniform $U(8; 10)$;
  3) uniform $U(5; 8)$;
  4) uniform $U(25; 35)$;
  5) uniform $U(4; 6)$.

Values for the reward parameters:

1) $REW\_FAILURE\_PENALTY = 100.0$
2) $REW\_EMERGENCY\_PENALTY = 200.0$
3) $REW\_REPAIR\_PENALTY = 10.0$

VI. Training results

The model was trained as described above during the 2000 simulation runs.

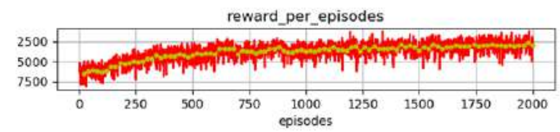


Figure 2. Total agent's reward for one simulation during training.

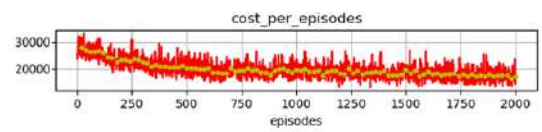


Figure 3. Total costs for running the technological cycle for one simulation during training

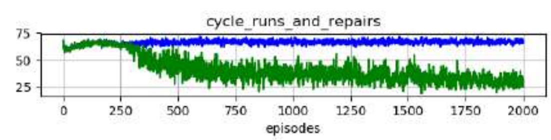


Figure 4. Total number of repairs (in green) and cycle restarts (in blue) for one simulation during training.

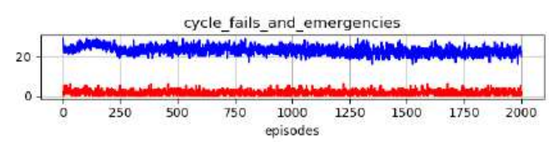


Figure 5. Total number of hardware failures (in blue) and emergencies (red) for one simulation during training.

According to the statistics gathered during training the following can be noticed:

- increase in the agent's reward (Fig.2);
- tendency towards the decrease in costs for running the cycle (Fig.3);
- decrease in the number of repairs for the hardware units (Fig.4);
- tendency towards the decrease in the number of hardware failures (Fig.5);
- increase in the average operation time for the cycle after start (Fig.6);

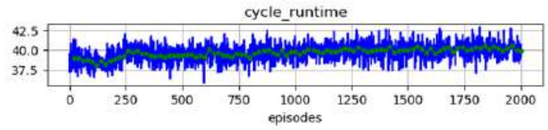


Figure 6. Average cycle operation time for one simulation during training.

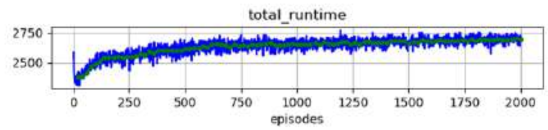


Figure 7. Total cycle operation time for one simulation during training.

- increase in the total cycle operation time (Fig.7).

It is interesting to compare the performance of the model by comparing it to some baseline models that do not implement complex action selection policies like the trained model does. The examples of such models can be a controller that never recommends maintaining the hardware units (Baseline-Zero) and a controller that implements random action selection policy (Baseline-Random). On Fig.8 and Fig.9 histograms for the distributions of costs and durations of normal cycle operation are given for 5000 test runs of the simulation.

For Baseline-Zero median costs value is 19455.82, median operation time - 2686.46. For Baseline-Random median costs value is 18002.09 and median operation time is 2737.25.

The same histograms for DQRN1 agent are shown on Fig.10. For this model the median costs value is 17115.98 and median operation time is 2697.59.

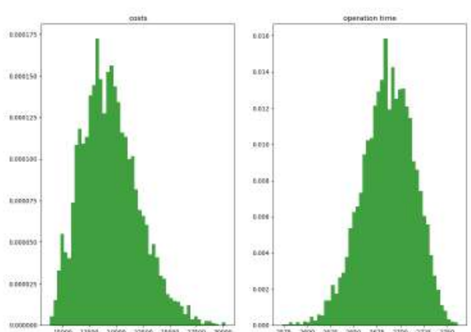


Figure 8. Histograms for the distributions of costs and cycle normal operation time for the controller Baseline-Zero during testing.

It is particularly effective to use the software and hardware system in the cases when time intervals τ_{SOB} between the emergencies in the slowly operating technological process are large enough to allow the operative control ($\tau_{SOB} > T_{kph}$, where T_{kph} - critical time of realization of the process, that was estimated using the simulation model).

Another important purpose of the system is realization of multiple series of simulation experiments implementing Monte-Carlo methods in order to determine how the global

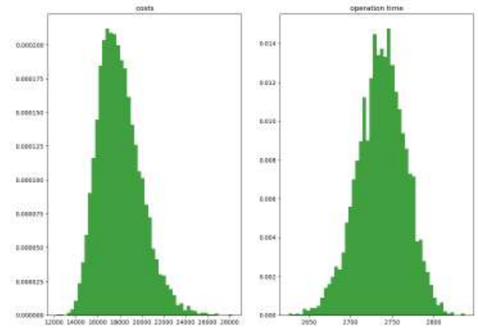


Figure 9. Histograms for the distributions of costs and cycle normal operation time for the controller Baseline-Random during testing.

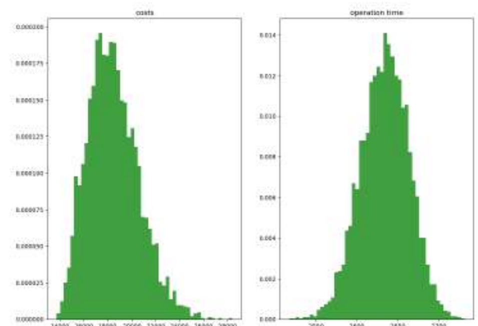


Figure 10. Histograms for the distributions of costs and cycle normal operation time for the controller DQRN1 during testing.

variables $Z_{fh}(t_0)$ and $U_{fh}(t_0)$ change with model time. Those dependencies can be then used to compare the model values to the real values: $Z_{fh}(t_0)$ to $Z^*_{fh}(t)$ and $U_{fh}(t_0)$ to $U^*_{fh}(t)$, where t_0 and t - respectively moments of model time and moments of real time in the technological production cycle.

In case when for all the elements of these vectors the absolute values of difference are within an acceptable error margin (5)(6), the simulation model is considered to be adequate in the dynamics of implementation of the control for the technological production cycle using the neuroregulator model.

$$\left|Z_{fh}(t_0) - Z^*_{fh}(t)\right| < \delta \tag{5}$$

$$\left|U_{fh}(t_0) - U^*_{fh}(t)\right| < \delta \tag{6}$$

VII. Conclusion

The proposed technology of adaptive control for the technological cycle using the neuroregulator models allows:

- to chose a rational set of resources for the technological cycle as well as a set of hardware units that will allow a required level of hardware reliability for the normal operation of the production process without emergencies;

- operatively adjust the characteristics of realization of the control process to reach the performance of the technological cycle within the acceptable margins of change for the technological parameters;
- to address the problem of estimating the costs of implementing the technological cycle with the given set of resources and hardware available at the facility.

References

[1] Maximey, I.V., Smorodin, V.S., Demidenko, O.M. *Development of simulation models of complex technical systems*, Gomel, F. Skorina State University, 2014. 298 p.

[2] Maximey, I.V., Demidenko, O.M., Smorodin, V.S. *Problems of theory and practice of modeling complex systems*, Gomel, F. Skorina State University, 2015. 263 p.

[3] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 1989, vol. 2, no 4, pp. 303–314/

[4] Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K. End to End Learning for Self-Driving Cars. Available at: http://arxiv.org/abs/1604.07316 (accessed 2019, Nov)/

[5] Francois-Lavet, V. et al.: An Introduction to Deep Reinforcement Learning. Available at: http://arxiv.org/abs/1811.12560 (accessed 2019, Nov)/

[6] Levine, S., Finn, C., Darrell, T., Abbeel, P. End-to-End Training of Deep Visuomotor Policies. *J. Mach. Learn. Res.*, 2015, vol. 17, pp.39:1-39:40/

[7] Hausknecht M., Stone P. Deep recurrent q-learning for partially observable mdps. *2015 AAAI Fall Symposium Series*, 2015./

[8] Lample, G., Chaplot, D. S. Playing FPS Games with Deep Reinforcement Learning. *AAAI'17 Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp.2140-2146/

[9] Mnih V., Kavukcuoglu K., Silver D., Rusu A., Veness J., Bellemare M., Graves A., Riedmiller M., Fidjeland A., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wiestra D., Legg S., Hassabis D. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, no 7540, pp.529–533/

[10] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Driessche, G.V., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T.P., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, vol. 529, pp.484-489/

[11] Hagan, M.T., Demuth, H.B. Neural networks for control. *Proceedings of the American Control Conference*, 1999, vol. 3, pp. 1642—1656/

[12] Narendra, K.S., Mukhopadhyay, S. Adaptive Control Using Neural Networks and Approximate Models. *IEEE Transactions on Neural Networks*, 1997, vol. 8, pp. 475-485/

[13] Narendra, K.S., Parthasarathy, K. Identification and Control of Dynamical Systems Using Neural Networks. *IEEE Transactions on Neural Networks*, 1990, vol. 1, pp. 4-27/

[14] Lin, L.-J. *Reinforcement learning for robots using neural networks*, Pittsburgh, Carnegie Mellon University, 1992./

[15] Sutton, R. S., Barto, A. G. *Reinforcement Learning: An Introduction*, Cambridge, The MIT Press, 1998./

[16] Watkins, C. J. C. H., and Dayan, P. Q-learning. *Machine Learning*, 1992, vol. 8, no 3-4, pp. 279–292/

[17] Watkins, C. J. C. H. *Learning from Delayed Rewards*, Cambridge, University of Cambridge, 1989/

[18] Bakker, B. Reinforcement Learning with Long Short-Term Memory, *NIPS*, 2001/

[19] Hochreiter, S., Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997, vol. 9, no 8, pp.1735–1780/

[20] Graves A. Generating Sequences With Recurrent Neural Networks. Available at: https://arxiv.org/abs/1308.0850 (accessed 2019, Nov)/

УПРАВЛЕНИЕ ТЕХНОЛОГИЧЕСКИМ ЦИКЛОМ ПРОИЗВОДСТВА НА ОСНОВЕ МОДЕЛИ НЕЙРОКОНТРОЛЛЕРА

Смородин В.С. - ГГУ им. Ф.Скорины, кафедра математических проблем управления и информатики

Прохоренко В.А. - ГГУ им. Ф.Скорины, кафедра математических проблем управления и информатики

Цель работы, результаты которой представлены в рамках данной статьи, состояла в разработке методики адаптивного управления технологическим циклом производства на базе программно-аппаратной системы, содержащей индикаторы состояния оборудования, параметры функционирования технологического цикла, имитационную модель вероятностного технологического процесса и встроенную систему принятия решений. Оперативное взаимодействие компонентов программно-аппаратной системы и построение обратных связей по управлению реализуется с помощью параметров управления и переменных имитационной модели на основе результатов работы модели нейрорегулятора. Для достижения поставленной цели были решены задачи, связанные с применением нейросетевых технологий при построении архитектуры и математической модели нейрорегулятора. При этом математическая модель нейрорегулятора разработана на основе параметров функционирования физического прототипа, а построение обратных связей по управлению в режиме реального времени (адаптивного управления) основано на процедуре обучения рекуррентной нейронной сети, построенной с использованием LSTM-блоков. С учетом полученных результатов установлено, что рекуррентные сети с LSTM-модулями могут успешно применяться в качестве аппроксиматора Q-функции агента для решения поставленной задачи в условиях, когда наблюдаемая область состояний системы имеет сложную структуру.

Principles for enhancing the development and use of standards within Industry 4.0

Valery Taberko
JSC «Savushkin Product»
Brest, Republic of Belarus
tab@pda.savushkin.by

Daniil Shunkevich
Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
shunkevich@bsuir.by

Dmitry Ivaniuk
JSC «Savushkin Product»
Brest, Republic of Belarus
id@pda.savushkin.by

Oleksandr Pupena
National University of Food Technologies
Kyiv, Ukraine
pupena_san@ukr.net

***Abstract*—In this paper, we propose an approach to automating the processes of creating, developing and applying standards based on OSTIS Technology. The problems of modern approaches to the organization of these processes are considered, the role of standards in the framework of the Industry 4.0 concept is shown, examples of formalization of standards within the framework of the proposed approach are given.**

***Keywords*—Standards, Ontologies, Industry 4.0, OSTIS, ISA-88**

I. Introduction

Each developed sphere of human activity is based on a number of standards formally describing its various aspects – a system of concepts (including terminology), a typology and sequence of actions performed during the application process of appropriate methods and means, a model of a production facility, type and composition of project documentation accompanying activities and more.

The presence of standards allows us to solve one of the key problems that is relevant for any technology, especially for rapidly developing computer information technologies, the **compatibility problem** [1]. Compatibility can be considered in various aspects, ranging from the consistency of terms in the interaction of process participants and ending with the consistency of actions performed in the process the technology application, which, in turn, ensures that the result coincides when the same actions performed by different performers.

Despite the development of information technology, at present the vast majority of standards are presented either in the form of traditional linear documents, or in the form of web resources containing a set of static pages linked by hyperlinks. This approach to the presentation of standards has a number of significant drawbacks, which finally lead to the fact that the overhead costs of maintaining and applying the standard actually exceed the benefits of its application [2].

II. Problems and proposed approach

Let's list the most important and general problems associated with the development and application of modern standards in various fields [2], [3]:

- duplication of information within the document describing the standard;
- the complexity of maintaining the standard itself, due, among other things, to the duplication of information, in particular, the complexity of terminology changing;
- the problem of internationalization of the standard – in fact, translating the standard into several languages makes it necessary to support and coordinate independent versions of the standard in different languages;
- inconvenience of applying the standard, in particular, the complexity of the search for the necessary information. As a result, the complexity of studying the standard;
- inconsistency of the form of various standards among themselves, as a result – the complexity of the automation of processes of standards development and application;
- the complexity of the automation of checking the conformity of objects or processes with the requirements of a particular standard;
- and others.

As an example, consider the standard ***ISA 88*** [4] – the fundamental standard for batch production. The standard is widely used at enterprises in America and Europe, it is actively implemented in the territory of the Republic of Belarus, however, it has a number of disadvantages due to the above problems:

- the American version of the standard – *ANSI/ISA-88.00.01-2010* – is already updated, the third edition from 2010;
- at the same time, the European version adopted in 1997 – *IEC 61512-1* – is based on the old version *ISA-88.01-1995*;
- Russian version of the standard – *GOST R IEC 61512-1-2016* – is identical to *IEC 61512-1*, that is, it is also outdated;
- Russian version of the standard - *GOST R IEC 61512-1-2016* - also raises a number of questions related to the not very successful translation of the original English terms into Russian;
- when translating, meaning is lost due to poorly written text, failed translation, difficulty in understanding;
- and many others.

Another popular standard in the context of industry 4.0 is ***PackML*** (Packaging Machine Language) [5]. ***PackML*** is an industry standard for describing control systems for filling machines. Its main purpose is to simplify the development of such systems, abstract from hardware implementation and provide a single interface for interaction with the SCADA and MES levels. This standard defines both software elements (software modules for industrial controllers) and general principles for the development of packaging lines.

These and other standards are currently distributed in the form of documents that are inconvenient for automatic processing and, as indicated above, are highly dependent on the language in which each document is written.

These problems are mainly related to the presentation of standards. The most promising approach to solve these problems is the transformation of each specific standard into a knowledge base, which is based on a set of ontologies corresponding to this standard [1]–[3]. This approach allows us to significantly automate the development processes of the standard and its application.

The task of any standard in the general case is to describe a agreed system of concepts (and related terms), business processes, rules and other laws, methods for solving certain classes of problems, etc. *ontologies* are successfully used to formally describe this kind of information. Moreover, at present, in a number of areas, instead of developing a standard in the form of a traditional document, the corresponding ontology is being developed [6]–[9]. This approach provides obvious advantages in terms of automating the standards agreeing and application processes.

However, the problem that remains relevant is not with the form, but with the essence (semantics) of standards – the problem of inconsistency of the system of concepts and terms between different standards, which is relevant even for standards within the same field of activity.

To solve this problem and the problems listed above, it is proposed to use OSTIS Technology, one of the key tasks of which is to solve the problem of syntactic and semantic compatibility of computer systems [1], in particular, the compatibility of various types of knowledge [10] and various problem solving models [11].

As part of this work, we will consider the experience of using this technology in building an information and reference system for the standard for batch production ISA-88 [4] together with employees of JSC "Savushkin Product". Initially, this system was considered as a reference and training system for employees of the enterprise, however, it is currently being transformed into an international open-source project, the purpose of which is to create an up-to-date, constantly developed and user-friendly knowledge base that describes the specified standard. Work [12] considered the initial version of the formal description of the standard in question using the OSTIS Technology, as well as fragments of the formal description of the departments of the enterprise taking into account the requirements of the standard.

Subsequently, part of the descriptions presented, not directly related to the enterprise, was allocated to the above-mentioned separate system dedicated to the ISA-88 standard. The current version of the system is available online at http://ostis.savushkin.by.

III. ISA-88 STANDARD

The ***ISA-88*** (short for ANSI/ISA-88) standard is based on the previously developed standard NAMUR N33 and helps in solving several fundamental problems, such as the lack of a single model of batch production, the difficulty of requirements agreeing, the complexity of integrating solutions from various suppliers, the complexity of the management of batch production. To solve these problems, it was necessary to define common models, terminology, data structure and process description language. The structure of the standard corresponds to the mentioned tasks and includes four parts:

- ANSI/ISA-88.00.01-2010, Batch Control Part 1: Models and Terminology – defines standard models and terminology for formalizing requirements for batch production control systems, its equivalent is IEC 61512 1;
- ANSI/ISA 88.00.02 2001, Batch Control Part 2: Data Structures and Guidelines for Languages – defines data models for production management, data structures for information exchange, as well as a form for recording recipes;
- ANSI/ISA 88.00.03 2003, Batch Control Part 3: General and Site Recipe Models and Representation – defines models for presenting generalized recipes and exchanging such recipes between departments of the enterprise, as well as between the enterprise and its partners;

- ANSI/ISA 88.00.04 2006, Batch Control Part 4: Batch Production Records – defines data models and an indicative system model for recording, storing, retrieving and analyzing data on the progress of periodic production.

IV. System architecture and use cases

It is proposed to use OSTIS Technology and the corresponding set of models, methods and tools for developing semantically compatible intelligent systems as a basis for automating the processes of creating, developing and applying of standards. The basis of OSTIS Technology is a unified version of the information encoding language based on semantic networks with a basic set-theoretic interpretation, called the SC-code [1].

The use of SC-code and models for the presentation of various types of knowledge built on its basis will provide such opportunities as:

- automation of standardization processes by a distributed team of authors;
- the ability to fix conflicting points of view on the same problem in the discussion process and even in the process of applying the developed standard;
- the possibility of evolution of the standard directly in the process of its application;
- lack of information duplication at the semantic level;
- independence of the system of concepts from terminology, as a result – from the natural language in which the standard was originally created;
- the ability to automate standard verification processes, including identifying inconsistencies, information holes, logical duplications, etc.;
- improving the efficiency of using the standard, providing the ability to solve various problems without the need to convert the standard to any other format, in particular, the ability to automate the process of verifying compliance with anything necessary standards;
- and others.

The architecture of each system based on OSTIS Technology (ostis-system) includes a platform for interpreting semantic models of ostis-systems, as well as a semantic model of ostis-system described using SC-code (sc-model of ostis-system). In turn, the sc-model of ostis-system includes the sc-model of the knowledge base, the sc-model of the problem solver and the sc-model of the interface (in particular, the user interface). The principles of organization and designing of knowledge bases and problem solvers are discussed in more detail in [10] and [11], respectively.

Let us consider in more detail the composition of each of the above components in the framework of the system under the ISA-88 standard.

A. System knowledge base

The sc-model of the knowledge base of any ostis-system is based on a hierarchical system of subject domains and corresponding ontologies. Within the framework of this system, the core of sc-models of knowledge bases is distinguished, including the family of top-level subject domains [10]. This approach provides the compatibility of various types of knowledge in any ostis-system, as well as significantly reduce the time to develop a specific knowledge base.

Within the framework of the system under consideration, the following set of subject domains that correspond to the ISA-88 standard (Fig. 1) is distinguished:

Each subject domain has a corresponding structural specification, which includes a list of concepts studied within this domain (Fig. 2):

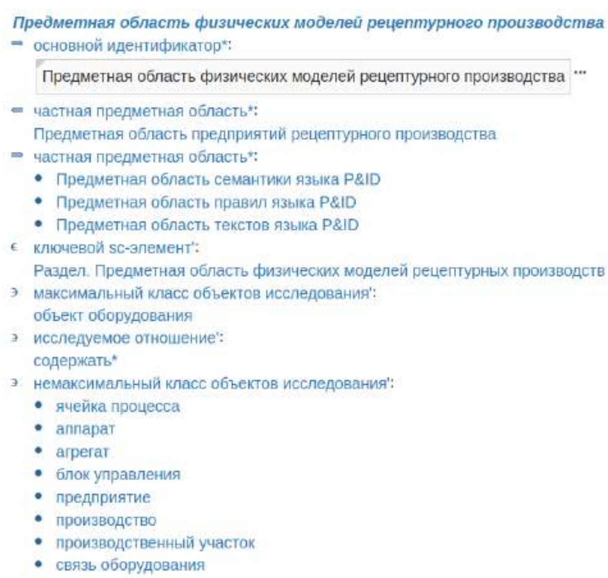


Figure 2. Subject domain specification.

In turn, each concept of the standard is described with the necessary level of detailing (Fig. 3):

In addition, an important advantage of using OSTIS Technology to develop this kind of system is the availability of methods and tools for the collective development of knowledge bases [10], which support the development of a standard presented in the form of a knowledge base distributed by a team of developers.

B. System problem solver

The problem solver of any ostis-system is a hierarchical system of agents interacting with each other exclusively through the semantic memory of this ostis-system (sc-agents) [11]. This approach provides the modifiability of the solver and the possibility of integrating various problem solving models within the solver.

In the context of the system according to the ISA-88 standard, an urgent task is to search for necessary

Figure 1. ISA-88 standard subject domains hierarchy.

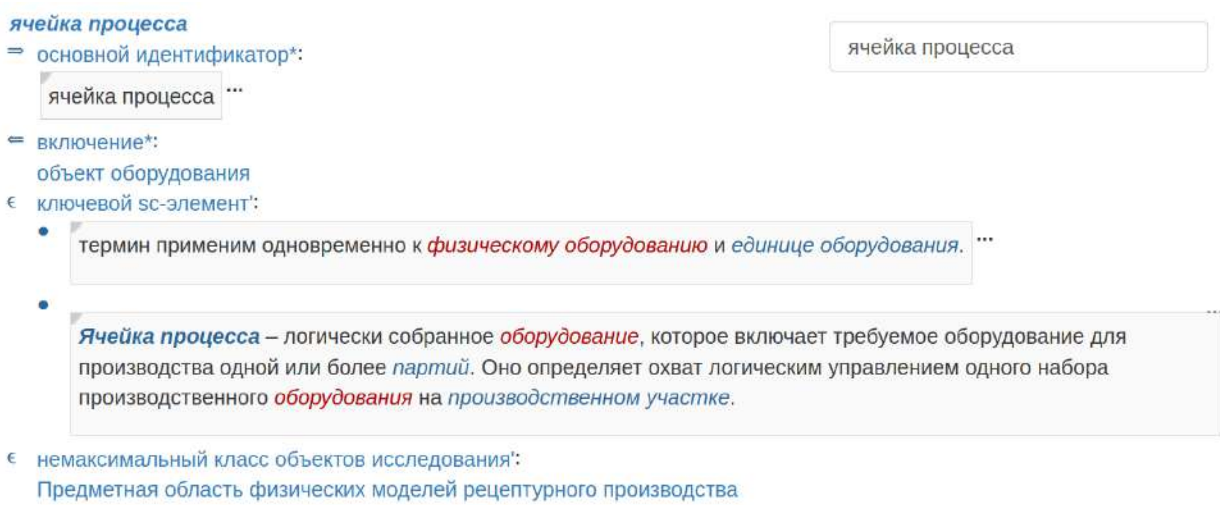


Figure 3. Concept specification within the standard.

information by the employees of the enterprise, from the definitions of certain concepts to the establishment of similarities and differences between the indicated entities. This task is relevant for both experienced employees and new employees who are first acquainted with the standard. An important aspect is the ability to ask questions in a way that is understandable to a user who is not necessarily familiar with the internal structure of an ostis system.

The approach used to build knowledge bases and problem solvers makes it possible to unify the principles of information processing, and, as a result, reduce the number of necessary sc-agents. So, for example, questions of the form

- "What subclasses does the given class have?";
- "What components is the specified entity divided into?";
- "What parts does the given object consist of?";
- "What are the varieties of a given entity?";
- etc.

can be reduced to the generalized question “What entities are particular with respect to a given entity?”, for the processing of which, therefore, it is enough to implement only one sc-agent. However, this does not prohibit having within the user interface a set of more specialized commands with names that are understandable to the end user, but within the framework of the system implemented in a universal key.

C. System user interface

The most important component for a system that describes a standard is the user interface, the quality of which largely determines the efficiency of using the system. Most often, users of the system are employees who are far from semantic technologies, and often information technologies in general, in connection with which the task of building an interface is urgent, which, on the one hand, would be easy to use and visual, on the other hand, would be universal (able to display different types of knowledge).

By default, system responses to the user are displayed in SCn-code, which is a hypertext version of the external representation of SC-code texts and can be read as linear text.

The information in the above figures is displayed as SCn-texts in a mode intended for the end user. For a more experienced user, a mode with a more detailed display of information is also provided.

An important feature of SC-code is a clear separation of the internal sign denoting some entity and the term or other external identifier corresponding to this entity. Due to this feature of each entity, an arbitrary number of external identifiers can be mapped, which, in turn, can easily provide a multilingual standard without the need for duplication of information.

So, for the current version of the system according to the ISA-88 standard, it turned out to be relevant to have versions of the system in Russian, English, Ukrainian and German. Fig. 4 presents the start page of the system in English. As can be seen from the presented examples, only the displayed terms change, the structure of the information remains unchanged.

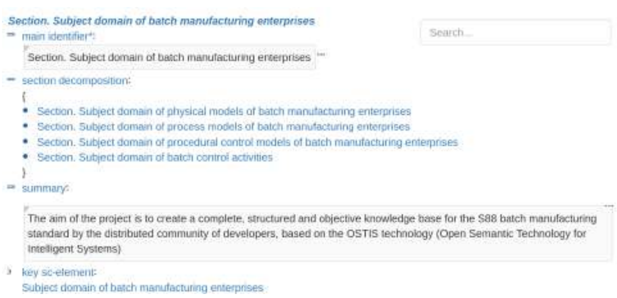


Figure 4. Start page in English.

Thus, the proposed approach also allows us to describe the syntax and semantics of external languages, which makes it possible to construct unified visualization tools not only for universal languages (variants for external representation of SC-code texts), but also for specialized languages, such as PFC. The figure 5 shows an image of a chart in PFC language, to the elements of which you can ask various questions, as well as to any other elements within the knowledge base.

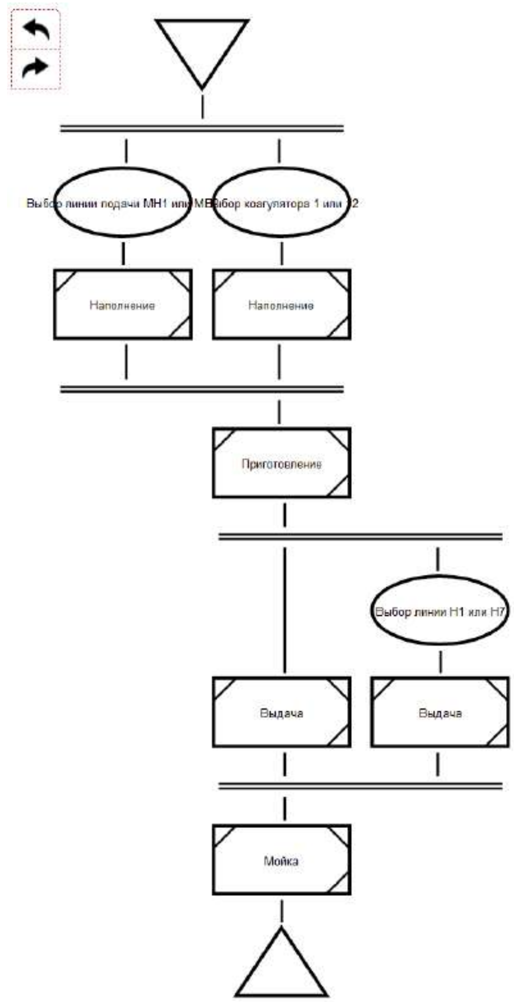


Figure 5. An example chart in PFC

In addition, the OSTIS Technology means allow to store and, very importantly, specify any external files, for example, images, documents, etc. within the framework of the knowledge base. The figure 6 shows the description in the knowledge base of the procedure image in the PFC language, which is part of the ISA-88 standard.

V. Integration of third-party solutions with a knowledge base

A standard system built on the basis of OSTIS Technology can be easily integrated with other systems in the workplace. To integrate the ISA-88 standard system with other systems running on JSC "Savushkin Product", a web-oriented approach is used – the ostis-system server is accessed with the use of the following queries:

```
http://ostis.savushkin.by?sys_id=procedure
```

where "sys_id=procedure" defines a term (the name of an entity) whose value we want to find out (in this example, in fact, the answer to the question“ What is a “procedure”?). This approach makes it relatively easy to

Figure 6. Image specification in the knowledge base.

add support of the knowledge base for current control systems projects, for this it is enough to indicate the names corresponding to the entities in the knowledge base within the control system. Thus, an interactive intelligent help system for control systems projects is implemented, allowing employees to simultaneously work with the control system and ask questions to the system directly during the work.

Fig. 7 shows an illustration of the display of information in the form of a mimic diagram of the HMI (from the control system project).

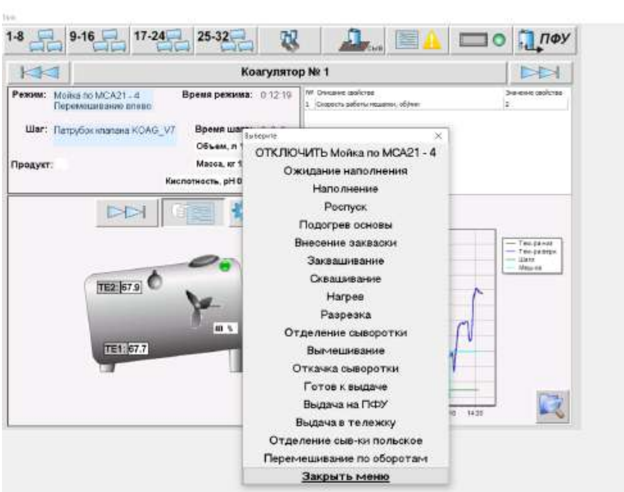


Figure 7. Example HMI from SCADA [12].

Fig. 8 shows a web page that displays the same information as a PFC chart (from the knowledge base).

VI. Conclusion

The paper considers an approach to automating the processes of creating, developing and applying standards based on OSTIS Technology. Using the example of the ISA-88 standard used at the Savushkin Product enterprise, the structure of the knowledge base, the features of the problem solver and the user interface of the support system for these processes are considered. It is shown that the developed system can be easily integrated with other enterprise systems, being the basis for building an information service system for employees in the context of Industry 4.0.

The approach proposed in the work allows us to provide not only the ability to automate the processes of creation, agreeing and development of standards, but also allows us to significantly increase the efficiency of the processes of applying the standard, both in manual and automatic mode.

As part of the further development of the obtained results, the authors plan to attract the international community to the development of the knowledge base and the entire prototype of the system according to the ISA-88 standard, for which the tools of collective development of knowledge bases proposed in the framework of OSTIS Technology will be used.

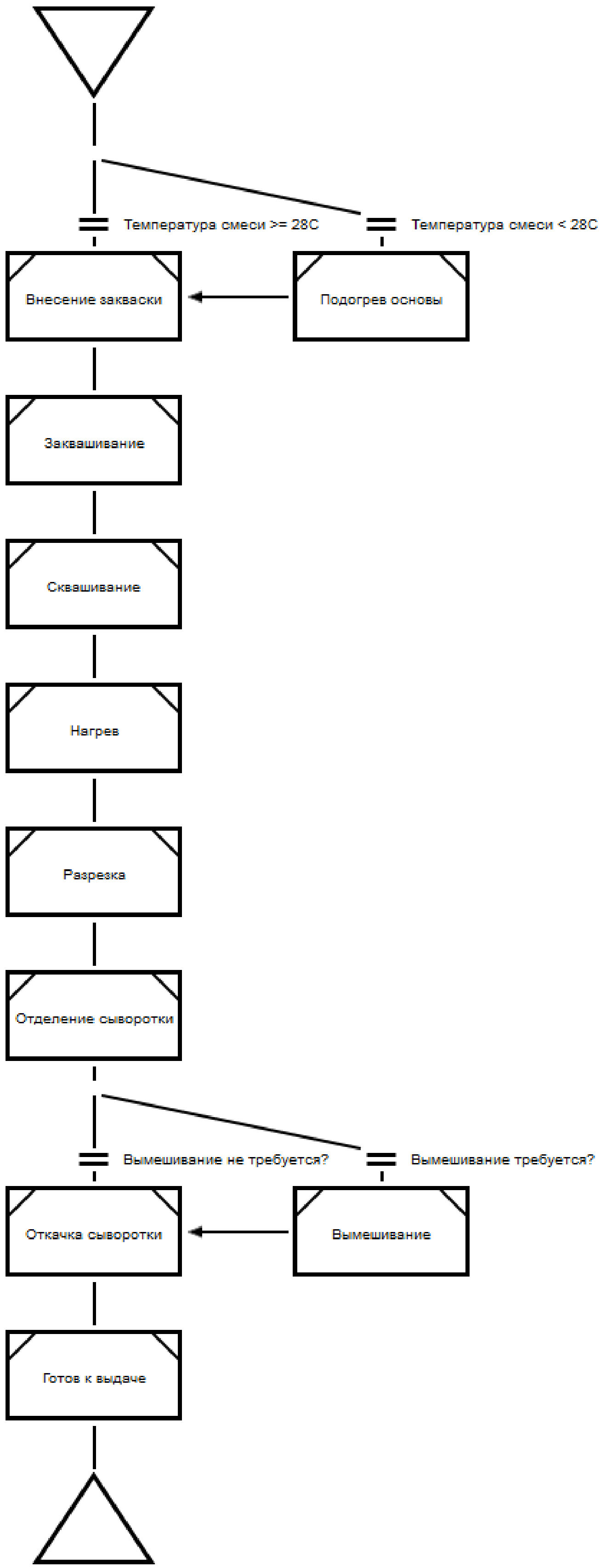


Figure 8. Corresponding PFC chart from OSTIS.

ACKNOWLEDGMENT

Authors would like to thank to the research teams of the Departments of intelligent information technologies of the Belarusian State University of Informatics and Radioelectronics and the Brest State Technical University.

REFERENCES

[1] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, “Methods and tools for ensuring compatibility of computer systems,” in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[2] P. Serenkov, V. Solomaho, V. Nifagin, and A. Minova, “Koncepcija infrastruktury standartizacii kak bazy znanij na osnove ontologij [the concept of a standardization infrastructure as an ontology-based knowledge base],” *Novosti. Standartizacija i sertifikacija. [News. Standardization and certification.]*, no. 5, pp. 25–29, 2004.

[3] V. Uglev, “Aktualizacija soderzhanija standartov proektirovanija slozhnyh tehnicheskih ob’ektov: ontologicheskij podhod [updating the content of design standards for complex technical objects: ontologic approach],” *Ontologija proektirovanija. [Ontology of designing]*, no. 1, pp. 80–86, 2012.

[4] “ISA-88 standard,” Available at: https://www.isa.org/isa88/, (accessed 2020, Jan).

[5] G. Mušič, “A low-cost PackML-based control solution for a modular production line,” *IFAC-PapersOnLine*, vol. 48, no. 10, pp. 184–189, 2015. [Online]. Available: https://doi.org/10.1016/j.ifacol.2015.08.129

[6] B. R. Heravi, M. Lycett, and S. de Cesare, “Ontology-based standards development: Application of OntoStanD to ebXML business process specification schema,” *International Journal of Accounting Information Systems*, vol. 15, no. 3, pp. 275–297, Sep. 2014. [Online]. Available: https://doi.org/10.1016/j.accinf.2014.01.005

[7] S. El-Sappagh, F. Franda, F. Ali, and K.-S. Kwak, “SNOMED CT standard ontology based on the ontology for general medical science,” *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, Aug. 2018. [Online]. Available: https://doi.org/10.1186/s12911-018-0651-5

[8] S. Schulz, R. Stegwee, and C. Chronaki, “Standards in healthcare data,” in *Fundamentals of Clinical Data Science*. Springer International Publishing, Dec. 2018, pp. 19–36. [Online]. Available: https://doi.org/10.1007/978-3-319-99713-1_3

[9] S. Schulz and C. Martínez-Costa, “How ontologies can improve semantic interoperability in health care,” in *Lecture Notes in Computer Science*. Springer International Publishing, 2013, pp. 1–10. [Online]. Available: https://doi.org/10.1007/978-3-319-03916-9_1

[10] I. Davydenko, “Semantic models, method and tools of knowledge bases coordinated development based on reusable components,” in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 99–118.

[11] D. Shunkevich, “Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems,” in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.

[12] V. V. Taberko, D. S. Ivanjuk, V. V. Golenkov, I. T. Davydenko, K. V. Ruseckij, D. V. Shunkevich, V. V. Zaharov, V. P. Ivashenko, and D. N. Koronchik, “Ontologicheskoe proektirovanie predprijatij recepturnogo proizvodstva na osnove ontologij [ontological design of prescription production enterprises based on ontologies],” in *Open semantic technologies for intelligent systems*, ser. 1, V. Golenkov, Ed. BSUIR, Minsk, 2017, pp. 265–280.

Принципы повышения эффективности процессов разработки и использования стандартов в рамках концепции Industry 4.0

Таберко В.В., Иванюк Д.С.,
Шункевич Д.В., Пупена А.Н.

В данной работе предлагается подход к автоматизации процессов создания, развития и применения стандартов на основе Технологии OSTIS. Рассмотрены проблемы современных подходов к организации указанных процессов, показана роль стандартов в рамках концепции Industry 4.0, приведены примеры формализации стандартов в рамках предлагаемого подхода.

На примере стандарта ISA-88, применяемого на предприятии ОАО "Савушкин продукт рассмотрена структура базы знаний, особенности решателя задач и пользовательского интерфейса системы поддержки указанных процессов. Показано, что разработанная система легко может интегрироваться с другими системами предприятия, являясь основой для построения системы информационного обслуживания сотрудников предприятия в контексте Industry 4.0.

Предложенный в работе подход позволяет обеспечить не только возможность автоматизации процессов создания, согласования и развития стандартов, но и позволяет значительно повысить эффективность процессов применения стандарта, как в ручном, так и в автоматическом режиме.

Received 08.02.2020

Implementation of an intelligent decision support system to accompany the manufacturing process

Vladimir Golovko
Alexsander Kroshchanka
Brest State Technical University
Brest, Belarus
gva@bstu.by
kroschenko@gmail.com

Mikhail Kovalev
Belarusian State University of Informatics an Raddioelectronics
Minsk, Belarus
michail.kovalev7@gmail.com

Valery Taberko
Dzmitry Ivaniuk
JSC "Savushkin product"
Brest, Belarus
id@pda.savushkin.by

***Abstract*—The article reviews the implementation of an intelligent decision support system in case of solving real manufacturing problems of JSC "Savushkin Product" for quality control of marking products. The proposed system integrates neural network and semantic models. Also, article proposes original method of localization and recognition of images from a high-speed video stream, as well as a method of intelligent processing of recognition results for developing decision to manufacturing problems.**

***Keywords*—deep neural networks, object detection, knowledge base, integration, inference, IDSS**

INTRODUCTION

The modern direction of industrial development is moving towards increasing automation and robotization of manufacturing. Tasks are set to build autonomous work not only of individual parts of manufacturing, but also of entire plants. Setting and solving such problems in the modern world is called the fourth industrial revolution [1], the development of which is associated with the concept of Industry 4.0 and the Internet of things(IoT) [2].

Under these conditions, the complexity of managing manufacturing [3] increases. There is a need to make decisions not only in the manufacturing process, but also in the process of its adjustment and debugging. Building intelligent systems that can make or offer decisions in the manufacturing process is a promising applied area of artificial intelligence.

The use of intelligent decision support systems(IDSS) can significantly simplify the decision-making process in the workplace for a person [4]. Such systems also allow to make decisions independently and, if necessary, to justify the reasons for the decision. However, implementation of this feature requires integration of different problems solving models [5].

In this article, we consider the modification and implementation of the intelligent decision support system based on the integration of neural network and semantic models described in [6]. The application of this system is considered on the example of the problem of JSC "Savushkin product": quality control of marking on products.

Figure 1. Example of a frame from a video stream

I. FORMULATION OF THE PROBLEM

We were given the task of checking the correctness of the marking printed on products manufactured by JSC "Savushkin product".

The source data is a video stream obtained by an RGB camera mounted above the manufacturing tape. In our example, this stream is characterized by a speed of 76 frames per second. An example of a frame from a stream is shown in Fig. 1.

The solution to this problem includes:

- marking detection and recognition;
- the identification of marking issues;
- search for the causes of these problems;
- search decision to these problems;
- applying a decision independently, if possible.

Problems with marking can be as follows:

- **no paint.** If empty bottles start to go, then the printer has run out of paint. The system can also refer to it (since the printer is connected to the network) and check the availability of paint.
- **camera shift** the system knows that the batch has started filling, but there are no positive recognition results from the camera.
- **incorrect marking.** The marking is recognized, passed to the system, but it does not match the standard – this means that an error occurred when setting the marking text and it is necessary to stop the filling process and notify the operator.
- **unreadable marking.** Marking is not recognized - several digits are not recognized, so the printer nozzles are clogged - it is needed to stop the filling process and notify the operator to clean the printer nozzles. In this case, it is needed to remove bottles with incorrect marking from the conveyor.

In general, we set the following requirements for the system that will solve the problem:

- **High-speed operation.** Manufacturing processes are very fast, so the search for incorrectly marked products and its rejection also must be very fast.
- **Ability to explain.** It is necessary to identify not only incorrect marking, but also to explain the reasons for this situation.
- **Autonomy.** The system should minimize human involvement in the quality control process. Of course, there may be situations when system can't handle problem without operator help, but even in such situations, the system must be able to instruct the operator and be able to explain the reasons for certain actions.
- **Adaptability.** It is necessary to have ability to adapt the system to recognize markings on any other products.

II. Existing approaches

The task of controlling the quality of marking at enterprises similar to the enterprise of JSC "Savushkin Product" is often solved manually. A human periodically selectively checks a part of the product. This approach has the following disadvantages:

- only part of products are checked;
- there is a possibility that a human will not notice a slight discrepancy between the checked marking and the standard one;
- the use of monotonous manual labor.

At the moment, developments to automate these activities are underway and are being implemented, but in most cases we are only talking about identifying problems with marking, but not at all about finding the causes and solutions to these problems.

For example, Omron [7] sensors are often used. These sensors are equipped with a camera and are able to recognize product markings on high-speed tape. Using these sensors allows to automate the work of a person for quality control, but there are the following disadvantages:

- Recognition quality isn't high enough. Sensors are based on Optical Character Recognition (OCR), an approach used for document recognition that is highly dependent on the image quality of the text. Due to the high speed of the manufacturing tape, it is not always possible to get a high-quality photo of the marking.
- Need to buy specialized software for system configuration.
- No built-in system for finding and fixing marking problems.

III. The proposed approach

From the formulation of the problem, it follows that the system that will solve this problem must be able to search for and propose decisions to emerging problems, as well as justify these decisions. Therefore, the proposed approach is based on the experience of building IDSS during designing this system [4].

Classic DSS can be defined as an interactive automated system that helps the decision maker (DSS) use data and models to identify and solve problems and make decisions [8]. Intelligent DSS is included to DSS concept that is characterized by using of models that are traditionally considered intelligent.

The proposed system goes beyond the definition of IDSS, since the system can not only offer, but also apply decisions. For this purpose, system uses the integration of neural network and semantic models.The proposed system goes beyond the definition of IDSS, since the system can not only offer, but also apply decisions. For this purpose, system uses the integration of neural network and semantic models.

Neural network models are responsible for localization and recognition of markings, which is a non-trivial task because the manufacturing tape has high speed.

Semantic models, represented as the knowledge base(KB) based on ontologies, are responsible for search and decision-making. Based on this, the proposed system can be defined as a knowledge-driven DSS [9].

The proposed system is based on OSTIS technology and its principles. OSTIS technology uses knowledge representation and processing models focused on unification and working with knowledge at the semantic level. The main principles and models used in the approach include:

- knowledge integration model and unified semantic-based knowledge representation model [5], based on the SC code [10];
- principles of situational control theory [11];
- ontological model of events and phenomena in knowledge processing processes [12];

- multi-agent approach [13];
- hybrid knowledge processing models [14].

The main components of the system are:

- **Machine vision module.** The task of the module is localization and recognition product markings on the image and transfer to the results of this recognition to the decision-making module. Also, this module stores all trained artificial neural networks (ANNs), the description of which is stored in the KB. In the future, the module should be able to switch between trained ANNs if the engineer sets the appropriate configuration of the KB.
- **Decision-making module.**. Consists of the KB and the problem solver. The KB contains all the necessary knowledge for making and implementing decisions, such as logical rules, statements, current markings, device states, and so on. The problem solver contains a set of internal and external agents that work with the KB. These agents are used for inference for finding decisions, calling external programs for implementing decisions, and preparing decision for the engineer's terminal.
- **Robotic subsystem control module** This module has access to the subsystem that directly carries out the marking of products. The task of this module is implementation of system decisions that can be taken without the involvement of the engineer, such as marking updates, switching the subsystem on and off, etc.
- **Engineer's terminal.** The user interface module that can be used to track decisions made by the system, including decisions that require the attention of an engineer. The terminal provides full access to the KB, so the engineer has the ability to manually configure the system for certain tasks. For example, he can indicate to the system the fallacy of its reasoning and correct the rule by which it made an wrong decision.

The figure 2 shows the general diagram of interaction of the marking quality control system modules.

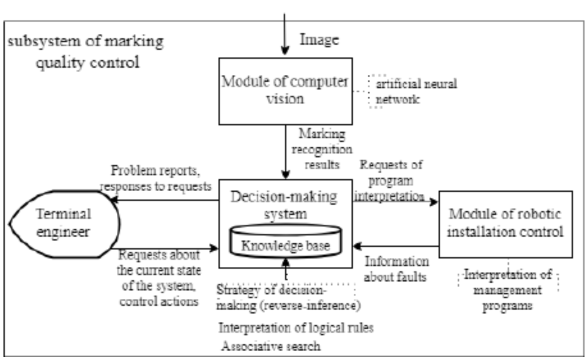


Figure 2. The general diagram of interaction of the marking quality control system modules [6]

Figure 3. Controller RFC 4072S with connected camera DFK 22AUC03

IV. Description of system hardware part

PLCnext. Testing of given program solution was produced based on controller PLCnext RFC 4072S (Fig. 3). This hardware platform is open solution of modern automation systems, which satisfy requirements of IIoT [15].

Controller PLCnext RFC 4072S has next technical parameters: processor Intel® Core™i5-6300U 2x2.4 GHz, USB-interface, 4 Ethernet ports with different values of data transmitting speed. This model has degree of protection IP20. Additionally can be acquired software, fan module and SD-cards 2/8 GB for store programs and configurations [16].

During testing of the this hardware platform, we faced with next problems:

- small amount of storage available to work in this model of controller (only 2 GB of free space on original SD card);
- limited version of OS;
- limitations in current firmware.

First and second problem has been solved by installing a larger non-original SD card with a pre-installed Debian OS.

V. Description of machine vision module module

A. Marking detection problem statement

Detection of objects is possible in real time, but appears problem of fast frame processing by neural network. In-time completition of processing is not possible in case of each frame processing. In our task time

Figure 4. Main marking defects

window for processing one frame from video stream is only 0.013 s or 13 milliseconds. At present, neural network models, which capable of objects detection for a such time interval using a mobile device or a specialized controller, are not exist. Therefore, it is necessary to evaluate the importance of separate frames for performing detection.

On the other hand, bottles move on the manufacturing tape with certain frequency (about 3 pieces per second), which means that the neural network can process not every frame of the video stream, but only some and ignore the rest. This circumstance increases the time interval during which the processing should be performed, to a value of 250-300 milliseconds.

The process of marking recognition includes additional tasks such as test of marking printing. The first task is to determine the presence of marking on the cap. The second task is to determine the presence of marking distortions, which arise during the printing process, the absence of marking parts, etc. And, finally, the third task is the actual detection of the numbers in the marking, the formation of output information (the date of manufacturing of the goods, numbers in the consignment, etc.) and the determination of the correctness of this data in accordance with a predefined pattern. The main marking defects are shown in Fig. 4.

The presence of several subtasks involves the use of a group of neural networks, each of which performs its own part of the work.

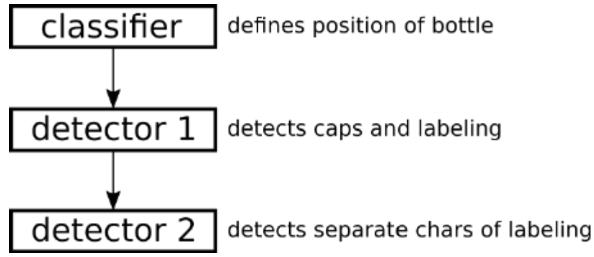


Figure 5. Structure of the marking recognition module

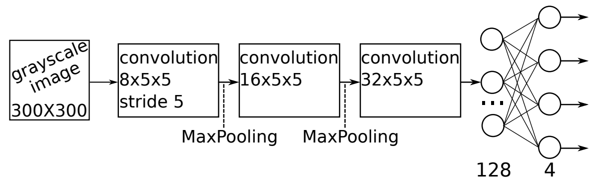


Figure 6. Structure of classifier for evaluating bottle position

B. Architecture of the recognition module

Our recognition module is an extension of the idea described in [17]. The developed marking recognition module consists of three neural networks (Fig. 5).

The first network is a simple classifier based on a convolutional neural network, which determines the position of the bottle in the frame. We have selected four classes of position by distance from the center of the frame. Class 1 describes the minimal distance from the center of the frame. Only frames of this class are transferred for further analysis to other models. Class 2 and 3 describe the average and maximal distance. Finally, class 4 is needed for the case when the cap with the marking is not in the frame (for example, an empty line is viewed).

The architecture of the classifier is shown in Fig. 6. It consists of 5 layers and has 4 output neurons according to the number of classes that determine the position of the bottle in the frame. All layers use the ReLU activation function except for the 3rd and last layers. They use linear and softmax activation functions, respectively. Also, max pooling is applied after the first and second convolutional layers with stride = 2.

In case the frame was classified as class 1, it's transmitted further to the second neural network.

The second model is a detector and searches for caps and markings in the frame. Here, an SSD network based on the MobileNet v1 [18] classifier was chosen as the architecture.

At the stage of detection the cap and marking, an analysis is made of the presence of a defect associated with the absence of marking. This is made trivial: if object **cap** has been detected without object **marking**, then it has assumed that a defect occurs. If a defect

Figure 7. Distance of the bottle from the center of the frame (4 classes)

has been identified, the module notifies the operator. If no defects were found in the marking, the frame is transferred to the next neural network.

Finally, the third neural network is also SSD-MobileNet v1, which detects individual digits in the marking. After that, the formation of the final recognition result is performed.

The use of two serial networks for the detection of individual objects is explained by the fact that the original image has a high resolution and is scaled to 300X300 pixels before being fed to the detector. This conversion makes the detection of individual digits almost impossible due to their relatively small size. To eliminate this drawback, the image of the marking in the original dimensions with the necessary scaling is fed to the input of the second detector.

C. Training datasets: preparing and main features

To create a training dataset for the classifier, we have used the neural network model Faster R-CNN (based on the pre-trained ResNet50 classifier). This model has better detection efficiency compared to the SSD-MobileNet model, but it is slower [19]. This network was used to detect the caps in the frame. A trained detector was used to split the available dataset to bottle position classes. The Euclidean distance from the center of the cap to the center of the frame was used as a measure of distance. By this way, classes 1-4 were formed for classifier (Fig. 7). The resulting dataset includes 6189 images, 1238 of which make up the testing dataset.

After classifier was trained, the final classification accuracy was about 92%.

Both detectors (for detection of caps / markings and individual digits) were trained based on pre-trained models.

When training the detector of caps and markings, a prepared dataset was used, which includes 637 images in the training part and 157 in the testing part.

The following learning procedure was used to train the digit detector. At the beginning, the SSD model was pre-trained on images from SVHN dataset of house numbers for 80 epochs (several images from this dataset are shown in Fig. 8). For pre-training, a variant of this dataset was used, which includes 33402 images in the training part and 13068 in the test part. After pre-training, we have continued to train the neural network using a dataset of markings. This dataset contains 347 images in the training part and 72 in the test part. Examples of images were used for training are presented in Fig. 9.

Figure 8. Examples of images from SVHN dataset [20]

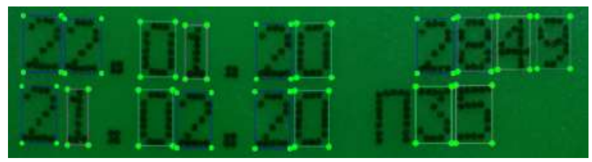


Figure 9. Image from training dataset of detector 2

D. Results

The use of the SSD model allows to achieve a detection efficiency of 99% (mAP = 0.99) for caps and markings detection and 87% (mAP = 0.87) for individual digits. Additionally, the processing speed allows to detect objects in the video stream at a speed of 76 frames per second. Efficiency of the detection of individual digits are presented in the table I.

Table I
DETECTION EFFICIENCY OF INDIVIDUAL CLASSES OF DIGITS

Class label	AP
0	0.8871
1.	0.8766
2.	0.8686
3.	0.8096
4.	0.8874
5.	0.8998
6.	0.8847
7.	0.8933
8.	0.8691
9.	0.8857
mAP	0.8762

The result of detection by the first and second detector is shown in Fig. 10 and 11.

VI. DECISION MAKING MODULE

A. Knowledge base

KB designed using OSTIS technology are divided into subject domains, which in turn have its own ontologies. The proposed system has next subject domains:

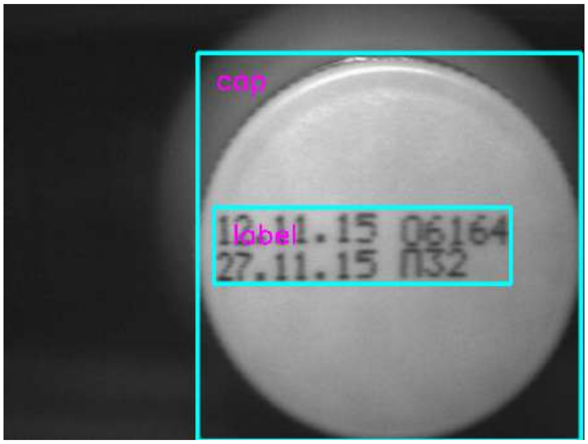


Figure 10. Example of cap/marking detection

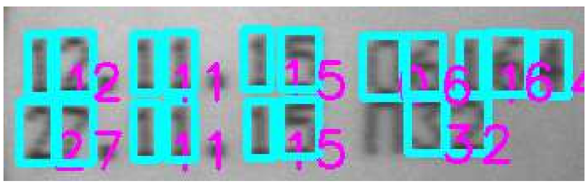


Figure 11. Example of digits detection

- Subject domain of product marking;
- Subject domain of neural network models;

Each subject domain that described in the KB has ontology in KB.

1) The subject domain of product marking: This subject domain contains a description of all necessary information about the product and quality control of its marking, such as:

- product brand;
- the recognized marking;
- the standard marking;
- number of consecutive unmarked products;
- the level of paint in the printer;
- and etc.

The figure 12 shows a fragment from this subject domain that describes the quality control point for marking the bottles of some yogurt. This fragment contains information that machine vision module didn't recognize the marking of bottle, which is currently being controled. Moreover, at this control point, this is the fourth bottle with an unrecognized marking. For this control point, there is a limit on the number of consecutive bottles with the wrong marking, which is equal to three. Also, the control point knows a printer, which prints the marking and the level of paint in it. In this case, this level is zero.

This information will be enough for a human to conclude on the reasons for the missing of marking. Below we will consider the mechanism by which the system will make a similar conclusion.

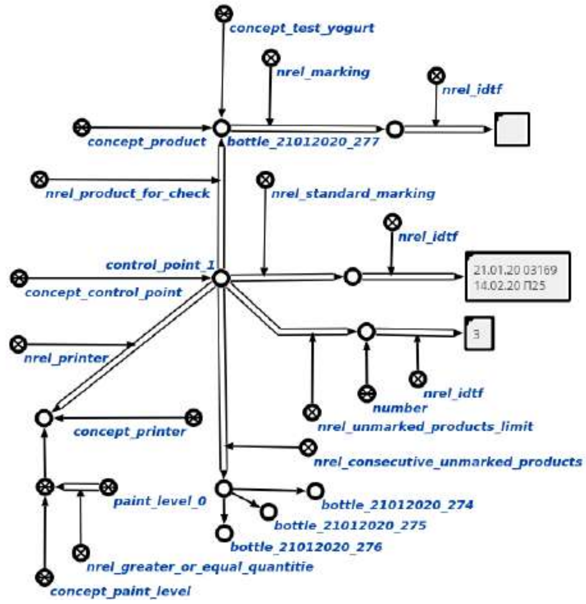


Figure 12. Fragment of subject domain of product marking for some yogurt bottles

This subject domain also contains a set of implicative and equivalence bundles, which we will call logical rules for short. According to these rules, the reverse inference [21] is used to make the decision or a set of decisions. Inference uses the ≪if≫ part as search patterns in the KB. When matching the ≪if≫ part of a statement was found, the system generates the knowledge described in the ≪then≫ part of the implicative bundle of the used logical rule. For logical rules, presented in the form of equivalence tuples, the mechanism of its using is similar, with the only difference that in place of the ≪if-then≫ parts there can be any part of the equivalence tuple.

It should be noted that the logical rule can also be the specifications of agents or programs. These specifications are represented in the form of the implicative tuple, in which "if" part describes the input data, and in "then" part describes the output data. When making the inference, the problem solver will use these logical rules on a par with the rest, but when using these logical rules, the appropriate agent or program will be called.

Each logical rule has the number of times it is used by inference. This technique will allow the system to self-learn and speed up the inference for trivial situations [6].

The figure 13 shows an example of a logical rule that can be written in natural language like this: *If the product is not marked, but it is critically bad marked, and the paint level in the printer is less than or equal to 0, then it is needed to stop the production tape and send a message that the printer has run out of paint.*

Applying of this logical rule to the KB which containing the fragment shown in the figure 12 will lead to the fact that the system will be able to conclude on the reasons for the missing of marking.

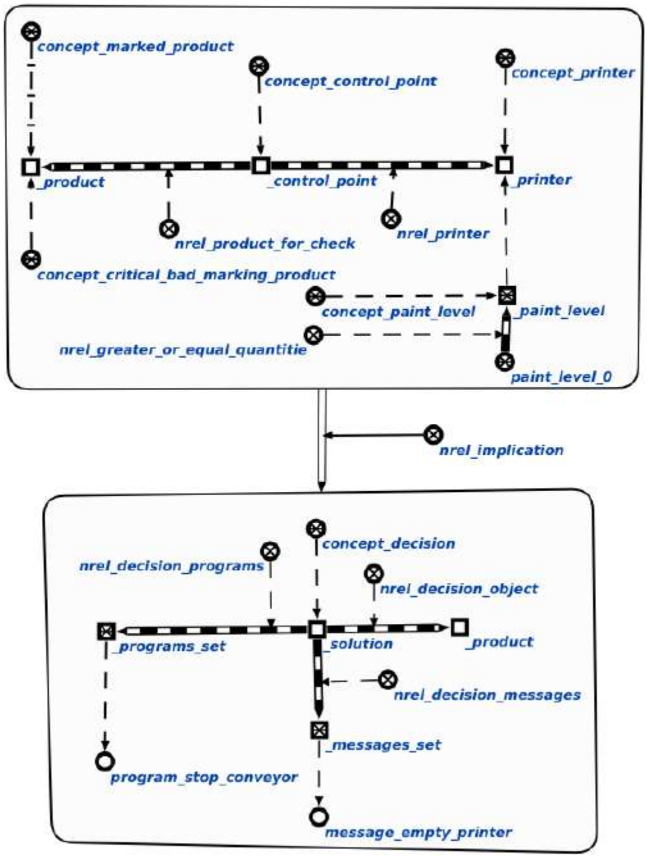


Figure 13. Logical rule for checking the reasons for missing a marking

One of the most important features of the system is the ability to explain made or proposed decisions. For this purpose, the inference makes a decision tree in the course. Decision tree stores the input recognition data, the recognized marking, the chain of applied logical rules and the applied (proposed for application by the engineer) decision.

2) Subject domain of neural network models: The description of this subject domain will help the specialist when setting up the system for a specific task. Based on this knowledge, the system will be able to offer one or another trained ANN available in the machine vision module. In this regard, there is a need to describe the entire hierarchy of subject domain of neural networks models, proposed in [5], in the KB. This subject domain contains describtion of the following information about the trained ANN:

- type of input data;
- set of recognized classes or output data type;
- class of tasks solved by ANN;
- architecture;
- average operating time;
- the quality of recognition.

In addition, this detailed description of the trained ANN in the KB can be used to provide information support to the engineer who will update the architecture or retrain the ANN.

In the future, this part of the KB can be used to expand the system's authority from supporting decision-making on marking control to supporting decision-making on selecting the configuration of machine vision module for a specific hardware and a specific hardware platform.

B. Problem solver

In OSTIS technology, problem solvers are constructed on the basis of the multi-agent approach. According to this approach, the problem solver is implemented as a set of agents called *sc-agents*. All sc-agents interact through common memory, passing data to each other as semantic network structures (sc-texts) [14].

The general tasks of the problem solver of IDSS are:

- access to knowledge in the KB;
- processing (analysis, verification, and generation)of knowledge; item interaction with other modules of the system.

The main task of the solver is implementation of reverse inference. In this way, the system knows in advance the set of reasons why incorrect markings may be printed.

At the start work, solver creates sets of logical rules(by pattern search), which applying will lead to the appearance of the necessary semantic construction in the KB. Next, it tries to apply the most frequently used logical rule. The logical rule can be applied when the KB contains semantic construction that isomorphic to the construction that was obtained by substituting nodes associated with the processed product into a template from a logical rule. This pattern is the first part of the logical rule, the second part describes the knowledge that will be generated after applying this logical rule.

If the rule can be applied, the system will generate knowledge from second part of rule and will add the rule and the result of its using to the decision tree.

In the case when there is not enough knowledge to apply a logical rule, the solver recursively initiates the work of itself, where it is already trying to find logical rules, the applying of which will lead to the appearance of the missing knowledge in the KB.

If the applying of any logical rule does not lead to the the appearance of the necessary semantic constructions in the KB, the agent reports that it can't find the decision for this problem.

In the future, we consider an option in which the system will be able to generate logical rules itself, based on patterns in the course of the system's operation or on cases of manual correction of problems.

VII. Conclusion

The implementation of the proposed IDSS significantly improves quality control in manufacturing, since such a system is able not only to identify a problem with products, but also to help find the causes of these problems, and even, in some cases, to solve them independently.

The quality of detection of individual digits can be improved with by increase of the training dataset, this is a subject of future work. The second aspect that affects the quality of detection is the location of the marking. It was noticed that when the marking is in a horizontal orientation, the quality of detection is higher than in case of rotated marking. Therefore a study of the possibility of using neural network models to select the correct marking orientation is required.

A perspective development of the system is the addition of software functionality to support the engineer when scaling the solution to any other product or on another hardware platform. The engineer should be able to set the configuration of the machine vision module, as well as receive recommendations on this configuration from the system. Recommendations will be based on the specification of the task and the hardware capabilities of the platform. This requires a description of the ontology of neural network models and hardware platforms.

References

[1] Schwab K. The fourth industrial revolution. Crown Publ., NY, 2016, 198 p.

[2] Tarasov V. B. Strategicheskii injiniring predpriyatii buduschego: massovoe sotrudnichestvo, internet veschei, iniciativa “industriya 4.0”, chto dalshe // Injiniring predpriyatii i upravlenie znaniyami (IP&UZ-2016): collection of scientific papers, XIX Scientific and practical conference, Moscow, April 26-27, 2016 / by scientific ed. Telnova. M.: PLEKHANOV Russian University of Economics, 2016 pp. 57–68., (in Russian).

[3] Schwab K., Davis N. Shaping the Fourth Industrial Revolution. World Economic Forum, Geneva, 2018, 287 p.

[4] Tariq A., Rafi K. Intelligent Decision Support Systems. Information and Knowledge Management, ISSN 2224-5758 (Paper) ISSN 2224-896X (Online) Vol 2, No.6, 2012

[5] Golovko V. A., Golenkov V. V., Ivashenko V. P., Taberko V. V., Ivaniuk D. S., Kroshchanka A. A., Kovalev M. V. Integration of artificial neural networks and knowledge bases. In: Open semantic technologies for designing intelligent systems (OSTIS2018): materials of the International Science and Technology Conf. Minsk, February 15-17, 2018) – Minsk: BSUIR, 2018. – P. 133–145

[6] Golovko V. A., Ivashenko V. P., Taberko V. V., Ivaniuk D. S., Kroshchanka A. A., Kovalev M. V. Principles of decision making building systems based on the integration of neural networks and semantic models. In: Open semantic technologies for designing intelligent systems (OSTIS2019): materials of the International Science and Technology Conf. Minsk, February 21-23, 2019 – Minsk: BSUIR, 2019 – P. 91–102

[7] (2020, Jan.) Omron documentation. [Online]. Available: https://robotics.omron.com/browse-documents

[8] Saraev A.D., Shcherbina O. A. System analysis and modern information technologies. Proceedings of the Crimean Academy of Sciences. - Simferopol: SONAT, 2006. - Pp. 47-59.

[9] Power D. J. Decision support systems: concepts and resources for managers, Westport, Conn., Quorum Books, 2002, 251 p.

[10] V. Golenkov and N. Gulyakina, "Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual’nykh sistem. Chast’ 2: Unifitsirovannye modeli proektirovaniya [project of open semantic technology of component design of intelligent systems. part 2: Unified design models] Ontologiya proektirovaniya [Ontology of design], no. 4, pp. 34-53, 2014, (in Russian).

[11] D. Pospelov, "Situacionnoe upravlenie. Teoriya i praktika Moscow: Nauka, 228p, 1986, (in Russian).

[12] V. Ivashenko, "Ontologicheskaya model’ prostranstvenno vremennykh otnoshenii sobytii i yavlenii v protsessakh obrabotki znanii Vestnik BrSTU, Brest, No5(107), pp. 13–17, 2017, (in Russian).

[13] V Tarasov, "Ot mnogoagentnih sistem k intellektualnim organizaciyam filosofiya psihologiya informatika Moscow: Editorial URSS, 352 p., 2002, (in Russian).

[14] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems,"Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems], vol. 2, pp. 119-132, 2018.

[15] PHOENIX CONTACT | Control technology and software – Intelligent automation [Online]. Available: https://www.phoenixcontact.com/assets/2018/interactive_ed/101_86135/index.html#4.

[16] PHOENIX CONTACT | Safety controller – RFC 4072S – 1051328 [Online]. Available: https://www.phoenixcontact.com/online/portal/us?uri=pxc-oc-itemdetail:pid=1051328&library=usen&tab=1.

[17] Golovko, V. Brands and Caps Labeling Recognition in Images Using Deep Learning / V. Golovko, A. Kroshchanka, E. Mikhno // International Conference on Pattern Recognition and Information Processing (PRIP 2019). – Springer, 2019. – P. 35–51.

[18] Howard A. G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H., MobileNets: Efficient ConvolutionalNeural Networks for Mobile Vision Applications [Online]. Available: https://arxiv.org/pdf/1704.04861.pdf.

[19] Hui J., Object detection: speed and accuracy comparison (FasterR-CNN, R-FCN, SSD, FPN, RetinaNet and YOLOv3) [Online]. Available: https://medium.com/@jonathan_hui/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359.

[20] Netzer Y., Wang T., Coates A., Bissacco A., Wu B., Ng A.Y. Reading Digits in Natural Images with Unsupervised Feature Learning, NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[21] V. Vagin, A. Zagoryanskaya, M. Fomina, "Dostovernii i pravdopodobnii vivod v intellektualnih sistemah Moscow: FIZMATLIT, 2008, 704 p., (in Russian).

Реализация интеллектуальной системы поддержки принятия решений для сопровождения производственного процесса

Головко В. А., Крощенко А. А., Ковалев М. В.
Таберко В. В., Иванюк Д. С.

В работе рассматривается реализация интеллектуальной системы поддержки принятия решений в рамках решения реальных производственных задачах ОАО «Савушкин Продукт» по контролю качества нанесения маркировки на продукцию. Предложенная система интегрирует в себе нейросетевые и семантические модели. Предложен оригинальный метод локализация и распознавания изображений из видеопотока высокой скорости, а также способ интеллектуальной обработки результатов распознавания для выработки решений производственных проблем.

Received 30.01.2020

Perspective Approaches to Semantic Knowledge Representation and their Applications in the Context of the Task of Automatic Identification of the Semantically Equivalent Fragments of the Text Documents

Aliaksandr Hardzei
Minsk State Linguistic University
Minsk, Belarus, 220034
Email: alieks2001@yahoo.com

Yury Krapivin
Brest State Technical University
Brest, Belarus, 224017
Email: ybox@list.ru

***Abstract*—The article presents the solution of the maintenance the trust in science in context of the problem of the automatic identification of the semantically equivalent fragments of the text documents. The solution refers to the implementation of the principles of unification of information representation in the memory of computer systems and semantic knowledge representation according to the Theory for Automatic Generation of Knowledge.**

Key words: **natural language, automatic text processing, research integrity, adopted fragment.**

I. Introduction

As a concept, "Ethics" deals with what is "good" or "bad" and also is concerned with the moral duty and obligations. For generations, every culture and society across the world, has established for their members' an ethical code which is to be adhered and practiced. It is even possible to treat such code as a standard which defines society's (people) values, beliefs and behaviors. The key for applying such standard for different situations of real life is a culture. It plays an important role in establishing the ethical standards in a specific society and to be practiced by all. A quite trivial example that illustrates an obvious fact of existence and ascendancy of ethical questions on the quality of the real life – is research integrity. The consequences of degradation processes that take place might lead to unfortunate results in different spheres from economics up to mental and physical health and the ultimately the failure of the business [1].

II. Reliable science and research integrity

Today high-quality, reliable science and research integrity go hand-in-hand. The scientific research is increasingly interdisciplinary, public-private partnerships are commonplace and there is large-scale global competition. These developments put integrity under pressure [2].

Dr. Robert Merton in his work "A Note on Science and Technology in a Democratic Order" [3] articulated an ethos of science. He argued that, although no formal scientific code exists, the values and norms of modern science can nevertheless be inferred from scientists' common practices and widely held attitudes. Merton discussed four idealized norms: universalism, communality, disinterestedness, and organized skepticism. The first one, universalism, refers to the idea that scientific claims must be held to objective and "preestablished impersonal criteria." This value can be inferred by the scientific method or the requirement of peer review before publication in the vast majority of academic journals. The second one, communality, means that the findings of science are common property to the scientific community and that scientific progress relies on open communication and sharing. According to the norm of disinterestedness the science should limit the influence of bias as much as possible and should be done for the sake of science, rather than self-interest or power. Disinterestedness can often be the most difficult norm to achieve, especially when one's job or academic status relies on publications or citations. Many scientists believe that lack of disinterestedness is a systemic issue – one that funders, publishers, and scientists alike should work to address. Who, for example, gets to prioritize research: those who fund, implement, or publish it? And the last one, organized skepticism, refers to the necessity of proof or verification subjects science to more scrutiny than any other field. This norm points once again to peer review and the value of reproducibility [4].

"We need to take a systematic and institutional approach to overcome challenges to integrity," says project coordinator Hub Zwart of Radboud University Nijmegen in the Netherlands. "Research integrity must be supported by the research culture. This means we don't want a campaign against misconduct. Instead, we want to promote a campaign for good science – research is teamwork and research integrity is teamwork"[2].

In a general way, the term "Education" can be referred to as the transmission of values and knowledge accumulated by a society to its members through the process of socialization or enculturation. Every individual will acquire education that will guide him to learn a culture, wherein he molds his behavior and directs it towards the eventual role in the society. In the context of the research integrity the education can be acquired through activities conducted by the scientific adviser and the entire research environment takes the place of the school.

As shown in [1] institutions of higher education have a major role to play in preparing the younger generation for a propitious future. Apart from imparting quality education, they need to instill high ethical values and practices amongst the young researcher fraternity. The rapid development of information technologies all over the world, which is reflected in the popularization of the electronic form of information storage, accumulation and processing in all areas of human activity, makes this task quite sophisticated. That's why the solving the task at different levels: from college up to research institution can ensure a bright future.

As an example, the wrong understanding of plagiarism and incorrect definition could cause academic society with of numerous cases of plagiarism. Students often unclear understood plagiarism and how to prepare properly a written work. Teachers also could interpret plagiarism differently. Plagiarism often becomes as a moral maze for students. They deal with a moral and ethical dilemma having unclear understanding what behavior is appropriate. Therefore the institutions and academic communities are suggested to discuss and formally provide a detailed and clear definition of plagiarism in process of implementation of plagiarism prevention systems of higher education [5]. In recommendations for plagiarism prevention policy formation is suggested creating a definition of plagiarism not only on institutional but also at national level. This definition should be appropriate to different fields of study and daily use, specifying what behavior and actions are acceptable [6].

The research community promotes integrity to maintain trust in science in different ways. According to [7] Japan Agency for Medical Research and Development (AMED) was launched in April 2015 to promote integrated medical research and development (R&D) ranging from basic research to practical applications, in order to smoothly achieve the nationwide application of research outcomes, and to establish an environment therefor. AMED consolidates budgets for R&D expenses, which had previously been allocated from different sources, such as the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Health, Labour and Welfare, and the Ministry of Economy, Trade and Industry. It provides funds strategically to universities, research institutions, etc. By promoting medical R&D, AMED aims to achieve the world's highest level of medical care/services to contribute to a society in which people live long and healthy lives. To achieve this mission, it is imperative that R&D funded by AMED is widely understood and supported. Maintaining and improving research integrity is a prerequisite to this end. AMED is taking various measures to ensure fair and appropriate R&D. It is asking researchers to participate in its responsible conduct in research (RCR) education program and to comply with its rules for managing conflicts of interest. In addition, AMED also conducts a grant program to create and distribute a variety of educational materials on RCR and other matters. Further, AMED is establishing a platform that allows researchers to exchange information about research integrity, and it is undertaking additional measures, such as holding meetings and international symposia on research integrity. This example demonstrates that research integrity is not only a located in the minds of individual researchers. Rather it is also a characteristic of organizations, such as research departments, universities, or networks [8]. Furthermore, it is not a stable trait of organizations but represents ongoing processes in organizations. This means that (organizational) integrity can be developed and nurtured – by managers as well as by the members of the organization [9]. The usage of the systems of adopted text fragments recognition [10] -– is one of such traits. The work [11] demonstrates a general algorithm of solving the task of automatic recognition of semantically adopted fragments of the text documents, which is based on the knowledge system, oriented to the recognition of the concepts, actions and their attributes in the text documents.

III. Perspective approaches to semantic knowledge representation for Artificial Intelligence and their applications

Quite bright example of trends in natural language processing – ACL 2019 – the 57th annual meeting of the Association for Computational Linguistics that took place in Italy and enrolled more than 3000 registered attendees (2900 submissions, 660 accepted papers). The application of knowledge graphs in solving natural language processing tasks was researched about 30 papers out of 660 which is a good 5% share for such a huge event [12].

Another one novel and perspective approach to semantic knowledge representation for Artificial Intelligence was formulated by prof. A. Hardzei, which proposed the theory having in the foundation semantic formalism for knowledge representation and knowledge inference.

The theory is based on Universal Semantic Code (USC) developed by prof. V.V. Martynov where there were proposed semantic primitives, i.e. semantically irreducible kernel words, and the rules of their combinations were defined. In general, semantic coding is a process of converting natural language phrases to a chain of semantic primitives or semantic formulas and back. This is the crucial difference with Semantic Web where there are no semantic formulas but semantic tags represented not formally but by means of natural language.

The essential difference of Theory for Automatic Generation of Knowledge Architecture (TAPAZ) vs USC is in the method of defining a structure of the semantic formula and operations of transformation of semantic formulas to each other.

USC operates with complex formulas consisting of two parts reflecting a consequence "if …than…" or "stimulus → reaction". TAPAZ proposes an alternative way of representing semantic formulas as extended formulas generated by adding parenthesis on the right margin according to determined rules. Each formula has a "semantic counterpart", coupled with a mathematical formalism a highlighted fragment of the Model of the World, representing an interpretation of the formula in natural language. Each formula has one, and only one semantic meaning. The meaning is not assigned but inferred from the structure of the formula.

Besides, the theory is supported by geometric model of the World demonstrating how one "individ", a kind of a pattern as a separate entity in the selected fragment of the Model of the World, which consists of the core, the shell and the surroundings, of the model transfers an impulse to another individ through an environment. The impulse direction depends on the role of the individ. There are only four roles: subject, object, instrument, and mediator. The roles indicate members of some action and may be strictly specified according to its purpose, for example, the role of the instrument may be specialized only as: activator, suppressor, enhancer or converter. The semantic formulas in TAPAZ represent actions surrounded by members of the action.

Accumulating all together a semantic classifier of actions has been proposed. The classifier has 112 semantic classes. Each name of the class represents a highest abstract level of the action and supported with a list of actions giving concrete implementation. For example, the class action "connect" may be implemented by: gluing, nailing soldering etc. Such a structure has an ontological nature and has a practical application for calculation of subject domains [13].

Thus, the system core (World Model) based on TAPAZ is a combination of 112 macro-processes (actions) of the TAPAZ Semantic Classifier with a series of specialized processes of the selected subject domain, for example, Remote Sensing of the Earth (ERS)[1]. Each process has 18 semantic fields in accordance with the TAPAZ Role List of Individs. The semantic weight of ERS-process in a synonymous series is determined by the completeness of the fields, the frequency index and the height of the Oriented Graph of the Semantic Classifier as its vertices are filled.

The construction of the system core is carried out in manual, semi-automatic and automatic modes. At the first stage, the formation of the Intellectual Knowledge Base by prescribing ontological relationships between the independent taxonomy of ERS-individs and the dependent taxonomy of ERS-processes is allowed.

The Specialized Intellectual Knowledge Base (SIKB) of actual space research and technology in the field of ERS combines the TAPAZ-based system core with the periphery generated by the Oriented Graph of the Semantic Classifier, complemented by stepwise recursion and expanded by the TAPAZ Role List of Individs.

The rules for constructing, restricting, reducing and transforming algebraic formulas, the rules for semantic reading of algebraic formulas and interpreting typical combinations of individs, the rules for constructing the Oriented Graph of the Semantic Classifier, the procedure for semantic coding and decoding of its vertices, the groups and rows of the Semantic Classifier, the TAPAZ Role List of Individs and step recursion form the Universal Problem Solver directly interacts with Web Interface.

The algorithm for extracting specialized terminology from ERS-content by an expert to fill the Knowledge Base involves the software ExpertTool version 2.3.4.7, developed by A. A. Matsko. The algorithm for extracting specialized terminology from ERS-content by an expert and filling in the TAPAZ semantic fields (the role list of individs in the ERS-domain) is similar to the TAPAZ Role List of Individs:

subject (initiator→spreader→inspirer→creator) →instrument(activator→suppressor→enhancer →converter)→mediator (landmark→locus →carrier→adapter→material→prototype →resource→source→chronotope→fund) →object →product.

Interactive filling the semantic fields for each specified process of the ERS-domain requires answers to typical questions:

Who? With which tool? In relation to whom / what? In what place? Arriving on what? Adjusting with what?

[1]Currently, TAPAZ technology is used as linguistic support for the event's Internet monitoring subsystem 5[1] "Development of an Intelligent Web-Based System for Searching and Processing Content in English and Russian Using the Semantic Coding Method for Information Support for Solving Managerial, Design-and-Search and Expert Tasks in the Field of Remote Sensing of the Earth", section 1 "Development of the Belarusian Space System for Remote Sensing of the Earth", sub-programme 7 "Research and Peaceful Use of Outer Space" of the State program "High Technology and Technics" for 2016–2020.

Making of what? Following what example? Spending what? Knowing what? In what period? Due to whose prompt? Affecting who / what? Produces whom / what?

All search engines currently operating in the world search either by keywords or by tuples of keywords (keywords in nominal and / or verb groups) using standard software tools for content preprocessing and automatic lexical (tagger) and syntax (parser) markup.

The main drawback of both types of search is its inaccuracy and, as a consequence, the immensity for the user of the huge number of URL (Uniform Resource Locator) found by the search engine, forcing search engine developers to limit the search area on the Internet to the region of the query place, which leads to the knowingly incomplete search results. Attempts to supplement the search for keywords and / or their tuples with contextual synonyms based on empirical ontologies that are incorrectly called "semantic"[2] only increase the inaccuracy and incompleteness of the search, overloading the search page with information noise and causing the user to feel unreliable of the received sample.

The TAPAZ technology offers a search by event fragments or technological cycles, which are described by special units that are macro-processes[3] in the assembly, when specialized ERS-processes are put in accordance with a certain algorithm in accordance with TAPAZ macro-processes and the roles of all participants in the events[4] are calculated.

This approach provides maximum accuracy and speed of search, relevance of search results. In addition, it allows you to find similar technological cycles in close (adjacent) and distant subject domains, thereby providing support to the user in analytical activities, which greatly expands the functionality of the search engine, shifting it to the side of inventing.

Within the framework of the "Fractal" Development Work, the software ExpertTool 2.3.4.2 was created, which allows manual semantic markup of content using the TAPAZ technology. A manual of the semantic preprocessing of ERS-content has been prepared for experts in order to unify manual conjugation of specialized ERS-processes with TAPAZ macro-processes. The following were developed: 1) an algorithm for the formation of TAPAZ-units; 2) an algorithm for an expert to extract specialized terminology from thematic ERS-content to fill the Knowledge Base; 3) the algorithm for updating the user request to the system at the request of the system. The properties of the TAPAZ-2 Semantic Classifier Graph as a constructor of the Knowledge Base Architecture of an Intelligent WEB-system and ways to reduce the number of its vertices for sequential generation, processing and storage of the Graph taking into account the capabilities of modern computing technology are determined. A prototype of the TAPAZ-2 dictionary was prepared in the form of an Excel table, including 416 of the most frequency specialized predicates of Remote Sensing and 516 of their uses, coupled with 112 macro-processes of the TAPAZ-2 Semantic Classifier. The semantic structure and semantic functionality of the System are proposed and justified, 18 typical roles of event participants were decoded.

The ultimate goal is automatic semantic markup of ERS-content in the TAPAZ technology, which allows to achieve maximum efficiency (accuracy, speed and completeness) of the search engine, as well as the automatic assembly of TAPAZ-units in the Knowledge Base for the analytical support of management, design-and-search and expert solutions remote sensing tasks.

One more trendy approach to semantic knowledge representation for Artificial Intelligence was formulated by prof. V. Golenkov. He proposed the concept of the Open Semantic Technology for Intelligent Systems Design (OSTIS Technology) [14, 15], focused on the development of computer systems of the new generation (first of all – hybrid intelligent systems [16]) which will be called semantic computer systems or ostis-systems, if it is necessary to emphasize their compliance with the standards of OSTIS Technology. The model of hybrid knowledge bases of ostis-systems and models for representing various types of knowledge within the framework of such a knowledge base [17], as well as model of a hybrid problem solver, which allows to integrate various problem solving models [18], were also proposed.

The main requirement for OSTIS Technology is to ensure the possibility of joint use within the ostis-systems of various types of knowledge and various problems solving models with the possibility of unlimited expansion of the list of knowledge used in ostis-system and problem solving models without significant labor costs. The consequence of this requirement is the need to implement the component approach at all levels, from simple components of knowledge bases and problem solvers to whole ostis-systems. To meet these requirements, the most important task is not only the development of appropriate ostis-systems models and their components, but also the development of an integrated methodology and appropriate tools for automating the construction and modification of ostis-systems [19].

Prof. Golenkov stated also the principle of unification

[2]"Functional concepts such as "Subject", "Predicate" must be carefully distinguished from categorical concepts, such as "Noun component", "Verb", and the difference between them should not be masked by the fact that sometimes the same term is used for concepts both types". Chomsky N. (1972) Aspects of the theory of syntax, Moscow, Moscow State University, p. 65 (in Russian).

[3]Macro-process is one of 112 extremely abstract processes that are isomorphic to any subject domain and are calculated and encoded by the TAPAZ-algebra.

[4]"There are such concepts as "culprit", "tool", "product of labor" <...> We are here in the field of various categories, apparently ontological, but essentially semantic". Kotarbinski T. (1975) Treatise on the Good Work. Moscow, Economics, p.31 (in Russian).

of information representation in the memory of computer systems and developing the standard that should not limit the creative freedom of the developer, but guarantee the compatibility of its results. As the standard of the universal sense representation of information in the memory of computer systems he proposed SC-code (Semantic Computer Code). Unlike USC of V.V. Martynov, who made an important step in creating a universal formal method of sense coding of knowledge, it, firstly, is non-linear in nature and, secondly, is specifically focused on coding information in the memory of computers of a new generation, focused on the development of semantically compatible intelligent systems and called semantic associative computers. Thus, the main leitmotif of the proposed sense presentation of information is the orientation to the formal memory model of a non-Von-Neumann computer designed for the implementation of intelligent systems using the sense representation of information. The features of this representation are as follows:

- associativity;
- all information is enclosed in a connections configuration, i.e. processing information is reduced to the reconfiguration of connections (to graph-dynamic processes);
- transparent semantic interpretability and, as a result, semantic compatibility [19].

Taking into account the trends of the AI approaches development, it is proposed to strengthen the linguistic database [20] with the help of approaches mentioned above in the context of solving the task of automatic recognition of semantically adopted fragments of the text documents. The examples of the ostis-knowledge base that describe the knowledge domain of grammatical categories in terms of TAPAZ, available to use in implementation of algorithms of the natural language processing at different levels of the deep of the analysis, are shown in Figures 1 – 3.

IV. Conclusion

An involvement of the well-developed linguistic text analysis that is based on the knowledge of natural language, together with the implementation of the principles of unification of information representation in the memory of computer systems and semantic knowledge representation according to the Theory for Automatic Generation of Knowledge Architecture might made a contribution to maintenance the trust in science and the development of national science in the sphere of the Artificial Intelligence.

References

[1] Ethics in Higher Education. Available at: https://www.igi-global.com/gateway/chapter/114343. (accessed 2019, Dec).

Figure 1. The class of base semantic categories

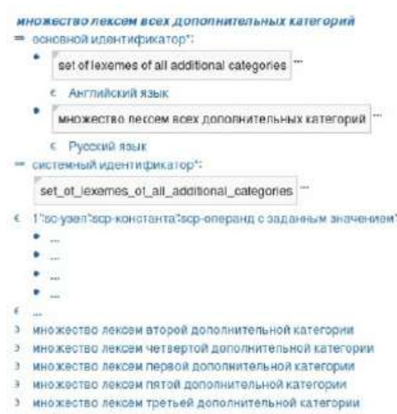


Figure 2. The class of supplementary semantic categories.

[2] Promoting research integrity for a ... - Information Centre - Research & Innovation - European Commission. Available at: https://ec.europa.eu/research/infocentre/article-en.cfm?artid=49608. (accessed 2019, Dec).

[3] Robert K. Merton, The Normative Structure of Science (1942). Available at: https://panarchy.org/merton/science.html. (accessed 2019, Dec).

[4] Merton's norms and the Scientific Ethos. Available at: https://www.futurelearn.com/courses/open-social-science-research/0/steps/31422. (accessed 2019, Nov).

[5] Carroll, J., & Appleton, J. Plagiarism: A good practice guide. Available at: http://www.jisc.ac.uk/uploaded-documents/brookes.pdf. (accessed 2019, Dec).

[6] Sarlauskiene, L., Stabingis, L. Understanding of plagiarism by the students in HEIs of Lithuania / *Procedia - Social and Behavioral*

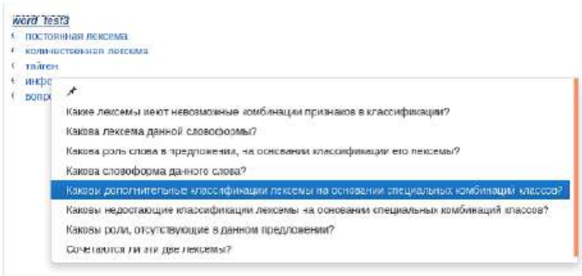


Figure 3. The way of communication with the help of proper intellectual agents of natural language text processing.

Sciences, 2014, pp.638 – 646.

[7] Yakugaku Zasshi.Promoting Research Integrity. Available at: doi: 10.1248/yakushi.17-00181-3. (accessed 2019, Dec).

[8] Palazzo, G. Organizational integrity—understanding the dimensions of ethical and unethical behavior in corporations, *Corporate ethics and corporate governance*, 2007, pp. 113-128.

[9] Paine, L. Managing for organizational integrity, *Harvard business review*, 1994. vol 72, no. 2, pp. 106-117.

[10] Krapivin, Y. System «PlagiarismControl» as the Tool for the Expertise of the Text Documents, *Informatics*, 2018, vol. 15, no. 1, pp. 103–109 (in Russian).

[11] Krapivin, Y.B. Metody i algoritmy avtomaticheskogo raspoznavanija vosproizvedennyh fragmentov tekstovyh dokumentov: *avtoref. dis. kand. teh. nauk: 05.13.17*, Bel. gos. un-t., - Minsk, 2019. - 24 p.

[12] ACL 2019 . Available at: http://www.acl2019.org/EN/index.xhtml. (accessed 2019, Dec).

[13] Hardzei, A. Theory for Automatic Generation of Knowledge Architecture: TAPAZ-2, Minsk, 2017, 50p.

[14] Golenkov, V. V., Guljakina, N. A."Proekt otkrytoj semanticheskoj tehnologii komponentnogo proektirovanija intellektual'nyh sistem. chast' 1: Principy sozdanija [project of], open semantic technology for component design of intelligent systems. part 1: Creation principles, *Ontologija proektirovanija [Ontology of design]*,2014, no. 1, pp. 42–64.

[15] Golenkov, V. [et al.] "From training intelligent systems to training their development tools," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, Ed., BSUIR. Minsk , BSUIR, 2018, pp. 81–98

[16] Kolesnikov, A. Gibridnye intellektual'nye sistemy: Teoriya I tekhnologiya razrabotki [Hybrid intelligent systems: theory and technology of development], A. M. Yashin, Ed. SPb.: Izd-vo SPbGTU, 2001.

[17] Davydenko, I. "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2018, pp.99–118.

[18] Shunkevich, D. "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2018, pp.119–132.

[19] Golenkov, V. [et al.] "Methods and tools for ensuring compatibility of computer systems," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2019, pp. 53–90.

[20] Krapivin, Y. Funktsional'nost' cross-language v zadache avtomaticheskogo raspoznavaniya semanticheski ekvivalentnykh fragmentov tekstovykh dokumentov [Cross-language Functionality in the Problem of the Automatic Identification of the Semantically Equivalent Fragments of the Text Documents], *Iskusstvennyi intellect [Artificial Intelligence]*, 2013, vol. 62, no. 4, pp. 187–194.

Перспективные подходы к семантическому представлению знаний и их приложения в контексте задачи автоматического распознавания семантически эквивалентных фрагментов текстовых документов

Гордей А.Н., Крапивин Ю.Б.

В статье представлено решение проблемы поддержания доверия к науке в контексте задачи автоматического распознавания семантически эквивалентных фрагментов текстовых документов. Решение основывается на реализации принципов унификации представления информации в памяти компьютерных систем и представления семантических знаний в соответствии с теорией автоматического порождения архитектуры знаний (ТАПАЗ).

Ключевые слова: естественный язык, автоматическая обработка текста, достоверность научного исследования, заимствованный фрагмент.

Received 06.02.2020

Principles of decision-making systems construction based on semantic image analysis

Natallia Iskra
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus
Email: niskra@bsuir.by

Vitali Iskra
Omnigon Communications LLC
New York, NY, USA
Email: iskra.vitaly@gmail.com

Marina Lukashevich
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus
Email: lukashevich@bsuir.by

***Abstract*—In this paper principles of decision-making systems construction are considered. An approach to image analysis based on semantic model is proposed and studied. The results show an improvement in processing speed and image captioning quality based on Visual Genome dataset.**

***Keywords*—decision-making, video surveillance, neural networks, semantic analysis, image captioning**

I. Introduction

The task of image analysis in decision-making systems using technical vision today is acute. Automatic image interpretation is a non-trivial task. For example, for a video surveillance system, it would be relevant not only to record and save video, but also to analyze what is happening, as well as to signal any suspicious situations - violations, incidents, actions that require an immediate response.

The approach to image analysis considered in this paper proceeds as follows:

Step 1. Individual objects detection. These can be the objects that are significant in the context of the system (for example, traffic participants, road markings and signs in traffic monitoring systems), areas that outline objects (bounding boxes), or more precise object-by-pixel selection.

Step 2. Building a semantic model. At this stage, relations between objects and / or attributes of individual objects are formalized.

Step 3. Model interpretation. According to the constructed model, a textual description of what is happening (an annotation of the image or image caption, for example, for keeping a surveillance log) can be obtained, or specific situations on the image that are of interest (for example, cases of traffic rules violation, traffic accidents, etc.) can be determined. In this case, the interpretation of the model will consist in highlighting only those relationships and attributes that can signal of an abnormal situation.

The most important part in a situational analysis implementing is the construction of an interpretable image model. Modern approaches to models construction have a large number of limitations. In this article, the main focus is on the methodology for constructing this model, the selection of an algorithm for detecting objects in an image, as a preliminary stage of building a model, as well as the principles of quality analysis of the constructed model and decision-making based on it. To represent the obtained model and implement the decision-making process on the basis of the obtained model, it is proposed to use the approaches developed in the framework of OSTIS Technology.

II. Methods overview

A. Object detection in images

The first step during the image analysis is to process the source image and detect the objects automatically. During this step one of the following tasks is performed as a rule [1]:

- Semantic Segmentation – for every pixel in the source image determine its class and category;
- Classification and Localization – determine the class of a single object in the image and its exact location;
- Object Detection – define a class and a rectangular area bounding each of the objects in the image;
- Instance Segmentation – on the image with multiple objects determine the contours (all visible pixels) and the class of each of the objects.

From the standpoint of a semantic model construction last two tasks are of the most interest.

Among the existing modern object detection algorithms, including those based on deep learning methods, the most relevant approaches are:

- Sliding Window [2];
- Region Proposals [3];
- Single Shot Detection [4].

Each of the approaches has its own advantages and disadvantages, in terms of their relevance to application in systems that include image analysis [5].

To construct the model, described in this paper, it seems to be the most promising to use methods, based on the group of neural networks architectures with region proposals, so-called R-CNN, and their development:

- R-CNN [3] – represents a model of sequential image processing pipeline: generation of a set of regional proposals, the use of a pre-trained convolutional neural network with a final layer of support vectors and linear regression for a more accurate regions estimation;
- Fast R-CNN [6] – a model in which, to speed up the performance of the previous processing pipeline, a selection of regions and the union of all neural network models into one are used;
- Faster R-CNN [7] – to accelerate the model even further, a selective search of regions is used;
- Mask R-CNN [8] – unlike previous models, this one uses a binary mask to determine not just a rectangular region - a candidate for objects, but specific pixels belonging to the object, which, in essence, is the solution to the image segmentation problem described above.

B. Semantic model

Within the framework of modern approaches as the basis of the semantic image model the so-called Scene Graph [9] is widely used. A scene graph is a data structure that describes the contents of a scene, which, in turn, can be specified by an image or its textual description. In the scene graph instances of objects their attributes and relationships between objects are encoded.

Formally, a scene graph is defined as follows: let C be a set of object classes, A – a set of attribute types, R – a set of relation types. A scene graph is defined as $G = (O, E)$, where $O = \{o_1, \ldots, o_n\}$ – a set of objects – nodes of a graph, $E \in O \times R \times O$ – a set of graph edges. Every object is represented by $o_i = \{c_i, A_i\}$, where $c_i \in C$ – the class of the object, and $A_i \in A$ – its attributes.

A scene graph can be grounded to an image. Let B be a set of rectangular areas, each of which delineates a certain object in the image (they are generally called Bounding Boxes), then the grounding of the scene graph *G = (O, E)* to the image is the function $\gamma : O \rightarrow B$, or γ_o.

To conduct the experiments the dataset Visual Genome [10] is commonly used. It consists of 108 077 labelled images, for which 5.4 millions of the textual descriptions and scene-graphs for the whole images and their sections (regions) were produced using crowd sourcing. All the scene graphs are grounded to either textual descriptions, or images (regions), or both.

The example of scene graph – region grounding in Visual Gnome is shown in Fig. 1.

Grounding scene graphs to a textual descriptions (each object, attribute and relation) in Visual Genome corresponds to WordNet synset [12]. WordNet – network word representation,that is structured according to semantic relations. In WordNet each word is represented as a set of its synonymous meanings, which are called synsets. Each synset comprises of a triplet <word>.<pos>.<number>, where word – is a word itself, pos – its part of speech (n – noun, v – verb, a – adjective, r – adverb), number – index of the meaning in the set. E.g. the term "person" in WordNet is represented by three meanings person.n.01, person.n.02 and person.n.03. Textual grounding of the object "person" in Visual Genome corresponds to the synset person.n.01. In WordNet there are relations of synonymy, antonymy, "part – whole" (meronym – holonym), "general – specific" (hypernym – holonym).

Using a graph representation to describe an image model has a number of significant advantages compared to more traditional approaches to image captioning aimed toward a natural language description (considered in [13] and other works). The graph representation is more unambiguous (invariant) and is much better suited for automatic processing, and, in particular, the interpretation of such a model.

However, despite these advantages, the currently used approach to the scene graph construction has a number of disadvantages that make it difficult to interpret image models presented in this form. The key disadvantage, in our opinion, is the lack of any semantic unification (standardization) in the principles of building scene graphs, in particular, in the principles of distinguishing relations and attributes (and, generally speaking, in this case, there is no clear boundary between the concepts of relation and attribute), in the framework of even one data set, as well as the lack of syntactic unification in the representation of scene graphs in various approaches. In addition, in modern approaches to the construction of scene graphs, as a rule, the problem of internationalization still remains.

In turn, the lack of unification in the representation of scene graphs makes it impossible to build universal processing tools for such graphs, in particular, means for verifying and decision making based on scene graphs.

An ontological approach is currently used as the basis for solving the problem of unification in various fields. In this paper, to implement this approach, it is proposed to use the OSTIS Technology, within the framework of which a unified standard for coding information (SC-code) is proposed [14], the problem of unification of the principles for representing different types of knowledge [15] and the problem of integrating various models for problem solving [16] are solved .

We list some of the advantages of OSTIS Technology that are relevant in the context of solving the problem posed in this paper:

- unification of the representation of semantic image models;
- ensuring the independence of the image model from the external language in which the preliminary description was made;
- the possibility of structuring the model according to various criteria, as well as the possibility of representing meta-information, which, in particular,

Figure 1. An example of an image from Visual Genome with grounding [11].

will allow us to select image regions that are more or less significant for solving the current problem;

- the availability of verification tools and adjustments to the image model, partially considered in [17], which make it possible to verify the model itself for its internal consistency and adequacy of the subject area, and in the future will automatically evaluate the degree of conformity of the automatically obtained model to standard models developed by experts;
- the presence of a large amount of source data that has been accumulated at the moment and which may be useful for further research makes the task of automating the transition from scene graphs, for example, from the Visual Genome dataset, to a unified semantic representation in relevant SC-code.

III. Semantic model construction technique

To build a semantic image model in the form of a scene graph, we must first detect the objects in the image, and then for each pair of objects decide whether they have a relations and which ones [18]. The selection of relations can be greatly simplified by using external knowledge bases (general purpose or specialized for a specific domain) [17]. In both cases, for the image on which n objects are found, it is necessary to consider (n^2 – n) relations. In this paper it is proposed to simplify the solution by identifying the so-called salient (significant or the most relevant) objects [19], and to further consider (n – 1) relationships. This approach corresponds to the scenario of tracking certain objects or situations in surveillance systems.

Frequency analysis of Visual Genome data shows that the most frequent relationships between objects in images are spatial relationships: the "on" relationship occurs 642,150 times, the "in" relation – 218,166, "behind" - 39,833. In addition, due to the hierarchical structure of WordNet grounding, spatial relationships can be described in more detail: for example, "car in a parking lot" or "car is parked in a parking lot". Indeed, when looking at an image, a person first of all notes how the objects are located relative to each other in space. In automatic processing it is also possible to determine semantic spatial relations between objects [20]. In addition, reducing the set of relations of the image model to spatial relations will allow at the current stage to significantly simplify the process of constructing and interpreting the model, while maintaining the ability to assess the anomalies and oddities.

The technique for automatic model construction for spatial relations is presented below.

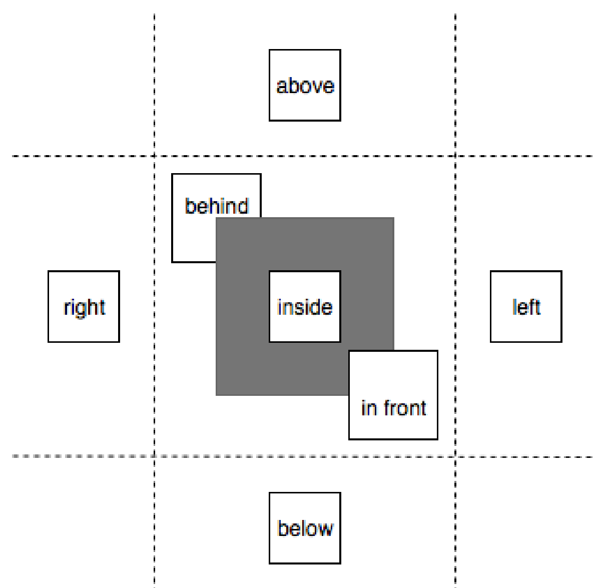


Figure 2. Spatial relations system.

In Fig. 2 the system of all possible spatial relations is visualized: the area of the salient object (subject) is filled, all the other areas are the options for the location of the object of interest (object), for which, using the decision tree in Fig. 3, the type of spatial relationship in the form “subject-relation-object” will be determined.

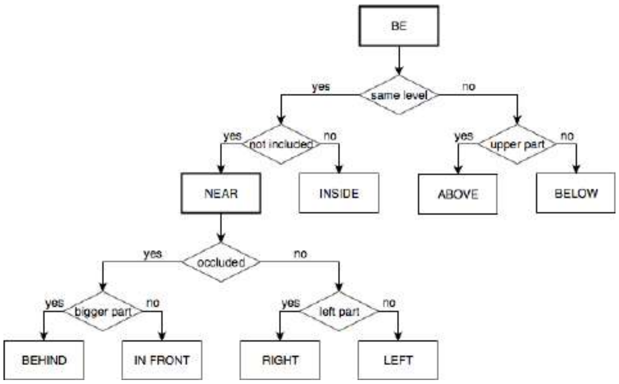


Figure 3. The decision tree for automatic spatial relations model construction.

It should be noted that for the names of types of relations in this model prepositions are used (and prepositions, as it was described above, are not represented in WordNet), i.e. at this stage, grounding to WordNet is not possible, but at the next step (for interpretation), synsets containing these prepositions and their meanings (be.v.01, along.r.01, etc.) will be used.

In the decision tree the rectangles show the blocks corresponding to a certain type of relationship, while more general relationships that need to be specified are highlighted (similar to hypernyms from WordNet). When constructing a tree to speed up the process of final decision-making, the rules (shown in the figure by rhombus’s) were formulated in accordance with the statistical data of the Visual Genome analysis, so that a more frequent case would be to the left of the tree. So, in the initial dataset, the“near” relationship is found more often than other spatial relationships (26,316 times), the “above” is significantly more common than the “below” – 13,767 times and 3,349 times respectively etc.

The implementation of the method used for the experiments described below first detects objects using the Faster R-CNN method, determining the classes of objects and their bounding boxes. The salient object is determined as the object with the largest bounding box.

In natural images the boundaries of the object regions, as a rule, intersect. If the intersection of the regions of the salient object and the object of interest is less than 50% of the area of the object of interest region, the relations corresponding to the decision rule are selected from the set “top”, “bottom”, “left”, “right” (that is, it is considered that there is no intersection). At an intersection of more than 50%, the ratio is selected based on a comparison of the pixel masks of the objects obtained by applying Mask R-CNN to the object regions: if there are more pixels of a significant object in the intersection zone, the “back” relation is selected, and the “in front” relation is the opposite case.

To describe spatial relationships in the framework of OSTIS Technology, a corresponding ontology was developed, the description of which and usage examples are presented in [20].

IV. Experiments

A. Experimental evaluation of model construction

For experimental evaluation of the semantic model construction technique from Visual Genome dataset the subset of images was selected. It is a sample of images in which each of the relations under consideration is represented by 50 regions with a grounding to the image (i.e. 50 regions for the relation “above”, 50 regions – “below”, etc. – the total of 350 regions). The examples of the images are given in Fig. 4 - 5.

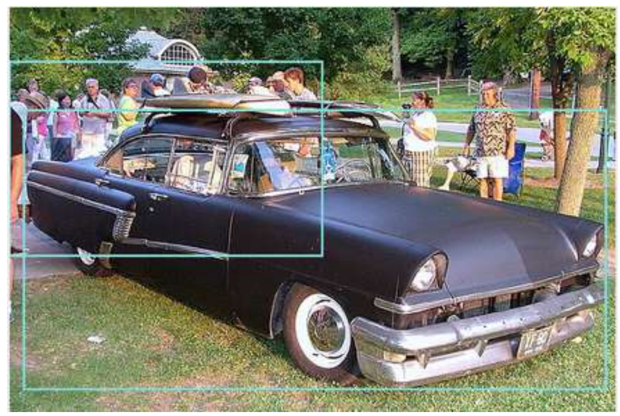


Figure 4. The example of the region grounding for “crowd behind car”.

Figure 5. The example of the region grounding for “car below trees”.

The size of the experimental set is relatively small, since it was planned to manually verify the results of determining the relations in order to evaluate not only the accuracy of the model, but also the “gameability” of the obtained results, i.e. to exclude situations where a high indicator of the quality assessment metric (the correct result) may correspond to an expression that a human expert considers “unnatural” (for example, “the sky is blue in color” instead of “the sky is blue”) [21].

In the experiment, relationships in the selected regions are automatically determined and the results are compared with the reference (given in the dataset) and evaluated by experts (see Table I).

B. Experimental evaluation of model interpretation

To experimentally evaluate the interpretation of the constructed model for the set of regions, textual descriptions of the regions are generated by replacing the relationships with the most common synonyms from WordNet (for example,“car below tre” turns into “car parked under tree”) and the resulting annotations are compared with the reference using the METEOR metric [13].

The annotation results are also compared with the results obtained using a combined neural network [22] and purely convolutional neural network [23] approaches to annotating image regions without constructing a semantic model (Table III).

As mentioned earlier, the description of the image model in the form of natural language text has a number of significant drawbacks. The rejection of such a description and the transition to graph representation leads to the need for a transition from classical text metrics (METEOR [13], etc.) to metrics that allow us to evaluate the similarity of graph models.

A graph representation makes it possible to simplify the comparison of two models at the syntactic level, however, problems related to the semantic data presented remain urgent, which in the textual presentation faded into the background due to the large number of problems associated with the presentation form.

In general, we can distinguish the following levels of complexity of situations that arise when comparing graph image models:

- the system of terms and the system of concepts (logical ontology) coincide. In this case, the comparison is reduced to the search for isomorphic fragments, however, the problem of assessing the significance of each fragment remains relevant;
- the system of terms does not coincide, but the system of concepts coincides, i.e. the same concepts are used, but they can be named differently (for example, in the context of a street situation, the meaning of the words “car” and “automobile” will coincide). In this case, the identification and gluing of the same concepts, named differently, is additionally required. In the general case, this problem concerns not only concepts, but also specific entities;
- the system of concepts does not match. In this case, the alignment of systems of concepts is additionally required, in this case involving the definition of concepts used in the evaluated model through the concepts used in the example model.

The indicated levels relate to the case when strict coincidence of models is evaluated to the level of specific objects and concepts, however, when interpreting a model it is often enough to use a generalized model (for example, in a situation “a person sits on a chair” and a “person sits in an armchair” it is often important that a person sits and it doesn’t matter where). Thus, the task of generalizing models with subsequent comparison is also relevant. In classical textual approaches, a similar problem is partially solved by identifying synonyms.

Using OSTIS Technology to represent image models and construct relevant metrics has several advantages, in particular, one of them is the availability of means for structuring information and representing meta-information. Fig. 6 shows an example representation of similarities and differences for the two pieces of information presented in SCg-code [24].

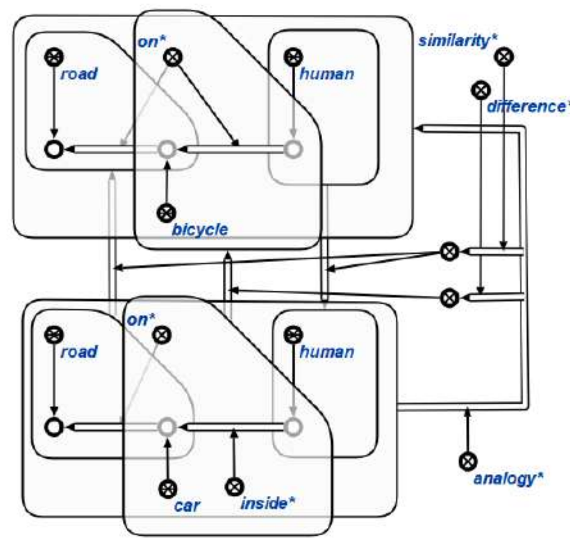


Figure 6. Representation of similarities and differences in SCg.

Note that the development of graph metrics based on OSTIS Technology will allow them to be used for other tasks, for example, to assess the quality and completeness of a student’s response in intelligent training systems, to assess the contribution of each developer to a common knowledge base, etc.

V. Results and discussion

In Table I the results of semantic model construction evaluation are shown. Object detection is considered correct, if the class labels match, the differences in bounding boxes estimation are considered insignificant in given context.

In Table II the results of spatial relations matching are shown.

Table I
Model construction evaluation.

	Number	%
Set size (relations/objects)	350/700	100
Object detection (RCNN-based)	687	98.1
Relations (dataset match)	335	95.7
Relations (visual analysis)	340	97.18

Table II
The analysis of spatial relations estimation.

Spatial relation	Visual analysis (for 50)	Model (for 50)
BEHIND	49	44
IN FRONT	48	45
RIGHT	50	50
LEFT	50	50
INSIDE	50	50
ABOVE	49	48
BELOW	49	48

In Table III the results of image captioning evaluation are shown.

Table III
Region captioning results evaluation

Coder model	METEOR
CNN + RNN [22]	0.305
TCN [23]	0.290
Semantic model	0.515

As shown in the table, the use of a semantic model for encoding information from an image significantly exceeds neural network models when constructing meaningful phrases that describe regions. According to the METEOR metric, which takes into account not only the structure of the annotation, but also its semantic variations, the proposed method shows the results by more than 60 % better than the neural network approaches.

VI. Decision making based on semantic model

To make decisions on the basis of the proposed model at this stage (with a small number of classes of objects and relations between them), a general mechanism can be used, which was examined in detail, in particular, in [25]. The specified technique assumes a reliable logical conclusion based on the logical rules and ontology of contingencies available in the knowledge base, where for each class some recommendations are assumed.

Let us consider in more detail the example of decision-making. The Fig. 7 shows the image from the surveillance camera, on which the regions of objects detected are highlighted. For convenience, some regions are omitted (in the current implementation, the detector on this image detected 25 people, 7 cars, 4 umbrellas, 2 backpacks and 4 bags).

According to the technique described above, a salient object, i.e. the key subject of the relationship, is an instance of the id1 class “pedestrian crossing” (label “crosswalk”, synset crossing.n.05). In the current implementation, this is due to the fact that it has the largest size, but subsequently the application of the ontological approach will also allow contextual information to be taken into account.

The following objects of the corresponding classes were detected in the image:

- id2, id5, id6 – class “car”
- id3, id4, id7, id8, id9 – class “person”

According to the technique for constructing a model based on existing intersections of regions, the following relationships between pairs of objects are estimated:

1) id2 ->id1: “on”
2) id3 ->id1: “on”
3) id4 ->id1: “below”
4) id5 ->id1: “on”
5) id6 ->id1: “above”
6) id7 ->id1: “inside”
7) id8 ->id1: “below”
8) id9 ->id1: “on” (detection error due to camera placement)

In the form of SCg language this model is presented as in Fig. 8.

Based on estimated relations the following captions can be generated:

1) car on crosswalk - car is parked on crosswalk
2) person on crosswalk - person is crossing the road on crosswalk

Based on the of detected objects and relations, decisions in this example are made in the context of “people cross the road, cars let pedestrians pass”. Thus, normal (regular) situations for “person” with respect to “crosswalk” – “on” and “inside”, for “car” with respect to “crosswalk” – the opposite.

The example of formal rule in SCg language is shown in Fig. 9.

Using a rule base, applying the simple inference mechanisms, the following contingencies can be distinguished:

1) traffic rules violation: car on the crosswalk – in pairs 1 and 4
2) traffic rules violation: a person is using a crosswalk – in pairs 4 and 7

Rule Violation in pair 8 will not be determined at the moment, due to the camera placement. To prevent this mistakes, it is possible to detect not regions, but masks, however, in this case, image processing will take much longer.

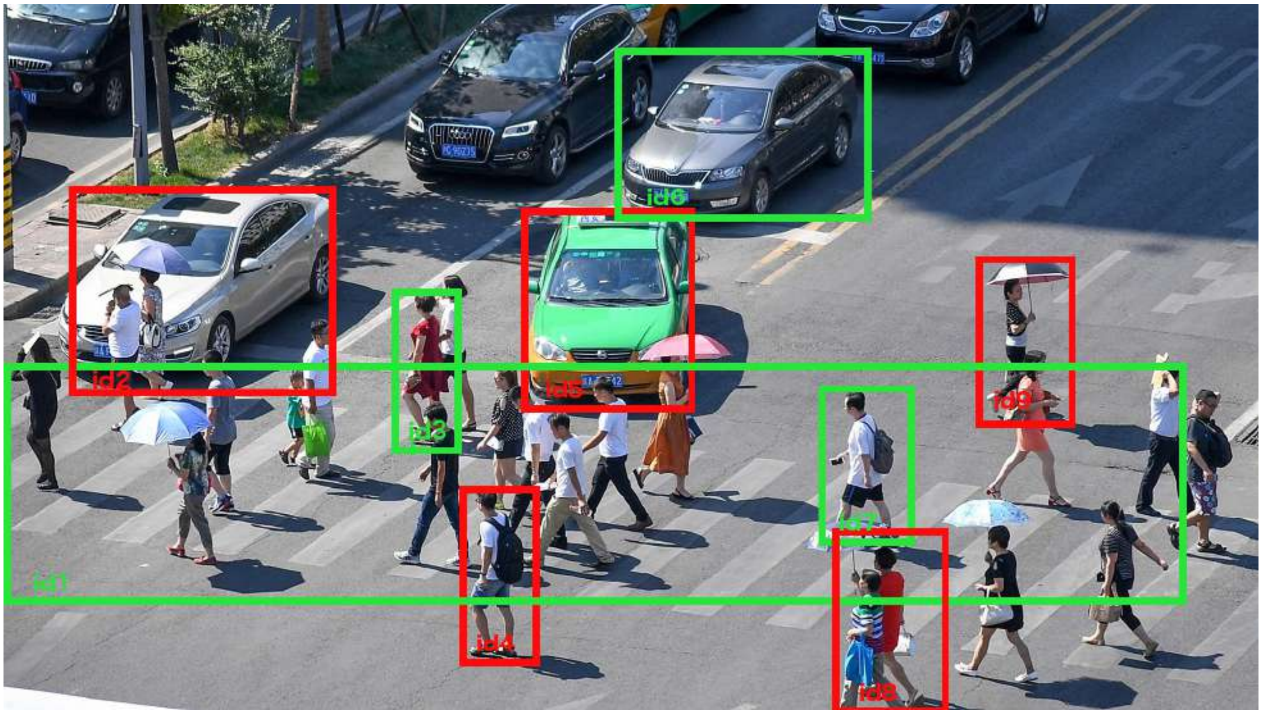


Figure 7. The example of the image for decision-making.

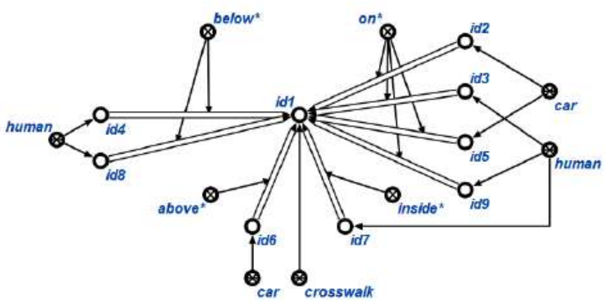


Figure 8. Semantic model of image in SCg.

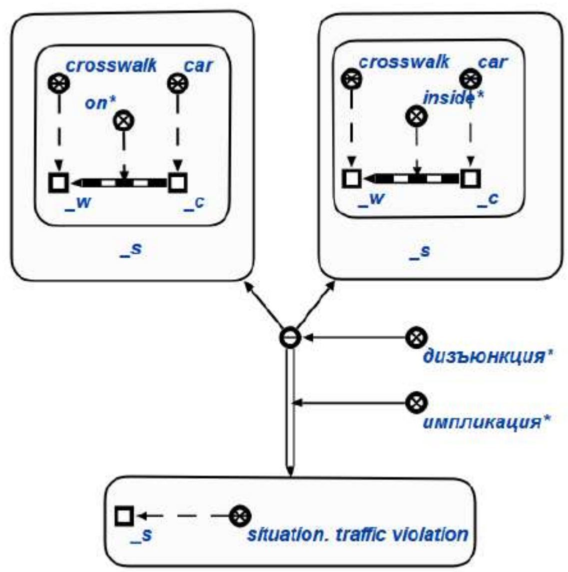


Figure 9. Example rule for decision making.

VII. Conclusion

Thus, the proposed method of constructing a semantic model analyzes less relationships between objects, which can significantly reduce the image processing time on test sets from the Visual Genome dataset and improve the quality of annotation.

It should be noted that this approach contains simplifications - the largest of the objects is considered salient and only relations between two objects are considered (i.e. only fragments of a scene-graph), also attributes of objects are not taken into account.

In further work, it is planned to use more complex approaches to determining a salient object (including based on specific subject area), the complete construction and analysis of graph scenes.

In turn, the use of OSTIS Technology to represent the model and implement the decision-making mechanism makes it possible to ensure the modifiability and scalability of the system, built on the basis of the approaches proposed in this paper, which in the future will allow to eliminate the described limitations.

Acknowledgment

The research presented in this paper was conducted in close collaboration with the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics. Authors would like to thank the research group of the Department of Intelligent Information Technologies for productive cooperation.

References

[1] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," *arXiv preprint arXiv: 1809.03193*, 2018.

[2] J. Müller, A. Fregin, and K. Dietmayer, "Disparity sliding window: object proposals from disparity images," in *IEEE/RSJ International conference on intelligent robots and systems*, 2018, pp. 5777–5784.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, Cham, 2016, pp. 21–37.

[5] P. S. Hursov and N. A. Iskra, "Algoritmy detektsii ob'ektov dlya analiza izobrazhenii [Object detection algorithms for image analysis]," in *Informatsionnye tekhnologii i sistemy: materialy mezhdunarodnoi nauchnoi konferentsii [Information Technologies and Systems: materials of the international scientific conference]*, Minsk, 2019, pp. 128–129, (In Russ).

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, pp. 91–99, 2015.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[9] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[10] "Visual Genome," Available at: https://visualgenome.org.

[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, and M. S. Bernstein, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, no. 123(1), pp. 32–73, 2017.

[12] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[13] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[14] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, "Methods and tools for ensuring compatibility of computer systems," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[15] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 99–118.

[16] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," in *Open semantic technologies for intelligent systems*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.

[17] N. Iskra, V. Iskra, and M. Lukashevich, "Neural network based image understanding with ontological approach," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system: materialy mezhdunarodnoj nauchno-tekhnicheskoj konferencii [Open semantic technologies for intelligent systems: materials of the international scientific and technical conference]*, Minsk, 2019, pp. 113–122.

[18] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the european conference on computer vision*, 2018, pp. 670–685.

[19] A. Borji, M. M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, pp. 1–34, 2014.

[20] N. A. Iskra, A. L. Mezhen', and D. V. Shunkevich, "Ontologiya predmetnoj oblasti prostranstvennyh sushchnostej dlya sistemy semanticheskogo analiza izobrazhenij [Ontology of the subject area of spatial entities for the system of semantic image analysis]," in *Informatsionnye tekhnologii i sistemy: materialy mezhdunarodnoi nauchnoi konferentsii [Information Technologies and Systems: of the international scientific conference]*, Minsk, 2019, pp. 112–113.

[21] D. Shunkevich and N. Iskra, "Ontological approach to image captioning evaluation," in *Pattern Recognition and Information Processing: Proceedings of the 14 international conference*. Minsk: Bestprint, 2019, pp. 219–223.

[22] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4565–4574.

[23] N. Iskra and V. Iskra, "Temporal convolutional and recurrent networks for image captioning," in *Pattern Recognition and Information Processing: Proceedings of the 14 international conference*. Minsk: Bestprint, 2019, pp. 346–349.

[24] "IMS metasystem," Available at: http://ims.ostis.net/, (accessed 2020, Jan).

[25] V. Golovko, A. Kroshchanka, V. Ivashenko, M. Kovalev, V. Taberko, and D. Ivaniuk, "Principles of decision-making systems building based on the integration of neural networks and semantic models," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 91–102.

Принципы построения систем принятия решений на основе семантического анализа изображений

Искра Н.А., Искра В.В., Лукашевич М.М.

В статье рассматриваются принципы построения систем принятия решений. Предложен и изучен подход к анализу изображений на основе семантической модели. Результаты показывают улучшение скорости обработки и качества аннотирования на основе набора данных Visual Genome.

Received 08.02.2020

The principles of building intelligent speech assistants based on open semantic technology

Vadim Zahariev, Daniil Shunkevich, Sergei Nikiforov, Timofei Lyahor, Elias Azarov
Belarussian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: zahariev@bsuir.by, shunkevich@bsuir.by, nikiforov.sergei.al@gmail.com, linoge@bsuir.by, azarov@bsuir.by

***Abstract*—The speech interface is one of the most natural and convenient ways of organizing user interactions with intelligent systems. Nowadays universal voice assistants are implemented by the largest world companies for various platforms and have gained popularity among users.**

The aim of this work is to develop the principles of building voice assistants for intelligent computer systems based on OSTIS Technology [8]. The main requirements in this case are ensuring maximum independence from the subject area and providing the possibility of adaptation (primarily automatic) of the universal assistant for a specific user and the features of a specific subject area.

***Keywords*—dialogue system, voice assistant, semantic network, spoken language understanding**

I. Introduction

The actual field of application of voice user interfaces in modern information and communication technologies is the development of voice assistants - dialogue systems that allow users to access knowledge and interact with agents of the [11] intelligent system in a user-friendly manner.

Universal voice assistants developed by the world's largest companies are widely used in modern smartphones and operating systems. Their development trends lie on the transformation of stand-alone applications into platforms that can be deployed on devices of various manufacturers, and on the basis of which various services built on the ground of intelligent information technologies [14], [15]. This fact allows manufacturers of both electronics and software to open new niches in consumer markets using the latest advances in artificial intelligence technology.

Analytical agencies forecast that the combined annual growth rate of the global market for products using speech assistant technology will be more than 30%, increasing from \$1.2 billion in 2018 to \$5.4 billion by 2024 [7]. According to their estimates, this trend will contribute to two factors, such as an increase in the total number of smartphones and the expansion of the standard functions of speech interfaces: managing the dialogue context, personalization, the ability to conduct dialogue in several languages or respond not only in voice but also in text mode.

II. Analysis of existing solutions and problem statement

Consider the architecture and principles of the voice assistant, peculiar to most modern solutions. In our description, we will focus on the voice assistant "Alexa" from the company "Amazon". This product is currently the most popular assistant (with the exception of mobile phones) in the speech technology market (about 65%) [1], based on a modern stack of speech technologies. Many other major players in the speech market such as "Google" ("Google Assistant"), "MicroSoft" ("Cortana"), "Yandex" ("Alice") are trying to adopt solutions specific to "Alexa" [10] in their products. Therefore, the architecture under consideration can be considered typical.

Modern voice assistant forms a distributed software and hardware system consisting of two main parts: front end and back end (Fig. 1).

The front end part is deployed on a specialized device or installed as an application on the user's device. The client is responsible for issues related to capturing and playing back audio information, pre-processing, wake-word system triggering activation, encoding and decoding data, and generating backend requests. Access to the server part is carried out through the corresponding program interface, most often REST [4].

The back end part includes following main components.

- spoken language understanding subsystem (SLU), which consists of automatic speech recognition module (ASR) and natural language understanding (NLU) module;
- dialogue management module (DM) includes a subsystem of "Skills" (general type like weather, music, navigation and specialized like web search, wiki, news etc.).
- natural language messages generator (NLG) module.

Let's review server components in detail. The Speech Understanding Subsystem (SLU) converts the input signal into some semantic equivalent of the spoken phrase - "intent", which reflects the meaning of the spoken phrase and the user's expectations. This subsystem includes:

The automatic speech recognition module (ASR) of speech implements the process of converting speech into text. Acoustic and linguistic models are used to describe this process. Algorithms that implement the corresponding models make it possible to determine fragments of the speech wave in the audio signal equal to the basic phonetic units of the target language and form a phonetic text from them. Then a spelling text is obtained which is based on the relevant rules of morphology and grammar. Modern ASR implementations involve the use of a combining statistical methods such as hidden Markov models (HMM) of neural network methods based on convolutional (CNN) and recurrent neural networks (LSTM-RNN) [5], [17].

The natural language understanding module (NLU) implements a natural language text processing sequence that includes the main stages: tokenization, defining the boundaries of sentences, parts of speech tagging, named entities recognition, syntactic and semantic analysis. At the output of this module, an entity is formed called "intent", which is a structured formalized user request. It conveys the main meaning of what was said and the wishes of the user. A frame model of utterances is used to form an "intention", as a result of which an object is formed which is transferred to the next module.

The dialogue management manager (DM) is a module that directly controls the dialogue, its structure and progress, fills the

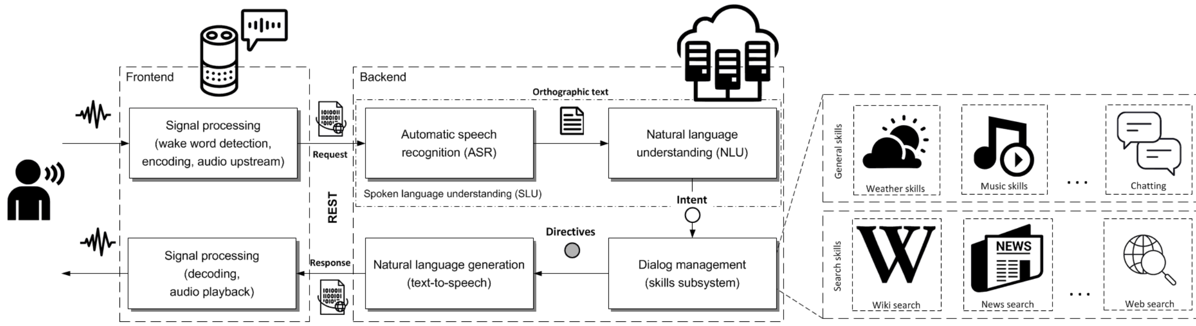


Figure 1. Dialogue system architecture

general context, contains knowledge about a specific subject area in the form of compact fragmented ontologies in the semantic network and the rules for responding to them which are known as “skills”. It receives the input data from the SLU component in the form of incoming "intentions" and must select the necessary "skill" block, then attains the global state of the dialogue process and transfers the output data to the generation module in the form of a "directive" object. The figure 2 shows examples of the description of intent and "directive".

The natural language generation module (NLG) synthesizes the voice response message in accordance with the available signal and speaker voice model and text of the response message which is located in the "directive" object. In connection with the steady tendency towards personalization of devices and programs, the possibility of adapting systems to the voice of a new speaker is one of the interesting features that we would like to consider in our work and which is not available in current solutions on the market. According to research [13] the voice of a person known to the listener is perceived 30% better than the voice of an unfamiliar person. Changing the speaker’s voice during the speech synthesis process allows you to attract the attention of system users to key information in a voice message, to emphasize especially important events or moments. It helps to improve the ergonomics of the system.


```
{
    "intent": "GetHoroscope",
    "sign": "libra",
    "date": "2019-12-31"
}
```


a) Intent object example


```
{
    "namespace": "Speech",
    "name": "speak",
    "text": "The horoscope
    for Libra."
}
```


b) Directive object example

Figure 2. Intent and directive object example

It should be noted that this architecture has proven itself in all modern solutions. However, the current situation in the market of voice assistants, from our point of view, has several unresolved problems and limitations:

- The main parts (semantic processing of information in the NLU and DM modules) of the system are proprietary closed. Developers and researchers do not have the ability to make changes to the knowledge base, supplement and modify existing ontologies. There is only the opportunity to use these modules as a service without directly participating in their development. Also, scientific and practical details related to the methods of formalization and processing of knowledge were not disclosed, which does not allow comparing the proposed solutions with alternative technologies and implementations.
- All common voice assistants have an exclusively distributed implementation, where the main part is located on the server side. There are no alternative, the so-called "on the eadge" solutions that allow you to deploy the system in an independent local environment, for example, on your servers in your own data center or directly on the client device. Such a method would make it possible to ensure stable operation of the system in cases where there is no stable Internet connection and could also be in demand if the user does not want to transmit their personal data to companies in the form of voice fragments and descriptions of their requests in the system in form of “intentions” objects (intent) and “directives”. This thesis is of particular relevance in connection with the increasing incidence of leakage of personal information from large companies [12], [6].

In this regard, from our point of view, the urgent task is to build a voice assistant based on open semantic technologies that allow a large group of developers to participate in the design and extend the knowledge base. To solve this problem, it is necessary to formulate a number of requirements for such an assistant.

Analysis of the user needs (including various companies) allows us to present the following set of functional requirements to the developed speech assistants:

- speaker independent recognition, the ability to correctly recognize messages from various interlocutors, possibly with speech dysfunctions;
- in a situation where the message is not understood by the system or clarification of any parameters is required, the system should be able to ask the interlocutor clarifying questions;
- the ability to recognize a speaker by voice, as a result - the ability to conduct a dialogue simultaneously with a group of interlocutors;
- the ability to work in conditions of noise interference;
- the ability to accumulate and take into account the history of the dialogue with the same interlocutor for a long time (to build and store a portrait of the interlocutor);
- the ability to take into account the current state of the

user, including his emotions, as well as such individual characteristics as gender, age, etc.;

- the speech assistant can receive information about the interlocutor not only directly from the dialogue, but also have predefined information about him that is of interest in the context of the current dialogue;
- the speech assistant can conduct a dialogue of an infotainment nature (to answer user questions or conduct a conversation without a specific goal), and to pursue a specific goal that affects the dialogue process (for example, to calm or amuse a person to talk to).

The development of speech assistants that meet these requirements is hindered by a number of problems. Some problems were considered and partially solved by the authors in previous works:

- in [18], the problem of identifying and eliminating ambiguities (including those associated with speech defects, noise, etc.) in a speech signal due to a knowledge base is considered;
- in [19], an approach to the description of the context of the dialogue with the possibility of its consideration in the analysis of voice messages is considered;

In this paper, the main attention will be paid to the principles of dialogue organization (situational dialogue management), description of the user model, as well as mechanisms for adapting the dialogue process to the characteristics of a specific user and specific subject area.

An important feature that distinguishes dialogue system (in specific domain area) from universal voice assistents is the lack of the need to understand the meaning of the message completely; more often, to generate the required response, it is enough to determine the type of message and select some keywords. This feature significantly reduces the complexity of the problem being solved and will be taken into account further when detailing the proposed approach.

III. Proposed approach

OSTIS Technology and the corresponding set of models, methods and tools for developing semantically compatible intelligent systems as the basis for building voice assistants are proposed here. The basis of OSTIS Technology is a unified version of the information encoding based on semantic networks with set-theoretic interpretation, called the SC code [8].

The architecture of each system built on OSTIS Technology (ostis-systems) includes a platform for interpreting semantic models of ostis-systems, as well as a semantic model of ostis-systems using SC-code (sc-model of ostis-systems). In turn, the sc-model of the ostis-system includes the sc-model of the knowledge base, the sc-model of the task solver and the sc-model of the interface (in particular, the user interface). The principles of engineering and designing knowledge bases and problem solvers are discussed in more detail in [9] and [16], respectively.

Models and tools application proposed by OSTIS Technology will provide, in addition to the advantages indicated in the above works, the opportunity to

- create, store and analyze the user's portrait, including both long-term information and its current state;
- save and analyze the history of the dialogue with each interlocutor;
- clearly distinguish the part of the dialogue management, depending only on the meaning of the messages and not depending on the language in which the dialogue is conducted, and the part depending on the language of dialogue;
- integrate the subject-independent part of the knowledge base with the subject-dependent part within each system, which will allow you to flexibly take into account the characteristics of the subject area when conducting dialogue.

Further in the text, we will assume that the dialogue is carried out in Russian, however, most of the models presented do not depend on the language in which the dialogue is conducted. To write formal texts within the framework of the work, we will use options for external display of SC-code constructions such as SCg (graphic version) and SCn (hypertext version).

IV. System architecture

The general dialogue scheme can be written as follows:

- The user delivers a voice message;
- The speech analysis module, based on the dictionary of speech identifiers available in the knowledge base of the system, selects keywords within the message and correlates them with entities in the knowledge base;
- Based on the rules available in the knowledge base, the system classifies the received message;
- Based on the rules available in the knowledge base, the system generates a new message addressed to the current interlocutor (possibly non-atomic);
- The speech synthesis module generates a fragment of the speech signal corresponding to the new message and voices it;

A. Speech analysis

To perform the processing of the speech signal inside the ASR module, it is necessary to fulfill it analysis and parametric description, i.e. represent as a sequence of characteristic vectors of the same dimension.

The proposed signal model and method for its evaluation has a number of distinctive features. They are a discrete Fourier transform or determination of the autocorrelation function of a signal in a short fragment. The method under consideration does not impose strict restrictions associated with the observance of the stationary conditions of the signal parameters on the analysis frame. This allows one to obtain a high temporal and frequency resolution of the signal, as well as a clearer spectral picture of the energy localization at the corresponding frequencies 3, and as a result, a more accurate estimate of the signal parameters (on average by 10-15 %).

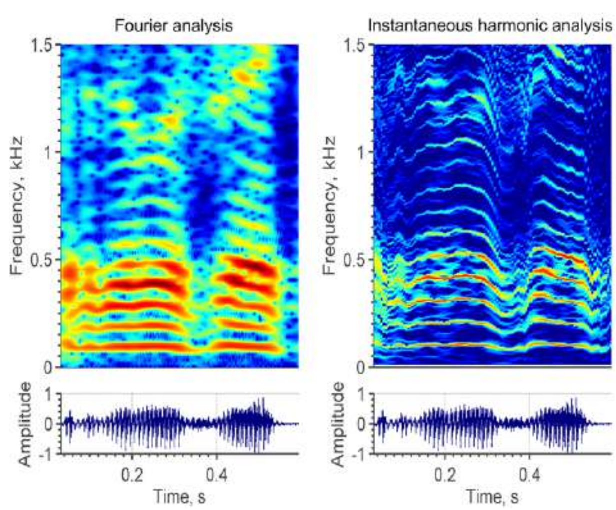


Figure 3. STFT and IHA based spectrograms

The speech signal is divided into overlapping fragments, each of which is described by a set of parameters: the spectral envelope, the instantaneous fundamental frequency (if the fragment is voiced) and the type of excitation, which can be voiced, unvoiced or mixed.

The quasiperiodic component of the speech signal is represented as the sum of sinusoids or the real part of complex exponentials with continuous amplitude, frequency and phase, and noise as a random process with a given power spectral density (PSD):

$$\begin{aligned} s(n) &= \sum_{p}^{P} A_p(n) cos\phi_k(n) + r(n) = \\ &= Re[\sum_{p}^{P} A_p(n) \exp j\phi_p(n)] + r(n) \end{aligned} \tag{1}$$

where P – number of sinusoids (complex exponentials), $A_p(n)$ – instantaneous amplitude of the p-th sinusoid, $\phi_p(n)$ – instantaneous phase of the p-th sine waves $r(n)$ – aperiodic component. The instantaneous frequency $F_p(n)$, located in the interval $[0, \pi]$ (π corresponds to the Nyquist frequency), is a derivative of the instantaneous phase. It is assumed that the amplitude changes slowly, which means limiting the frequency band of each of the components. Using the obtained harmonic amplitudes of the voiced and PSD unvoiced components, a common spectral envelope is formed.

This set of parameters is extracted from a speech signal using an algorithm consisting of the following steps:

- estimation of the instantaneous fundamental frequency using the error-resistant algorithm for tracking the instantaneous fundamental frequency "IRAPT (Instantaneous Robust Algorithm for Pitch Tracking)" cite Azarov2012;
- deformation of the time axis of the signal to ensure the stationary frequency of the fundamental tone;
- estimation of instantaneous harmonic parameters of a speech signal using a DFT modulated filter bank - each harmonic of the fundamental tone of voiced speech falls into a separate channel of the filter bank, where it is converted into an analytical complex signal from which the instantaneous amplitude, phase and frequency are extracted;
- based on the analysis of the obtained instantaneous frequency values, various regions of the spectrum are classified as periodic and aperiodic;
- harmonics belonging to periodic spectral regions are synthesized and subtracted from the original signal;
- the remainder is transferred to the frequency domain using the short-term Fourier transform;
- parameters of the synthesized harmonics and the PSD of the remainder are combined into one common spectral envelope and translated into a logarithmic scale;
- adjacent spectral envelopes are analyzed to determine how to excite the entire analyzed fragment of the signal.

Each spectral envelope is represented as a vector of logarithmic energy values equally spaced on the chalk scale. For a speech signal with a sampling frequency of 44.1 kHz, a 100-dimensional vector is used. The characteristic vector consists of the fundamental tone frequency values, the spectral envelope and the sign of vocalization of the current speech fragment. The dimension of the vector determines a compromise between the quality of signal reconstruction and computational complexity. Based on practical experiments, it was found that the selected dimension is sufficient for the reconstruction of natural speech.

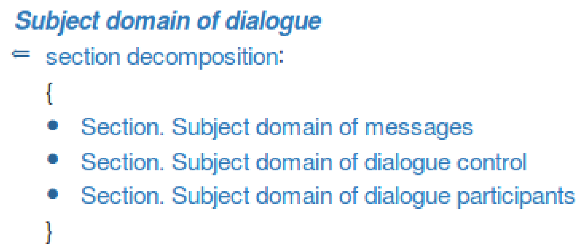


Figure 4. The hierarchy of subject areas.

B. Knowledge base

The basis of the knowledge base of any ostis-system (more precisely, sc-models of the knowledge base) is a hierarchical system of subject areas and their corresponding ontologies. The figure 4 shows the upper hierarchy of the knowledge base part that relates directly to voice assistants.

Consider in more detail the concepts studied in each of these subject areas and examples of their use.

C. Message subject area

The figure 5 shows the top level of message classification according to various criteria, independent of the subject area.

An atomic message refers to a message that does not include other messages, in turn, Non-atomic message is a message that includes other messages. At the same time, a non-atomic message can consist of one sentence, but have several semantic parts, for example, "Can I log in if I am not 16 years old?" (age is reported and a question is asked) or "Hello, my name is Sergey." (the name is communicated and a greeting is expressed).

In turn, the presented classes can be further divided into more private ones. The figure 6 shows the classification of interrogative sentences.

D. Subject area of dialogue participants

To present information about the participants in the dialogue, an appropriate domain model and ontology have been developed.

Figure 7 shows a fragment of the description in the knowledge base of a specific known user system. The above description contains both long-term information about the user, which will be saved after the dialogue is completed (gender, name, etc.) and short-term, which can be updated with each new dialogue - information on age, date of the last visit, mood, etc..

E. Dialog management area

Within the subject area of dialogue management, rules are presented according to which the analysis of user messages and the generation of response messages are carried out.

In accordance with the above general plan of work, the system distinguishes several categories of rules:

- voice message classification rules;
- rules for generating new messages within the current dialogue;
- voice message generation rules;

To simplify processing, some rules can be written not in the form of logical statements, but with the help of additional relations (for example, keywords that define the class of messages) and their corresponding knowledge processing agents. This hybridization of declarative and imperative recording options is widely used within the framework of OSTIS Technology in order to increase the efficiency of information processing while maintaining consistency of presentation at a basic level.

message
⇒ inclusion*:
- incentive message
- interrogative message
- declarative message

⇐ subdividing*:
- {
  - daytime message
  - evening message
  - morning message

  }
- {
  - message with respectful treatment
  - message with standard treatment

  }
- {
  - exclamatory message
  - non-exclamatory message

  }
- {
  - non-atomic message
  - atomic message

  }
- {
  - undefined language message
  - english language message
  - russian language message

  }

Figure 5. Typology of messages.

interrogative message
⇐ subdividing*:
{
- complete dictal question
- partial dictal question
- complete modal question
- partial modal question

}

Figure 6. Typology of interrogative messages.

Figure 8 shows an example of a simple rule for classifying messages based on keywords. The shown rule systemizes a message as a welcome class if it contains the appropriate words.

Figure 9 provides a formal definition of an atomic message.

Figure 10 shows a rule requiring you to find out the name of the interlocutor, if it is still not known to the system.

F. Subject-dependent fragments of the knowledge base

If necessary, the subject-independent part of the knowledge base can be supplemented with any information clarifying the specifics of a particular subject area, if it is necessary to improve the quality of the dialogue. The figure 11 shows a fragment of the knowledge base for the speech assistant-waiter of a cafe. The given fragment of the knowledge base includes a description of some drinks composition and confectionery products available in the menu, as well as information about the order made by a specific client.

G. Problem solver

The task solver of any ostis-system (more precisely, the sc-model of the ostis-system task solver) is a hierarchical system of knowledge processing agents in semantic memory (sc-agents) that interact only by specifying the actions they perform in the specified memory.

The top level of the agents hierarchy for the speech assistant task solver in SCn is as follows:

Voice Assistant Problem Solver
⇐ *abstract sc-agent decomposition**:
{
- *Logical Rule Enforcement Agent*
- *Voice Message Analysis Agent*
  ⇐ *abstract sc-agent decomposition*:*
  {
  - *Keyword selection agent in a voice message*
  - *Keyword selection agent from a set of words in a speech message*
  - *Translation agent of fragments of a speech message to a knowledge base*

  }
- *Message processing agent within the framework of semantic memory*
  ⇐ *abstract sc-agent decomposition*:*
  {
  - *Message classification agent*
  - *Agent for the formation of the order of atomic messages in the framework of non-atomic*
  - *Agent decomposing non-atomic messages into atomic messages*

  }
- *Voice Concatenation Agent*
- *Voice Message Generation Agent*

}

H. Speech synthesis

One of the requirements for the developed voice assistant indicated in this article is adaptation to a specific speaker. Changing the speaker's voice in the process of speech synthesis allows you to attract the attention of system users to key information in a voice message, emphasize important events or moments, according to studies. The voice of a person known to the listener is perceived 30% better than the voice of an unknown person [44, 45]. In this paper, we would like to show the applicability of the developed methods for signal synthesis with personalized speaker properties, namely, building a personal

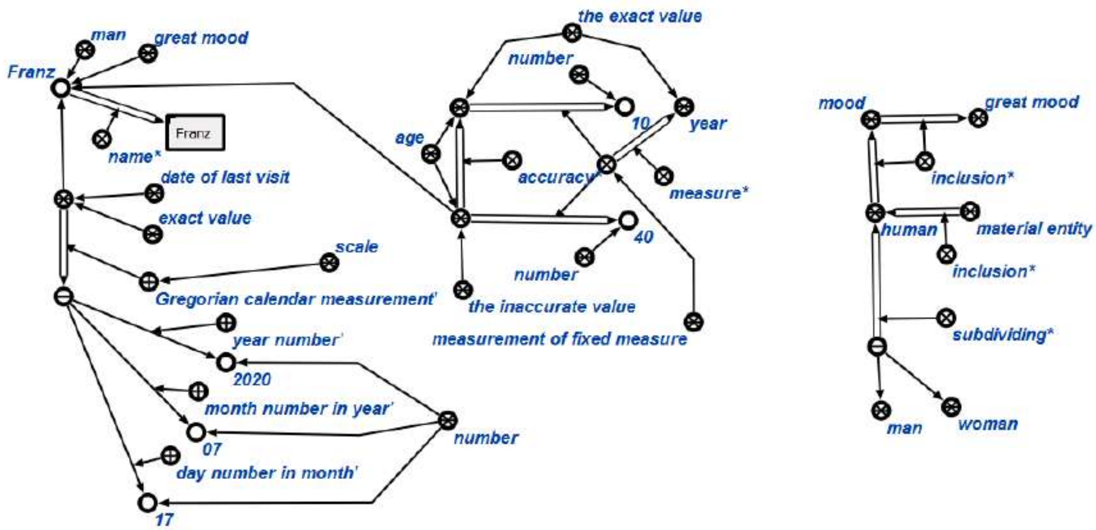


Figure 7. Portrait of a famous user system.

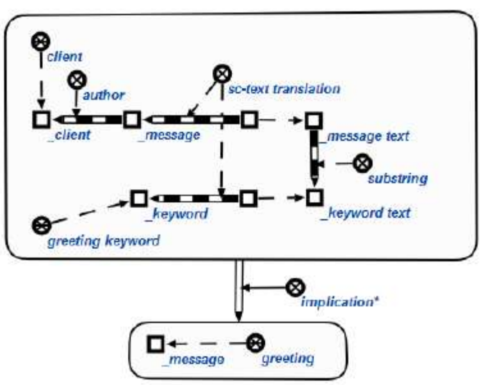


Figure 8. Example rule for classifying a message.

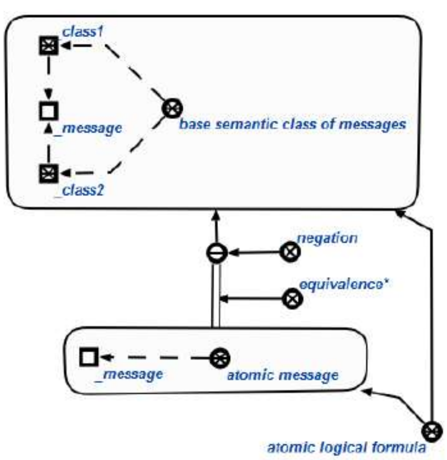


Figure 9. Definition of an atomic message.

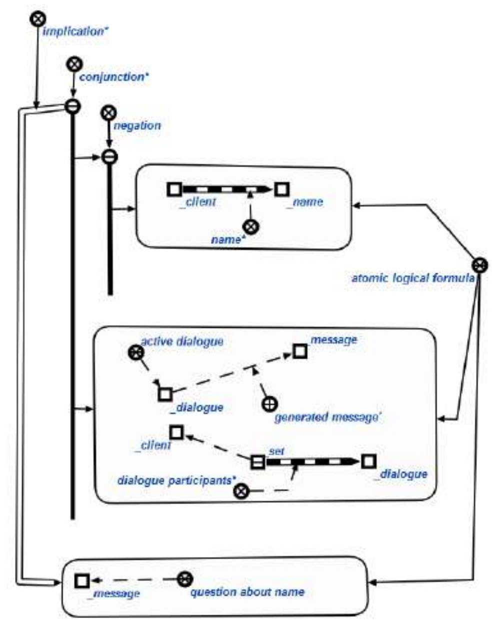


Figure 10. Rule of generating a question about a name.

speaker model that will get you around to synthesize speech with the voice of the selected target speaker.

The voice model of the speaker is based on a neural network, built on the principle of an automatic encoder. An automatic encoder is a multilayer neural network that converts multidimensional data into lower dimensional codes and then restores them in their original form. It was shown in [?] that data reduction systems based on neural networks have much broader capabilities, because, unlike the principal component analysis method, they permit nonlinear transformations to be performed.

The used artificial neural network configuration is shown in 12.

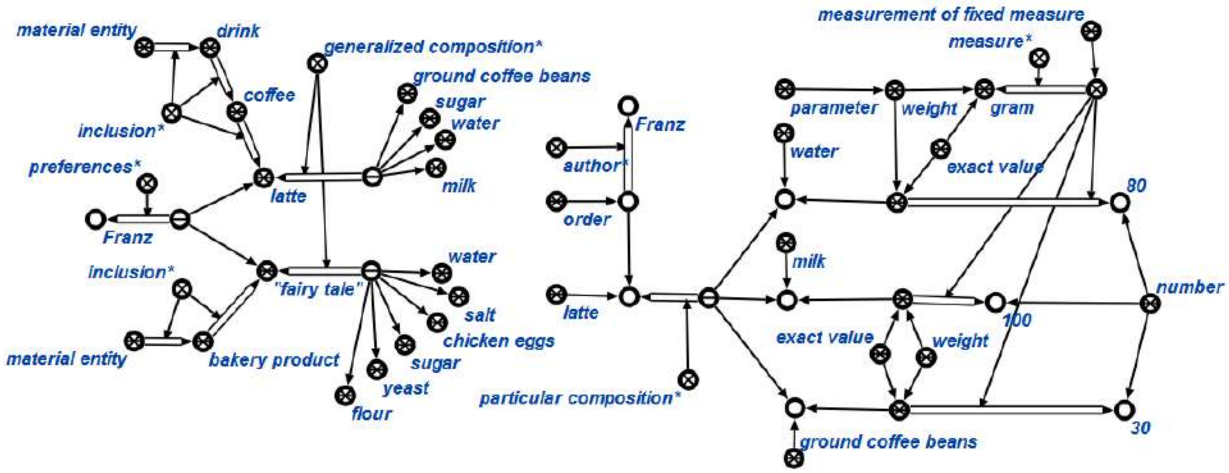


Figure 11. An example of a knowledge base fragment, depending on the subject area.

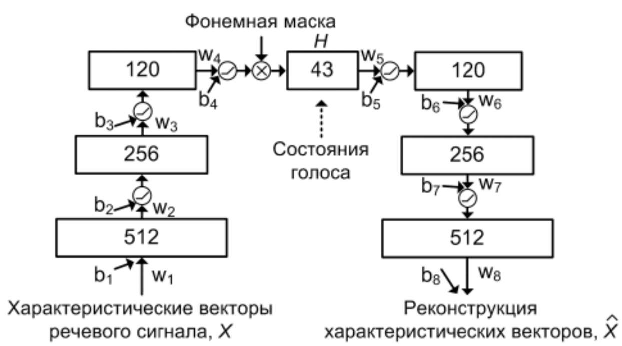


Figure 12. Speaker model based on auto-encoder

The auto-encoder performs next mapping function:

$$H = (w_4 RL(w_3 RL(w_2 RL(w_1 X + b_1) + b_2) + b_3) + b_4) \otimes M \quad (2)$$

where X is characteristic vector of the speech signal, H – vector of reduced dimension, M – phonetic mask vector, w_{1-4} and b_{1-4} – weight coefficients and offsets of the corresponding network signals, $\otimes$ – stands for elementwise multiplication. The network uses a piecewise linear activation function $RL(x) = max(0, x)$, since it is shown that it provides a more effective internal representation of speech data compared to the logistic one and allows you to speed up the learning process [**?**]. At the output of the encoder, lower dimensional codes are generated, which are constrained in order to perform phonetic binding. The restriction is imposed by multiplying the signal H by a phoneme mask, which is a sparse matrix, and formed on the basis of the phonetic marking of the speech corpus.

The decoder reconstructs the reduced-dimensional codes into the characteristic vectors $\hat{X}$. The corresponding display function is as follows:

$$\hat{X} = (w_8 RL(w_7 RL(w_6 RL(w_5 H + b_5) + b_6) + b_7) + b_8) \quad (3)$$

The next number of neurons in each hidden layer of the neural network was used: 512-256-120-43-120-256-512. Network training involves several steps:

- preliminary segmentation of the teaching speech corps into phonemes;
- initialization of network parameters and preliminary training;
- training of the encoder / decoder system;

As a result of training, a voice model is formed, which includes a model of each individual phoneme and the transitions between them contained in the training sample. For more details, the process of model formation is presented in [2].

V. Example of working

The scenario of the system is proposed:

1) The user asks a question like "What is X?" (the typology of questions is still limited to one class of questions);
2) The speech signal analysis module selects a fragment corresponding to the name of entity X in the request and finds an entity with that name (textit exactly how - see the question above) in the knowledge base;
3) The module for analyzing a speech signal in a formal language (in SC-code) forms a query to the knowledge base of the form "What is X?" for the found entity;
4) Ostis-system generates a response that is displayed to the user visually (in SCn, SCg). A subset of the answer (natural language definition or explanation for a given entity) goes to the speech synthesis module;
5) The speech synthesis module voices the received natural language text;

This dialog can be used as an example: - "Welcome. What is your name?"
–"Hello Andrey."
- "How is it going?"
- "Great!"
- "We can offer drinks and rolls. What do you prefer?"
- "I would like to order a latte."
- "Great choice, expect. Have a nice day."

Figures 13 – 19 show fragments of the knowledge base that sequentially reflect changes in it after processing each message in the dialog.

VI. Conclusion

An approach to the development of intelligent speech assistants based on the integration of modern approaches to

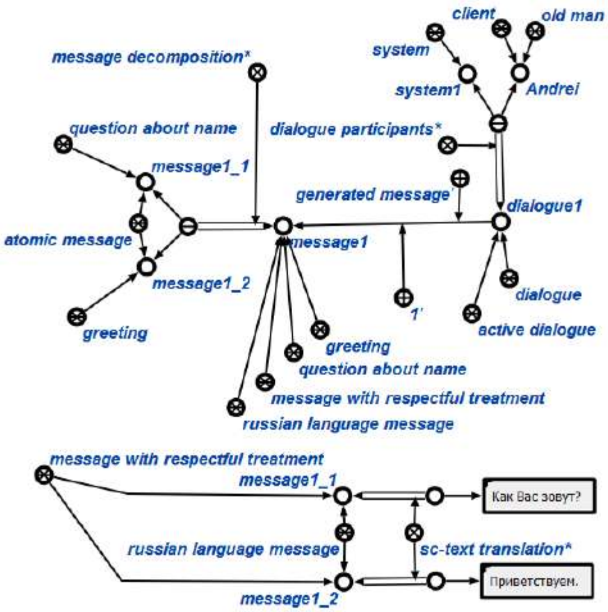


Figure 13. The portion of the knowledge base after the system generates a greeting and a question about the name.

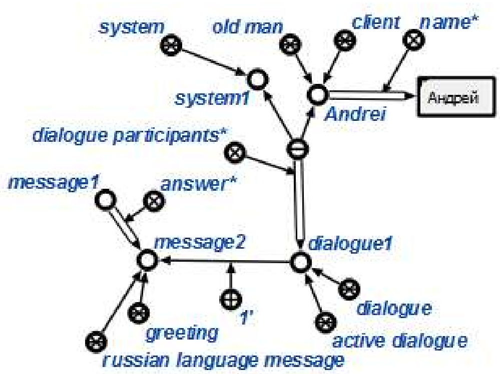


Figure 14. A part of the knowledge base after receiving and processing the user's answer to the question about the name.

speech signals processing and semantic dialogue management models is proposed. Obtained results, in particular, developed set of ontologies, can be applied to the lay-out of speech assistants for various purposes wit the posibility to be adapted to the characteristics of a specific subject area.

ACKNOWLEDGMENT

The authors would like to thank the scientific teams from the departments of intellectual information technologies, control systems and electronic computing facilities of the Belarusian State University of Informatics and Radio Electronics for their help and valuable comments.

REFERENCES

[1] Amazon Echo & Alexa Stats [Electronic resourse]. Access mode: https://voicebot.ai/amazon-echo-alexa-stats Date of access: 08.01.2020.

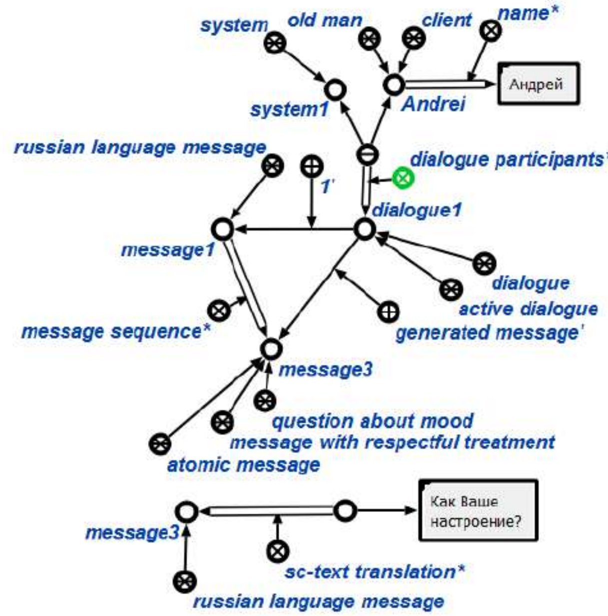


Figure 15. Section of the knowledge base after the system generates a question about the user's mood.

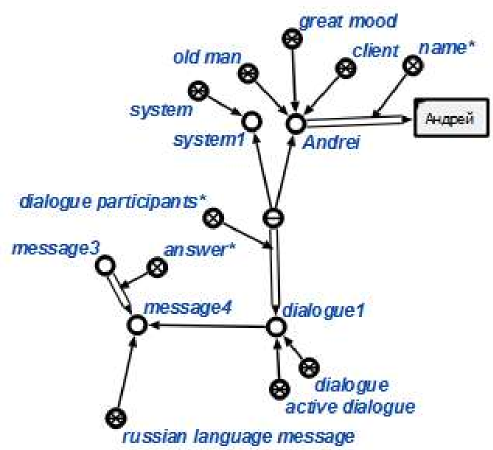


Figure 16. Section of the knowledge base after receiving and processing the user's response to a question about mood.

[2] E. Azarov, M. Vashkevich, A. Petrovsky, "Instantaneous harmonic representation of speech using multicomponent sinusoidal excitation". INTERSPEECH 2013: proceedings of 12th Annual Conference of the International Speech, Lyon, France, 2013. P. 1697–1701.

[3] E. Azarov, A. Petrovsky, "Formirovanie personal'noj modeli golosa diktora s universal'nym foneticheskim prostranstvom priznakov na osnove iskusstvennoj nejronnoj seti [The formation of a personal model of the speaker's voice with a universal phonetic space of signs based on an artificial neural network]". Trudy SPIIRAN. 2014. Vol. 5. No. 36. P. 128-150. (in Russian)

[4] F. Bülthoff, M. Maleshkova "RESTful or RESTless – Current State of Today's Top Web APIs." arXiv:abs/1902.10514. 2019.

[5] R. Corona, J. Thomason, R. Mooney "Improving Black-box Speech Recognition using Semantic Parsing". Proceedings of

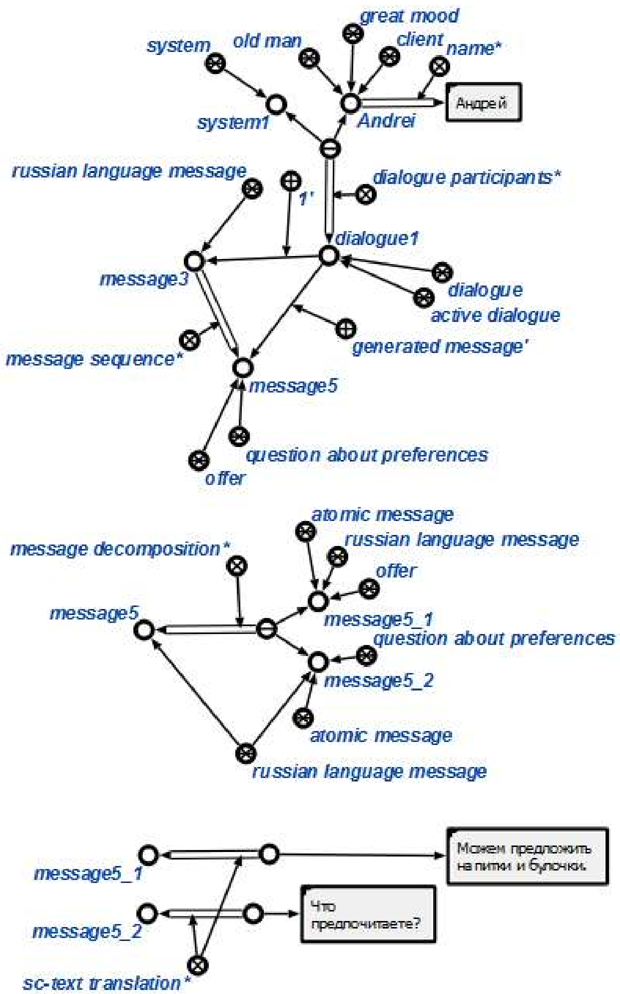


Figure 17. Knowledge base part after generating a system message containing a suggestion and a question about preferences.

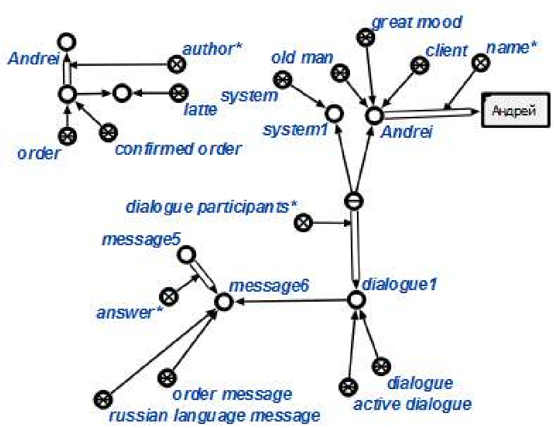


Figure 18. Knowledge base part after receiving and processing a user message containing order information.

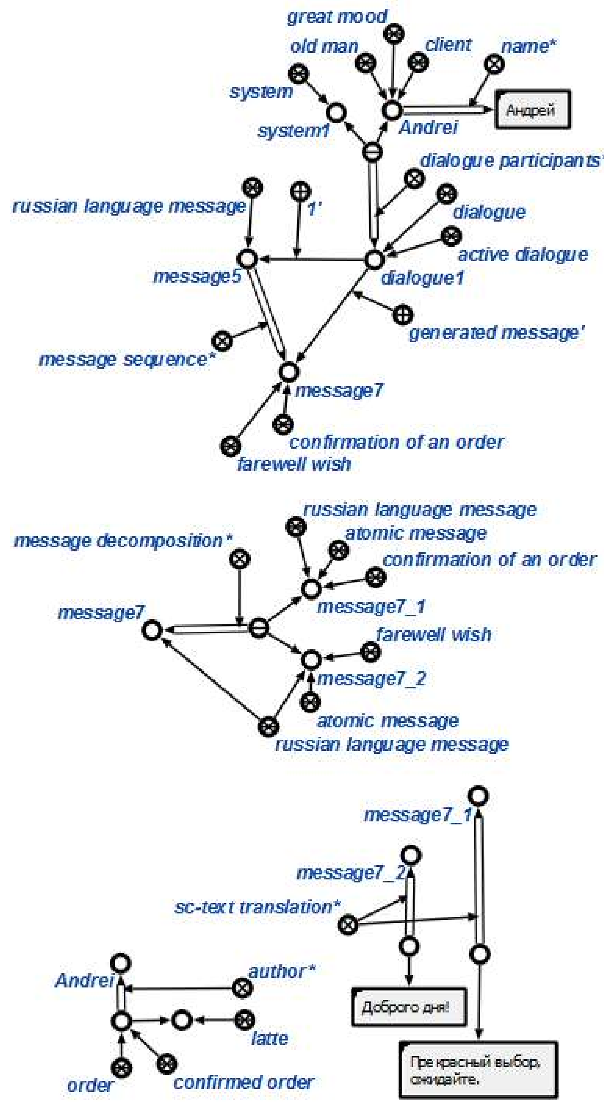


Figure 19. Knowledge base part after the system generates a message containing an order confirmation and a farewell message.

the Eighth International Joint Conference on Natural Language Processing. 2017. Vol. 2. P. 122–127.

[6] K. OḞlaherty, "Data Leak Warning Issued To Millions Of Google Chrome And Firefox Users" [Electronic resourse]. Access mode: https://www.forbes.com/sites/kateoflahertyuk/2019/07/19/data-leak-warning-issued-to-millions-of-google-chrome-and-firefox-users Date of access: 22.12.2019.

[7] Global Voice Assistant Market By Technology, By Application, By End User, By Region, Competition, Forecast & Opportunities, 2024 [Electronic resourse]. Access mode: https://www.businesswire.com/news/home/20190916005535/en/Global-Voice-Assistant-Market-Projected-Grow-1.2. Date of access: 20.12.2019.

[8] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, “Methods and tools for ensuring compatibility of computer systems,” in Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems], V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[9] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components”, in Open semantic technologies for intelligent systems, V. Golenkov,

Ed. BSUIR, Minsk, 2018, pp. 99–118.
[10] M. B. Hoy "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants". Medical reference services quarterly. 2018. V. 37. No. 1. P. 81-88.
[11] M. MacTear, Z. Callejas, D. Griol, "The Conversational Interface: Talking to Smart Devices". Springer, 2016. 422p.
[12] A. Matyus, "Facebook faces another huge data leak affecting 267 million users" [Electronic resourse]. Access mode: https://www.digitaltrends.com/news/facebook-data-leak-267-million-users-affected Date of access: 22.12.2019.
[13] T. Polzehl, "On Speaker-Independent Personality Perception and Prediction from Speech", INTERSPEECH. 2012 proceedings of 13th Annual Conference of the International Speech Communication Association. Portland, 2012. P. 258–261.
[14] F. Rubin, "Amazon's Alexa and Google Assistant try making themselves the talk of CES 2020" [Electronic resourse]. Access mode: https://www.cnet.com/news/amazon-alexa-and-google-assistant-try-making-themselves-talk-of-ces-2020/. Date of access: 02.01.2018
[15] R. Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components". IEEE Signal Processing Magazine, 2017, No. 1 (34). P. 67-81.
[16] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," in Open semantic technologies for intelligent systems, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.
[17] H. Tang, L. Lu, L. Kong, K. Gimpel, K. Livescu, C. Dyer, "End-to-End Neural Segmental Models for Speech Recognition", IEEE Journal of Selected Topics in Signal Processing. 2017. V. 11. No. 8. P. 1254–1264.
[18] V. A. Zahariev, E. S. Azarov, K. V. Rusetski, "An approach to speech ambiguities eliminating using semantically-acoustical analysis", Open Semantic Technologies for Intelligent Systems (OSTIS-2018). Minsk: BSUIR, 2018. P. 211 – 222.
[19] V. A. Zahariev, T. Lyahor, N. Hubarevich, E. S. Azarov, "Semantic analysis of voice messages based on a formalized context", Open Semantic Technologies for Intelligent Systems (OSTIS-2019). Minsk: BSUIR, 2019. P. 103 - 112.

Принципы построения интеллектуальных речевых ассистентов на основе открытой семантической технологии

Захарьев В.А., Шункевич Д.В., Никифоров С.А., Ляхор Т.В., Азаров И.С.

Целью данной работы является разработка принципов построения речевых ассистентов для интеллектуальных систем на базе Технологии ОСТИС.

Приведено описание существующих систем, их типовой архитектуры и принципов работы. В соответсвии со сравнительной характеристикой систем данного типа, а также результатами анализа актуальных потребностей пользователей, сформулирован ряд требований, которые показывают потенциальную эффективность применения Технологии ОСТИС для разработки диалоговых систем и голосовых асистентов в часности. Ключевыми условиями в данном случае являются обеспечение, с одной стороны, максимальной независимости от предметной области, с другой стороны – обеспечение возможности адаптации (в первую очередь – автоматической) универсального ассистента под конкретного пользователя и особенности конкретной предметной области.

В работе приведен пример реализации интеллектуального речевого ассистента, основанный на интеграции современных подходов к обработке речевого сигнала и семантических моделей управления диалогом. Приведены особенности реализации этапов анализа и синтеза речевого сигнала для удовлетворения реализации вышеобозначенных требований. Полученные результаты, в частности, разработанный набор онтологий, могут быть применены для разработки речевых ассистентов различного назначения, в том числе адаптированы с учетом особенностей конкретной предметной области.

Received 12.02.2020

Ontology Method for Chinese Language Processing

Longwei Qian, Mikhail Sadouski, Wenzu Li
Department of Intelligent Information Technology
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
qianlw1226@gmail.com, mishachess@mail.ru, wzzggml@gmail.com

***Abstract*—This paper describes a method for computer processing of Chinese language texts based on Chinese linguistic ontology. The ontology is used to build a unified semantic model of Chinese linguistic knowledge and effectively organize the domain knowledge for processing of Chinese language texts. The described method of texts processing could be used further for Chinese language user interface implementation.**

***Keywords*—Chinese language processing, ontology, knowledge-driven, Chinese language generation**

I. Introduction

Natural language processing has always been one of the core contents of research in the field of artificial intelligence. It is also the key basic technology for implementing natural language user interfaces. Natural language user interface understands users' intent through learning and understanding natural language, then uses the representation form of natural language to serve users through the interface.

A. Problems of Chinese Language Processing

Due to the diversity and openness of natural language, the computer processing (understanding) of different natural languages has own characteristics and special difficulties. Chinese language processing (CLP) is an exclusive field dedicated to conducting relevant research on Chinese languages and characters [1]. Researches on Chinese language processing show that machine translation systems from foreign languages (especially English) to Chinese language are far superior in terms of performance compared to machine translation systems from Chinese language to foreign languages. Even if application systems such as information retrieval and information extraction, which do not require deep syntax analysis, must overcome the obstacles of word segmentation and part-of-speech tagging in Chinese language processing [2].

Compared with other natural language processing technologies, the reasons that make Chinese language processing difficult can mainly be divided into the following aspects:

- the same part-of-speech can serve multiple syntactic components without morphological changes, i.e. in Chinese language texts regardless of any syntactic component served by each part-of-speech its morphology does not change;
- the construction principles of Chinese sentences are basically consistent with the construction principles of phrases. According to the different composition method, the structure of Chinese phrases can be divided into various types such as subject-predicate, verb-object, modifier-noun, joint and so on. However, if Chinese sentences is directly composed of words, the structure of Chinese sentences are also basically these;
- the function of functional words in Chinese language is complicated. The function words generally refer to words that do not have a full meaning but have grammatical meaning or function. Due to the lack of morphological changes in Chinese language texts, sentence patterns, and even tenses in sentences need to be represented by functional words;
- the writing habits in Chinese language. European languages such as English, Russian are basically written in words, with natural spaces between words. However, in Chinese language text there is no natural space to separate each word.

B. Analysis of Existing Approaches for Chinese Language Processing

The current natural language processing methods are divided into the following directions [3]:

- the methods of establishing logical reasoning system based on rules and knowledge base and based on Noam Chomsky's grammar theory;
- the machine learning methods based on large-scale corpus and based on mathematical statistics and information theory.

The statistical methods avoid the error situation caused by the subjective factors of manual writing rules in the rule-based methods, and reduce the dependence of the system on human. However, it inevitably has the defects that the performance of the model is too dependent on training samples, and it lacks domain adaptive capabilities.

At present, whether the rule-based methods or the statistical-based methods, the methods and evaluation standards used in Chinese language processing are almost borrowed from European language processing methods such as English. The analysis process lacks the characteristics of Chinese language texts [4].

C. Analysis of Existing Approaches for Chinese Language Generation

After the natural language interface analyses the natural language, the intelligent system needs to process these information, then to generate natural language texts and display it for users. This paper studies the process of generating the natural language texts from the semantic structure in the knowledge base. The semantic structure is constructed using a formal semantic network language. In the field of natural language processing, there are many researches on analyzing natural language texts to transform them into semantic networks. But there are few researches on generating natural language from formal semantic network structure.

The traditional natural language generation process is to transform a tree structure to natural language texts. However, the information structure formed by the semantic network language is a network topology. Simmons discoursed the use of augmented transition networks to generate English sentences [5]. Shapiro used extended transformation network syntax to implement sentences generation from semantic network [6]. These methods use fixed templates for sentence generation and sentence planning, which lack flexibility and adaptability. Dai Yintang studied the semantic network language generation, transformed a complex semantic network into multiple semantic trees, then serialized into language nodes [7]. This method improves the flexibility of sentence generation, but lacks the constraints of language knowledge, so the generated language fluency is insufficient.

D. Analysis of Existing Knowledge Base for Chinese Language Processing

Faced with the problems of Chinese language processing, Liu Zhiyuan from Tsinghua university proposed adding knowledge to the data-driven natural language processing model, and researched the natural language processing model driven by knowledge and data [8]. The knowledge of human can be divided into domain knowledge, common sense knowledge and linguistic knowledge. These complex knowledge help intelligent systems to achieve the deep understanding of natural language texts. The various knowledge can reasonably explain natural language processing. The linguistic knowledge can provide basic features and basic data resources for algorithms. The well-known knowledge bases for the natural language processing such as Freebase, DBpedia and so on, in the file of Chinese language processing, there are CN-DBpedia [9], zhishi.me, et al.

So far in the field of Chinese language processing, many excellent **language knowledge bases** have been developed.

The "Grammatical Knowledge-base of Contemporary Chinese (GKB)" [10], developed by Institute of Computational Languages of Peking University. It contains 73,000 Chinese words in accordance with the criteria of combining grammatical function and meaning. In accordance with the principle of grammatical function distribution, the various grammatical attributes of each word are classified and described.

The Chinese FrameNet ontology [11], developed by Shanxi University, the Chinese FrameNet ontology is constructed by drawing on the ontological thought of the Framenet project and its strong semantic analysis capabilities. The semantic frame is used to identify the relationships between lexis, and a valence mode is provided to reflect the syntactic-level relationships. So that the computer can perform a deeper semantic analysis of natural language texts.

The "HowNet" [12], developed by the professor Dong Zhengqiang. "HowNet" considers the lexis or Chinese characters is the smallest linguistic unit, but not the smallest semantic unit. It considers that the lexis can be described in a smaller semantic unit - sememe.

The GKB is organized according to the database model, and knowledge does not constitute semantic relationships. The theoretical basis of Chinese FrameNet Ontology is frame semantics, which faces narrow lexis coverage rate and insufficient description of syntactic information. The "HowNet" focuses on the description of semantic of Chinese lexis and concepts, lacks the description of grammatical and semantic information of the Chinese language texts.

E. The Proposed Approach

This paper takes a basis of these excellent Chinese language knowledge bases, proposes to construct the language ontology for Chinese language processing based on open semantic technology for intelligent systems (OSTIS technology). The OSTIS technology is aimed for presentation and processing of various knowledge, it's focused on the development of knowledge-driven computer systems [13]. The ontological approach of OSTIS for Chinese language processing is to construct a hierarchical system of subject domains of Chinese linguistic and their respective ontologies.

The Chinese linguistic knowledge could be described from various levels of Chinese linguistics such as lexis, syntax and semantics. This method uses the ontological thought to reasonably organize the knowledge and structure the knowledge in the file of Chinese linguistics. Based on OSTIS technology, the subject domains and respective ontologies of various levels of Chinese linguistics are constructed for Chinese language processing.

With the help of different levels of Chinese linguistic knowledge, it is beneficial to ambiguity resolution in Chinese language processing and content selection in Chinese language generation. In the process of machine translation, Chinese linguistic ontology can be used as a conceptual source for intermediate languages. The knowledge-driven Chinese language processing model can effectively improve the performance of Chinese language processing.

II. The structure of the knowledge base of the Chinese language interface

The knowledge base of natural language processing developed based on OSTIS technology is a part of the common linguistic knowledge base for all ostis-systems. Based on the general structure of linguistic subject domains, the development of Chinese linguistic subject domains and specific knowledge processing is the proposed method in this paper to implement the Chinese language interface that could process Chinese language texts.

According to the principle of using the OSTIS technology to construct the knowledge base, the construction of the knowledge for Chinese linguistics is a hierarchical system of the various subject domains (SD) of the Chinese linguistic knowledge and their corresponding ontology. The knowledge formalization can be considered as a formalization and specification of a subject domain, including a structural specification of a subject domain, terminological ontology, theory-set ontology, logical ontology and so on [14].

Based on OSTIS technology Chinese linguistic ontology classifies knowledge according to the subject domains. The various subject domains contain the different aspects of corresponding Chinese linguistic knowledge, as well as the rules and methods for operating the knowledge.

For Chinese language the grammatical features are not very obvious. A reasonable description of lexical semantic knowledge is more important for Chinese language processing. In addition to the previously described knowledge base. there are excellent lexis knowledge bases such as WordNet, VerbNet, et al. They have important reference significance for constructing lexical ontology. The domain ontology of Chinese linguistics is constructed by reusing existing ontologies and linguistic knowledge oriented to Chinese language processing. In the construction process of linguistic ontology, drawing on the reuse of existing Chinese linguistic ontology can improve the efficiency of ontology construction.

In this paper the process of constructing the domain ontology of Chinese linguistics can be roughly divided into the following 4 steps:

- determine that the requirements of the constructed ontology are oriented to Chinese language processing;
- research domain knowledge and reusable ontology;
- use specific examples to clarify subject domains, specific concepts and various relationships between these concepts;
- evaluation and improvement of ontology.

The goal of the proposed method is to construct syntactic ontology and semantic ontology for Chinese language processing using the bottom-up approach of domain ontology construction.

The development of subject domain of Chinese linguistic and their corresponding ontology using OSTIS technology uses semantic networks representation - SC-code (Semantic Code). Semantic networks representation is in the form of semantic network language with the basic set-theoretic interpretation.

The following is the general structure of the Chinese language subject domain represented in SCn-language [13].

SD of Chinese language texts
$\Rightarrow$ *particular SD**:
- *SD of Chinese language syntax*
- *SD of Chinese language semantic*

The *Subject domain of Chinese language syntax* describes the characteristics of the Chinese language syntax, the functional characteristics of the syntactic components. The *Subject domain of Chinese language semantic* describes the semantic characteristics of lexis, the semantic relationships and the semantic structure in the Chinese language texts.

Thus the key to the proposed method is the description of these specific subject domain and its corresponding ontology. The key elements of a specific subject domain are the main concepts studied in the subject domain and relationships. The relationships between these specific subject domains also need to be considered in the ontology. In general, the entire Chinese linguistic ontology forms a hierarchical and multilevel semantic model.

The focus of current research is on the subject domain of Chinese syntax, which will be described in detail below.

In the *Subject domain of Chinese language syntax*, the syntactic and semantic information of sentences, phrases and lexis in Chinese language texts need to be considered. These specific subject domains could provide linguistic knowledge for Chinese language processing.

The following is the structural fragment of the subject domain of Chinese language syntactic represented in SCn-language.

SD of Chinese language syntax
⇒ *private SD**:
- ***SD of Chinese sentence pattern***
- ***SD of Chinese language phrase***
- ***SD of Chinese grammatical information of the lexis***
- ***SD of Chinese function words***
- ***SD of non-lexis of Chinese language***

The *SD of Chinese sentence pattern* indicates types of the sentence, syntactic and semantic relationships between sentences. The *SD of Chinese language phrase* indicates types of the phrase, the internal structure and external functions of basic Chinese language phrases, syntactic and semantic relationships between phrases. The *SD of Chinese grammatical information of the lexis* indicates classifications of Chinese language lexis, and their grammatical and semantic functions of lexis. The *SD of Chinese function words* describes the types, functional characteristics of functional words, as well as grammatical attributes of functional words. The *SD of non-lexis of Chinese language* describes the features and syntactic functions of non-lexis in Chinese language texts.

The description of *Subject domain of the non-lexis components* in the Chinese language texts conforms to the characteristics of Chinese hieroglyph, some language components don't conforms to the definition of words in Chinese language. In European languages the language component smaller than word is the letter, but in the Chinese language texts the non-lexis components not only include Chinese characters, but also the idioms and the abbreviations.

Sentences have always been the smallest research unit in the field of natural language processing. Sentence analysis is an important intermediate stage connecting discourse analysis and lexis analysis. The detailed description of sentence knowledge is an important basic stage in the entire natural language processing. The *Subject domain of Chinese sentence pattern* studies the types of sentence patterns in Chinese language, determination of sentence components and their relationships, i.e. the division of sentence components in a simple sentence and an accurate functional description of sentence components.

The following is the structural fragment of the *Subject domain of the sentence pattern* represented in SCn-language.

SD of Chinese sentence pattern
∋ *maximum studied object class'*:
sentence pattern
∋ *not maximum studied object class'*:
- *simple sentence*
- *compound sentence*

∋ *explored relation'*:
- *subject'*
- *object'*
- *head predicate'*
- *attribute'*
- *center word'*
- *adverbial modifier'*
- *complement'*

Simple sentence
⇐ *subdividing**:
{
- *subject-predicate sentence*
- *non-subject-predicate sentence*
- *special sentence*

}

The sentence structure of a simple subject-predicate sentence can be described by the ontology. For example, it can be composed of a noun phrase and a verb phrase according to subject predicate relationship (Fig. 1).

As a highly deterministic partial analysis of Chinese language texts, the analysis of phrases can solve most of the problems of local ambiguity structure. As an intermediate result of sentence analysis, phrase analysis are also the basis of deeper chunk analysis and complete syntax analysis. The *Subject domain of Chinese phrases* studies the types of Chinese phrases and the relations between the internal structure of these phrases.

The following is the structural fragment of the *Subject domain of the phrases* represented in SCn-language:

SD of Chinese phrase
∋ *maximum studied object class'*:
phrase
∋ *not maximum studied object class'*:
- *noun phrase*
- *verb phrase*
- *adjective phrase*
- *numeral classifier phrase*

∋ *explored relation'*:
- *subject'*
- *object'*
- *head predicate'*
- *attribute'*
- *center word'*

Verb phrase
⇐ *subdividing**:
{
- *transitive verb phrase*
- *intransitive verb phrase*

}

Various basic phrases structures can be described by ontology according to syntactic relationships. For example, a verb phrase can be composed of verb and other any a

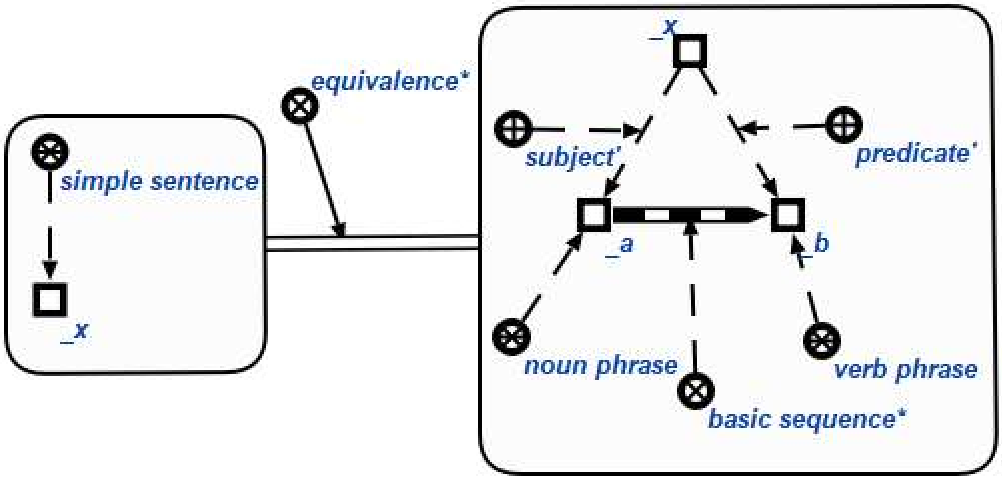


Figure 1: Logical statement about subject predicate sentence

basic phrase or verb and a noun according to predicate-object relationship (Fig. 2).

Through the structure of the ontology of phrase knowledge and the sentence pattern described above, the construction of the ontology of Chinese language knowledge conforms to the structural consistency of phrases and sentences in Chinese language texts.

Due to the problem of Chinese writing habits, the standardization of words in Chinese language processing is the first theoretical problem. In the field of Chinese language processing, the "Contemporary Chinese Language Word Segmentation Standard Used for Information Processing" has been proposed. In this standard, the word are represented by "segmentation unit". The precise definition is: "the basic unit for Chinese language processing with certain semantic or grammatical functions". According to the knowledge of Chinese linguistics, the Subject domain of Chinese lexis studies types, syntactic functions and semantic functions of lexis. The following is the structural fragment of the *Subject domain of the lexical grammatical information* represented in SCn-language:

SD of lexical grammatical information
∋ *maximum studied object class'*:
  lexical grammatical information
= the grammatical information of the segmentation unit in Chinese language
∋ *not maximum studied object class'*:
- *noun*
- *verb*
- *adjective*

Verb
⇒ *inclusion**:
- *verb that do not take object*
- *verb that take object*
- *verb that take double object*

Based on the ontology method of OSTIS technology, the Chinese linguistic knowledge is structured for processing of Chinese language texts. The goal of constructing knowledge base based on ontology is to provide a common understanding of Chinese linguistic knowledge, determine commonly recognized concepts. By constructing the knowledge base of Chinese linguistic knowledge, the clear definition of the concepts and their interrelationships in Chinese linguistic knowledge is determined. The ontology also contains the main theories and basic principles of Chinese language, as well as the methods and rules of knowledge processing. The knowledge base divides Chinese linguistic knowledge into various subject domain, it's beneficial to the management and application of knowledge. Thus the knowledge base of Chinese language has the ability for Chinese language processing. The Chinese language interface can use the concepts, semantic relationships, inference rules and other knowledge in the knowledge base to achieve Chinese language processing.

III. Implementation of Chinese language processing

Automatic analysis of natural language is a hierarchical process, which is roughly divided into lexical analysis, syntactic analysis and semantic analysis. Generally, there are some stages for the process of the natural language text understanding with the aid of the OSTIS-system [14]. Based on these steps the proposed method uses discourse analysis (context) of the input text. Context analysis is performed at each specific level of natural language automatic analysis and interaction with the content in the internal knowledge base or external information resources during the analysis process. The analysis process of the

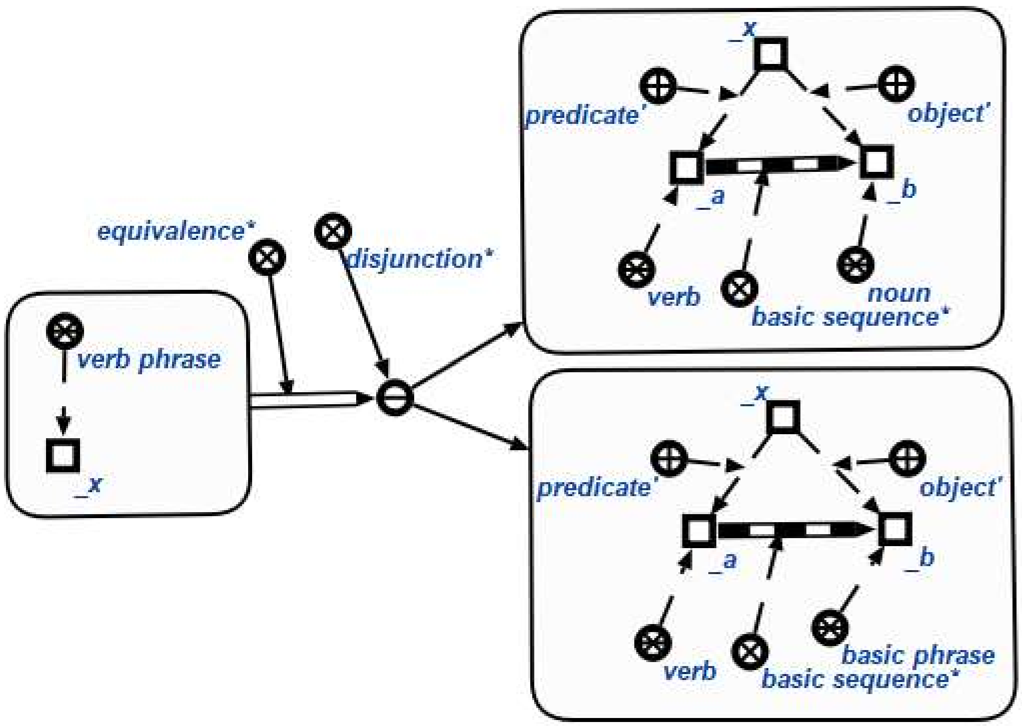


Figure 2: Logical statement of a verb phrase

natural language texts based on OSTIS technology [15] can be roughly divided into the following sections:

- Linguistic (graphematic) analysis generates a series of independent words in a given order;
- Lexical morphological reduction, lexical analysis, context analysis (comparative analysis with the content in the knowledge base of a specific domain);
- Syntactic tree generated, the specific agent transforms it into a semantically equivalent structure in the knowledge base;
- Semantic Analysis: through semantic ontology in linguistics and ontology knowledge in the specific subject domain, the agent correctly analyzes the semantics of the sentence.

The analysis of natural language texts based on OSTIS technology needs to be adjusted specifically to suit the characteristics of Chinese language texts.

Due to the writing habits of Chinese language text, there are no natural spaces between words. So the graphematic analysis need determine the basic text components: segmentation units, e.g. with the help of third-party software [16], an ordered sequence of Chinese characters is segmented into correct independent words with a given order. The lexical context analysis process is combined with the Chinese language texts segmentation process. Comparing the results of Chinese language texts segmentation with the content in the knowledge base of a specific domain, the accuracy of the Chinese language texts segmentation results can be improved. The agent transforms the results of the syntactic analysis into a semantically equivalent fragment in the knowledge base, the structure of fragment in the knowledge base is constructed in the semantic network language. By combining with the knowledge of linguistic ontology and the knowledge base of the specific domain, the information of natural language texts is integrated into the existing knowledge base to realize the deep semantic understanding of Chinese language texts.

The Chinese language processing described in the above section belongs to the study of natural language understanding. Another part of natural language processing is natural language generation. Natural language generation and automatic analysis of natural language are not simple and mutually inverse processes. There are obvious differences in both the implementation model and the basic process. In the field of natural language generation, the classic pipeline model consists of three parts: document (content) planning, sentence planning and surface realization. Corresponding to classic model, the natural language generation process is divided into the following three parts:

- Content selection: Choose a suitable and complete semantic structure, which can represent a complete theme;

- Content unit division: A complete semantic structure is divided into one or more semantic substructures, these substructures is interrelated with semantic relationships. Each semantic substructure is suitably represented by a natural sentence;
- Sentence generation: Transform the semantic substructure into a natural language sentence, then the complete semantic structure is transformed into natural language texts.

The process of selecting content for a concept and presenting it to the interface through various output forms is called semantic presentation. The application of semantic presentation in this paper is to use a unified method for content planning and sentence planning for the related knowledge of a central concept to generate natural language texts. According to the user's intention, the user interface identifies the central concept of solving user problems and determines the boundaries of the semantic structure in the knowledge base. Based on the knowledge of the Chinese language ontology and the content of the specific domain knowledge base, the confirmed semantic structure could be transformed into the highly readable natural language texts on the interface for users.

The quality of Chinese language sentences generated from semantic substructures is the basis for the entire semantic structure to be transformed into Chinese language texts. The specific process of generating Chinese language sentence from semantic substructure is divided into: determining the grammatical information of lexis, phrase generation and sentence generation.

The grammatical information of the lexis is determined by the constructed ontology (Fig. 3), i.e. the syntactic functions of the lexis itself and the ability to combine with other notional words and function words and so on.

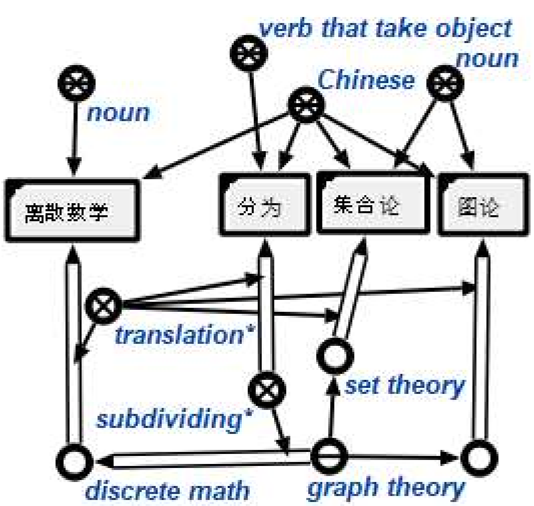


Figure 3: Grammatical information of lexis

Lexis constitutes phrases or sentences through their syntactic and semantic relations. According to the principles of Chinese linguistics, in the process of Chinese language generation, firstly lexis constitute the smaller language units — phrases. The syntactic and semantic function of the lexis in the phrase are determined by the phrase knowledge ontology, then the lexis constitute the phrase (Fig. 4).

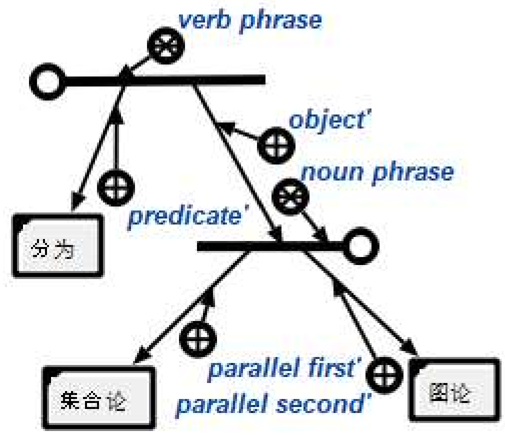


Figure 4: Phrase generation

In the Figure 4 the noun phrase consists of two nouns according to parallel relation. The noun phrase is generated using the rules of Chinese linguistics described in the logical ontology. In the Figure 4 the noun phrase could be composed by the following rule: noun plus conjunction, then plus another noun. The verb phrase are composed of the verb and the noun phrase according to the predicate-object relation. The verb serves as predicate in the verb phrase, and the noun phrase serves as object.

The ontology of sentence pattern knowledge confirms the syntactic and semantic functions that the phrases serve in a sentence, and then phrases "achieve" the sentence (Fig. 5).

Structured knowledge in the knowledge base of Chinese language is applied in the process of Chinese language processing. In the process of Chinese language understanding, the knowledge in the knowledge base of Chinese linguistic provides analysis of lexis, phrases, and sentences in Chinese language texts. The Chinese language texts could be transformed into semantically equivalent structures of specific domain. The semantically equivalent structure is fused with the knowledge in knowledge base of specific domain to achieve the deep understanding of Chinese language texts. In the process of natural language generation, the knowledge in the knowledge base of Chinese linguistics is used to determine the syntactic and semantic function of Chinese language for each semantic node in the semantic structure. The knowledge of Chinese linguistics is used to explain each part of the language generation process, making the

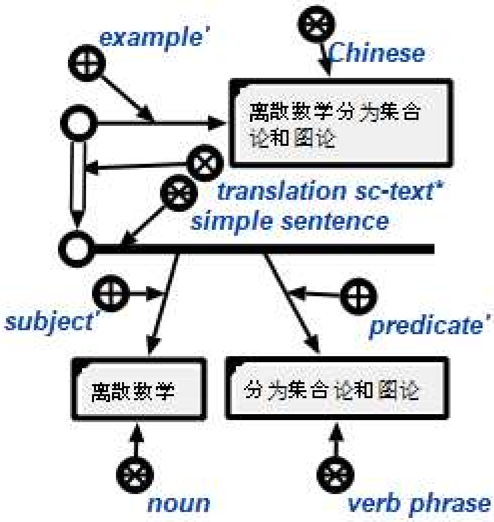


Figure 5: Sentence generation

generated Chinese language texts more accurate. Thus the Chinese language interface is implemented using the Chinese language processing method designed based on ontology.

IV. Conclusion

The proposed method of Chinese language processing is intended to developing the semantic model of Chinese language knowledge. The semantic model can be implemented on different platforms without changing the model itself. In the process of Chinese language automatic analysis, the domain knowledge base can be used as a word segmentation dictionary to achieve more accurate Chinese word segmentation. Based on the ontology of Chinese language knowledge Chinese language processing can be explained with reasonable linguistic knowledge. Obviously, the proposed method in this paper is the preliminary result of the research work of Chinese language processing based on ontology. The improvement of Chinese syntactic ontology, the development of Chinese semantic ontology is the focus of the next work. The more comprehensive context analysis in Chinese language automatic analysis, the problem of the semantic structure planning in Chinese language generation need be more in-depth researched.

References

[1] Zong C.Q. Chinese Language Processing: Achievements and Problems. Chinese Journal of Language Policy and Planning, 2016, vol. 01, No 06, pp. 19-26. (in Chinese)

[2] Yu S.W. Zhu X.F. Wang H. New Progress of the Grammatical Knowledge-base of Contemporary Chinese. Journal of Chinese Information Processing, 2001, vol. 15, No 01, pp. 59-65. (in Chinese)

[3] 2018 Natural Language Processing Research Report. Available at: https://aminer.org/ (in Chinese)

[4] Zong C.Q. Cao Y.Q. Yu S.W. Sixty Years of Chinese Information Processing. Applied Linguistics, 2009, vol. 01, No 04, pp. 53-61. (in Chinese)

[5] R. F. Simmons Generating English discourse from semantic networks. Communications of the ACM, 1972, vol. 15, No. 10, pp. 891-905.

[6] S. C. Shapiro, Generalized Augmented Transition Network Grammars for Generation from Semantic Network. American Journal of Computational Linguistics, 1982, vol. 08, No 01, pp. 12-25.

[7] Dai Y.T. Research on the Key Techniques of Semantic Networks based on Knowledge Collaboration. CA: Fudan University, 2009. (in Chinese)

[8] Liu Z.Y. Knowledge guided natural language understanding. China Conference on Knowledge Graph and Semantic Computing, 2018, pp. 199-206. (in Chinese)

[9] Xu B. Xiao Y.H. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 428-438. Springer, Cham, 2017.

[10] Yu S.W. Zhu X.F. Modern Chinese Grammar Information Dictionary, Peking University Open Research Data Platform. [online]. Available at: http://dx.doi.org/10.18170/DVN/EDQWIL.(in Chinese)

[11] Jia J.Z. Dong G. The Study on Integration of CFN and VerbNet, WordNet. New Technology of Library and Information Service, 2008, No 06, pp. 06-10. (in Chinese)

[12] Hownet knowledge database. Available at: http://www.keenage.com/. (in Chinese)

[13] Golenkov V.V. Gulyakina N.A. Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem. Chast' 1 Printsipy sozdaniya [Project of open semantic technology of component designing of intelligent systems. Part 1 Principles of creation]. Ontologiya proektirovaniya [Ontology of designing], 2014, no 1, pp. 42-64. (in Russian)

[14] Maksimov V.Yu., Klyshinsky ES. The problem of understanding in artificial intelligence systems. New information technologies in automated systems, 2016, pp. 43-60

[15] Hubarevich N. Boyko I. Knowledge acquisition based on natural language texts. Open semantic technologies for intelligent systems, 2018, pp. 199-206

[16] Language Technology Platform. Available at: http://www.ltp-cloud.com/

Онтологический подход к обработке китайского языка

Цянь Лунвэй, Садовский М.Е., Ли Вэньцзу

В статье рассматриваются существующие методы обработки китайского языка, а также методы построения онтологии предметной области. Был проведен анализ проблем, возникающих при обработке китайского языка в настоящее время.

На основании различных рассмотренных методов построения баз знаний английского и китайского языка был предложен онтологический подход к обработке китайского языка. Предложенный метод направлен на разработку семантической модели знаний о китайском языке. Как один из этапов реализации подхода была создана онтология китайского языка, которую можно использовать в дальнейших этапах для обработки китайского языка.

Таким образом, для дальнейшей обработки и реализации естественно-языкового пользовательского интерфейса в работе предлагается модель обработки китайского языка, основанная на знаниях.

Received 10.01.2020

Ontological approach to automating the processes of question generation and knowledge control in intelligent learning systems

Wenzu Li, Natalia Grakova, Longwei Qian
Belarussian State University Informatics and Radioelectronics
Minsk, Belarus
lwzzggml@gmail.com, grakova@bsuir.by, qianlw1226@gmail.com

***Abstract*—This article proposes an approach for designing a general subsystem of automatic generation of questions and automatic verification of answers in intelligent learning systems. The proposed approach is based on the existing approaches and ontology theory, using this approach it is possible to generate various types of questions from the knowledge bases automatically and automatically verify the correctness and completeness of user answers. Compared with the existing approaches, the approach proposed in this article has certain advantages, and the subsystem designed using this approach can be used in systems built using OSTIS Technology.**

***Keywords*—question generation, answer verification, knowledge base, ontology, learning systems**

I. Introduction

With the rapid development of computer technologies and Internet, the way people acquire and store information has changed greatly. Various multimedia software and examination systems have been used by many educational institutions, which has greatly changed the way users learn. However, with the rise of artificial intelligence technologies in recent years, learning methods have moved from the traditional multimedia training mode to the era of intelligent education [5].

The use of advanced learning methods provided by artificial intelligence technologies in the learning process can form a new learning mode, stimulate the user's learning interest and improve the user's learning efficiency. Artificial intelligence methods are used in the field of education, although they can not change the nature of the learning process, but can produce new learning methods and greatly improve the learning efficiency.

Compared with traditional multimedia training systems (MTS), intelligent learning systems (ILS) have the following advantages:

- the learning materials of the MTS are pre-selected by experienced teachers and then stored in the database, so the test question database will appear redundant and duplicated. ILS uses semantic network and ontology approach to store learning materials, which are rich in content and easy to expand and modify;
- users of the MTS cannot ask questions to the system, and the system cannot answer questions to users. But this is a basic function in the ILS;
- when testing users, ILS can use the knowledge in the knowledge base to automatically generate various types of questions, and can also verify the user answers including correctness and integrity checks. ILS can give prompt steps of the test questions according to user requirements, and finally give system reference opinions based on user test results;
- ILS allows users to use natural language to ask questions and answer various types of test questions automatically generated by the system.

Using the knowledge base to automatically generate various types of questions and the automatically verify the user answers is one of the most important features of ILS, but the approaches of question generation proposed by most researchers only allow to generate very simple questions (choice questions; fill in the blank questions and etc.), and the correlation between the correct options and the incorrect options (disturbing options) of the generated choice questions is not high. Therefore, most of the approaches for automatic generation of questions do not meet the requirements of practical applications, and the proposed approaches are implemented on specific systems, so they are not universal. At present, most approaches of answer verification are based on keyword matching and probability statistics and these approaches rely heavily on the quality of the corpus and do not consider the semantic similarity between the answers. Therefore, this article proposes an approach for automatic generation of questions and automatic verification of user answers to solve the above problems. The approach proposed in this article is aimed at developing a general subsystem for automatic generation of questions and automatic verification of answers for systems built on OSTIS technology [1]. The results of the work of the subsystem are implemented in an intelligent learning system for discrete mathematics.

II. Existing approaches and problems

A. Automatic question generation

Automatic question generation method (AQGM) studies how to automatically generate test questions from electronic documents, corpus or knowledge bases through computer technologies.

Unlike traditional approaches of using database, AQGM does not directly extract pre-stored test questions from the database, but uses various types of knowledge sources (including electronic documents, corpus or knowledge bases) to automatically generate questions that meet test requirements. Compared with the traditional approach of using database to extract questions, AQGM is a knowledge-based question generation method, so the generated questions are more flexible [5], [6], [8].

Approaches for automatic generation of questions can be divided into the following categories:

- based on electronic documents;
- based on conceptual corpus;
- based on knowledge base.

Within the approach of extracting questions using electronic documents it is easy to obtain knowledge resources, so the approach has great freedom and the scope of knowledge resources involved is also wide [11]. The knowledge in the corpus is selected by professionals, and this type of knowledge has been filtered, so the knowledge in the corpus is of high quality, which ensures that the quality of the automatically generated questions is high, too. Knowledge in the knowledge base is well structured and verified. Knowledge base is developed after the domain experts analyze the knowledge, this knowledge is filtered and has a certain structure, so the questions automatically generated using the knowledge base are more flexible and diverse, and the quality is the highest [5]. The current research on AQGM is still in its infancy, here are some research results:

- Ding Xiangmin developed an automatic generation system of multiple choice questions, firstly the system summarizes some common Chinese sentence pattern templates by using statistical method, and then uses these sentence pattern templates to extract multiple choice questions from the electronic documents in the aviation field, finally the incorrect (disturbing) options of multiple choice questions are automatically generated by the ontology base in the aviation field developed by Protégé [7], [8];
- The automatic generation system of choice questions developed by Andreas Papasalouros mainly uses the relation between parent class and subclass, class and element, element and attribute in OWL ontology to automatically generate choice questions [10];
- Based on the approaches proposed above, Li Hui used Protégé to create an ontology in the field of computer theory, using various relations between parent class and subclass, element and attribute in the ontology can automatically generate choice questions, fill in the blanks questions and judgment questions, where incorrect options (disturbing) of choice questions are also automatically generated through these relations [5], [7].

Although these approaches discussed above have many advantages, there are also many problems:

- although the approach of using electronic documents and sentence templates to automatically generate questions has abundant sources of knowledge, due to the many types of electronic documents, a large number of sentence templates are required and the types of questions generated are fixed and of low quality;
- the approach of using corpus to automatically generate questions requires the combination of electronic documents and English dictionaries to generate complete questions. The scope and quality of the generated questions depend heavily on the size and quality of the corpus, and the correlation between the incorrect and correct options of the generated choice questions is not high;
- at present, there is no unified standard for the development of most knowledge bases, so different knowledge bases have different knowledge structures and they are not compatible with each other. Because the knowledge bases are not compatible with each other, the approach of using the knowledge base to automatically generate questions can only be used in the corresponding knowledge base, and for the knowledge base developed by other approaches, only new question generation approaches can be developed;
- the approaches introduced above allow to automatically generate only the simplest objective questions, and the generated choice questions answer options are not highly correlated or independent of each other, so these approaches have not reached the conditions for practical application.

Based on the considered existing approaches and in accordance with the OSTIS Technology [1], [2], [3], [4], an approach is proposed for the automatic generation of various types of questions for ILS. This approach allows automatic generation of several types of questions from existing knowledge bases, and then save them to the ontology of questions according to the strategies for automatic generation of questions. The proposed approach will be implemented in the discrete mathematics learning system.

B. Automatic verification of answers

Answer verification is divided into subjective question answer verification and objective question answer verification. Objective questions refer to a type of question with a unique standard answer, and the user answers for this type of question can only be right or wrong, so the answer verification of the objective question only needs to compare whether the user answers and the standard answers are the same. Objective questions include: choice questions, judgment questions, etc. Most subjective questions do not have a unique standard answer, as long as the user's answer conforms to the logic and meets the requirements of the question [17]. Common subjective questions include definition explanation questions, theorem proving questions, etc. In order to verify the answers of subjective questions, it is necessary to compare the similarity between the standard answer and the user's answer [9], [19], [20]. If the similarity is higher, the user's answer is closer to the standard answer, and the score the user gets is higher. The essence of similarity comparison of subjective questions is text similarity comparison. Text similarity comparison is currently divided into two directions:

1) comparison of text similarity based on natural language;
2) comparison of text similarity based on semantic graph.

Approach for comparing text similarity based on natural language:

1) Based on keywords and keyword combinations:

- N-gram similarity
  The N-gram approach divides two texts or sentences according to N-tuples, and determines the similarity between the texts or sentences by calculating the ratio of the same number of N-tuples and the total number of N-tuples between the texts or sentences [12], [19];
- Jaccard similarity The Jaccard approach uses the idea of set theory to determine the similarity between texts or sentences based on the ratio of the number of identical words or word groups to the number of all non-repeating words or word groups between texts or sentences [12], [20].

2) Based on vector space model (VSM):

The core idea of VSM is to first convert the text into a vector in space by mathematical modeling, and then calculate the similarity value between the spatial vectors through cosine similarity, Euclidean distance, etc [12], [15]. VSM includes the following approaches:

- TF-IDF
  TF-IDF regards text after the participle as an independent feature group, and gives a certain weight according to its importance in the whole document, then TF-IDF transforms the feature group into a space vector and calculates the similarity between the space vectors to get the text similarity [13], [14].
- Word2vec
  Word2vec is a word vector training tool that uses neural networks and corpora to predict the probability of a context based on the current word or the probability of the current word based on the context. The trained word vector can be combined with other approaches (WMD, Sentence2Vec, etc.) to obtain the sentence vector and then used to calculate the text similarity [16], [19].
- Doc2Vec
  Doc2vec tool is an extension of word2vec tool, which is used to learn the distribution vector representation of any length of text [16].

3) Based on deep learning

- In recent years, many researchers have begun to use deep learning for natural language processing. This approach mainly uses DSSM, ConvNet, Tree-LSTM, Siamese LSTM and other multi-layer neural networks to model words or sentences to obtain word vectors or sentence vectors, and then calculate the text similarity [17], [18].

Similarity comparison approach based on semantic graph:

The core idea of the text similarity comparison approach based on semantic graphs is to first convert natural language text into semantic graphs through tools such as syntactic dependency trees or natural language interfaces, and then calculate the text similarity by comparing the similarities between semantic graphs [21].

- SPICE
  SPICE (Semantic Propositional Image Caption Evaluation) approach is mainly used to evaluate the quality of automatically generated image caption. The main working principle of this approach is to compare the similarity between the automatically generated image caption (candidate caption) and the image caption (reference caption) manually labeled by the staff [21]. The main feature of SPICE is the comparison of similarities through semantic content, and its working process mainly includes the following three steps:
  - first, the candidate caption and the reference caption are labeled with grammatical dependency relation. The annotation contents include: subject, predicate, preposition and other grammatical information;
  - secondly, the candidate caption and reference caption with grammatical annotation information are converted into scene graphs by relying on the parse tree, and the scene graphs are encoded according to the relations between objects, attributes and objects;
  - finally, the candidate caption scene graph and reference caption scene graph are decomposed into set tuples according to the objects, the attributes of the objects, and the relations between the objects. The system calculates the similarity by calculating the ratio between the number of identical tuples

in the candidate set tuples and the reference set tuples and the total number of tuples.

Although the approaches discussed above can compare text similarity to some extent, these approaches also have many shortcomings:

- the text similarity comparison approach based on keywords only compares the similarity between texts by words or word groups, and cannot distinguish the synonymy and polysemy of words or word groups. This approach is now mainly used for spell checking and error correction;
- TF-IDF approach assumes that each feature word in the text exists independently, and does not consider the relation between words and their positions in the sentence. When the corpus is large, this method will generate a high-dimensional sparse matrix, resulting in increased computational complexity;
- although the approach based on deep learning has greatly improved the accuracy compared with other approaches, it is also a main research direction now, but this approach relies heavily on the quality of corpus, and when the corpus changes and updates, it needs to retrain the neural network model. This approach also does not have language independence, when using corpus of different languages (such as English, Russian, etc.) to express the same semantic content, the neural network model also needs to be redesigned;
- although the SPICE approach compares text similarity from the semantic level, this approach can only describe simple semantic relations such as the attributes of objects and the connection relation between objects, so many knowledge structures in real life cannot be described.

Based on the SPICE approach, this article proposes an approach for comparing text similarity using OSTIS Technology [1] and unified knowledge coding language SC-code [4]. The approach proposed in this article is to decompose the semantic graph of various types of knowledge represented by SCg-code into various substructures according to certain rules, and calculate the ratio between the same number of substructures and the total number of substructures between different semantic graphs to get the similarity between semantic graphs. Because the approach proposed in this article uses system external language SCg as the representation language of natural language text semantic graph, so it has language independence. The system uses natural language interfaces to convert various natural language (English, Chinese, Russian, etc.) answers into the system's external language SCg representation.

III. Proposed approach

The main task of this article is to introduce the design approach of subsystem of automatic generation of questions and automatic verification of user answers for ILS. Because the subsystem needs to complete two basic functions, so it can be divided into two parts: automatic generation of questions and automatic verification of answers. Next, we will consider the specific approaches of implementing this subsystem using a discrete mathematical system as an example.

A. Proposed question generation approach

Combining the previously discussed approaches for automatic generation of questions and the structural characteristics of the OSTIS learning system knowledge base, this article proposes an approach for automatically generating various types of questions for the ILS and a design approach for the ontology of questions. By using the approaches for automatic generation of questions and the OSTIS technology [1], [2], [3], subjective and objective questions can be automatically generated from the OSTIS learning system knowledge base. The generated subjective questions include: the questions of definition interpretation; the questions of axiomatic interpretation; the questions of proof; The objective questions generated include: choice questions; fill in the blank questions; judgment questions; The generated questions are stored in the ontology of questions in the form of semantic graph using the system external language SCg. Because the system external language SCg-code does not rely on any external natural language, these generated questions can be transformed into corresponding natural language questions through different natural language interfaces.

Consider in more detail the strategies for generating questions:

1) Question generation strategy based on classes

This strategy uses various relations satisfied between the classes to automatically generate objective questions.

- Based on inclusion relation
  In the knowledge base of the OSTIS learning systems, many classes satisfy the inclusion relation, and some classes contain many subclasses, so the inclusion relation between classes can be used to automatically generate objective questions. The set theory expression form of inclusion relation between classes is as follows: $S_i \subseteq C (i \geq 1)$, ($S$-subclass, i-subclass number, C-parent class). Taking the generated judgment questions as an example, its set theory expression is: $S_i \subseteq C$ is "TRUE" or $S_i \subseteq C$ is "FALSE".
- Based on subdividing relation
  Subdividing relation is a quasi-binary oriented relation whose domain of definition is a family of all possible sets. The result of set subdivision is to get pairs of disjoint sets, and the union of these disjoint sets is the original set. There are also many classes in the knowledge base that satisfy the subdivision relation, so this relation can be used to automatically

generate various types of objective questions. The expression form of set theory of subdividing relation between classes is as follows: $S_1 \cup S_2 \cup S_3 ... \cup S_i = C$ $(i > 1, S_i \cap S_j = \phi)$. Taking the generated fill in the blank questions as an example, its set theory expression is: Set C is subdivided into S_1, S_2, ____ and S_i.

- Based on strict inclusion relation
  Strict inclusion relation is a special form of inclusion relation, it is also a very important relation in knowledge base. Using strict inclusion relation to automatically generate objective questions is similar to using inclusion relation. The expression form of set theory of strict inclusion relation between classes is as follows: $S_i \subset C (i \geq 1)$, (S-subclass, i-subclass number, C-parent class). Taking the generated choice questions as an example, its set theory expression is: Set C strictly contains_____? The correct options of choice questions are the subclasses strictly contained in set C, and the incorrect options (disturbing) are the disjoint sibling and parent classes of set C.

2) Question generation strategy based on elements

- Based on role relation
  Role relation is a type of relation often used when building a knowledge base, so role relation between elements can be used to automatically generate objective questions;
- Based on binary relation
  There are many kinds of binary relations between elements in knowledge base, so system can use these binary relations to generate objective questions automatically.

3) Question generation strategy based on identifiers

Usually some sets and relations in the knowledge base have multiple identifiers in addition to the system identifier and the main identifier, so multiple identifiers of sets and relations can be used to automatically generate objective questions.

4) Question generation strategy based on axioms

Many axioms and their mathematical expressions are stored in the discrete mathematical knowledge base, so these axioms and their mathematical expressions can be used to generate objective questions automatically.

5) Question generation strategy based on multiple relations

Many relations in the knowledge base satisfy the attributes of reflexivity, symmetry and transmission, so system can use these relations and their attributes to generate objective questions. The following is a fragment in the knowledge base that satisfies this type of relation in the SCn-code:

maximal clique*
∈ *reflexive relation*
∈ *transitive relation*
∈ *binary relation*

6) Question generation strategy based on image examples

This approach uses some concepts, relations and theorems in the knowledge base and their explanatory image examples to automatically generate some objective questions.

7) Subjective question generation strategy

First, the system stores various definitions and statements interpretation questions automatically generated by this approach into the ontology of questions according to a certain structure, and then the definitions and statements using natural language expressions are manually converted into SCL-code (a special sub-language of the SC language intended for formalizing logical formulas) expressions and stored in the logical ontology.

Using these approaches proposed above can automatically generate choice questions, fill in the blank questions, judgment questions and definition explanation questions, etc. These questions, which are automatically generated using the discrete mathematical knowledge base, are stored in the ontology of questions [1], [2]. When the user needs to be tested, the system extracts specific types of questions from the ontology of questions according to the user's requirements, and then transforms them into natural language form through natural language interface.

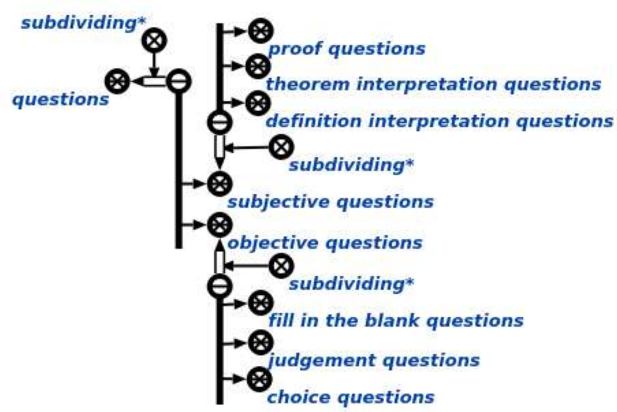


Figure 1. Hierarchy of the ontology of questions

As the carrier of various types of questions generated by storage, the ontology of questions is an important part of the automatic generation of questions and automatic verification of answers subsystem. The structure of the ontology of questions determines the efficiency of extracting the questions from the ontology of questions. Because the generated questions are divided into subjective questions and objective questions, and objective questions and subjective questions are divided into specific types, so the structure of ontology of questions can be constructed according to the type and hierarchy of the generated

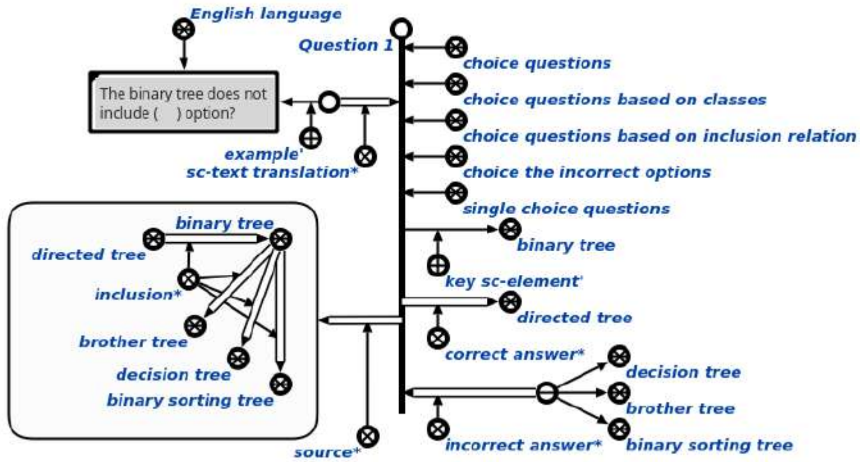


Figure 2. The semantic graph structure of choice question

questions. (Fig. 1) shows the hierarchy of the ontology of questions.

The ontology of questions can be divided into the following types according to the stored question types: choice question ontology, fill in the blank question ontology, judgment question ontology and definition interpretation question ontology, etc., and each type of ontology is used to store the corresponding types of questions. Each type of ontology is created according to the question generation strategies and the characteristics of each type of question [5]. For example, the characteristics of choice questions include:

- choice questions are single choice questions or multiple choice questions;
- for example, when using question generation strategy based on classes, if the number of subclasses that meet the conditions exceeds two, a type of question that selects options that do not meet the requirements of the question can be generated.

Lets consider the choice question ontology of the SCn-code syntax:

choice questions
$\Leftarrow$ *subdividing**:
{
- *choice questions based on multiple relations*
- *choice questions based on axioms*
- *choice questions based on image examples*
- *choice questions based on identifiers*
- *choice questions based on elements*
  $\Leftarrow$ *subdividing*:*
  {
  - *choice questions based on role relation*
  - *choice questions based on binary relation*
  }
- *choice questions based on classes*
  $\Leftarrow$ *subdividing*:*
  {
  - *choice questions based on subdividing relation*
  - *choice questions based on inclusion relation*
  - *choice questions based on strict inclusion relation*
  }
}
$\Leftarrow$ *subdividing**:
{
- *multiple choice questions*
- *single choice questions*
}
$\Leftarrow$ *subdividing**:
{
- *choice the incorrect options*
- *choice the correct options*
}

Various types of automatically generated questions are stored in the ontology of questions in the form of semantic graph. (Fig. 2) shows the semantic graph structure of choice question in the SCg-code. (Fig. 3) shows the semantic graph structure of definition interpretation question in the SCg-code [1], [4].

The automatically generated various types of questions have a structure similar to the above two types of questions, and they are stored in the corresponding ontology according to this type of structure.

The approach for automatic generation of questions and the approach of using ontology to storage the generated questions, which proposed in this article have the following advantages:

- because the knowledge bases developed using OSTIS technology have the same knowledge storage structure, so only a simple modification to the approach for automatic generation of questions proposed in

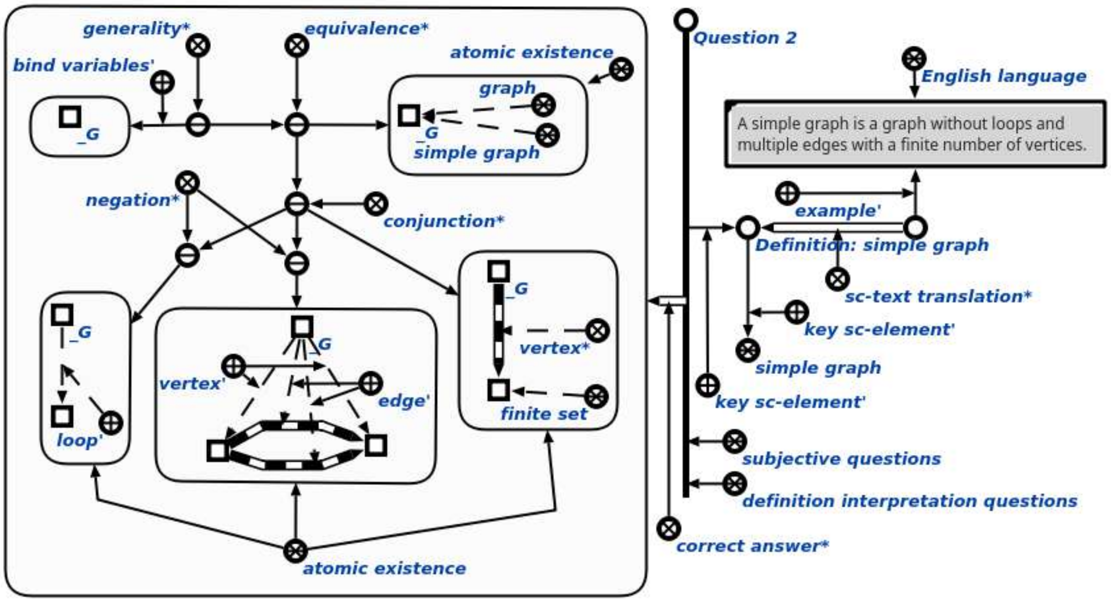


Figure 3. The semantic graph structure of definition interpretation question

this article can be used in other OSTIS systems;

- the generated questions are expressed using a unified knowledge coding language SC-code, so they do not depend on natural language, that is, the generated questions can be converted into different natural language forms only through different natural language interfaces;
- the generated questions are stored in the ontology of questions according to the types of the questions and the generation strategies, so when the questions need to be extracted from the ontology of questions, the efficiency of question extraction can be greatly improved;
- using the approach proposed in this article it is possible not only generate subjective and objective questions, but also the quality of the generated questions is very high.

The approaches for automatic generation of questions and using ontology to storage the generated questions proposed in this article solved the existing problems of the approaches for automatic generation of questions in the previous section, so the approaches proposed in this paper has certain advantages.

B. Proposed answer verification approach

In this article, the answer verification is divided into subjective question answer verification and objective question answer verification. Because objective questions have definite standard answers, so it only needs to directly compare standard answers with user answers. There are no definite answers to the subjective questions, as long as the user answers conform to the logic and meet the requirements of the questions, so it is necessary to compare the similarity between the standard answers and the user answers. According to the types of knowledge, subjective question answer verification can be divided into:

- factual knowledge answer verification;
- logical knowledge answer verification.

Factual knowledge refers to knowledge that does not contain variable types, and this type of knowledge expresses facts. Logical knowledge usually contains variables, and there are logical relationships between knowledge. Most of the answers to subjective questions are logical knowledge [1], [4], [21].

Based on the SPICE approach [21], this article proposes an approach for automatically verifying the answers to subjective questions using OSTIS technology. According to the task requirements, the approach proposed in this article needs to verify the correctness and completeness of user answers (for example, the answer is correct but incomplete, and the answer is partially correct, etc.). The answer verification approaches of factual knowledge and logical knowledge are similar, the answers described by the semantic graph are divided into sub-structures according to certain approaches and then the similarity is compared. The implementation process of the answer verification approach proposed in this article is as follows:

1) First, the knowledge base developers transform the natural language answers of subjective questions into SC-code forms and store them in the ontology of questions. We use s to represent standard answers in the form of semantic graphs.

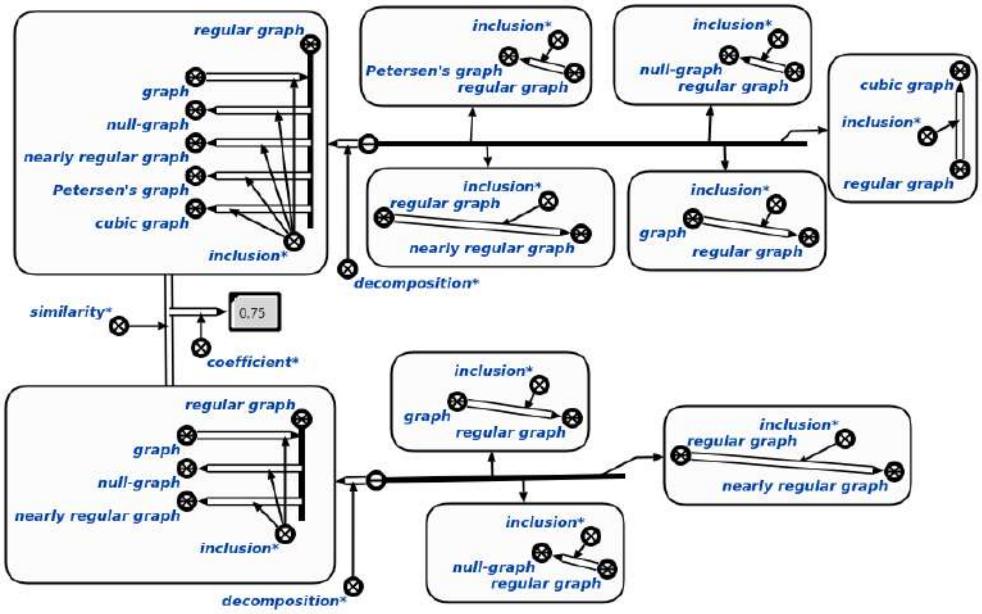


Figure 4. An example of similarity calculation of knowledge of factual types

2) Then the standard answers s and user answers u expressed using SCg-code/SCL-code [1], [2] are decomposed into sub-structures according to the rules of knowledge representation. u refers to the expression form after the natural language user's answers are transformed to SC-code through the natural language interface. The set $T_{sc}(s)$ represents all the sub-structures after the decomposition of the standard answers s, and the set $T_{sc}(u)$ represents all the sub-structures after the decomposition of the user answers u.
3) Finally, the similarity is calculated by comparing the ratio between the number of identical sub-structures and the total number of sub-structures between the standard answer and the user answer. The main calculation parameters include: precision P_{sc}, recall R_{sc}, and similarity F_{sc}. Their specific calculation process is shown in formulas (1), (2), (3).

$$P_{sc}(u,s) = \frac{|T_{sc}(u) \otimes T_{sc}(s)|}{|T_{sc}(u)|} \tag{1}$$

$$R_{sc}(u,s) = \frac{|T_{sc}(u) \otimes T_{sc}(s)|}{|T_{sc}(s)|} \tag{2}$$

$$F_{sc}(u,s) = \frac{2 \cdot P_{sc}(u,s) \cdot R_{sc}(u,s)}{P_{sc}(u,s) + R_{sc}(u,s)} \tag{3}$$

(Fig. 4) shows an example of similarity calculation of knowledge of factual types in the SC-code.

Although the verification process of answers of logical knowledge and factual knowledge is basically similar, it is necessary to consider some unique characteristics of logical semantic graphs when performing substructure decomposition, such as:

- there is a strict logical order between the nodes in the logical semantic graph, so the integrity of the logical formula cannot be broken when sub-structure decomposition is performed;
- each substructure of the decomposed logical semantic graph must have logical meaning.

The logical expression of inclusion of set is: $\forall a, a \in A \rightarrow a \in B$. (Fig. 5) shows an example of the decomposition of definition of inclusion of set into substructures in the SC-code.

Compared with existing answer verification approaches, the proposed approach in this article has many advantages:

- using semantics to compare the similarity between answers;
- because the OSTIS technology [1] has a unified way of knowledge expression, so the approach proposed in this article can calculate the similarity between texts with more complex content than the SPICE approach;
- compared with the approach of deep learning, the approach proposed in this article can judge the completeness and find the incorrect parts of the user answers through the sub-structures;
- the approach proposed in this article uses a unified knowledge coding language SC-code, so it can be used in other OSTIS systems only by modifying a few rules, and does not depend on natural language.

These advantages make up for the existing problems of answer verification in the previous section, so the answer

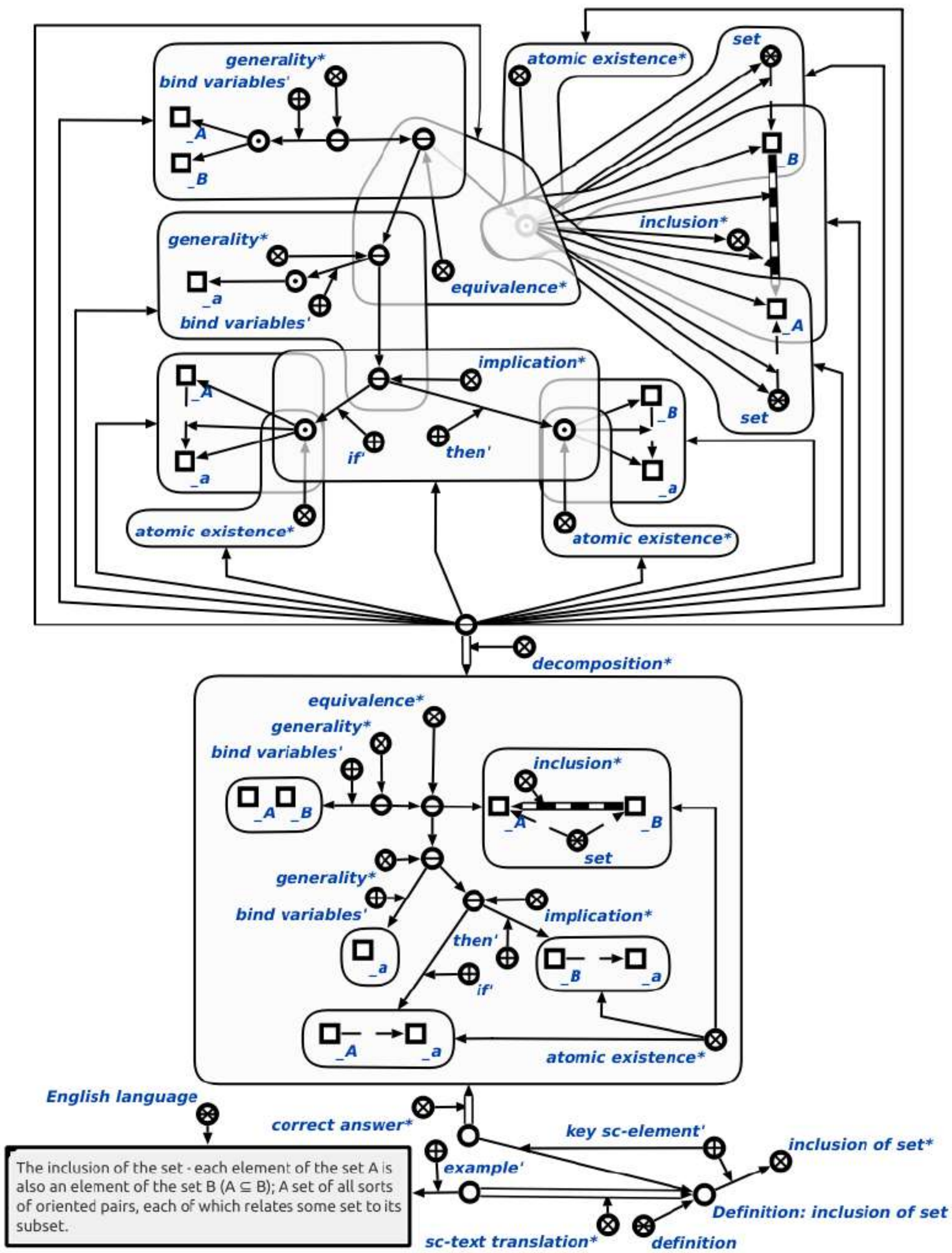


Figure 5. An example of the decomposition of definition of inclusion of set

verification approach proposed in this article has certain advantages.

The automatic question generation approach and automatic answer verification approach proposed in this article have many advantages, but there are also many problems to be solved, such as:

- how to control the quality and repetition rate of automatically generated questions in the ontology of questions, and the similarity threshold setting between standard answers and user answers;
- usually, a logical proposition may have several equivalent propositions, so the answer verification of between equivalent propositions needs to be solved;
- how to verify the answers when the concepts expressed between the standard answers and the user answers are the same but the terms are different;

IV. Conclusion and further work

This article first analyzes the development status and advantages of ILS, and then discusses the problems of the existing automatic question generation and answer verification approaches, finally, combining existing methods and OSTIS technology, this article proposes an approach for automatic generation of questions and automatic verification of answers. The proposed approach solves some existing problems well, and the subsystem of question generation and answer verification designed using this approach has good compatibility and natural language independence, so it can be used well with other OSTIS learning systems. Although the approaches proposed in this article have several advantages, there are also quite a few problems. For example, how to control the quality and rate of repetition of questions and etc. The next step is to introduce the proposed approach into the discrete mathematics learning system.

ACKNOWLEDGMENT

The work in this paper was done with the support of research teams of the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics. Authors would like to thank every researcher in the Department of Intelligent Information Technologies.

REFERENCES

[1] V. V. Golenkov and N. A. Guljakina, "Proekt otkrytoj semanticheskoj tehnologii komponentnogo proektirovanija intellektual'nyh sistem. chast' 1: Principy sozdanija project of open semantic technology for component design of intelligent systems. part 1: Creation principles]," *Ontologija proektirovanija [Ontology of design]*, no. 1, pp. 42–64, 2014.

[2] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, "Methods and tools for ensuring compatibility of computer systems," in *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2019, pp. 25–52.

[3] V. Golenkov, D. Shunkevich, I Davydenko, and N. Grakova, "Principles of organization and automation of the semantic computer systems development," in *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk, BSUIR, 2019, pp. 53–90.

[4] (2018, Dec) Ims.ostis metasystem. [Online]. Available: https://ims.ostis.net

[5] Li H. Research on item automatic generation based on DL and domain ontology. *Journal of Changchun University of Technology (Natural Science Edition)*, 2012, Vol. 33(04), pp. 461-464.

[6] Ruslan M., Le A. H., Nikiforos K. A computer-aided environment for generating multiple—choice test items. *Natural Language Engineering*, 2006, Vol. 12(02), pp. 177-194.

[7] "Protégé [Electronic resource]," mode of access: http://protege.stanford.edu. Date of access: 01.12.2019.

[8] Din X. M., Gu H. B. Research on automatic generation of Chinese multiple choice questions based on ontology. *Computer Engineering and Design*, 2010, Vol. 31(06), pp. 1397-1400.

[9] Yuan X. L. Comparison and analysis of commonly used intelligent test paper grouping algorithms. *Computer Education*, 2016, Vol. 2016(05), pp. 121-123.

[10] Andreas P., Konstantinos K., Konstantinos K. Automatic generation of multiple-choice questions from domain ontologies. *IADIS International Conference e-Learning*, 2008.

[11] Arjun S. B., Manas K., Sujan K. S. Automatic Generation of Multiple Choice Questions Using Wikipedia. *International Conference on Pattern Recognition and Machine Intelligence*, 2013, Vol. 8251, pp. 733-738.

[12] Wan C. L., Yang Y. H., Deng F. A review of text similarity calculation methods. *Information science*, 2019, Vol. 37(33), pp. 158-168.

[13] "Application of TF-IDF and cosine similarity [Electronic resource]," mode of access: http://www.ruanyifeng.com/blog/2013/03/tf-idf.html. Date of access: 01.12.2019.

[14] "NLP text similarity (TF-IDF) [Electronic resource]," mode of access: https://www.cnblogs.com/liangjf/p/8283519.html. Date of access: 05.09.2019.

[15] "Vector Space Model (VSM) Algorithm [Electronic resource]," mode of access: https://zhuanlan.zhihu.com/p/31886934. Date of access: 02.11.2019.

[16] "Principle and practice of Doc2vec [Electronic resource]," mode of access: https://blog.csdn.net/u010417185/article/details/80654558. Date of access: 02.12.2019.

[17] Li X. J. Realization of automatic scoring algorithm for subjective questions based on artificial intelligence. *Journal of Jiangnan University (Natural Science Edition)*, 2009, Vol. 08(03), pp. 292-295.

[18] Yang M., Li P. F., Zhu Q. M. Sentence similarity calculation method based on Tree-LSTM. *Journal of Peking University (Natural Science Edition)*, 2018, Vol. 54(03), pp. 481-486.

[19] Wan H. R., Zhang Y. S. Review on Research Progress of Text Similarity Calculation. *Journal of Beijing Information Science (Technology University)*, 2019, Vol. 34(01), pp. 68-74.

[20] Qin C. X., Zhu T., Zhao P. M., Zhang Y. Research progress in semantic analysis of natural language. *Library and information work*, 2014, Vol. 58(22), pp. 130-137.

[21] Peter A., Basura F., Mark J. SPICE: Semantic Propositional Image Caption Evaluation. *Computer Vision and Pattern Recognition (cs.CV)*, 2016.

Онтологический подход к автоматизации процессов генерации вопросов и контроля знаний в интеллектуальных обучающих системах

Ли Вэньцзу, Гракова Н.В., Цянь Лунвэй

В данной работе предложен подход к проектированию общей подсистемы автоматической генерации вопросов и автоматической верификации ответов в интеллектуальных обучающих систем. Данный подход предлагается на основе существующих подходов и теории онтологий, с помощью предлагаемого подхода можно автоматически генерировать различные типы вопросов из баз знаний и автоматически проверять правильность и полноту ответов пользователей. В сравнении с существующими подходами, предложенный в данной работе подход имеет определенные технические преимущества, а подсистема, разработанная с использованием предложенного в данной работе подхода, может быть применена в различных обучающих системах, построенных по технологии OSTIS.

Received 10.01.2020

Optimizing the concurrency level of network artificial intelligence computations

Anatoly Prihozhy
Information Technologies and Robotics Department
Belarusian National Technical University
Minsk, Republic of Belarus
Email: prihozhy@yahoo.com

***Abstract*—Recent achievements in development of artificial intelligence systems would be impossible without the use of high performance distributed computing platforms. This paper presents a graph model of concurrent network schedules and a technique that estimates the execution time and implementation cost over maximum weight cliques in task graphs. It proposes an algorithm for recalculating clique sets after changing the concurrency level of a schedule by adding an edge to the concurrency graph and removing the edge from the complement graph. Since the set of pairs of concurrent tasks has been found, it treats a schedule existence problem as solving a combined logical equation. The proposed model and technique are a basis for the development and implementation of network algorithms.**

***Keywords*—artificial intelligence systems; distributed computations; computing schedule; parallelization; optimization**

I. INTRODUCTION

Distributed artificial intelligence systems and semantic networks aim at solving complex knowledge acquisition, reasoning, learning, recognition, planning, decision-making and other problems in the case they process data on large-scale distributed computing platforms [1] - [3]. Distributed intelligence systems consist of numerous autonomous agents that perform learning, processing and communication tasks. Large-scale graph mining is an important technology in modeling, analyzing and solving artificial intelligence problems on large computer networks [4] - [6]. Processing such graphs is possible only by developing distributed algorithms for parallel systems. Deep neural networks are an important branch in artificial intelligence [1]. In deep machine learning, the neural networks trained by observing big data provide good solutions for problems thought to be unsolvable: autonomous driving, image classification, speech recognition, medical diagnosis, complex games and others. Parallelization strategies and concurrent algorithms for evaluation, implementations and training of deep neural networks make such networks the field of high performance parallel computing. In this paper, we extend the net scheduling techniques [7] - [9] for optimizing the concurrency level of solving intelligent problems.

II. GRAPH MODEL OF NETWORK SCHEDULE

Network scheduling determines the precedence and concurrency relations between tasks, which conserve both time and resources [7] - [9]. Let $N = 1, ...n$ be a set of task numbers. A directed graph $G_H = (N, H)$ can describe the net schedule without constraints on resources, where H is the tasks direct precedence relation. If tasks $i_1, ..., i_k$ are direct predecessors of task j in the network schedule, then j may execute when all of its predecessors have finished executing.

A binary matrix Q (Figure 1) describes data dependences between the tasks, which element $q_{i,j}$ equals 1 if i is a predecessor of j, and equals 0 otherwise. In a triple matrix W, element $w_{i,j}$ equals 0 if the tasks i and j may not execute on the same processor, equals 1 if the tasks may execute on the same processor sequentially, and equals 2 if the tasks may execute on the same processor concurrently. The last case applies when the tasks $T1$ and $T2$ are mutually exclusive [7]; that is the tasks are under conditions $if\ c1\ then\ T1\ end\ if$ and $if\ c2\ then\ T2\ end\ if$, and test variables $c1$ and $c2$ are orthogonal, i.e. they cannot take value true simultaneously. We can equivalently transform the tasks behavior to the single basic block dataflow model presented in [9].

$$Q = \begin{bmatrix} 0&1&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&1&1&0&0&1&0&0\\ 0&0&0&0&1&1&0&0&1&0&0\\ 0&0&0&0&0&1&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1&1&0&1\\ 0&0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \qquad W = \begin{bmatrix} 0&0&1&0&0&1&0&1&1&1&1\\ 0&0&0&0&0&0&0&0&0&0&0\\ 1&0&0&0&0&1&0&1&1&1&1\\ 0&0&0&0&1&0&1&0&0&0&0\\ 0&0&0&1&0&0&1&0&0&0&0\\ 1&0&1&0&0&0&0&1&1&1&1\\ 0&0&0&1&1&0&0&0&0&0&0\\ 1&0&1&0&0&1&0&0&1&1&1\\ 1&0&1&0&0&1&0&1&0&1&1\\ 1&0&1&0&0&1&0&1&1&0&1\\ 1&0&1&0&0&1&0&1&1&1&0 \end{bmatrix}$$

Figure 1. Matrices Q and W of tasks precedence and compatibility

III. SCHEDULE EXECUTION TIME AND COST

Let t_j and s_j be the execution time and implementation cost respectively of a task on processor of type j. Time t_j can be constant or variable. If $type(i)$ denotes the type of processor that executes task i, the network schedule execution time is

$$T = \max_{u \in U_{\overline{D}}} \sum_{i \in u} t_{type(i)} \tag{1}$$

where $U_{\overline{D}}$ is the set of cliques of graph $G_{\overline{D}} = (N, \overline{D})$ constructed on a set N of the nodes that represent tasks, and on a set $\overline{D}$ of edges that represent precedence of the tasks. The clique of $G_{\overline{D}}$ that gives the maximum sum of the tasks execution time defines the schedule execution time. The network schedule implementation cost is

$$S = \sum_{j=u}^{Types} S_j \times (\max_{v \in V_D} m_{jv}) \tag{2}$$

where $Types$ is the number of processor types; V_D is the set of cliques of graph $G_D = (N, D)$ constructed on the set D

of edges that represent concurrency of the tasks. The number of processors of type j needed to execute the tasks of clique v concurrently is m_{jv}. The sum of costs of all type processors defines the overall cost. Clique set V_D provides the number of all type processors. Sets $U_{\overline{D}}$ and V_D determine the path and section of maximum weight in graph G_H.

IV. OPTIMIZATION OF NETWORK SCHEDULES

The optimization of a network schedule aims at: (1) minimizing the schedule execution time, given constraints on resources; (2) minimizing the resources, given constraints on the execution time. While set D_M determines the most concurrent (and thus fastest) network schedule, a subset D of D_M determines a network schedule of less concurrency, yet lower system cost. Set D also defines execution time T and cost S. We can find up to 2^r different network schedules, where r is the cardinality of set D_M. As the tasks in any pair of set D_O of orthogonal tasks execute concurrently and does not require additional resources, we can always include D_O in D. For instance, the sample matrix Q (Figure 1) is potentially a source for generating 2^{39} network schedules. Synthesizing a net schedule involves solving one of two optimization problems, depending on the optimization criteria selected:

$$\min_{D \in D_M} \{T_D | S_D \leq S_O\} \tag{3}$$

or

$$\min_{D \in D_M} \{S_D | T_D \leq T_O\} \tag{4}$$

where T_O and S_O are constraints on execution time and implementation cost. We account for the estimates of execution time (1) and implementation cost (2) to calculate the value of T_O and S_O .

Two techniques let us generate D while solving problems (3) and (4) by consecutively adding pairs to D and by consecutively removing pairs from D. The first technique solves problem (3) and starts with set $D = D_O$. The second technique solves problem (4), starts with set D_M, and never removes orthogonal pairs of D_O from D. Because of the concurrency of orthogonal tasks, the pairs of D_O do not require additional execution time and implementation cost.

Both techniques select a pair for including in or removing from D by analyzing the maximum-weight cliques of sets $U_{\overline{D}}$ and V_D. First of all, they select pairs that decrease the execution time and not increase the implementation cost.

V. CALCULATION OF GRAPH CLIQUE SETS

Adding or removing a pair from D changes the clique set according to four rules. Two rules transform $U_{\overline{D}}$ into $U_{\overline{D"}}$ when we add pair $d = (i, j) \in D_M$ to set D creating new set $D" = D \cup \{d\}$. The first rule splits a clique containing tasks i and j into two new cliques of less cardinality; the second rule allows the removal of cliques from the new set $D"$:

- Rule 1 (splitting) - If element $u \in U_{\overline{D}}$ satisfies the condition that $\{i, j\} \in u$, then the elements $u \backslash \{i\}$ and $u \backslash \{j\}$ are added to set $U_{\overline{D"}}$; otherwise element u is
- Rule 2 (absorbing) - If in set $U_{\overline{D"}}$ two elements u' and $u"$ exist for which $u' \supseteq u"$, then element $u"$ is removed from the set

Two additional rules recalculate set V_D as new set $V_{D"}$. The third rule combines two cliques containing both tasks i and j into a new clique that is included into set $V_{D"}$. The fourth rule removes the absorbed cliques from the set:

- Rule 3 (merging) - If $v' \cup v" \supseteq \{i, j\}$ is true for $v', v" \in V_D$ then element $v = (v' \cap v") \cup \{i, j\}$ is added to $V_{D"}$. All elements of V_D are also included in $V_{D"}$
- Rule 4 (absorbing) - If in set $V_{D"}$ two elements v' and $v"$ exist for which $v' \supseteq v"$ then element $v"$ is removed from the set

If we remove pair d from set D and $D' = D \backslash \{d\}$ is the new set, then rules 1 and 2 transform set V_D into the set $V_{D"}$, and rules 3 and 4 transform set $U_{\overline{D}}$ into the set $U_{\overline{D"}}$.

Solving problem (3) to minimize the execution time for matrix Q with one processor of type $p1$ and two processors of type $p2$ ($t_{p1} = 100ms, t_{p2} = 40ms$) produces set D, which contains 31 pairs, as described by the zero elements of the top right part of matrix Q_D^x (Figure 2). The markings along the column heads indicate the tasks, executed on processors of type $p1, p2$ and control type c. No pair is added to D without increasing the number of processors (Figure 3). For each clique of set $U_{\overline{D}}$, the execution time is the sum of the clique tasks' execution time. The overall time is 340 ms (Figure 4).

$$Q_D^x = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & x & 0 & x & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & x \\ 0 & 0 & 0 & 0 & 1 & 1 & x & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & x & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & x \\ 0 & 0 & 0 & \overline{x} & \overline{x} & 0 & 0 & 1 & 1 & 0 & 1 \\ \overline{x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & x & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \overline{x} & 0 & 0 & 0 & 0 & 0 & 0 & \overline{x} & 0 & 0 & x \\ 0 & 0 & \overline{x} & 0 & 0 & \overline{x} & 0 & 0 & 0 & \overline{x} & 0 \end{bmatrix}$$

Figure 2. Matrix Q_D^x

No	1	c	1	2	2	1	2	1	1	1	1	p1	p2
1	0	1	0	0	0	0	0	0	1	1	0	2	0
2	0	1	0	0	0	1	0	1	0	0	0	2	0
3	1	0	0	0	0	0	0	0	1	0	0	2	0
4	0	1	1	0	0	0	1	0	0	1	0	2	1
5	0	1	1	1	0	0	0	0	0	1	0	2	1
6	0	1	1	1	0	0	0	1	0	0	0	2	1
7	1	0	0	0	0	1	1	0	0	0	0	2	1
8	0	0	0	0	0	0	0	0	1	0	1	2	0
9	0	1	0	0	0	1	1	0	0	1	0	2	1
10	1	0	1	1	0	0	0	0	0	0	0	2	1
11	1	0	1	0	0	0	1	0	0	0	0	2	1
12	0	0	0	1	0	0	0	1	0	0	1	2	1
13	0	0	0	0	1	0	0	1	0	0	1	2	1
14	0	1	0	0	1	0	0	0	0	1	0	1	1
15	0	1	0	0	1	0	0	1	0	0	0	1	1
16	1	0	0	0	1	0	0	0	0	0	0	1	1
											max=	2	1

Figure 3. Clique set V_D for matrix Q_D^x

If we add pair (i, j) to set D, tasks i and j are concurrent; if (i, j) is not included in set D_M, task i precedes task j. For pair (i, j) of set D_M not included in set D, we know that tasks i and j are not concurrent, but do not know whether i should precede j or j should precede i.

Introducing Boolean variable x_{ij} into matrix Q_D^x for pair (i, j) and its negation $\overline{x}_{ij}$ for pair (j, i) solves this problem. If x_{ij} equals 1, task i precedes task j. If the value equals 0, j precedes i. Thus, while many net schedules possible for a given D, for some set D no net schedule exists.

No	1	c	1	2	2	1	2	1	1	1	1	T
1	1	0	0	0	0	0	0	1	0	1	0	120
2	0	0	1	0	0	1	0	0	0	0	1	120
3	1	0	0	0	0	0	0	0	0	1	1	120
4	0	0	0	0	0	0	1	1	1	0	0	180
5	0	0	0	1	1	1	0	0	1	0	0	280
6	0	0	0	1	1	0	1	0	1	0	0	340
7	0	0	1	0	1	1	0	0	1	0	0	220
8	1	1	0	0	0	0	0	0	0	0	1	280
9	0	0	0	0	0	0	0	0	0	0	1	140
											max=	340

Figure 4. Clique set $U_{\overline{D}}$ for matrix Q_D^x

VI. SOLVING THE EXISTENCE PROBLEM

For set D and the given values of variables x_{ij}, a net schedule exists if the matrix derived from the matrix Q_D^x by substituting the variable values describes a transitive relation. This transitivity condition expresses the requirement that the net schedule must have the level of concurrency the set D defines. The relation is transitive if the logical equations (5)-(7) have at least one solution for x_{ij}. In the equations, variables z_{ij} are intermediate. Equation (5) describes the transitivity condition for the elements of set D, and equation (6) describes the transitivity condition for the elements out of D. One algorithm effectively solves the equations (5)-(7) by constructing a graph G_D^x and searching for its non-conflicting labeling (Figure 5).

$$\sum_{\substack{i,j,k \in N, (i,j) \in D \\ (i,k) \notin D, (k,j) \notin D}} [(z_{ik} \wedge z_{kj}) \vee (z_{jk} \wedge z_{ki})] \tag{5}$$

$$\sum_{\substack{i,j,k \in N, (i,j) \notin D \\ (i,k) \notin D, (k,j) \notin D}} [(z_{ij} \wedge z_{ik} \wedge z_{kj}) \vee (z_{ji} \wedge z_{jk} \wedge z_{ki})] \tag{6}$$

$$z_{ij} = \begin{cases} 1, & \text{if } (i,j) \notin D_M \text{ and } i < j \\ 0, & \text{if } (i,j) \notin D_M \text{ and } j < i \\ x_{ij}, & \text{if } (i,j) \in D_M \backslash D \text{ and } i < j \\ \overline{x_{ij}}, & \text{if } (i,j) \in D_M \backslash D \text{ and } j < i \end{cases} \tag{7}$$

The graph nodes are variables x_{ij} that correspond to non-concurrent pairs of tasks. The algorithm introduces edge (x_{ij}, x_{ik}) if tasks j and k are concurrent and pair (j,k) belongs to set D. It labels the graph nodes with 0 and 1. The initial label 1 is assigned to the nodes which belong to set $\{x_{ij} | (i,j) \notin D_M, i,j = 1,...,n, i < j\}$ of the Boolean variables that correspond to the non-concurrent task pairs not included in set D_M. If an edge connects two variables x_{ij} and x_{jk} that satisfy constraint $i < j < k$, it is labeled +, otherwise it's labeled -. Labeling two variables and the edge connecting them creates one type of conflict if the variable labels are the same and the edge label is +, or the variable labels are different and the edge label is -. If the graph has at least one of this first type of conflict, equation (5) has no solution. For variables x_{ij}, x_{ik}, and x_{kj} where $i < k < j$, a second type of conflict occurs if variable x_{ij}'s value equals 0 (1) and the values of x_{ik} and x_{kj} equal 1 (0). If the graph has at least one such conflict, equation (6) has no solution. To generate a net schedule, the algorithm must label the nodes in such a way as to avoid the conflicts of both types.

Removing nine pairs from D updates matrix Q_D^x (Figure 6) in such a way as to satisfy equations (5)-(7). Figures 7 and

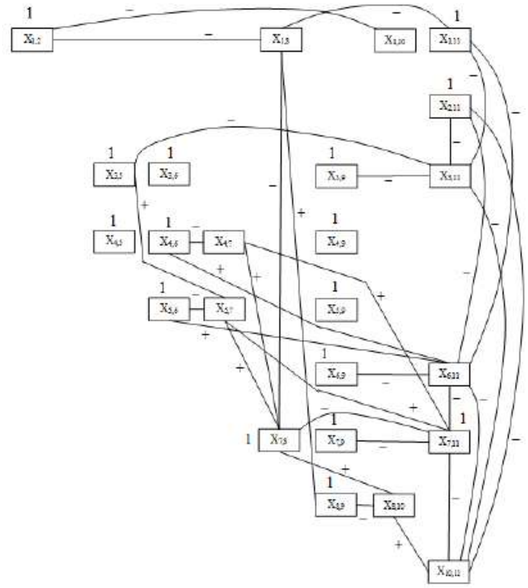


Figure 5. Conflicts graph G_D^x for logical equation $L1$

8 show the updated clique sets V_D and $U_{\overline{D}}$. The concurrent schedule execution time has increased from 340 ms to 380 ms.

$$Q_D^x = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & x & 0 & x & x & x & 1 \\ 0 & 0 & x & 0 & 0 & x & 0 & 0 & x & 0 & 1 \\ 0 & \overline{x} & 0 & 0 & 1 & 1 & x & 0 & 1 & 0 & x \\ 0 & 0 & 0 & 0 & 1 & 1 & x & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & x & 0 & 1 & 0 & 0 \\ \overline{x} & \overline{x} & 0 & 0 & 0 & 0 & x & 0 & 1 & x & x \\ 0 & 0 & 0 & \overline{x} & \overline{x} & \overline{x} & 0 & 1 & 1 & 0 & 1 \\ \overline{x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & x & 0 \\ \overline{x} & \overline{x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & x \\ \overline{x} & 0 & 0 & 0 & 0 & \overline{x} & 0 & \overline{x} & \overline{x} & 0 & x \\ 0 & 0 & \overline{x} & 0 & 0 & \overline{x} & 0 & 0 & \overline{x} & \overline{x} & 0 \end{bmatrix}$$

Figure 6. Updated matrix Q_D^x

VII. RESULTS

We have used the proposed model and technique of optimizing the concurrency level of network schedules as the basis of a graph language and a tool for development and execution of network algorithms. The graph vertices correspond to tasks. The graph edges correspond to data dependences between the tasks, and correspond to a control flow that implements constraints on resources. Tokens mark edges of the graph. The graph operates over firing vertices and moving tokens from input to output edges. The vertex firing can be conditional and cyclic, thus representing the behavior of all type of control structures in concurrent algorithms. The optimization of the concurrency level of a network schedule changes algorithm graph by adding or removing edges and tokens. Such a procedure of transformation is capable of taking into account the amount of available computational resources, while preserving the mapping of the input data onto the output data.

VIII. CONCLUSION

This paper has represented the concurrency level of a computing network schedule in an artificial intelligence system with

No	1	c	1	2	2	1	2	1	1	1	1	p1	p2
1	0	1	0	0	0	0	0	0	1	1	0	1	1
2	0	1	0	0	0	1	0	1	0	0	0	1	1
3	1	0	0	0	0	0	0	0	1	0	0	1	1
4	0	1	1	0	0	0	1	0	0	1	0	2	0
5	0	1	1	1	0	0	0	0	0	1	0	2	1
6	0	1	1	1	0	0	0	1	0	0	0	2	1
7	1	0	0	0	0	1	1	0	0	0	0	2	1
8	0	0	0	0	0	0	0	0	1	0	1	1	0
9	0	1	0	0	0	1	1	0	0	1	0	2	1
10	1	0	1	1	0	0	0	0	0	0	0	2	1
11	1	0	1	0	0	0	1	0	0	0	0	2	1
12	0	0	0	1	0	0	0	1	0	0	1	2	1
13	0	0	0	0	1	0	0	1	0	0	1	1	1
14	0	1	0	0	1	0	0	0	0	1	0	1	1
15	0	1	0	0	1	0	0	1	0	0	0	1	1
											max=	2	1

Figure 7. Updated clique set V_D for matrix Q_D^x

No	1	c	1	2	2	1	2	1	1	1	1	T
1	0	0	0	0	0	1	1	0	1	0	1	220
2	1	0	0	0	0	1	0	0	1	1	1	200
3	1	1	0	0	0	1	0	0	1	0	1	160
4	0	1	1	0	0	1	0	0	1	0	1	160
5	1	0	0	0	0	0	0	1	1	1	0	160
6	0	0	0	1	1	1	1	0	1	0	0	380
7	0	0	0	0	0	0	1	1	1	0	0	180
8	0	0	1	0	1	1	0	0	1	0	0	220
											max=	380

Figure 8. Updated clique set $U_{\overline{D}}$ for matrix Q_D^x

a set of pairs of tasks executed in parallel. We have estimated the execution time and implementation cost of the schedule over the clique set of the concurrency graph and over the clique set of the complement graph. Our optimization algorithm minimizes either the schedule execution time or resources. It finds the optimal concurrency level of network computations by solving the schedule existence problem at each step of updating the set of pairs of concurrent tasks.

REFERENCES

[1] Ben-Nun, T. Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis / T. Ben-Nun, T. Hoefler // ETH Zurich, Department of Computer Science, Zürich. – 2018, 8006, Switzerland, pp. 1-47.

[2] Lee, W-P. Designing a parallel evolutionary algorithm for inferring gene networks on the cloud computing environment / W-P. Lee, Y-T. Hsiao, W-C. Hwang // BMC Syst Biol. – 2014, Vol. 8, No. 5, doi:10.1186/1752-0509-8-5.

[3] Qiao, S. A. Fast Parallel Community Discovery Model on Complex Networks Through Approximate Optimization / S. Qiao, N. Han, Y. Gao, R-H. Li, J. Huang, J. Guo, L. A. Gutierrez, X. Wu // IEEE Trans. On Knowledge and Data Eng. – 2018, Vol.30, No. 9, pp. 1638-1651.

[4] Lattanzi, S. Distributed Graph Algorithmics: Theory and Practice / S. Lattanzi, V. Mirrokni. // Proc. Eighth ACM Int. Conf. on Web Search and Data Mining – WSDM'15, Shanghai, China, 2015, pp. 419-420.

[5] Sakr, S. Large-Scale graph processing using Apache Giraph / S. Sakr, F. M. Orakzai, I. Abdelaziz, Z. Khayyat. // Springer, 2016. – 197 p.

[6] Khayyat, Z. Mizan: Optimizing Graph Mining in Large Parallel Systems / Z. Khayyat, K. Awara, H. Jamjoom, A. Alonazi, D. Williams, P. Kalnis // Proc. 8th ACM European Conference on Computer Systems. – ACM, 2013 pp. 169-182.

[7] Prihozhy, A. A. Analysis, transformation and optimization for high performance parallel computing / A. A. Prihozhy. – Minsk : BNTU, 2019. – 229 p.

[8] Prihozhy, A. Techniques for optimization of net algorithms / A. Prihozhy, D. Mlynek, M. Solomennik, M. Mattavelli // Proc. Int. Conf. PARELEC'2002. – IEEE CS, Los Alamitos, California, 2002, pp.211-216.

[9] Prihozhy, A. Net scheduling in High-Level Synthesis / A. Prihozhy // IEEE Design and Test of Computers, 1996, No.1, pp.26-35.

Оптимизация уровня параллелизма сетевых вычислений в системах искусственного интеллекта

Прихожий А. А.

Последние достижения в системах искусственного интеллекта невозможны без использования распределенных вычислительных платформ. В статье рассматривается графовая модель сетевых планов, а параметры планов оценивается посредством клик на графах предшествования и распараллеленности. Предлагается алгоритм пересчета клик при изменении уровня параллелизма вычислений посредством добавления или удаления ребра в графе. Для найденного множества пар распараллеленных задач решается проблема существования плана. Алгоритм распараллеливания минимизирует либо время выполнения плана, либо используемые ресурсы. Предлагаемые модель и алгоритм реализованы в программном обеспечении.

Received 23.12.2019

An Architecture of Semantic Information Extraction Tool Text2ALM

1st Yuliya Lierler
University of Nebraska Omaha
Omaha,USA
ylierler@unomaha.edu

2nd Craig Olson
University of Nebraska Omaha
Omaha,USA
craig.olson19@gmail.com

***Abstract*—In this work we design a narrative understanding tool TEXT2ALM. This tool uses an action language $\mathcal{ALM}$ to perform inferences on complex interactions of events described in narratives. The methodology used to implement the TEXT2ALM system was originally outlined by Lierler, Inclezan, and Gelfond [11] via a manual process of converting a narrative to an $\mathcal{ALM}$ model. It relies on a conglomeration of resources and techniques from two distinct fields of artificial intelligence, namely, natural language processing and knowledge representation and reasoning. The effectiveness of system TEXT2ALM is measured by its ability to correctly answer questions from the bAbI tasks by Facebook Research in 2015. This tool matched or exceeded the performance of state-of-the-art machine learning methods in six of the seven tested tasks.**

I. Introduction

The field of Information Extraction (IE) is concerned with gathering snippets of meaning from text and storing the derived data in structured, machine interpretable form. Consider a sentence *BBDO South in Atlanta, which handles corporate advertising for Georgia-Pacific, will assume additional duties for brands like Angel Soft toilet tissue and Sparkle paper towels, said Ken Haldin, a spokesman for Georgia-Pacific from Atlanta.* A sample IE system that focuses on identifying organizations and their corporate locations may extract the following predicates from this sentence: $locatedIn(BBDOSouth, Atlanta)$ and $locatedIn(GeorgiaPacific, Atlanta)$. These predicates can then be stored either in a relational database or a logic program, and queried accordingly by well-known methods in computer science. Thus, IE allows us to turn unstructured data present in text into structured data easily accessible for automated querying. In this paper, we focus on an IE system that is capable of processing simple narratives with *action verbs*, in particular, verbs that express physical acts such as *go*, *give*, and *put*. Consider a sample narrative, named *JS*, discourse:

John traveled to the hallway. (1)

Sandra journeyed to the hallway. (2)

The actions *travel* and *journey* in the narrative describe changes to the narrative's environment, and can be coupled with the reader's commonsense knowledge to form and alter the reader's mental picture for the narrative. For example, after reading sentence (1), a human knows that (i) *John* is the subject of the sentence and *traveled* is an action verb describing an action performed by *John*; and (ii) *traveled* describes the act of motion, and specifically that *John*'s location changes from an arbitrary initial location to a new destination, the *hallway*. Lierler et al. [11] outline a methodology for constructing a Question Answering (QA) system by utilizing IE techniques. Their methodology focuses on performing inferences using the complex interactions of events in narratives. Their process utilizes an action language $\mathcal{ALM}$ [7] and an extension of the VERBNET lexicon [10], [16]. Language $\mathcal{ALM}$ enables a system to structure knowledge regarding complex interactions of events and implicit background knowledge in a straight-forward and modularized manner. The represented knowledge is then used to derive inferences about a given text. The proposed methodology assumes the extension of the VERBNET lexicon with interpretable semantic annotations in $\mathcal{ALM}$. The VERBNET lexicon groups English verbs into classes allowing us to infer that such verbs as *travel* and *journey* practically refer to the same class of events.

The processes described in [11] are exemplified via two sample narratives that were completed manually. The authors translated those narratives to $\mathcal{ALM}$ programs by hand and wrote the supporting $\mathcal{ALM}$ modules to capture knowledge as needed. A narrative understanding system developed within this work, TEXT2ALM, automates the method from [11]. When considering the *JS* discourse as an example, system TEXT2ALM produces a set of facts in spirit of the following:

$$move(john, hallway, 0) \quad move(sandra, hallway, 1) \quad (3)$$

$$loc\_in(john, hallway, 1) \quad loc\_in(john, hallway, 2) \quad (4)$$

$$loc\_in(sandra, hallway, 2) \quad (5)$$

where $0, 1, 2$ are time points associated with occurrences of described actions in the *JS* discourse. Intuitively, time point 0 corresponds to a time prior to utterance of sentence (1). Time point 1 corresponds to a time upon the completion of the event described in (1). Facts in (3)-(5) allow us to provide grounds for answering

This is an abstract of the paper titled "Information Extraction Tool Text2Alm: From Narratives to Action Language System Descriptions" that appeared at Proceedings of the 35th International Conference on Logic Programming in 2019. The work was partially supported by the NSF 1707371 grant.

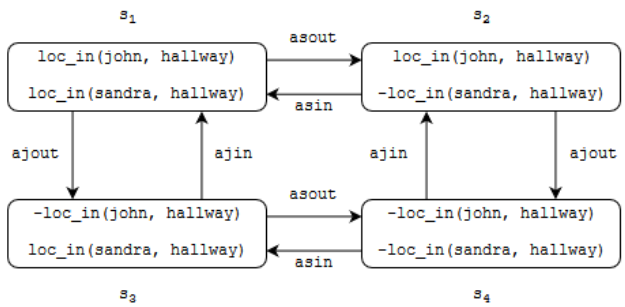


Figure 1. Sample transition diagram capturing the *JS* discourse.

questions related to the *JS* discourse such as: Is *John* inside the *hallway* at the end of the story (time 2)? Indeed, given the fact $loc\_in(john, hallway, 2)$. Who is in the *hallway* at the end of the story? John and Sandra constitute an answer given $loc\_in(john, hallway, 2)$ and $loc\_in(sandra, hallway, 2)$.

II. Background

a) NLP Resource VERBNET*:* VERBNET is a domain-independent English verb lexicon organized into a hierarchical set of verb classes [10], [16]. The verb classes aim to achieve syntactic and semantic coherence between members of a class. Each class is characterized by a set of verbs and their thematic roles. For example, the verb *run* is a member of the VERBNET class RUN-51.3.2. This class is characterized by (i) 96 members including verbs such as *bolt, frolic, scamper,* and *weave*, (ii) four thematic roles, namely, *theme*, *initial location*, *trajectory* and *destination*.

b) Dynamic Domains, Transition Diagrams, and Action Language $\mathcal{ALM}$: Action languages are formal KRR languages that provide convenient syntactic constructs to represent knowledge about dynamic domains. The knowledge is compiled into a transition diagram, where nodes correspond to possible states of a considered dynamic domain and edges correspond to actions/events whose occurrence signal transitions/changes in the dynamic system.

The *JS* discourse exemplifies a narrative modeling a dynamic domain with three entities *John, Sandra, hallway* and four actions, specifically:

1) *ajin* – *John* travels into the *hallway*,
2) *ajout* – *John* travels out of the *hallway*,
3) *asin* – *Sandra* travels into the *hallway*, and
4) *asout* – *Sandra* travels out of the *hallway*.

The transition diagram capturing the possible states of this domain is given in Figure 1. State $s1$ designates the state where the location of *John* and *Sandra* is the *hallway*. Likewise, state $s2$ characterizes the state where *John*'s location is the *hallway*, but *Sandra*'s location is not the *hallway*. Occurrence of action $asout$ is responsible for the transition from state $s1$ to state $s2$.

Scenarios of a dynamic domain correspond to *trajectories* in the domain's transition diagram. Trajectories are sequences of alternating states and actions. A trajectory captures the sequence of events, starting with the initial state associated with time point 0. Each edge is associated with the time point incrementing by 1. Consider a sample trajectory $\langle s4, ajin, s2, asin, s1\rangle$ for the transition diagram in Figure 1. It captures the following scenario:

- *John* and *Sandra* are not in the *hallway* at the initial time point 0,
- *John* travels into the *hallway* at time point 0, resulting in a new state of the dynamic system to be $s2$ (*John* is in the *hallway*, while *Sandra* is not) at time 1,
- *Sandra* travels into the *hallway* at time 1, resulting in a new state of the dynamic system to be $s3$ (*John* and *Sandra* are both in the *hallway*) at time 2.

It is easy to see how this sample trajectory captures the scenario of the *JS* discourse.

In this work we utilize an advanced action language $\mathcal{ALM}$ [7] to model dynamic domains of given narratives. This language provides an ability to capture the commonalities of similar actions. We illustrate the syntax and semantics of $\mathcal{ALM}$ using the *JS* discourse dynamic domain by first defining an $\mathcal{ALM}$ "system description" and then an $\mathcal{ALM}$ "history" for this discourse.

In language $\mathcal{ALM}$, a dynamic domain is described via a *system description* that captures a transition diagram specifying the behavior of a given domain. An $\mathcal{ALM}$ system description consists of a theory and a structure. A *theory* is comprised of a hierarchy of modules, where a module represents a unit of general knowledge. A module contains declarations of sorts, attributes, and properties of the domain, together with axioms describing the behavior of actions and properties. There are four types axioms, namely, (i) dynamic causal laws, (ii) executability conditions, (iii) state constraints, and (iv) function definitions. The properties that can be changed by actions are called fluents and modeled by functions in $\mathcal{ALM}$. The *structure* declares instances of entities and actions of the domain. Figure 2 illustrates these concepts with the $\mathcal{ALM}$ formalization of the *JS* discourse domain. The resulting formalization depicts the transition diagram that we can obtain from the diagram in Figure 1 by erasing the edges annotated with $ajout$ and $asout$.

The *JS* discourse theory is composed of a single module containing the necessary knowledge associated with the domain. The module starts with the declarations of sorts (*agents*, *points*, *move*) and fluents (`loc_in`). Sorts *universe* and *actions* are predefined in $\mathcal{ALM}$ so that any entity of a domain is considered of *universe* sort, whereas any declared action/event is considered of *actions* sort. While declaring actions, the $\mathcal{ALM}$ user specifies its attributes, which are roles that entities participating in the

```
system description JS_discourse
  theory JS_discourse_theory
    module JS_discourse_module
    sort declarations
     points, agents :: universe
     move :: actions
      attributes
       actor : agents -> booleans
       origin : points -> booleans
      destination : points -> booleans
    function declarations
     fluents
     loc_in : agents * points -> booleans
    axioms
    dynamic causal laws
      occurs(X) causes loc_in(A,D)
       if instance(X,move), actor(X,A),
            destination(X,D).
    executability conditions
     impossible occurs(X) if
       instance(X,move), actor(X,A),
       loc_in(A,P), origin(X,O), P!=O.
     impossible occurs(X) if
       instance(X,move), actor(X,A),
       loc_in(A,P), destination(X,D), P=D.
    structure john_and_sandra
    instances
      john, sandra in agents
      hallway in points
        ajin in move
         actor(john) = true
         destination(hallway) = true
         asin in move
         actor(sandra) = true
         destination(hallway) = true
```

Figure 2. An $\mathcal{ALM}$ system description formalizing the *JS* discourse dynamic domain

action take. For instance, the attributes of *move* include *actor*, *origin*, and *destination*. Here we would like a reader to draw a parallel between the notions of an attribute and a VERBNET thematic role.

There are two types of axioms in the *JS* discourse theory: dynamic causal laws and executability conditions. The only dynamic causal law states that if a *move* action occurs with a given *actor* and *destination*, then the *actor*'s location becomes that of the *destination*. The executability conditions restrict an action from occurring if the action is an instance of *move*, where the *actor* and *actor*'s location are defined, but either (i) the *actor*'s location is not equal to the *origin* of the *move* event or (ii) the *actor*'s location is already the *destination*.

An $\mathcal{ALM}$ structure in Figure 2 defines the entities of sorts *agents*, *points*, and *actions* that occurred in the *JS* discourse. For example, it states that *john* and *sandra* are *agents*. Then, the structure declares an action $ajin$ as an instance of *move* where *john* is the *actor* and *hallway* is the *destination*. Likewise, $asin$ is declared as an instance of *move*, where *sandra* is the *actor* and *hallway* is the *destination*.

In $\mathcal{ALM}$, a *history* is a particular scenario described by observations about the values of fluents and events that occur. In the case of narratives, a history describes the sequence of events by stating occurrences of specific actions at given time points. For instance, the *JS* discourse history contains the events

- *John moves* to the *hallway* at the beginning of the story (an action $ajin$ occurs at time 0) and
- *Sandra moves* to the *hallway* at the next point of the story (an action $asin$ occurs at time 1).

The following history is appended to the end of the system description in Figure 2 to form an $\mathcal{ALM}$ program for the *JS* disocurse. We note that $happened$ is a keyword that captures the occurrence of actions.

```
happened(ajin, 0).
happened(asin, 1).
```

c) An $\mathcal{ALM}$ Solver CALM*:* System CALM is an $\mathcal{ALM}$ solver developed at Texas Tech University by Wertz, Chandrasekan, and Zhang [18]. It uses an $\mathcal{ALM}$ program to produce a "model" for an encoded dynamic domain. Behind the scene system CALM (i) constructs a logic program under stable model/answer set semantics [5], whose answer sets/solutions are in one-to-one correspondence with the models of the $\mathcal{ALM}$ program, and (ii) uses an answer set solver SPARC [1] for finding these models. The $\mathcal{ALM}$ program in Figure 2 follows the CALM syntax. However, system CALM requires two additional components for this program to be executable. The user must specify (i) the computational task and (ii) the max time point considered.

System CALM can solve temporal projection and planning computational tasks. Our work utilizes *temporal projection*, which is the process of determining the effects of a given sequence of actions executed from a given initial situation (which may be not fully determined). In the case of a narrative, the initial situation is often unknown, whereas the sequence of actions are provided by the discourse. Inferring the effects of actions allows us to properly answer questions about the domain. To perform temporal projection, we insert the line following statement in the $\mathcal{ALM}$ program prior to the history:

```
temporal projection
```

Additionally, CALM requires the max number of time points/steps to be stated. Intuitively, we see this number as an upper bound on the "length" of considered trajectories. In temporal projection problems, this information denotes the final state's time point. To define the max step for the *JS* discourse $\mathcal{ALM}$ program, we insert the following line in the $\mathcal{ALM}$ program:

```
max steps 3
```

For the case of the temporal projection task, a model of an $\mathcal{ALM}$ program is a trajectory in the transition system captured by the $\mathcal{ALM}$ program that is "compatible" with the provided history. For example, a trajectory $\langle s4, ajin, s2, asin, s1\rangle$ is the only model for the *JS* discourse $\mathcal{ALM}$ program. For the *JS* discourse $\mathcal{ALM}$ program, the CALM computes a model that includes the following expressions:

```
happened(ajin, 0),  happened(asin, 1),
loc_in(john, hallway, 1),
loc_in(sandra, hallway, 2),
loc_in(john, hallway, 2)
```

d) $\mathcal{ALM}$ Knowledge Base COREALMLIB*:* The COREALMLIB is an $\mathcal{ALM}$ library of generic commonsense knowledge for modeling dynamic domains developed by Inclezan [6]. The library's foundation is the Component Library or CLib [2], which is a collection of general, reusable, and interrelated components of knowledge. CLib was populated with knowledge stemming from linguistic and ontological resources, such as VERBNET, WORDNET, FRAMENET, a thesaurus, and an English dictionary. The COREALMLIB was formed by translating CLib into $\mathcal{ALM}$ to obtain descriptions of 123 action classes grouped into 43 reusable modules. The modules are organized into a hierarchical structure and contain action classes.

III. SYSTEM TEXT2ALM ARCHITECTURE

Lierler, Inclezan, and Gelfond [11] outline a methodology for designing IE/QA systems to make inferences based on complex interactions of events in narratives. This methodology is exemplified with two sample narratives that were completed manually by the authors. System TEXT2ALM automates the process outlined in [11]. Figure 3 pretenses the architecture of the system. It implements four main tasks/processes:

1) TEXT2DRS Processing – Entity, Event, and Relation Extraction
2) DRS2ALM Processing – Creation of $\mathcal{ALM}$ Program
3) CALM Processing – $\mathcal{ALM}$ Model Generation and Interpretation
4) QA Processing

In Figure 3, each process is denoted by its own column. Ovals identify inputs and outputs. Systems or resources are represented with white, grey, and black rectangles. White rectangles denote existing, unmodified resources. Grey rectangles are used for existing, but modified resources. Black rectangles signify newly developed subsystems. The first three processes form the core of TEXT2ALM, seen as an IE system. The QA Processing component is specific to the bAbI QA benchmark that we use for illustrating the validity of the approach advocated by TEXT2ALM. The system is available at `https://github.com/cdolson19/Text2ALM`.

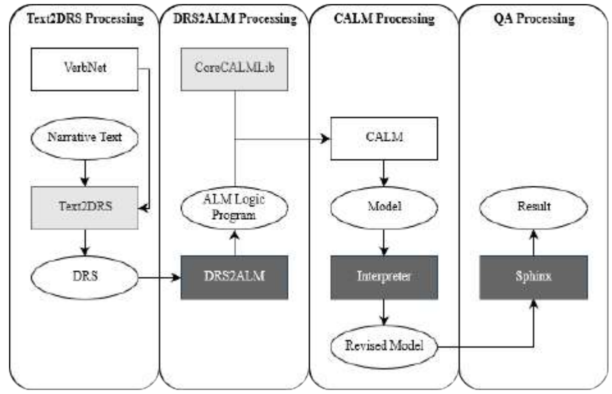


Figure 3. System TEXT2ALM Architecture

a) TEXT2DRS *Processing:* To produce $\mathcal{ALM}$ system descriptions for considered narratives, the method by Lierler et al. [11] utilizes NLP resources, such as semantic role labeler LTH [8], parsing and coreference resolution tools of CORENLP [13], and lexical resources PROPBANK [17] and SEMLINK [3]. System TEXT2DRS [12] was developed with these resources to deliver a tool that extracts entities, events, and their relations from given narratives. The TEXT2DRS tool became a starting point in the development of TEXT2ALM. The output of the TEXT2DRS system is called a discourse representation structure, or DRS [9]. A DRS captures key information present in discourse in a structured form. For example, for the *JS* discourse its DRS will include information that there are three entities and two events that take part in the *JS* narrative. It will also identify both of the events of the discourse with the VERBNET class RUN-51.3.2-1.

b) DRS2ALM *Processing:* The DRS2ALM subsystem is concerned with combining commonsense knowledge related to events in a discourse with the information from the DRS generated by TEXT2DRS. The goal of this process is to produce an $\mathcal{ALM}$ program consisting of a system description and a history pertaining the scenario described by the narrative. The system description is composed of a theory containing relevant commonsense knowledge, and a structure that is unique for a given narrative. One of the key components of the DRS2ALM Processing is the COREALMLIB knowledge base, which was modified to form CORECALMLIB to suit the needs of the TEXT2ALM system. In particular CORECALMLIB adds a layer to COREALMLIB that maps the entires in VERBNET ontology with the entries in COREALMLIB ontology. This layer allows us to properly convert the DRS of a given narrative into an $\mathcal{ALM}$ system description.

c) CALM *and QA Processing:* In the CALM Processing performed by TEXT2ALM, the CALM system is invoked on a given $\mathcal{ALM}$ program stemming from a narrative in question. The CALM system computes a model. We then perform post-processing on this model to make its content more readable for a human.

A model derived by the CALM system contains facts about the entities and events from the narrative supplemented with basic commonsense knowledge associated with the events. We use the bAbI QA tasks to test the TEXT2ALM system's IE effectiveness and implement QA capabilities within the SPHINX subsystem (see Figure 3). The SPHINX component utilizes regular expressions to identify a kind of question that is being asked in the bAbI tasks and then query the model for relevant information to derive an answer. The SPHINX system is not a general purpose question answering component.

Additional information on the components of system TEXT2ALM are given in [15].

IV. TEXT2ALM EVALUATION

a) Related Work:: Many modern QA systems predominately rely on machine learning techniques. However, there has recently been more work related to the design of QA systems combining advances of NLP and KRR. The TEXT2ALM system is a representative of the latter approach. Other approaches include the work by Clark, Dalvi, and Tandon [4] and Mitra and Baral [14]. Mitra and Baral [14] use a provided training dataset of narratives, questions, and answers to learn the knowledge needed to answer similar questions. Their approach posted nearly perfect test results on the bAbI tasks. However, this approach doesn't scale to narratives that utilize other action verbs, which are not present in the training set, including synonymous verbs. For example, if their system is trained on bAbI training data that contains verb *travel* it will process the *JS* discourse correctly. Yet, if we alter the *JS* discourse by exchanging *travel* with a synonymous word *stroll*, their system will fail to perform inferences on this altered narrative (note that *stroll* does not occur in the bAbI training set). The TEXT2ALM system does not rely upon the training narratives for the commonsense knowledge. If the verbs occurring in narratives belong to VERBNET classes whose semantics have been captured within CORECALMLIB then TEXT2ALM is normally able to process them properly.

Another relevant QA approach is the work by Clark, Dalvi, and Tandon [4]. This approach uses VERBNET to build a knowledge base containing rules of preconditions and effects of actions utilizing the semantic annotations that VERBNET provides for its classes. In our work, we can view $\mathcal{ALM}$ modules associated with VERBNET classes as machine interpretable alternatives to these annotations. Clark et al. [4] use the first and most basic action language STRIPS for inference that is more limited than $\mathcal{ALM}$.

b) Evaluation: We use Facebook AI Research's bAbI dataset [19] to evaluate system TEXT2ALM. These tasks were proposed by Facebook Research in 2015 as a benchmark for evaluating basic capabilities of QA systems in twenty categories. Each of the twenty bAbI QA tasks is composed of narratives and questions, where 1000 questions are given in training set and 1000 questions are given in a testing set. The goal of the tasks are to process testing sets properly while also minimizing the number of questions used from the training set to develop a solution. We evaluate the TEXT2ALM system with all 1000 questions in the testing sets for tasks 1, 2, 3, 5, 6, 7, and 8. These tasks are selected because they contain action-based narratives that are of focus in this work. Figure 4 provides an example of a narrative and a question from the training set of bAbI task 2-Two Supporting Facts. For this task, a QA system must combine information from two sentences in the given narrative. The narrative in Figure 4 consists of six sentences. A question is given in line 7, followed by the answer and identifiers for the two sentences that provide information to answer the question.

```
1 Mary moved to the bathroom.
2 Sandra journeyed to the bedroom.
3 Mary got the football there.
4 John went to the kitchen.
5 Mary went back to the kitchen.
6 Mary went back to the garden.
7 Where is the football? garden 3 6
```

Figure 4. Example entry from bAbI task 2 training set

The bAbI dataset enables us to compare TEXT2ALM's IE/QA ability with other modern approaches designed for this task. The left hand side of Figure 5 compares the accuracy of the TEXT2ALM system with the machine learning approach AM+NG+NL MemNN described by Weston et al. [19]. In that work, the authors compared results from 8 machine learning approaches on bAbI tasks and the AM+NG+NL MemNN (Memory Network) method performed best almost across the board. There were two exceptions among the seven tasks that we consider. For the Task 7-Counting the AM+N-GRAMS MemNN algorithm was reported to obtain a higher accuracy of 86%. Similarly, for the Task 8-Lists/Sets the AM+NONLINEAR MemNN algorithm was reported to obtain accuracy of 94%. Figure 5 also presents the details on the Inductive Rule Learning and Reasoning (IRLR) approach by [14]. We cannot compare TEXT2ALM performance with the methodology by [4] because their system is not available and it has not been evaluated using the bAbI tasks.

System TEXT2ALM matches the Memory Network approach by Weston et al. [19] at 100% accuracy in tasks 1, 2, 3, and 6 and performs better on tasks 7 and 8. When compared to the methodology by Mitra and Baral [14], the Text2ALM system matches the results for tasks 1, 2, 3, 6, and 8, but is outperformed in tasks 5 and 7.

	Accuracy		
bAbI Task	AM+NG+NL MemNN	IRLR	TEXT2ALM
1-Single SF	100	100	100
2-Two SF	100	100	100
3-Three SF	100	100	100
5-Three AR.	98	100	22
6-Yes/No	100	100	100
7-Counting	85	100	96.1
8-Lists/Sets	91	100	100
	Training Size		
1-Single SF	250	1000	100
2-Two SF	500	1000	100
3-Three SF	500	1000	100
5-Three AR.	1000	1000	100
6-Yes/No	500	1000	100
7-Counting	1000	1000	100
8-Lists/Sets	1000	1000	100

Figure 5. System Evaluation and Training Set Sizes

The right hand side of Figure 5 presents the number of questions in training sets used by each of the reported approaches in considered tasks. The TEXT2ALM training set comprised of 100 questions per QA bAbI task, for a total of 700 questions. These training questions and their associated narratives were used to develop the CORECALMLIB knowledge base (recall *Library* CORECALMLIB adopts the earlier COREALMLIB knowledge base). As a result of this process, the CORECALMLIB covers 20 first-level VERBNET classes out of its 274.

V. CONCLUSION

System TEXT2ALM matched or outperformed the results of modern machine learning methods in all of these tasks except task 5. It matched the results of another KRR approach [14] in tasks 1, 2, 3, 6, and 8. However, our approach adjusts well to narratives with a more diverse lexicon due to its architecture. Additionally, the ability of the CORECALMLIB to represent the interactions of events in the bAbI narratives serves as a proof of usefulness of the original COREALMLIB endeavor.

REFERENCES

[1] E. Balai, M. Gelfond, and Y. Zhang, “Towards answer set programming with sorts,” in *Logic Programming and Nonmonotonic Reasoning*, P. Cabalar and T. C. Son, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 135–147.

[2] K. Barker, B. Porter, and P. Clark, “A Library of Generic Concepts for Composing Knowledge Bases,” *Proceedings of the 1st International Conference on Knowledge Capture - K-CAP*, pp. 14–21, 2001.

[3] C. Bonial, K. Stowe, and M. Palmer, “SemLink,” `https://verbs.colorado.edu/semlink/`, 2013.

[4] P. Clark, B. Dalvi, and N. Tandon, “What happened? Leveraging VerbNet to predict the effects of actions in procedural text,” *CoRR*, vol. abs/1804.05435, 2018. [Online]. Available: http://arxiv.org/abs/1804.05435

[5] M. Gelfond and V. Lifschitz, “The stable model semantics for logic programming,” in *Proceedings of International Logic Programming Conference and Symposium*, R. Kowalski and K. Bowen, Eds. MIT Press, 1988, pp. 1070–1080.

[6] D. Inclezan, “CoreALMlib: An ALM library translated from the Component Library,” *Theory and Practice of Logic Programming*, vol. 16, no. 5-6, pp. 800–816, 2016.

[7] D. Inclezan and M. Gelfond, “Modular action language ALM,” *TPLP*, vol. 16, no. 2, pp. 189–235, 2016. [Online]. Available: http://dx.doi.org/10.1017/S1471068415000095

[8] R. Johansson and P. Nugues, “LTH: Semantic structure extraction using nonprojective dependency trees,” in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

[9] H. Kamp and U. Reyle, *From discourse to logic*. Kluwer, 1993, vol. 1,2.

[10] K. Kipper-Schuler, “VerbNet: A broad-coverage, comprehensive verb lexicon,” Ph.D. dissertation, University of Pennsylvania, 2005.

[11] Y. Lierler, D. Inclezan, and M. Gelfond, “Action languages and question answering,” in *IWCS 2017 - 12th International Conference on Computational Semantics - Short papers*, 2017.

[12] G. Ling, “From Narrative Text to VerbNet-Based DRSes: System Text2DRS,” 2018. [Online]. Available: https://www.unomaha.edu/college-of-information-science-and-technology/natural-language-processing-and-knowledge-representation-lab/_files/papers/Text2Drses_system_description.pdf

[13] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[14] A. Mitra and C. Baral, “Addressing a Question Answering Challenge by Combining Statistical Methods with Inductive Rule Learning and Reasoning,” pp. 2779–2785, 2016.

[15] C. Olson, “Processing Narratives by Means of Action Languages,” Master’s thesis, University of Nebraska Omaha, 5 2019.

[16] M. Palmer, “VerbNet,” Boulder, 2018. [Online]. Available: https://verbs.colorado.edu/verb-index/vn3.3/

[17] M. Palmer, D. Gildea, and P. Kingsbury, “The proposition bank: An annotated corpus of semantic roles,” *Computational Linguistics*, vol. 31, no. 1, pp. 71–106, Mar. 2005.

[18] E. Wertz, A. Chandrasekan, and Y. Zhang, “CALM: a Compiler for Modular Action Language ALM,” unpublished draft, 2018.

[19] J. Weston, A. Bordes, S. Chopra, and T. Mikolov, “Towards AI-complete question answering: A set of prerequisite toy tasks,” *CoRR*, vol. abs/1502.05698, 2015.

Архитектура приложения Text2ALM для семантической обработки языка

Юлия Лирлер и Крэйг Олсон

Приложения Text2ALM ориентируется на семантическую обработку текста с глаголами действия. Эта система использует язык программирования действий под названием ALM для выполнения выводов о сложных взаимодействиях событий, описанных в тексте. Система опирается на ресурсы и методы из двух различных областей искусственного интеллекта, а именно: обработка естественного языка и представление знаний. Эффективность приложения Text2ALM измеряется по ее способности правильно отвечать на вопросы из задач babi (Facebook Research, 2015). Text2ALM соответствовал или превышал производительность современных методов машинного обучения в шести из семи протестированных заданий.

Received 04.01.2020

A subsystem of the “IntonTrainer” for learning and training Belarussian intonation

Evgeniya Zenovko
The State Scientific Institution
«The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus»
Minsk, Belarus
evgeniakacan@gmail.com

Boris Lobanov
The State Scientific Institution
«The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus»
Minsk, Belarus
lobanov@newman.bas-net.by

***Abstract*—The article presents Belarusian module of program complex "IntonTrainer" for training learners in producing a variety of intonation patterns of Belarusian speech. The system is designed for visual learning basic Belarusian tonal types, comparing melodic portraits of a reference phrase and a phrase spoken by a learner. It involves an active learner-system interaction, where the user can not only listen to an audio, but also try to repeat correctly this phrase and achieve the best result. The main principle of the program is to analyse and compare student’s results with standard variants without teacher’s involvement.**

***Keywords*—program complex, intonation, tone, melodic contour, melodic portrait, linguaacoustic database, module, reference phrase**

I. Introduction

Today experimental studies of oral speech are of particular relevance, since they reveal the qualitative characteristics of the phonetic, and directly, prosodic means in the implementation of the speaker’s communicative intention and the level of information perception. The prosodic level is the least studied and developed in comparison with phonetic, phonematic or other levels of the language. However, in recent decades it has acquired an increased interest in this area of the language due to advances of information society and the need to update computer technology. Modern software market offers a wide selection of computer-based training systems with an emphasis on vocabulary, grammar and syntax. These applications offer not only explanations of grammatical material, examples, texts, tests, but also exercises for listening and pronunciation. Some of them include tasks for listening to a speech and sound recording.Another use the interval repetition method, when at certain periods you are asked to repeat the words that have already been studied for their final fixation in memory. It is important that educational applications include various methods of computer monitoring and error correction, thereby individualizing the learning process, avoiding the subjectivity of assessment and expanding the possibilities of independent work for students. Unfortunately, we should notice, that all the variety of programs are produced for different languages including Russian but not for the Belarusian language. We can only find bilingual thematic vocabularies, interactive phrasebooks ans some spacial programs. This fact demonstrates the necessity to create new applications for our native language for the purpose of its promotion and possibility to learn it in more detail.

Rhythmic-intonational skills directed on intonationally and rhythmically correct design of speech (the correct statement of stress, the observance of the necessary melody, rhythm, etc.) call for special attention. For such skills we need special software complexes for studying prosody, in particular intonation, and its components. Computer programs and systems designed for effective teaching of intonation must meet the following requirements and be able to:

1) detect deviations in the pronunciation of the learner from the codified pronunciation variants (words and phrases);
2) classify these deviations;
3) show the learner the difference in pronunciation and correct pronunciation;
4) have the means (video, graphics, exercises) to bring the pronunciation implementation to the necessary degree of correct pronunciation;
5) act as a simulator of pronunciation skills;
6) to have a mechanism for evaluating the correctness and correctness of achieved results.

The IntonTrainer fills all the bill of the same. It is designed to analyze, display and compare intonation (pitch) contours of the reference and spoken phrases, as well as a numerical evaluation of their intonation similarity. Evaluation of intonation similarity is carried out on the basis of the representation of intonation in the form of Melodic Portraits. For now the software includes subsystem modules with the sets of reference phrases that represent the main intonation models of Russian, English (British and American versions), German and Chinese speech. In addition, modules for emotional intonation of the Russian and English languages have been developed. Also in the program you can find song modules of these two languages. The main advantage of this program is an opportunity to create new modules for different languages, including Belarusian.

This work is a follow up study to the previously introduced model of universal melodic portraits (UMP) of accentual units (AU) for the representation for Belarusian intonation [1, 2]. The article depicts Belarusian module of IntonTrainer, its components and principles of its usage. The initial window of the Belarusian module is shown in Figure 1. For more detailed information you need to turn to official cite of the complex (https://intontrainer.by).

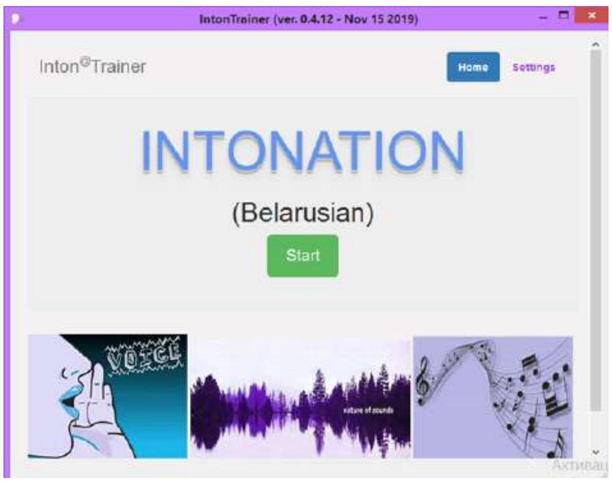


Figure 1. The initial window of the Belarusian module

II. Lingua-acoustic database for Belarusian module

Belarusian module is based on a lingua-acoustic database (LAD), which includes a large set of standard phrases. These reference phrases cover a greater number of Belarusian intonational patterns of speech according to the type of expression, stylistic features and their emotional coloring. The Belarusian module contains audio database of a female and a male voicies. They allow you to get acquainted with the four main melodic tones of the Belarusian literary language (falling, rising, falling-rising and rising-falling) on the examples of reference phrases, which reflect melodic contour of each phrase (Figure 2). In addition to information about the tone and the communicative type of sentence, the database depicts prosodic markup of sentences and the number of syntagmas in a phrase. The user can also listen to the audio file by clicking on the link.

Furthermore, LAD displays 26 groups of statements which provide a description of intonational features of Belarusin language in terms of the syntactic structure of a sentence, emotional coloring and a number of Syntagmas. Today Belarusian linguaacoustic database consists of 200 syntactic-accentual units. The module is

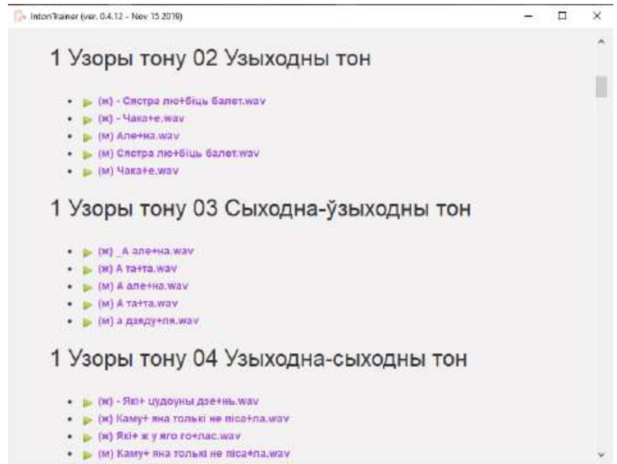


Figure 2. Database window of Belarusian module

in the process of refining, so in the future it will be added universal melodic portraits of phrases related to prose texts, poems and dialogues. The Belarusian module performs all the functions of IntonTrainer complex that allows you to learn the intonation of the language in its entirety.

III. A structure and functions of the Belarusian Module

The Belarusian module consists of 7 sections. First of all the program offers to get acquainted with basic tones (falling, rising, rising-falling and falling-rising) on the examples of sentences with different communicative types (figure 2). Then the user can learn intonation patterns of the phrases with different emotional colouring and types of intentional focus (for example sentences with neutral answer, alternative questions, appeals, forms of greeting and farewells, etc). Each example provides audio and visual representation of UMP, pairwise comparison of different Tone patterns (TP), explains peculiarities of TP usage.

When the student chooses the pattern for learning, he opens "Analysis results window" where the red column on the left shows the range of the melody change, i.e. frequencies of the pitch (F0), expressed in octaves (figure 4). On the right, a linear graph of the UMP is displayed in red, the nuclear syllable of which is marked with frequent vertical lines. Below the graphs, the minimum and maximum values of F0 for the selected phrase are listed, as well as the text of the phrase in which the nuclear vowel is indicated by the “+ ” sign. Listening to the selected phrase is carried out by pressing the «Play Record” button. An additional function of the program is the ability to select the markup of the studied phrase. At early version of the program it was necessary manually processed audio, highlighting pre-nuclear, nuclear and post-nuclear syllables. To date, it is also possible automatically to produces a tonal contour of the phrase. Therefore, the user can choose the view mode of UMP: "Manually" or "Automatically" (see Figure 3, 4).

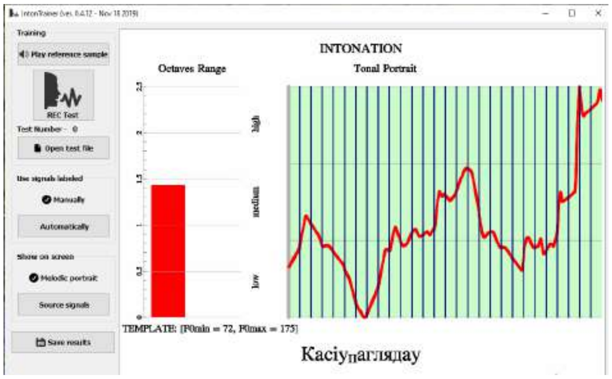


Figure 3. Melodic contour of the phrase "Kamu+ yana tolki ne pisa+la" marked manually

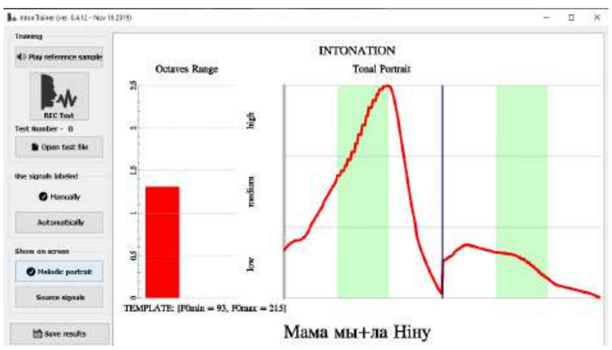


Figure 4. Melodic contour of the phrase "Kamu+ yana tolki ne pisa+la" marked automatically

If one wants to see graphical display of intonation contour with prenuclear, nuclear and post-nuclear marks and the oscillogram of the phrase, then needs to click “Show original” (figure 5).

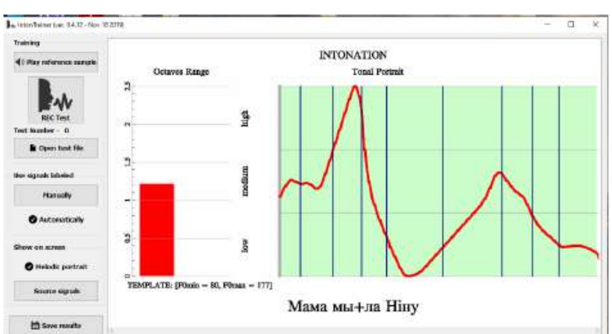


Figure 5. Additional window displaying F0 in real time

The next step is to get acquainted how to listen to reference phrase, repeat,learn it and receive the best result of learning.The system evaluates the student’s pronunciation on a ten scale with percentage. The result of the similarity of his pronunciation with the standard phrase is displayed on top of the screen (figure 6).

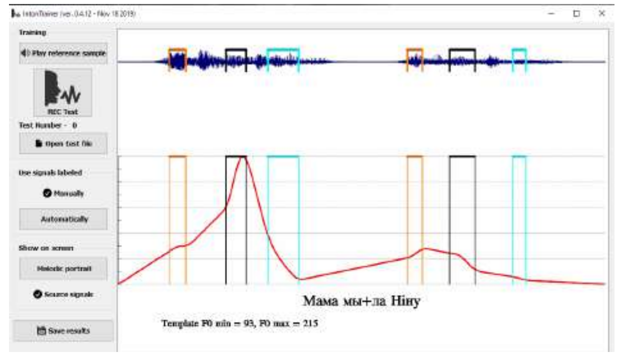


Figure 6. Comparison of reference phrase and pronounced phrase "Kamu+ yana tolki ne pisa+la"

The maim advantage of this version is the ability to set the desired number of test files (pronunciations) and to obtain an average rating of their similarity with the reference phrase. The left part of the figure 7 (Octaves Range) graphically shows the results of calculating the similarities in the pitch frequency range for each pronunciation (thin columns, the last pronunciation corresponds to black) and the resulting averaged score (thick column). It shows the numerical value of the average score - the Range is 49 per cent. The right side (Tonal Portrait) shows the results of calculating the UMP (or NMP) of the spoken phrases (thin lines) and the resulting averaged curve (thick line). Above is the numerical value of the average score of similarity with the reference curve - Shape (58 per cent). Therefore, the user can see his progress from the very beginning of studying to the end.

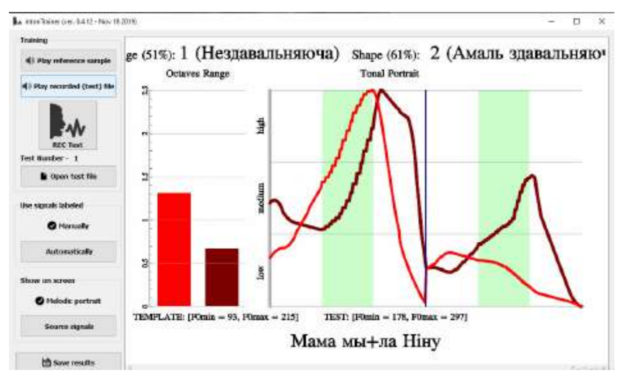


Figure 7. Comparison of reference phrase and multiple pronounced phrases "Mama my+ła Ninu"

When you click the “Save Results” button, an additional icon appears and a page opens in EXCEL, on which a complete set of 10 prosodic features of the reference phrase is written (see Table 1). The obtained data is stored in the same folder where the reference phrase being studied is stored.In table 1, in addition to data on the parameters of the original NMP curve, data on the value of its time derivative, d / dt (NMP), are also presented. Additional analysis of the parameters of the derivative of the NMP turns out to be useful for taking into account the dynamic characteristics of the CHO movement, which are characteristic of certain types of emotions.

Table 1 - Qualitative measures of the reference and pronounced phrase "Мама му+ła Nim"

The data on the proximity	Proximity	Distance	Relative data on templates and n	Relation	Reference data on Records	
Proximity thru curve 'Correlation'	70	30	F0 max [Hz]	1,575200185	F0 max [Hz]	300
Proximity thru curve 'Integral'	71	29	F0 min [Hz]	2,915894154	F0 min [Hz]	182
Proximity thru curve 'Local'	33	67	Diapason F0	1,346684299	Diapason F0	1,697802198
Average of the three above proximities	58	42	Register F0 [Hz]	2,025937736	Register F0 [Hz]	245,5
Proximity thru 'Range'	49	51	Mean Value of the curve NMP	1,169801074	Mean Value of the curve NI	26,70458178
			Center of the curve NMP	0,039090785	Center of the curve NMP	46,04527809
Reference data on templates			Width of the curve NMP	0,344736717	Width of the curve NMP	43,22284679
F0 max [Hz]	215		Mean Value of the Derivative cur	-0,233357351	Mean Value of the Derivativ	15,57381802
F0 min [Hz]	93		Center of the Derivative curve NI	1,188800893	Center of the Derivative curv	50,12033330
Diapason F0	2,531827957		Width of the Derivative curve NN	1,7038999	Width of the Derivative curv	47,59794965
Register F0 [Hz]	154		Voiced Sounds Level	0,944857046	Voiced Sounds Level	0,234547845
Mean Value of the curve NMP	20,39885101		Voiced Sounds Duration	0,7786784	Voiced Sounds Duration	450
Center of the curve NMP	41,61750005					
Width of the curve NMP	35,5240151					
Mean Value of the Derivative curve NMP	16,22123249					
Center of the Derivative curve NMP	38,234178					
Width of the Derivative curve NMP	32,44245767					
Voiced Sounds Level	1,180613933					
Voiced Sounds Duration	383					

IV. Conclusion

The software package IntonTrainer is recommended for the initial acquaintance and study of the main intonational constructions of the Belarusian literary language, comparison of its different patterns, as well as their implementation in dialogue, prose and poetry. As a means of training intonation, the Intontrainer gives an apportunity to improve intonation skills in various fields of activity. Today it is the only free program that offers to study several languages at once. The complex also allows to create new language modules based on its platform. Thus, the authors of this article have developed the first Belarusian language module, which makes it possible to learn how to use the Belarusian intonation correctly. It can be used as a means for teaching Belarusian as a foreign language, get practical results without involving a teacher and compare its melodic structure with another languages. The significance of such a development for the Belarusian language lies in the universality of its use by both students and linguists in scientific and practical research.

References

[1] Intontrainer. Available at: https://intontrainer.by/ (accessed 12.03.2019).

[2] Anne Bonneau, Vincent Colotte.: Automatic Feedback for L2 Prosody Learning. Speech and Language Technologies, (2011), pp.55-70.

[3] Juan Arias, Nestor Yoma, Hiram Vivanco.: Automatic intonation assessment for computer aided language learning. Speech Communication 52 (2010), pp. 254–267.

[4] Lobanov, B.M. The program model of analysis and visualization of emotional intonation in oral speech / B.M. Lobanov, V.A. Zhitko // Open semantic technologies for designing intelligent systems: a collection of scientific papers. Issue 3. / BSUIR; Editorial: V.V. Golenkov [et al.]. - Minsk: BSUIR, 2019 . C. 297-300.

[5] Lobanov B. M. Komp'yuternyy sintez i klonirovanie rechi [Computer synthesis and speech cloning]. Minsk, Belorusskaya nauka Publ., 2008. 357 p.

[6] Lobanov, B. A Prototype of the Software System for Study, Training and Analysis of Speech Intonation / B. Lobanov, V. Zhitko, V. Zahariev // Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings / Eds. Alexey Karpov, Oliver Jokisch, Rodmonga Potapova. – Springer, 2018. – P. 337-346.

[7] Lobanov B. Software Subsystem Analysis of Prosodic Signs of Emotional Intonation / B. Lobanov, V. Zhitko // Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings / Eds. Albert Ali Salah, Alexey Karpov, Rodmonga Potapova. – Springer, 2019. – P. 280-288.

[8] Nikolaeva T.M. Frazovaya intonaciya slavyanskih yazykov [Phrasal intonation of Slavic languages]. Moskva, Nauka Publ., 1977. 278 p.

[9] Vigonnaya L.T. Intanacyja. Nacisk. Arfaepija [Intonation. Stress. Orthoepy]. Minsk, Nauka i tehnika Publ., 1991/ 215 p.

[10] Zianouka Ja.S. Linguaacousticresources for Belarusian speech synthesizers. Open semantic technologies for designing intelligent systems: a collection of scientific papers. Issue 3. / BSUIR; Editorial: V.V. Golenkov [et al.]. - Minsk: BSUIR, 2019 .– C. 301-304.

Подсистема «IntonTrainer» для обучения и тренировки белорусской интонации

Зеновко Е.С., Лобанов Б.М.

В статье представлен белорусский модуль программного комплекса «IntonTrainer» для обучения учащихся выработке разнообразных интонационных паттернов белорусской речи. Система предназначена для визуального изучения основных белорусских тональных типов, сравнения мелодических портретов референтной фразы и фразы, произносимой учеником. Он включает в себя активное взаимодействие ученика и системы, где пользователь может не только слушать аудио, но и пытаться правильно повторить эту фразу и достичь наилучшего результата. Основным принципом программы является анализ и сравнение результатов студента со стандартными вариантами без участия учителя.

Received 14.01.2020

Current State of the Development and Use of the “IntonTainer” System

1st Prof. Boris Lobanov
The United Institute of Informatics Problems of National Academy of Sciences of Belarus
Minsk, Belarus
lobanov@newman.bas-net.by

2nd Vladimir Zhitko
The United Institute of Informatics Problems of National Academy of Sciences of Belarus
Minsk, Belarus
zhitko.vladimir@gmail.com

***Abstract*—The purpose of this work is to describe a current state of the development and use of the “IntonTainer” system that provides a variety of ways to analyze and visualize the human intonation. The processing algorithms of a separate library, which became the core of the next version of the IntonTrainer application is described. Information on the use of the developed system is provided.**

***Keywords*—Speech intonation, melodic portrait, intonation analysis, software model, IntonTrainer users**

I. Introduction

The history of computer aided learning of pronunciation of a foreign learned language traces back at least to 1964, with the first attempt to display the learner voice pitch on a screen in order to improve the perception and the realization of sentence intonation [1]. Works in the last two decades show that the Computer-Assisted Pronunciation Teaching (CAPT) systems are useful, flexible tools for giving pronunciation instructions and evaluating at subject’s speech [2]–[8]. However, pronunciation teaching has many issues. There are pedagogical, technological questions. From a technological viewpoint it is hard to provide understandable, accurate feedback to pronunciation mistakes. It is almost impossible to give a one hundred percent accurate and automatic diagnosis of speech production errors.

Earlier we proposed a software system for the analysis and training of speech intonation IntonTrainer [9], [10]. The system was designed to train learners in producing a variety of recurring intonation patterns of speech in foreign language. The system is based on comparing the melodic (tonal) portraits of a reference phrase and a phrase spoken by the learner. The main algorithms used in the training system proposed for analysing and comparing intonation features are considered. The proposed computer trainer provides additional visual feedback, as well as a quantitative estimation of the correctness of speech intonation in the process of foreign languages learning.

II. Current state of the development of the “IntonTainer”

This version provides many new possibilities for the analysis and training of intonation, namely:

- The ability to set the desired number of test files (pronunciations) and to obtain an average rating of their similarity with the reference phrase is added
- An auxiliary software module (Multi-Lingual Launcher) for launching any desired set of IntonTrainer modules is added
- The new IntonTrainer modules: Belorussian and Singing Lessons are added
- Added displaying the position of the boundaries of voice sections and the boundaries of accent units of phrases.
- The ability to analyze and compare the melodic portraits without the need for preliminary manual marking of it into pre-nuclear, nuclear and nuclear sections is added.
- The possibilities of parametric display of melodic intonation portraits are expanded.
- The ability to analyze and store in numerical form a set of prosodic signs of the analyzed phrases has been added.
- The accuracy of segmentation and labeling of the analyzed speech signals is increased.

“Fig. 1” show the view of the starting window of the Multi-Lingual Launcher program in the case when the user would need a full set of software modules. Using this service program, the user has an ability to promptly call one or more modules of interest to it.

The IntonTrainer application can be used not only as an intonation simulator, but also as an environment for conducting intonation studies. However, to date, its integration into other speech processing facilities has been difficult due to the lack of an external software interface for the IntonTrainer system. In this regard, it was decided to separate the processing algorithms into a separate library, which became the core of the next version of the IntonTrainer application. The structure of the core is shown in the diagram in “Fig. 2”.

Figure 1. The starting window of the Multi-Lingual Launcher.

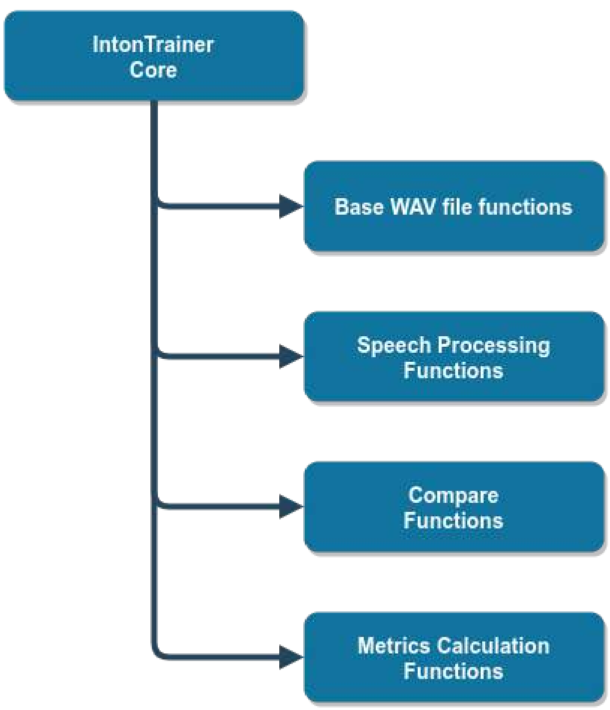


Figure 2. The structure of the system core.

The core includes the following main tasks for calculating speech flow parameters:

Pitch processing:

- get_pitch - the vector of the fundamental frequency (F0)
- get_pitch_log - logarithm vector from F0
- get_pitch_cutted - the frequency response vector taking into account the excess energy level
- get_pitch_min - minimum F0
- get_pitch_max - maximum F0
- get_pitch_interpolated – F0 vector with interpolated gaps
- get_pitch_norm - normalized vector of recovered F0
- get_pitch_smooth - smooth F0 vector

UMP processing:

- get_ump - UMP vector constructed by F0
- get_ump_mask - mask for mapping UMP to the original signal
- get_p - set of pre-nuclea UMP segments
- get_t - a set of segments of the post-nuclea UMP
- get_n - set of UMP nuclea segments

Signal energy processing:

- get_intensive - speech signal energy vector
- get_intensive_norm - normalized energy vector of a speech signal
- get_intensive_smooth - smoothed energy of a speech signal

Spectrum processing:

- get_spectrum - spectrum of the speech signal
- get_spectrum_norm - normalized spectrum of a speech signal
- get_cepstrum - speech signal cepstrum
- get_cepstrum_norm - normalized speech signal cepstrum

The calculation tasks are arranged in a certain hierarchy, where functions of a higher level use the results of calculations of the underlying functions. For each speech stream, the core creates a new object that caches the calculation results, which allows you to get different data on the same speech stream with less overhead, because Interim results will already be calculated.

The structure of the methods for processing speech flows is presented in the diagram of “Fig. 3”.

Each method in the core can be configured using the core tuning mechanism. To store the settings, it is possible to download and upload to an external file. Also, one of the main parts of the core are methods for comparing two or more speech signals and calculating signal metrics and comparing them.

These methods include:

- Evaluation of the similarity of the ranges of variation of F0

  Range Similarity = (F0max / F0min) user * 100% / (F0max / F0min) template

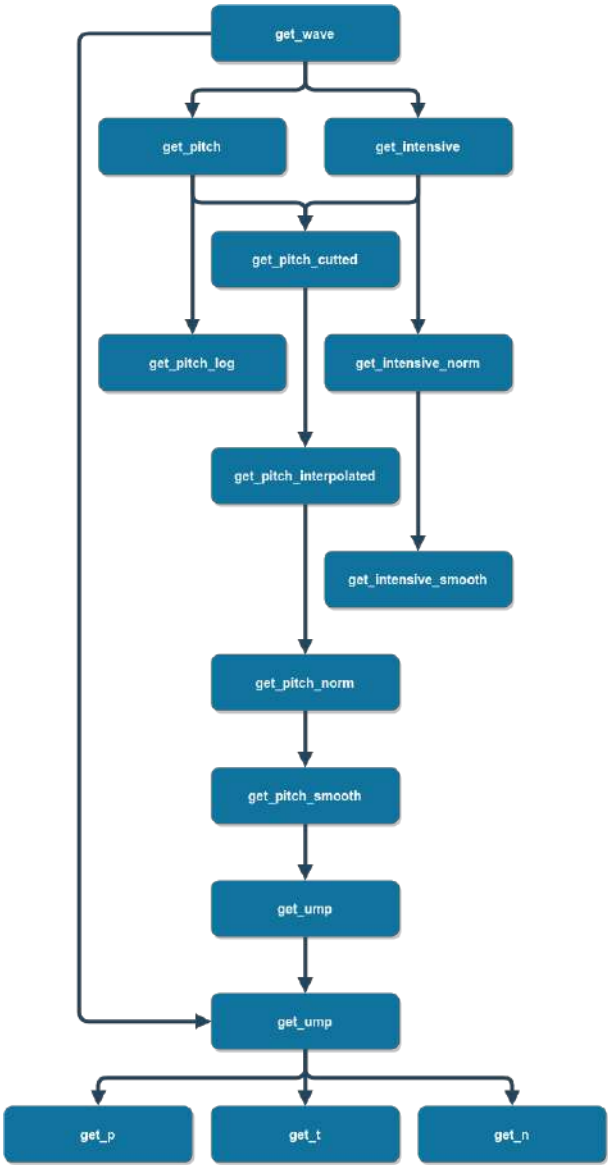


Figure 3. The structure of the methods for processing speech flows.

- Correlation assessment of similarity of UMP Shape Similarity (thru correlation) = [(r + 1) / 2] * 100%
- The average value of the similarity assessment UMP Shape Similarity (thru average distance) = (1 - D) * 100%
- Minimum assessment of the similarity of UMP Shape Similarity (thru max local distance) = (1 - d max) * 100%

Because the core library is written in pure C, this will allow you to create wrappers for other programming languages, such as python, java. Creating wrappers for other languages can be automated using SWIG. Highlighting the main algorithms of the application will also allow IntonTrainer to be used as a plug-in in other research in automatic signal processing environments.

III. Current state of use of the "IntonTainer" software package

The "IntonTainer" software package is recommended for use in the following popular fields:

- In linguistic education used as a means of visualizing intonation.
  (Primary introduction and study of the basic tone patterns of oral speech and songs, their pairwise comparisons, peculiarities of their usage as well as their actualization in dialogues prose and poetry).
- In individual intonation training for correct pronunciation used as a means of feedback.
  (Individual training for correct pronunciation of tone patterns when studying a foreign language, improving intonation skills of one's native language or singing in some professions, such as call center operators, radio and TV announcers, singers etc).
- In scientific and practical research used as a means of comparing intonation from different sources.
  (Study of individual, emotional and stylistic features of intonation. Comparative evaluation of speech intonation in norm and pathology. Estimation of the intonational quality of synthesized speech).

The "IntonTrainer" software package has been available for free download from the site https://intontrainer.by since the end of 2018. The world map ("Fig. 4") shows the geographical points, from which the requirements for the site for downloading the IntonTrainer came. The information in "Fig. 4" obtained using the site https://www.google.com/analytics/web/ and presented at the beginning of 2020.

Figure 4. The world map of the "IntonTrainer" downloading.

"Fig. 5" shows the distribution of downloads for the countries (except Belarus) with more than 5 download.

IV. Conclusion

The present prototype of the IntonTrainer system is implemented as a separate SWM under Windows OS (xp, 7, 8, 10) and Linux (Debian, Ubuntu). In order to make it possible to significantly increase the number of users,

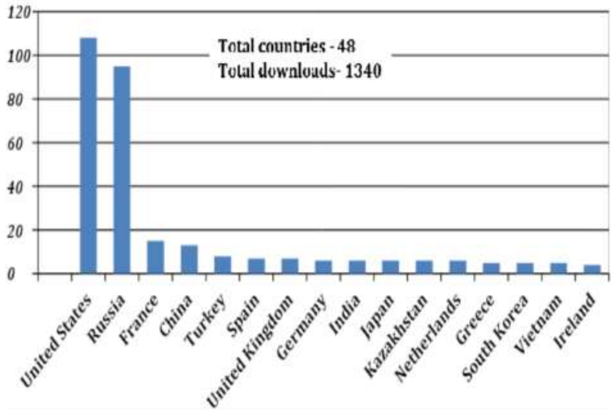


Figure 5. The distribution of downloads for the countries.

we plan to develop an advanced version. This version should be easily portable to mobile and cloud platforms.

References

[1] Vardanian, R. M. Teaching English through oscilloscope displays. Language Learning, 3/4, 109-118, 1964.

[2] Juan Pablo Arias a, Nestor Becerra Yoma a, Hiram Vivanco, “Automatic Intonation Assessment for Computer Aided Language Learning”, Speech Communication, vol. 52, no. 3, pp. 254-267, Mar. 2010.

[3] Akinori Ito, Tomoaki Konno, Masashi Ito and Shozo Makino, ”Evaluation of English Intonation based on Combination of Multiple Evaluation Scores”, INTERSPEECH-2009, pp. 596-599. 2009.

[4] Motoyuki Suzuki, Tatsuki Konno, Akinori Ito and Shozo Makino, “Automatic Evaluation System of English Prosody for Japanese Learner’s Speech”, Proc. 5th Int. Conf. Education and Information Systems, Technologies and Applications (EISTA), 2007.

[5] Kim and W. Sung. “Implementation of intonational quality assessment system”, INTERSPEECH-2002, pp. 1225–1228, Sept. 2002.

[6] Anthony Fox, “Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals”, Oxford: Oxford University Press, 2000.

[7] Kazunori Imoto, Yasushi Tsubota, Antoine Raux, Tatsuya Kawahara, Masatake Dantsuji, “Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system”, Acoustical science and technology 24(3), 159-160, May 2003.

[8] Konno, T. Ito, A. Ito, M. Makino, S. Suzuki, M. Grad. Sch. Of Eng., Tohoku Univ., Sendai, “Intonation evaluation of English utterances using synthesized speech for Computer-Assisted Language Learning”, Int. Conf. Natural Language Processing and Knowledge Engineering, pp. 1-7, Oct. 2008.

[9] Lobanov, B. A Prototype of the Software System for Study, Training and Analysis of Speech Intonation / B. Lobanov, V. Zhitko, V. Zahariev // Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings / – Springer, 2018. – P. 337-346.

[10] Lobanov B. Software Subsystem Analysis of Prosodic Signs of Emotional Intonation / B. Lobanov, V. Zhitko // Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings / – Springer, 2019. – P. 280-288.

Современное состояние разработки и использованиясистемы «ИнтонТайнер»

Лобанов Б. М., Житко В. А.

Цель данной работы - описать текущее состояние разработки и использования системы «IntonTainer», которая предоставляет различные способы анализа и визуализации интонации человека. Описаны алгоритмы обработки отдельной библиотеки, ставшей ядром следующей версии приложения “IntonTrainer”. Предоставляется информация об использовании разработанной системы.

Received 15.01.2020

Examples of intelligent adaptation of digital fields by means of the system GeoBazaDannych

Valery B. Taranchuk
Department of Computer Applications and Systems
Belarusian State University
Minsk, Republic of Belarus
taranchuk@bsu.by

***Abstract*—Examples of interactive formation of digital models of geological objects in computational experiments that meet the intuitive requirements of the expert are given and discussed. Special tools of the system GeoBazaDannych are noted, and the results of comparison with standard solutions in the complex "The Generator of the geological model of the deposit" are presented. Examples of approximation and reconstruction of the digital field and its interactive adaptation by means of the system GeoBazaDannych are given and discussed; the obtained solutions and their accuracy are illustrated by maps of isolines.**

***Keywords*—digital geological model, system GeoBazaDannych, interactive graphical visualization, intelligent adaptation of digital fields, “smart” methods of computer model adaptation**

I. Introduction

Geological modeling is an independent direction that includes the improvement of mathematical methods and algorithms; development of computer programs that provide a cycle of model construction, forming, filling and maintenance of databases [1]. The corresponding software includes the loading from different sources and data preprocessing, correlation, creation of digital cubes of reservoir properties, interactive data analysis, visualization with the help of any type graphics, mapping. The task of developing and implementing various computer-based geological models with self-tuning tools is one of the priorities. Herewith, an important component is the task of evaluating the adequacy and accuracy of the proposed digital models. The key issues are automation, adaptation of models taking into account continuously incoming additional data, as well as a revision of the results of processing the initial information using new interpretation methods [2].

The data used in geological and geoecological models are a representative part of the geodata, which classify, summarize information about processes and phenomena on the earth’s surface. This information becomes really useful when integrated into a single system. Geodata, as a generalization of accumulated information, include information not only from the field of Earth sciences, but also from others, such as transport, economics, ecology, management, education, analysis, artificial intelligence. Technological, system, and information features of geodata are noted in [3].

Geodata volumes are growing at a very high rate. Accordingly, it is natural to use "big data" technologies (the specifics for geodata are described in [4]), including automated data mining. One of the main aims of data mining is to find previously unknown, non-trivial, practically useful and understandable interpretations of knowledge in “raw” (primary) data sets [5], [6]. At the same time, following [5], “data mining does not exclude human participation in processing and analysis, but significantly simplifies the process of finding the necessary data from raw data, making it available to a wide range of analysts who are not specialists in statistics, mathematics or programming. Human participation is expressed in the cognitive aspects of participation and the application of informational cognitive models”.

Geodata mining tools are the same as for usual data; the basis is the theory, methods, and algorithms of applied statistics, databases, artificial intelligence, and image recognition. There are many different active and applied software tools for data mining, for example, 8 classes of data mining systems are identified in [5]. The variety of proposed methods and software tools make it necessary to assess the quality of geodata and determine their main characteristics. Criteria for evaluating the quality of geodata are discussed in [7].

A number of issues related to the analysis and evaluation of spatial data quality can be solved using the computer system GeoBazaDannych [8] – [10]. Possible options, methodological solutions, and software tools that allow you to confirm the validity of interpretations, visualize and obtain numerical values of errors calculated by different methods of intellectual data processing results included and used in computer geological models are discussed below. For illustrations, the key task of forming and processing digital fields used in computer models is selected. In particular, we discuss the proposed methods that have been tested for solving different applied problems, as well as implemented in the interactive computer system GeoBazaDannych specialized algorithms for calculating approximating digital fields.

The interactive computer system GeoBazaDannych is the complex of intelligent computer subsystems, mathematical, algorithmic and software for filling, maintaining and visualizing databases, input data for simulation and mathematical models, tools for conducting computational experiments, algorithmic tools and software for creating continuously updated computer models. By means of the system GeoBazaDannych, it is possible to generate and visualize digital descriptions of spatial distributions of data on sources of contamination, on the geological structure of the studied objects; graphically illustrate solutions to problems describing the dynamic processes of multiphase filtration, fluid migration, heat transfer, moisture, and mineral water-soluble compounds in rock strata; design and implement interactive scenarios for visualization and processing the results of computational experiments. GeoBazaDannych's subsystems allow you to calculate and perform expert assessments of local and integral characteristics of ecosystems in different approximations, calculate distributions of concentrations and mass balances of pollutants; create permanent models of oil production facilities; generate and display thematic maps on hard copies. The main components of the system GeoBazaDannych [8] – [12]:

- the data generator Gen_DATv;
- the generator and editor of thematic maps and digital fields Gen_MAPw;
- modules for organizing the operation of geographic information systems in interactive or batch modes;
- the software package Geo_mdl – mathematical, algorithmic and software tools for building geological models of soil layers, multi-layer reservoirs; modules for three-dimensional visualization of dynamic processes of distribution of water-soluble pollutants in active soil layers;
- software and algorithmic support for the formation and maintenance of permanent hydrodynamic models of multiphase filtration in porous, fractured media;
- the integrated software complex of the composer of digital geological and geoecological models (GGMD).

II. Examples of interactive adaptation of digital fields

Integration of the capabilities of various geographic information systems (GIS) and GeoBazaDannych is provided by a wide range of tools of the system for importing and exporting data, images, and functions. Several non-standard solutions that are recognized as difficult for all geodata processing packages are discussed below. The examples of approximation and reconstruction of the digital field, its interactive adaptation by means of the system GeoBazaDannych and evaluation of the accuracy of results using the tools of the GGMD complex illustrate the unique capabilities of the developed methods and software. The task of reconstruction of the grid function involves calculating the values of the approximating function at regular grid points from the values of randomly located experimental data points (observations), i.e. creating a regular array of Z values of node points from an irregular array of (X,Y,Z) values. The term "irregular array of values" means that the X, Y coordinates of data points are distributed irregularly across the function definition area. The procedure for constructing a regular network of level values and restoring the grid function is an interpolation or extrapolation of values from a collection of scattered sets of source points and values of surface levels in them to uniformly distributed nodes in the study area.

Methods for restoring grid functions and the corresponding algorithms are implemented in several specialized computer graphics and GIS packages. They can be divided into two classes: exact and smoothing interpolators [13], [14]. In fact, the method falls into a particular class depending on the user-defined settings when performing value calculations. Most methods for restoring the function and constructing a digital field are based on calculating weighted coefficients that are used to weigh the values of the measurement data at the nearest points. This means that, all other things being equal, the closer the data point is to a network node, the more weight it has in determining the value of the function being restored at that node.

It should be understood that restoring a grid function does not imply finding a single solution to a certain mathematical problem. Subjective opinion and expert qualifications are factors that are always present in such activities [15]. Therefore, for building digital models, it is important to have tools for interactive data processing, simulation of possible situations for obtaining and correcting input information, and modules for mathematical processing and statistical analysis [13], [16]. The multi-format data exchange tools mentioned above is required for simultaneous work in multiple software environments. It is important for the user to have tools that allow them to "play" with the source data and compare the results with the prepared reference models. How this is implemented in the system GeoBazaDannych is described below, the corresponding preparatory actions in the GGMD complex are described in detail in [11], [12], [17].

The reference surfaces in the following examples are objects described by the expressions (1) and (2). The functions *fOrigin, fHill, fPyramid, fTrench*, and *zBasic* included in these expressions are described in [11], [17].

$$
\begin{gathered}
zSurfB(x,y) = fOriginB(x,y)+ \\
+45 \cdot fHillZ(0.004 \cdot (x-400), 0.002 \cdot (y-1900))+ \\
+37 \cdot fPyramidA(0.004 \cdot (x-350), 0.002 \cdot (y-600))+ \\
+15 \cdot fTrenchB(0.003 \cdot (x-1100), 0.0015 \cdot (y-400))+ \\
+25 \cdot fHillYX(0.003 \cdot (x-1100), 0.0014 \cdot (y-1700))+ \\
+15 \cdot fHill(0.005 \cdot (x-2350), 0.003 \cdot (y-950))- \\
-15 \cdot fHill(0.006 \cdot (x-1900), 0.003 \cdot (y-1300))- \\
-32 \cdot fHill(0.0012 \cdot (x-3000), 0.0025 \cdot (y-2000))+ \\
+35 \cdot fPyramidB(0.002 \cdot (x-3300), 0.002 \cdot (y-650)), \\
fOriginB(x,y) = zBasicB(x).
\end{gathered} \tag{1}
$$

$$
\begin{gathered}
zSurfT(x,y) = fOriginT(x,y)+ \\
+42 \cdot fHillZ(0.004 \cdot (x-400), 0.002 \cdot (y-1900))+ \\
+37 \cdot fPyramidA(0.004 \cdot (x-350), 0.002 \cdot (y-600))+ \\
+10 \cdot fTrenchB(0.004 \cdot (x-900), 0.0015 \cdot (y-500))+ \\
+15 \cdot fHill(0.004 \cdot (x-1100), 0.0014 \cdot (y-1700))+ \\
+15 \cdot fHill(0.005 \cdot (x-2150), 0.003 \cdot (y-1050))- \\
-15 \cdot fHill(0.006 \cdot (x-1900), 0.003 \cdot (y-1300))- \\
-13 \cdot fHill(0.0012 \cdot (x-3000), 0.0025 \cdot (y-2000))+ \\
+25 \cdot fPyramidB(0.002 \cdot (x-3300), 0.002 \cdot (y-650)), \\
fOriginT(x,y) = zBasicT(x).
\end{gathered} \tag{2}
$$

An example of creating a model of the reference surface, which is obtained from the base surface by adding elements of the listed types (2 pyramids, 5 hills, 1 trench), is shown below in Fig. 1, Fig. 2. It's important that in the resulting equations (1), (2) the coefficients in the formulas of perturbation elements fHill, fPyramid, fTrench are chosen by user while visual construction. Shown in figures images of surfaces *zSurfB*(x,y), *zSurfT*(x,y) give an aggregate vision of these surfaces. Individual details can be studied on the maps of isolines Fig. 3, Fig. 4. We emphasize attention on fragments of the zSurfB surface, namely, perturbations *fHillZ*(400,1900), *fHillYZ*(1100,1700). The element *fHillZ*(400,1900) simulates the presence of a plateau (clipping with a horizontal plane), *fHillYZ*(1100,1700) simulates a split (in geology – shift-overshoot), and is mathematically described by clipping a smooth hill-type shape with a vertical plane. Similar perturbations are specified for the zSurfT surface, but instead of *fHillYZ*(1100,1700), the expression *fHill*(1100,1700) is accepted – without clipping (no split). Such fragments of fHillZ and fHillYZ cannot be reproduced using standard methods for restoring digital fields. Also note that in the following illustrations, the isolines corresponding to the plateau levels (for zSurfB – level 9) are not displayed. In the zSurfB illustration, level 10 is shown to further emphasize the boundaries of disturbances placed on the plateau section.

Illustrations Fig. 5 – Fig. 10 show examples of possible solutions using the methods of "adjustment" distributions in the system GeoBazaDannych, and for

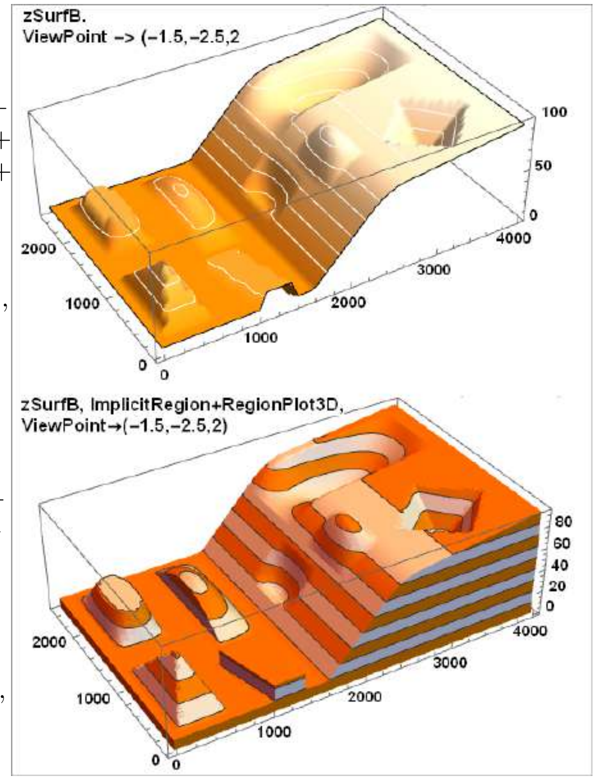


Figure 1. ZsurfB surface and volume graphs.

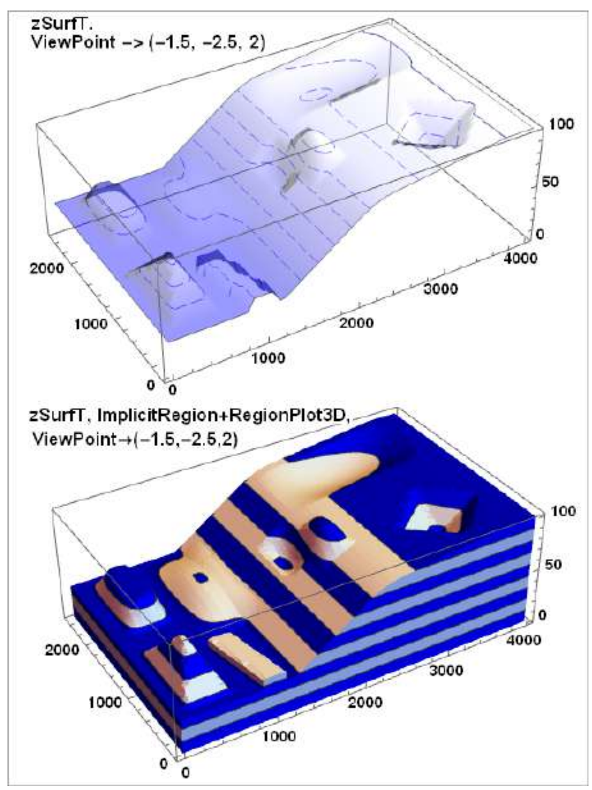


Figure 2. ZsurfT surface and volume graphs.

this purpose, the steps of simulating observations and restoring the distribution over a scattered set of points with measurements are performed. The corresponding results obtained in the [11] are shown on the isolines map in Fig. 4. The measurement points (163 dots) are placed on the profiles, they are given for different observation profiles by different primitives, and the centers of the perturbation forms are marked with circles (red circles, 10 dots). On maps, the isolines of the reference field are given as solid lines; to the restored field, they are dot-dashed green lines. There was no goal to choose a good system of profiles, they are formed sketchily, about how often designers do this. The reconstructed field is obtained by the 2nd-order interpolation method in Wolfram Mathematica. The implementation features are described in [12], note that in Mathematica, when working with irregularly placed data points, interpolation can only be performed using the 1st order method. Using the mentioned initial data, in Gen_MAPw obtained the results of digital field reconstruction by spline approximation of GeoBazaDannych's method, the results are shown in Fig. 6.

How to understand the results shown in Fig. 7 – Fig. 10?

In this case, the task is to correct the digital fields obtained by the approximation algorithm via using such GeoBazaDannych's elements as subdomain, selected, split, and boundary conditions [10]. In fact, it is an interactive intellectual adaptation.

To get results in the system GeoBazaDannych, the data of measurements on profiles and control points in GGMD are exported to an Xls file that is imported into Gen_MAPw. The corresponding illustration is given in Fig. 5, where the profiles, points of observation are shown; the values of measurements are displayed near their primitives. In the database, values are stored with machine precision, for brevity, the output of the map is made in the format of a single significant number after a decimal point. You should also note that the "overclocking numbers" command is applied to the image in Gen_MAPw – where there was an overlap, the program shifted the digits of the values, and turned some of them.

The possibilities of adapting digital fields in Gen_MAPw are illustrated in Fig. 7 – Fig. 10. The boundary of the area, the calculated grid is not changed, the boundaries of subdomains, inclusions, and the contour-split are entered (Fig. 7). "Allocated" sections (they're 4, different boundary conditions) provide autonomous calculation in nodes within these subdomains. Points that belong to these regions with measurements outside the "Selected" sections are not taken into account. Two sections marked with horizontal hatching have attributes that identify the areas bounded by them as inclusions. In inclusions, an autonomous calculation of the digital field is also performed, and points with measurements that get inside "give" their values to the border, and they are taken into account in the external part.

The isolines of the distribution calculated in this way and the isolines of the reference field (Fig. 8 – Fig. 10) are very close. It should be noted that the initial capabilities of the system GeoBazaDannych have been significantly expanded in recent years. This was achieved by integrating and addition the system with the tools of the GGMD complex and the functions of the computer algebra system. In the current state, GeoBazaDannych provides users not only with the means to solve specific industrial tasks, but also with the possibility of scientific research on new methods of analysis and processing of initial data and used computer models. In particular, for the above problem, error estimates are obtained using the method described in [12]. We do not give specific digits here, because the errors of the so-calculated adapted field are comparable to the accuracy of obtaining grid functions, and these errors are an order of magnitude lower than in the unadapted field.

III. Conclusion

The article discusses the questions of instrumental filling and use of the interactive computer system GeoBazaDannych. The results of intellectual processing of data included and used in computer geological models are presented and discussed.

References

[1] P. Bedrikovetsky, Mathematical Theory of Oil and Gas Recovery: With Applications to ex-USSR Oil and Gas. Springer Science & Business Media, 2013.

[2] 13th International Symposium on Software Engineering for Adaptive and Self-Managing Systems. Gothenburg, Sweden, May 28-29, 2018. Available at: https://conf.researchr.org/track/seams-2018/seams-2018-papers (accessed 2019, Dec).

[3] V. P. Savinyh, V. Ya. Cvetkov, "Geodannye kak sistemnyi informacionnyi resurs", Vestnik Rossiiskoi akademii nauk, vol. 84, no. 9, pp. 826–829, 2014, (in Russian).

[4] V. Ya. Tsvetkov, A. A. Lobanov, "Big Data as Information Barrier", European Researcher, vol. 78, no. 7–1, pp. 1237–1242, 2014.

[5] S. V. Shaitura, "Intellektual'nyi analiz geodannyh", Perspektivy nauki i obrazovaniya, no. 6 (18), pp. 24–30, 2015, (in Russian).

[6] V. V. Golenkov, "Ontology-based design of intelligent systems", Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], pp. 37–56, 2017.

[7] S. G. Dyshlenko, "Analiz i razrabotka harakteristik kachestva geodannyh", Perspektivy nauki i obrazovaniya, no. 2 (20), pp. 23–27, 2016, (in Russian).

[8] V. B. Taranchuk, "Chislennoe modelirovanie processov dvuhfaznoi mnogokomponentnoi fil'tracii v neodnorodnoi poristoi srede", Doktorskaya dissertaciya, Kazan' State University, 1992, (in Russian).

[9] S. A. Barvenov, M. F Kibash, V. B. Taranchuk, "Metodika, instrumentarii adaptacii matematicheskih modelei processov podzemnoi gidrodinamiki", Vybranyya navukovyya pracy BDU, t. "Matematika", Minsk, pp. 34–65, 2001, (in Russian).

[10] V. B. Taranchuk, "Postroenie cifrovyh geologo-ekologicheskih modelei v sisteme GBD-e", Geoekologiya Belarusi. Minsk : Pravo i ekonomika, Nauchnoe izdanie, pp. 72–96, 2006, (in Russian).

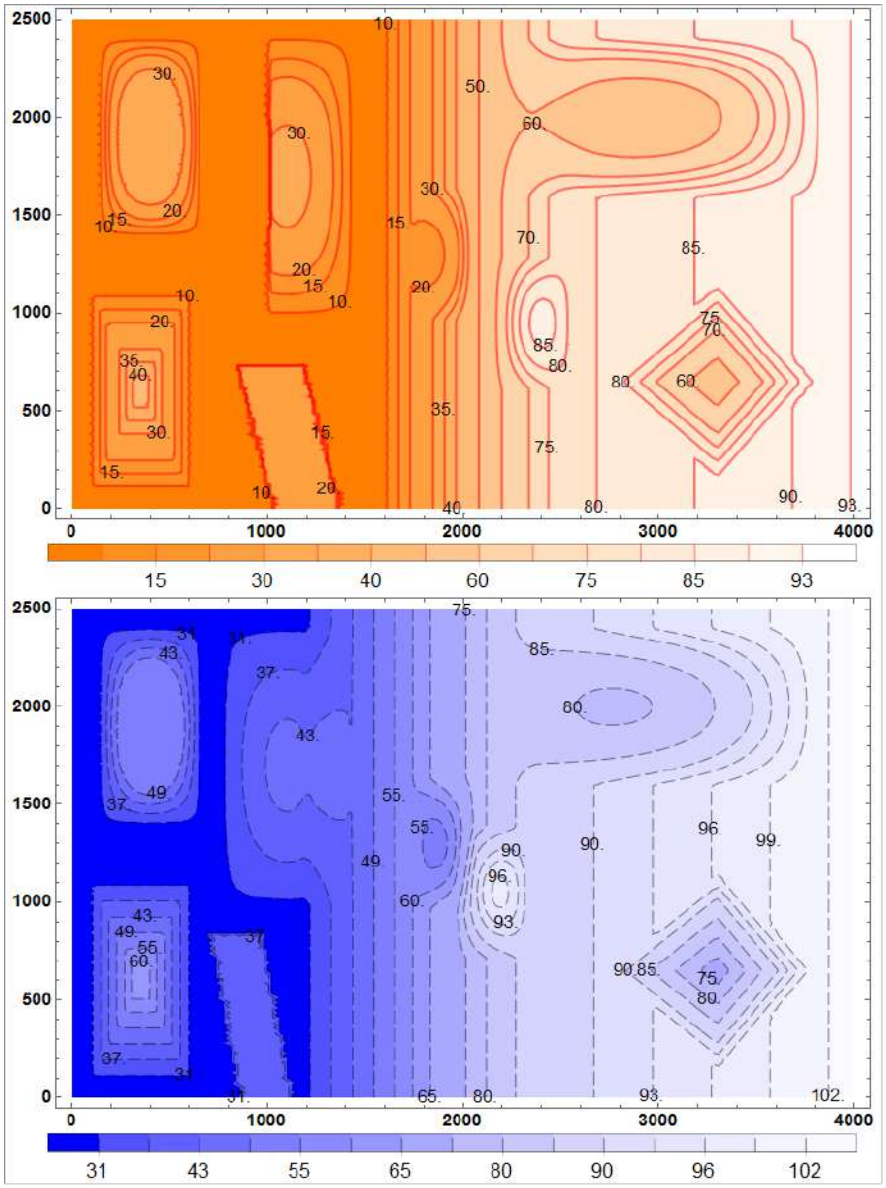


Figure 3. Isolines and density maps (color filling of intervals) of digital fields of *zSurfB*(x,y), *zSurfT*(x,y) functions. Also, color intervals are explained by legends and value labels, which are the same as the contour levels.

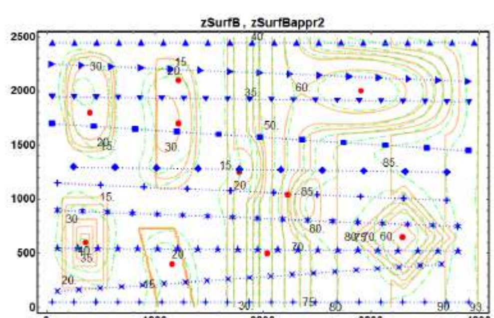


Figure 4. Reference centers of perturbations of the base surface (tape) with measurements and measurement points on the profiles. Isolines of the digital fields of the reference and restored distributions in GGMD.

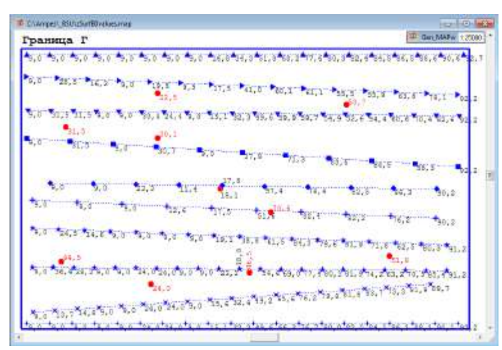


Figure 5. Results of exporting points with measurements from GGMD to Gen_MAPw.

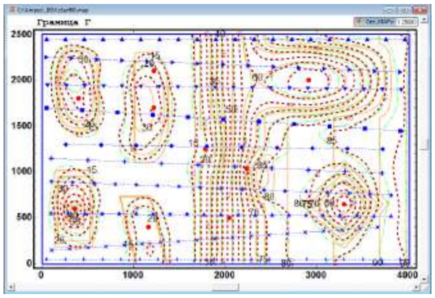


Figure 6. Isolines of reference and restored fields in GGMD and in Gen_MAPw (spline approximation algorithm).

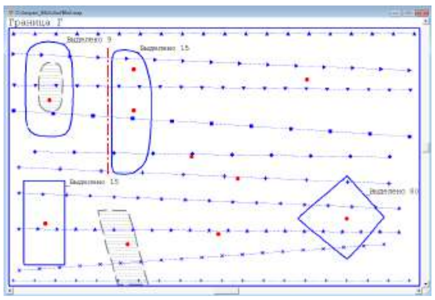


Figure 7. Input map for Gen_MAPw: subdomains, inclusions, splits, boundary conditions. Version with expert additions for calculating the adapted digital distribution.

[11] V. Taranchuk, “The integrated computer complex of an estimation and adapting of digital geological models”, Studia i Materiały, no. 2 (14), pp. 73–86, 2017.

[12] V. B. Taranchuk, “Metody i instrumentarii ocenki tochnosti komp’yuternyh geologicheskih modelei”, Vestnik BrGTU, no. 2, pp. 64–69, 2018, (in Russian).

[13] J. C. Davis, Statistics and data analysis in geology. New York: John Wiley & Sons Inc., 1986.

[14] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, “Design and Analysis of Computer Experiments”, Statistical Science, no. 4, pp. 409-435, 1989.

[15] E. M. Halimov, “Detal’nye geologicheskie modeli i trehmernoe modelirovanie”, Neftegazovaya geologiya. Teoriya i praktika, no. 3(7), pp. 1–10, 2012, (in Russian).

[16] WOLFRAM MATHEMATICA. Available at: http://www.wolfram.com /mathematica/ (accessed 2019, Dec).

[17] V. B. Taranchuk, “Examples of the use of artificial neural networks in the analysis of geodata”, Open semantic technologies for intelligent systems, vol. 3, pp. 225–230, 2019.

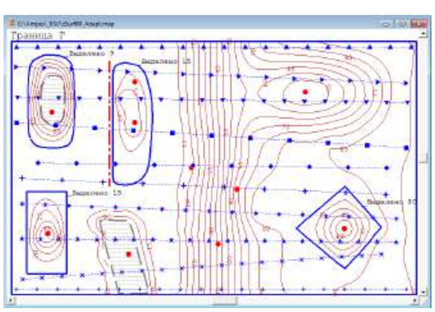


Figure 8. Result of Gen_MAPw-adapted distribution.

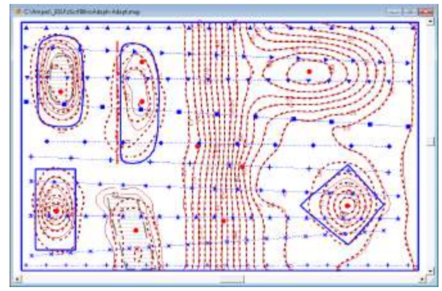


Figure 9. Result of Gen_MAPw-no_adapted and Gen_MAPw-adapted distributions.

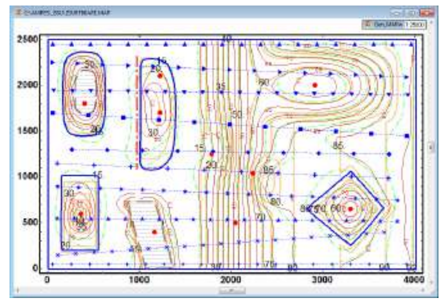


Figure 10. Result of Gen_MAPw-adapted distribution and fields in GGMD.

Примеры интеллектуальной адаптации цифровых полей средствами системы ГеоБаза Данных

Таранчук В. Б.

В статье обсуждаются вопросы инструментального наполнения и использования интерактивной компьютерной системы ГеоБаза Данных. Представлены и обсуждаются результаты интеллектуальной обработки данных, включаемых и используемых в компьютерных геологических моделях.

Received 28.01.2020

Experience of Transformation the Remote Earth Sensing Taxonomy into Ontology

Igor Boyko
Geoinformation Systems of the National Academy of Sciences of Belarus
Minsk, Belarus
igor_m_boyko@hotmail.com

Sergei Pisarevskiy
Geoinformation Systems of the National Academy of Sciences of Belarus
Minsk, Belarus
pi_sar@mail.ru

***Abstract*—This article represents an example of ontology development in the field of Remote Earth Sensing (RES). Initial taxonomy was transformed into ontology. The power of the ontology, its flexibility and logic of data structuring makes the application much more effective in comparison with the taxonomy.**

***Keywords*—axonomy, ontology, remote Earth sensing**

I. Introduction

There is a lot of literature about building taxonomies. No needs to repeat importance of the taxonomic models and approaches to build it. A very comprehensive and deep literature review about the topic is represented by A. Pellini and H. Jones [1]. Meanwhile, sometimes in parallel and sometimes as an alternative to taxonomies another data structure which is ontology may be used. It is happening that initially collected data organized in taxonomy should be reorganized into ontology. R. Iqbal and others [2] give detail literature overview about ontology development.

In the article the preference is given to ontology though initial data in the field of Remote Earth Sensing were represented with taxonomy.

II. RES taxonomy

The RES taxonomy was developed for information search and navigation purposes. An expert of the RES domain collected the data and built a hierarchy of the RES terminology. The hierarchy was represented as taxonomy where parent-child relations were defined and definitions to each term were given. As a tool for visual presentation the Stanford Protégé was used.

The combined fragment of the taxonomy shows how terns and their definitions represented in the tool (Picture 1).

For navigation such a structure works well. However, for search the structure provides very limited possibilities. If a user could try finding some equipment, for example, by the term 'stabilization' no result would be given because there is no any information about the function of the item in the taxonomy.

Each definition for any item in taxonomy is a source of information to extract particular concepts and populate it into ontology in appropriate position. So, additional concepts should be introduced to accumulate specific data.

For example, the concept 'function' reflects a property of the item and will be a classification category for all possible functions of items mentioned in the category 'Earth remote sensing spacecraft'. Similar, the concept 'parameter' will be a category comprising parameters of the items. The Picture 2 shows the list of concepts including different properties of the RES spacecraft components.

So, the new structure provides the list of properties which can be linked with the items having such properties. The definitions for each concept will be used or added and kept without any changes. For example, five properties for the concept 'Earth remote sensing spacecraft' were added: Function, Information, Operation stage, Parameter. The names of the properties were chosen by the expert in the RES field. Picture 3 the shows detailed fragment of taxonomy.

III. RES ontology

Eventually, to convert RES taxonomy into RES ontology relations between properties and items should be created. Picture 4 shows the relation 'implement_function' between the item 'Damping system' and the function 'Dump', which means: damping system implement_function damp. This is a regular triple: subject — predicate -– object. The same item can have more than one relation, so the item 'Damping system' has the relation 'change_parameter' with another property under 'Parameter' -– 'Angular velocity' (Picture 5), which means: damping system change_parameter angular velocity', and it is a regular triple: subject -– predicate -– object too.

The relations between items and their properties were consistently created according to definitions of the concepts in initial taxonomy. The detailed information structure in ontology allows to provide more extended search in comparison with taxonomy having the same information as ontology but without detailed structuring.

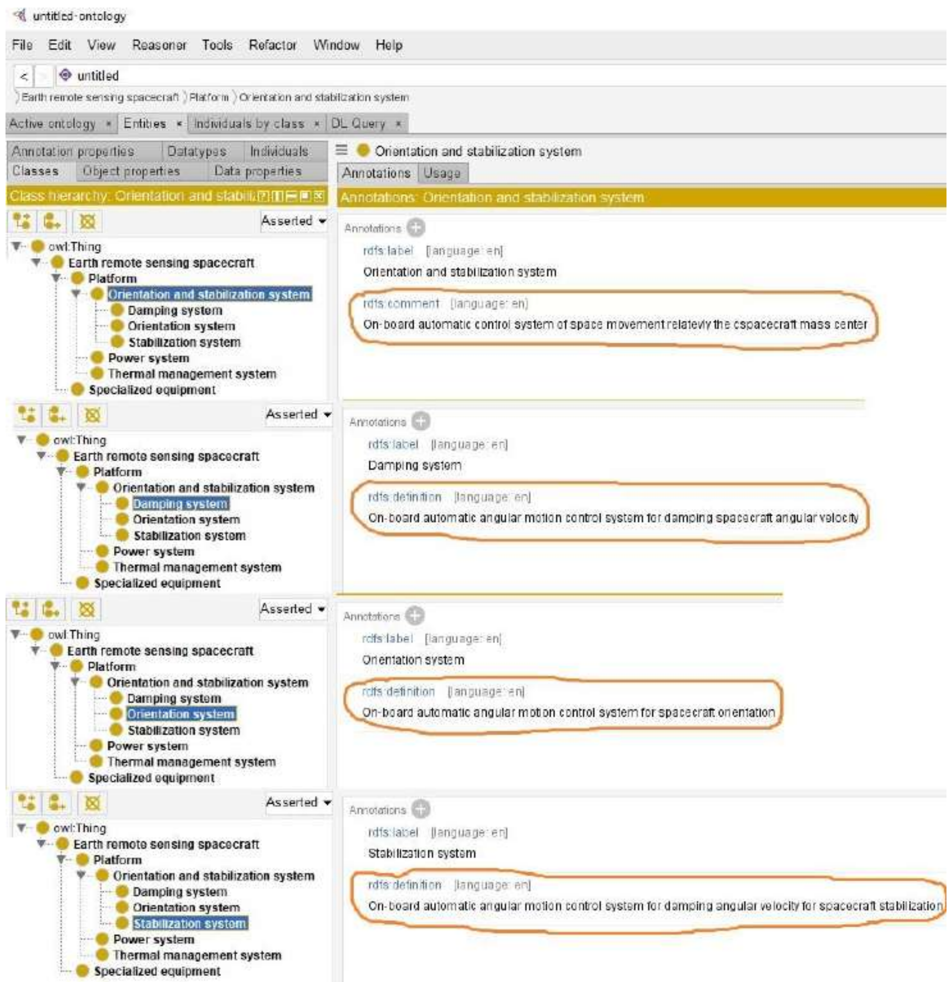


Figure 1. RES taxonomy fragment for items

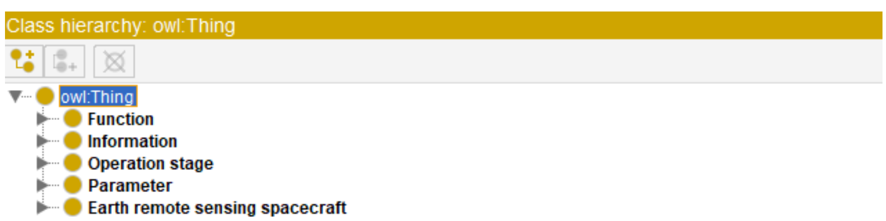


Figure 2. RES taxonomy fragment for items and properties

Now it is easy to find necessary information in ontology by query comprising as the name of some item so the name of some property. For example, for the query 'damp' Protégé finds all triples with the concept 'damp' or for the query 'angular velocity' it finds all triples with concept 'angular velocity'.

IV. Conclusion

Structuring the data, it is important to make a choice between taxonomy and ontology. The choice depends on the purpose of the project. If the search feature is crucial then better to give a preference to ontology, if the navigation is enough the preference may be given to taxonomy. Ontology is more flexible but demands more work, time and physical space.

One more choice is in the tool for data visualization and representation. It was a choice between Stanford Protégé editor [3] and OSTIS technology [4]. The preference was given to Protégé because of a very convenient visualization support and usability. OSTIS provides more comprehensive and developed approach to define relations between items in the ontology and provides unified programming language SC-code to manipulate with the

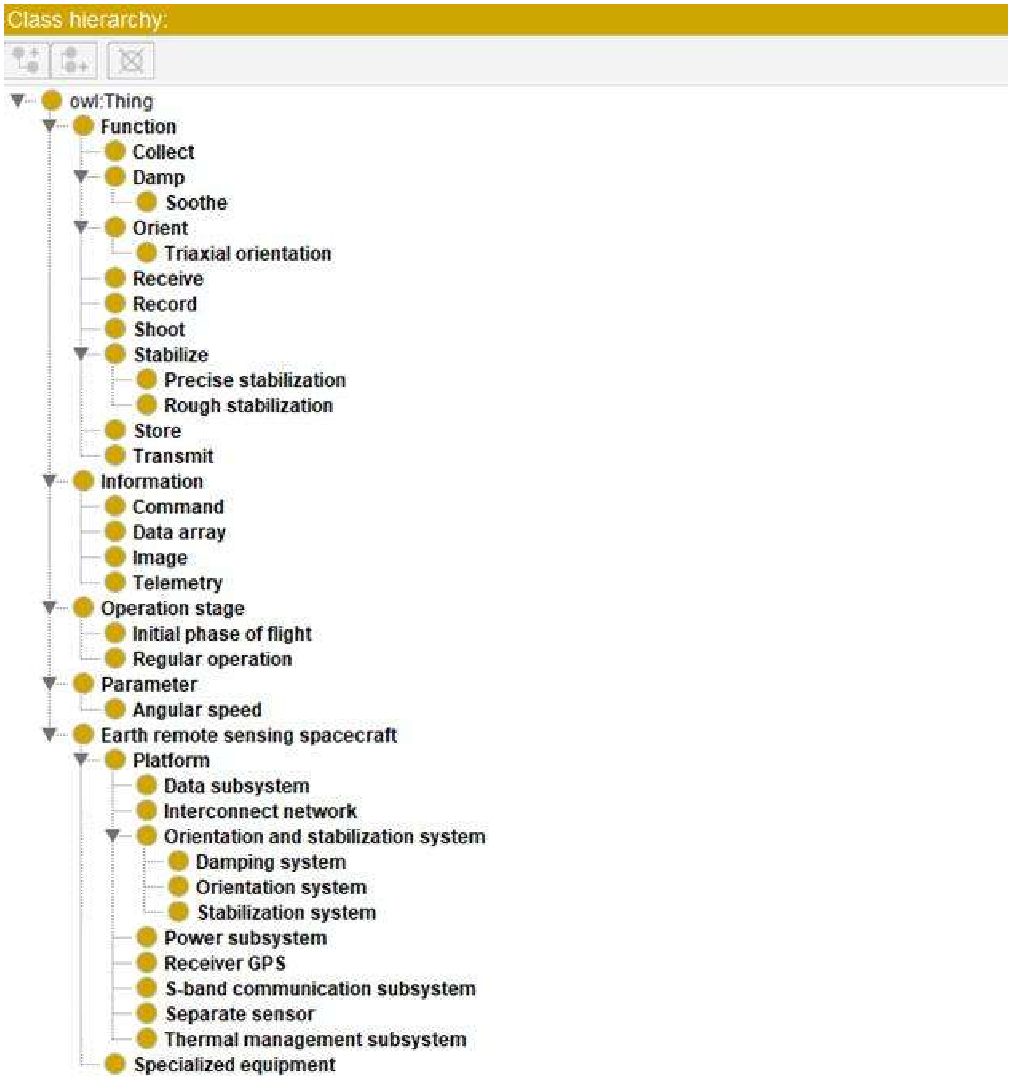


Figure 3. Detailed fragment of RES taxonomy for items and properties

stored data, which is a strong feature of the technology. The OSTIS tool does not provide the effective visual editor to control data yet, but if the tool would be completed it would be applied and illustrated in the article.

References

[1] Arnaldo Pellini, Harry Jones. Knowledge taxonomies. A literature review. ODI, London, May, 2011.

[2] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha and Nurfadhlina Mohd Sharef. An Analysis of Ontology Engineering Methodologies: A Literature Review. Research Journal of Applied Sciences, Engineering and Technology 6(16): 2993-3000, September, 2013.

[3] https://protege.stanford.edu/

[4] Golenkov V. V., Gulyakina N. A., Grakova N. V., Davydenko I. T., Nikulenka V. Y. From training intelligent systems to training their development tools. *Open Semantic Technologies for Intelligent Systems*, Minsk, 2018, Iss. 2, pp. 81-98.

Опыт преобразования таксономии по дистанционному зондированию Земли в онтологию

Бойко И.М., Писаревский С.Д.

В статье показан практический пример преобразования таксономии, содержащей информацию по дистанционному зондированию Земли в онтологию с целью обеспечить не только функцию навигации в иерархической структуре данных, но и обеспечить эффективный поиск требуемых данных по запросу пользователя.

Received 14.01.2020

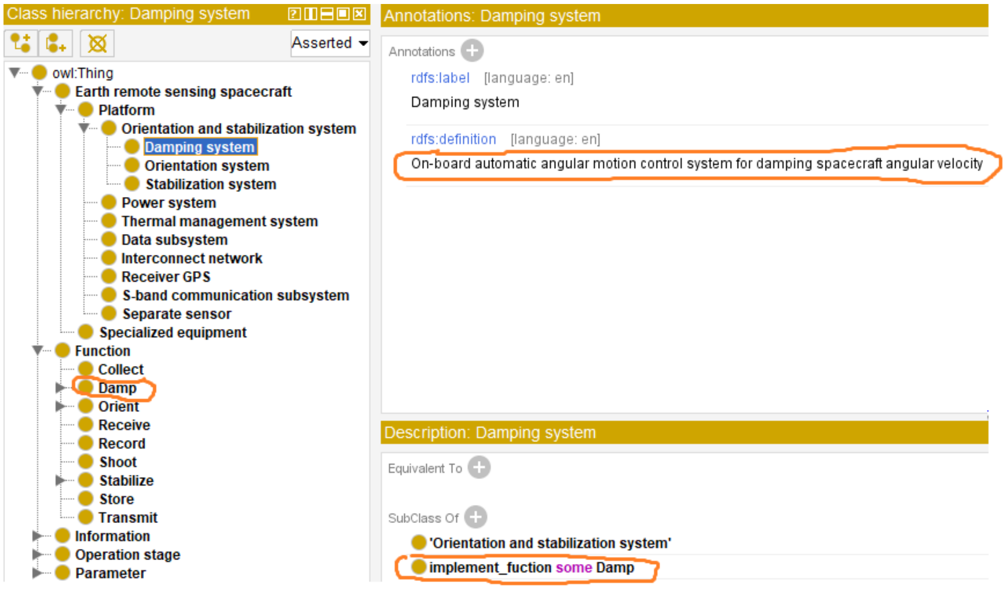


Figure 4. Example of RES ontology with the relation between item and property

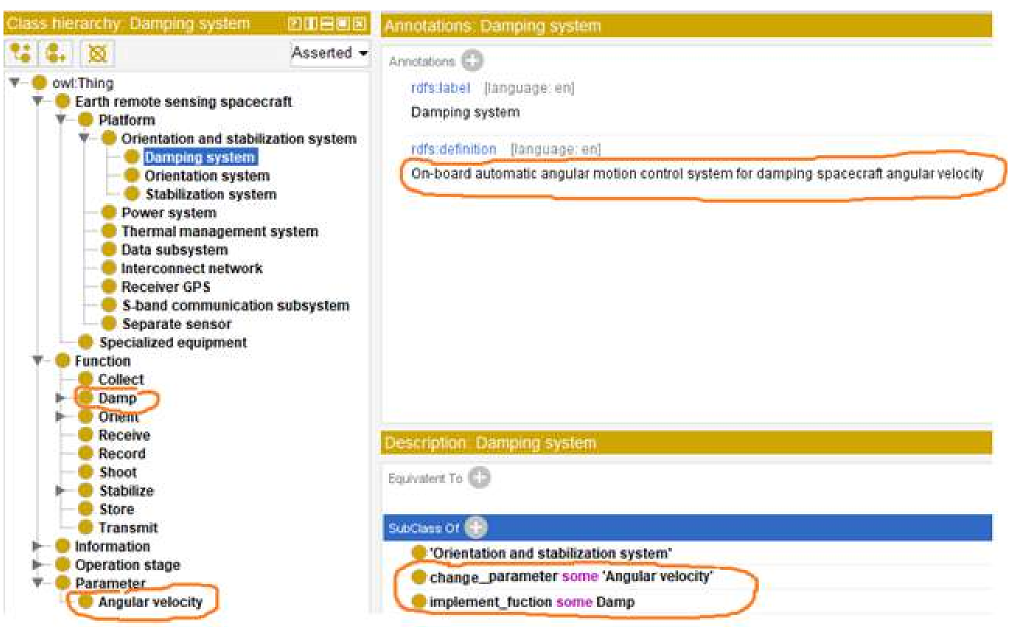


Figure 5. Extended example of RES ontology with the relations

Supervised Learning: from the allocation algorithm to the identification of hidden interpretable patterns

Vadim G. Rodchenko
Yanka Kupala State University of Grodno
Grodno, Republic of Belarus
rovar@grsu.by

***Abstract*—In Supervised Learning, traditionally, training is aimed at constructing an algorithm that in the future should carry out the correct classification of objects. The initial training sample is formed on the basis of classes alphabet and an a priori features dictionary, and then, in the process of training, separating surface classes are constructed. A practically useful classifier is constructed in result, but nothing can be learned about the properties of classes. An alternative approach to Supervised Learning is proposed, which is aimed at studying of the of classes properties at revealing the hidden interpreted patterns.**

***Keywords*—machine learning, data mining, classification, supervised learning, instance-based learning, training dataset**

I. Introduction

The development and implementation of new methods and algorithms of machine learning in order to increase the efficiency of processing the accumulated data arrays is one of the most important tasks of computer science [1].

Today, the progress in practical use of artificial intelligence technologies is completely depends on an increase in the effectiveness of Machine Learning (ML) methods and tools [2]. The most important and determining stage in ML is training, which is implemented on the basis of the data training samples and is aimed at identifying empirical patterns [3].

As part of a traditional approach, the result of Supervised Learning is the classification algorithm, which is actually a practically useful "black box", but which can hardly be interpreted.

The article proposes an alternative approach to Supervised Learning, which is based on study of the class properties and identification of the informative features combinations that provide class distinction. The results of the practical application of the approach based on a real dataset are described.

II. Implementing Machine Learning Based on Supervised Learning

Formally, the Machine Learning process based on Supervised Learning can be represented as the following chain of transformations:

$$S \xrightarrow{F_1} C \xrightarrow{F_2} A \xrightarrow{F_3} T \xrightarrow{F_4} I \xrightarrow{F_5} P \xrightarrow{F_6} R$$

where S is the alphabet of classes; C is a set of observed characteristics; A – a priori dictionary of features; T – training samples; I – a refined dictionary of informative features combinations to construct the decision spaces; P – classifier built on the basis of class patterns; R is the set of solutions; F1 – algorithm for obtaining the observed characteristics; F2 – algorithm to construct an a priori features dictionary; F3 – training samples generation algorithm; F4 – an algorithm for identifying the informative attributes combinations of an a priori dictionary to construct the decision spaces; F5 – an algorithm for constructing a classifier based on class patterns in decision spaces; F6 – P-based classification.

The alphabet of classes S, the dictionary C of observable characteristics, and the a priori dictionary of attributes A are developed by experts. The training samples T is formed on the basis of the states of objects observations taking into account S, C and A as a result of algorithms F1 and F2 execution.

Based on the analysis of the training data samples, a study of features combinations from the a priori dictionary is made and a refined dictionary I is formed. Combinations of features that turned out to be uninformative in terms of the class patterns separation are not included in the refined dictionary I.

We exclude from the training set the lines corresponding to the characteristics that did not fall into the updated dictionary I. Using the new training data set, we first construct the class patterns in the form of cluster structures and then use them to build a classifier.

The decision to assign a recognized object to a certain class is based on the study of the object belonging to the class pattern.

III. An Alternative Approach to Supervised Learning

In Machine Learning, based on Supervised Learning, actually two tasks are solved sequentially: 1) training and 2) classification. As a result of solving the first problem, a classifier constructed, which then is used to solve the second problem.

The results of existing methods and tools in the area of Supervised Learning analysis suggest two approaches to the implementation of the learning process.

At present, such an approach in which training is reduced to constructing decision rules that provide an extremum for a certain criterion is traditionally accepted and is universally used. The class of decision rules is a priori specified up to parameters, and training involves finding such parameter values that provide an extremum for the selected criterion.

On use of a priori dictionary of features applied to construct the training set attention does not payed. It is believed that the space to describe objects is given, and it is only necessary to build a separating surface in this space within the criterion framework.

An alternative approach to learning is based on the idea of finding such features combinations from the a priori dictionary that are most informative from the point of distinguishing classes. The identification of these features combinations occurs as a result of the feature spaces construction in which class patterns do not intersect [4]. After this, the procedure for constructing a classifier becomes trivial.

In the framework of the traditional approach to Supervised Learning, it is believed that a space for describing objects is given and it is necessary to build a separating surface in this space.

The practical application of the traditional approach has effectively solved a large number of different applied problems.

Especially exciting results were obtained in the area where the artificial neural network technologies are used. However, although neural network technologies provide for virtually automatic training, the problem of interpreting the revealed patterns has not yet been solved. Then, it turns out that the useful result obtained from resource-intensive process of preparing the training samples (up to 80% of all costs) is only the classification algorithm, which is actually a "black box".

Note that at present, in machine learning, the focus is done on all methods to construct the classification algorithms is the "Achilles heel", since they allow us to separate class patterns, but do not find out anything about the properties of classes.

In order to study and identify the properties of classes, an alternative approach to learning can be proposed, which is based on the assumption of the compactness hypothesis that compact sets in the attribute space correspond to class patterns.

Obviously, in the a priori dictionary there may be features that are not informative from the separating class patterns point of view, and then objects of one class are either placed non-compactly in the attribute space or are scattered among objects of another class.

Based on the analysis of data from the training samples, it is proposed to identify such combinations of features that provide separation of class patterns. In this case, the training will actually be aimed at identifying feature spaces in which the compactness hypothesis is confirmed [5].

Note that, firstly, found features combinations can be interpreted within the framework of the subject area, and, secondly, they can be used to construct the classification algorithms [6].

IV. Supervised Learning Based on Analysis of Class Pattern Properties

The classical formulation of the Supervised Learning problem assumes that there are many descriptions of objects X and many acceptable answers for their classification Y. There is an unknown target dependence $y^*: X \to Y$, values $X^m=\{(x_1,y_1),\dots,(x_m,y_m)\}$ which are known only for objects of the training set. It is necessary to construct an algorithm $a: X \to Y$ that would approximate this target dependence, not only on the objects of the finite sample, but also on the whole set X.

To solve the problem, a certain class of algorithms is preliminarily specified up to parameters, and training is reduced to finding the values of the parameters providing an extremum for the selected criterion.

When solving the tusk, a number of problem points arise:

1) The choice of the model $A = \{a : X \to Y\}$ is a non-trivial task and requires the participation of a qualified specialist, which ultimately allows to implement only an automated, but not automatic learning mode.
2) The class-separating surface is constructed on the basis of the data of the training sample X^m, and the question of the information content of the used features from the a priori dictionary remains open.
3) The constructed algorithm $a : X \to Y$ approximates the unknown target dependence, but is actually a "black box" that cannot be interpreted.

If an alternative approach to Supervised Learning is used, one can avoid these disadvantages. The statement of the learning problem in this case is as follows: let there be a lot of descriptions of objects X and a set of valid answers for their classification Y. There is an unknown target dependence $y^*: X \to Y$, values $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$ which is known only for the objects of the training set. Then it is required to find the feature subspaces in which class patterns do not intersect.

Suppose that training samples $X^m = \{(x_1, y_1),\dots, (x_m, y_m)\}$ is formed on the basis of the dictionary features $F = \{f_1,\dots, f_n\}$. Denote by $V = \{v_1,\dots, v_q\}$ the set of all possible combinations of features from F. Then V contains q=$\sum_{i=1}^{n} C_n^i$=2^n-1 subsets.

The search algorithm of feature subspaces in which class patterns do not intersect on data set V = {v_1,..., v_q} consists of the following nine steps:

STEP 1. Choose from V a subset of V^+={$v^+{}_1$,...,$v^+{}_n$}, where $v^+{}_i$ contains only one attribute.

STEP 2. For each $v^+{}_i$ build the class patterns and evaluate their relative placement.

STEP 3. Include $v^+{}_i$ in result set V^*={$v^*{}_1$,...,$v^*{}_k$} if the class patterns do not intersect.

STEP 4. Exclude from the set V={v_1,..., v_q} the subset V^+={$v^+{}_1$,...,$v^+{}_n$} and get $V^\wedge$ ={$v^\wedge{}_1$,...,$v^\wedge{}_p$}.

STEP 5. Exclude from $V^\wedge$ all combinations $v^\wedge{}_i$, that contain any combination from V^*={$v^*{}_1$,...,$v^*{}_k$}.

STEP 6. Take the next combination $v^\wedge{}_i$ from $V^\wedge$ and based on it construct a feature subspace.

STEP 7. In this feature subspace, we build class patterns and evaluate their relative positioning.

STEP 8. If the class patterns do not intersect, then include the combination of features $v^\wedge{}_i$ in the resulting set V^* and exclude from $V^\wedge$ all combinations that contain $v^\wedge{}_i$.

STEP 9. The process is repeated until $V^\wedge$ becomes empty.

As a result of the analysis of all elements V={v_1,..., v_q}, the set V^*={$v^*{}_1$,...,$v^*{}_t$} will be constructed, where $0 \leq t \leq q$. Based on the combinations $v^*{}_i \in V^*$, we formulate previously hidden and empirically revealed patterns: "in the feature space of the subset $v^*{}_i$ the classes do not intersect".

Since the combinations of features $v^*{}_i$ can be interpreted within the framework of a specific applied problem, then all the revealed patterns can be interpreted.

The combination of features $v^*{}_i \in V^*$ defines the solutions spaces in which class patterns do not intersect. For class patterns inside such spaces, the compactness hypothesis is confirmed, and therefore the construction of classification algorithms is straightforward.

V. Results of dataset analysis

We demonstrate the results of applying the proposed algorithm to detect patterns based on cluster structures. We have used the Mushroom dataset, which is hosted in the UCI Machine Learning Repository [7].

The Mushroom dataset contains data on 8124 instances, 3916 of which belong to the Poisoned class, and 4208 to the Eatable class. 22 attributes were used to describe instances: *odor*, *gill-color*, *ring-type*, *stalk-color-below-ring*, *stalk-color-above-ring*, *spore-print-color*, *ring-number*, *veil-color*, *cap-surface*, *cap-shape*, *cap-color*, *gill-attachment*, *gill-spacing*, *stalk-shape*, *bruises*, *stalk-root*, *gill-size*, *veil-type*, *stalk-surface-above-ring*, *stalk-surface-below-ring*, *population*, *habitat*.

Table 1 shows the results of a study of the intersection of class patterns based on a single attribute. The table shows that the odor attribute provides a good separation of the Poisoned and Eatable classes.

Table I
Results of the first numerical experiment

Attribute	Intersection (%)
odor	3.06
gill-color	55.26
ring-type	65.99
stalk-color-below-ring	87.44
stalk-color-above-ring	87.85
spore-print-color	98.16
ring-number	99.08
veil-color	99.80
cap-surface	99.90
cap-shape	99.90
cap-color	100
gill-attachment	100
gill-spacing	100
stalk-shape	100
bruises	100
stalk-root	100
gill-size	100
veil-type	100
stalk-surface-above-ring	100
stalk-surface-below-ring	100
population	100
habitat	100

Table 2 presents the results of two attributes combinations analysis. Table shows data on the 10 most suitable for separating the **Poisoned** and **Eatable** classes from the 230 possible combinations.

Table II
Results of the second numerical experiment

Combination attributes	Intersection (%)
odor, spore-print-color	1.23
odor, habitat	2.15
odor, cap-color	2.45
odor, gill-color	2.45
odor, stalk-color-below-ring	2.45
odor, stalk-root	2.86
odor, stalk-color-above-ring	2.86
odor, veil-color	2.86
odor, cap-surface	2.96
odor, cap-shape	2.96

It turns out that patterns of the **Poisoned** and **Eatable** classes stopped to intersect only in feature spaces formed by combinations of 4 features.

Table 3 shows the results of a numerical experiment based on combinations of four attributes. Only 13 features from 7315 possible combinations were identified in which the intersection of the **Poisoned** and **Eatable** classes patterns have not intersected.

Table III
RESULTS OF THE THIRD NUMERICAL EXPERIMENT

Combination attributes	Intersection (%)
odor, habitat, population, stalk-color-below-ring	0.0
odor, habitat, population,spore-print-color	0.0
bruises, gill-spacing, spore-print-color, stalk-root	0.0
bruises, cap-color, spore-print-color, stalk-root	0.0
bruises, gill-size, spore-print-color, stalk-root	0.0
bruises, ring-number, spore-print-color, stalk-root	0.0
bruises, population, spore-print-color, stalk-root	0.0
odor, bruises, habitat, stalk-surface-above-ring	0.0
odor, habitat, ring-type, stalk-root	0.0
odor, population, spore-print-color, stalk-root	0.0
bruises, habitat, spore-print-color, stalk-root	0.0
odor, cap-color, habitat, stalk-root	0.0
odor, habitat, stalk-color-below-ring, stalk-root	0.0

CONCLUSION

This paper presents an alternative approach for executing the Supervised Learning procedure, which is based on an analysis of the attributes properties of an a priori dictionary.

The purpose of training is to search for feature subspaces in which class patterns do not intersect. Class patterns are represented as cluster structures.

The learning algorithm allows to automatically analyze the training data samples and identify the most informative features in terms of class separation.

The results of the learning algorithm application to solve the classification problem are presented. The Mushroom dataset from the UCI Machine Learning Repository has been used.

ACKNOWLEDGMENT

The author thanks Boris Assanovich for his assistance in preparing the paper final version.

REFERENCES

[1] J. T. Tou and R. C. Gonzales, “Printsipy raspoznavaniya obrazov [Pattern recognition priceples],” Moscow: Mir, 1978, 412 p.

[2] P. Flach, “Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh [Machine learning. Science and art of constructing algorithms that extract knowledge from data],” Moscow: DMK Press, 2015, 400 p.

[3] N. G. Zagoruiko, “Prikladnye metody analiza dannykh i znanii [Applied methods of data analysis and knowledge],” Novosibirsk: IMSO RAN, 1999, 270 p.

[4] V. V. Krasnoproshin and V. G. Rodchenko, “[Cluster structures and their applications in Data Mining],” Informatika [Informatics], 2016, vol. 50, no 2, pp. 43-57.

[5] MachineLearning. Available at: http://www.machinelearning.ru/wiki/index.php?title (accessed 2019, Dec).

[6] Data mining. Available at: https://en.wikipedia.org/wiki/Data_mining (accessed 2019, Dec).

[7] UCI Machine Learning Repository. Mashroom Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/mushroom (accessed 2019, Dec).

Обучение с учителем: от построения алгоритма классификации к выявлению скрытых интерпретируемых закономерностей

Родченко В.Г.

Обучение с учителем направлено на построение алгоритма, который в дальнейшем должен осуществлять правильную классификацию объектов. Исходная обучающая выборка формируется на основе алфавита классов и априорного словаря признаков, а затем в процессе обучения строятся разделяющие классы поверхности. В результате получается практически полезный классификатор, но о свойствах классов ничего узнать не удается. Предложен альтернативный подход к обучению с учителем, который направлен на исследование свойств классов и на выявление скрытых интерпретируемых закономерностей. Целью обучения является поиск признаковых подпространств, в которых паттерны классов не пересекаются. Паттерны классов представляются в виде кластерных структур. Алгоритм обучения позволяет автоматически провести анализ данных обучающей выборки и выявить наиболее информативные признаки с точки зрения разделения классов. Представлены результаты применения алгоритма обучения для решения задачи классификации.

Received 27.01.2020

Software technology for deep learning of belief neural networks

1st Victor V. Krasnoproshin
Belarusian State University
Minsk, Belarus
Krasnoproshin@bsu.by

2nd Vadim V. Matskevich
Belarusian State University
Minsk, Belarus
Matskevich1997@gmail.com

***Abstract*—The paper provides the framework structure and contents description for solving applied problems using deep belief networks. Original network architecture, focused on parallel data processing, set of algorithms implementing training processes based on the annealing method and solving problems are proposed.**

The effectiveness of the described framework is demonstrated by the example of solving the problem of compressing color images.

***Index Terms*—framework, annealing method, deep belief network, parallel computations, training, dataset**

I. Introduction

Currently, a wide range of applied problems is being solved using neural network technologies implemented in the form of frameworks. A framework is a software package that implements a specific architecture of a neural network to solve a specific range of tasks.

The most difficult stage of neural network data processing is the network training process [1] [2]. The existing today frameworks use mainly gradient training methods [3]. With visible popularity, gradient methods have certain disadvantages. Therefore, the problem of training is still relevant [4].

The paper offers a description of the original framework implementing the architecture of deep belief networks, for the training of which the annealing method is used. This method is lack of the main disadvantages of gradient methods, but it works much slower [5]. An approach to solving this problem is proposed, and the efficiency of the proposed framework is demonstrated by example of solving the problem of compressing color images.

II. Problem Analysis

Deep belief networks are used to solve a number of applied problems such as: medical diagnostics, pattern recognition, image processing, selection of semantically significant features, etc. [6].

In order to describe the architecture of a deep belief network, it is necessary first to describe the architecture of the restricted Boltzmann machine. It is known that any deep belief network always contains layers of this type of machine.

At the heart of the machine is the concept of a stochastic neuron.

Formally, a restricted Boltzmann machine can be represented as a fully connected bipartite graph $G = (X, U)$,

$$\begin{cases} X = X_1 \cup X_2, X_1 \cap X_2 = \emptyset \\ U = \{u = (x_1, x_2) | \forall x_1 \in X_1, \forall x_2 \in X_2\}, \end{cases} \tag{1}$$

where X – vertex set – stochastic neurons, U – edges set – synaptic connections, while vertices of subset X_1 -– set the neurons of the input layer, X_2 – output layer neurons.

The number of neurons in the input layer is determined by the size of the input image, and the number of neurons in the output layer is determined based on the requirements for data compress ratio.

The output signals of layers of a restricted Boltzmann machine implement some laws of the probability distribution. Different types of machines are built depending on the laws of distribution used. In this paper, we will talk about machines types of Gauss-Bernoulli and Bernoulli-Bernoulli, because they are the most common.

For restricted Boltzmann machine of Gauss-Bernoulli type to each vertex of the input layer we assign a set of parameters $VB = \{b\}$ – vertex offsets and $\sigma = \{\sigma\}$ – vertex variances, and to the vertices of the output layer – set of parameters $HB = \{g\}$ – vertex offsets. The sizes of the sets are equal respectively

$$|VB| = |\sigma| = |X_1|, |HB| = |X_2| \tag{2}$$

Each edge connecting a pair of vertices of the input and output layers will be assigned a set of parameters $W = \{w\}$ – the weights of the edges.

The size of the set is equal to the following value

$$|W| = |X_1||X_2| \tag{3}$$

Thus, the described family of neural networks can be defined by four types of parameters:

$$RBM = (W, VB, \sigma, HB) \tag{4}$$

Note. A restricted Boltzmann machine of Bernoulli-Bernoulli type does not have set of parameters σ.

A deep belief network contains several layers consisting of restricted Boltzmann machines and, in addition, for generating the output signal may contain a multilayer perceptron (depending on the problem being solved).

A deep belief network in layers consisting of restricted Boltzmann machines solves the problem of data compression, which can be formally described as follows.

Let X be the space of input images of some fixed dimension, Y – the space of compressed images of much smaller dimension than the space X. I.e:

$$\begin{cases} dimX = fix \\ dimY \ll dimX \end{cases} \tag{5}$$

Then the task of data compression is to build compression functions f and recovery g, such that:

$$\begin{cases} f : X \to Y, g : Y \to X \\ d : X \times X \to \Re \\ d(x, g(f(x))) \to min, \forall x \in X, \end{cases} \tag{6}$$

where d is a function that evaluates the differences between two given vectors.

Note. In practice, data compression is carried out for a specific subject area. This, in turn, imposes certain restrictions on the input data and, therefore, reduces the dimension of the space X.

As noted, the most time-consuming step in the use of neural networks is the training process. Since a deep belief network always contains layers of restricted Boltzmann machines, the effectiveness of training the entire network as a whole depends on the effectiveness of solving the problem. Network training can be written as an optimization problem for each of the layers.

Let a training dataset x and a functional for evaluating the quality of data compression d (6) be given. It is needed to find the values of the parameters $(w^*, b^*, g^*, \sigma^*)$, giving a minimum of functional F, i.e.

$$F(x, d, w^*, b^*, g^*, \sigma^*) = \min_{w,b,g,\sigma} F(x, d, w, b, g, \sigma) \tag{7}$$

Note. A restricted Boltzmann machine of Bernoulli-Bernoulli type does not contain the parameter σ and the quality functional F, therefore, does not depend on σ.

To solve optimization problems, you can use either the gradient descent method or random search.

The gradient descent method has fast convergence, but at the same time has several disadvantages:

1) converges to local minimum points [5], which significantly reduces the quality of the solution;
2) requires differentiability of the objective function, which significantly reduces the class of problems to be solved.

The random search method is not widespread [7], however, it has some advantages:

1) does not require objective function differentiability, which significantly expands the class of applied problems;
2) under certain conditions [8] and from any initial approximation [9] it has convergence to the global minimum.

Given the above, we obtain the following training task.

Let a training dataset of N input images of dimension $dimX$ be given and requirements for data compression be fixed, i.e. $dimY = fix$.

It is necessary to develop a deep belief network architecture and a training algorithm (based on the annealing method) so that the following conditions are met:

1) training time should be acceptable (no more than a day);
2) the quality of training should be as high as possible, while the algorithm should require as little data as possible for training.

III. Framework Description

To solve this problem, software was developed in the form of a framework that includes all the necessary algorithms that implement the functioning of deep belief networks.

The framework proposed in the work consists of five main modules: trainDeepNN, compressImages, decompressImages, loadFromFileDeepNN, buildDeepNN.

The compressImages module compresses color images. The decompressImages module - restoring original images from their compressed representation. The loadFromFileDeepNN module - loading the network from the hard drive.

The buildDeepNN module builds a deep belief network with the following architecture.

Since the input data are usually images, the first layer of a deep belief network is formed as an ensemble of M_1 restricted Boltzmann machines of Gauss-Bernoulli type. This made it possible to "cover" the entire numerical range of input data. All machines forming one layer have the same architecture within the layer, so dimX must be a multiple of M_1. All subsequent ones are represented by ensembles of restricted Boltzmann machines of Bernoulli-Bernoulli type. Therefore, for each network layer, the following restriction must be satisfied. The product of the number of machines in the layer by the size of the input layer of each should be equal to the product of the number of machines in the previous layer and the size of the hidden layer of each. The output layer of the network is represented by an ensemble of M_s restricted Boltzmann machines of Bernoulli-Bernoulli type. To complete the data compression requirement, the total number of neurons in the hidden layers of machines must be strictly equal to dimY. The number of adjustable parameters in each of the machines must be strictly less than N. This is necessary to ensure the generalizing ability of the network.

Let's show architecture with an example with four layers (see Fig. 1).

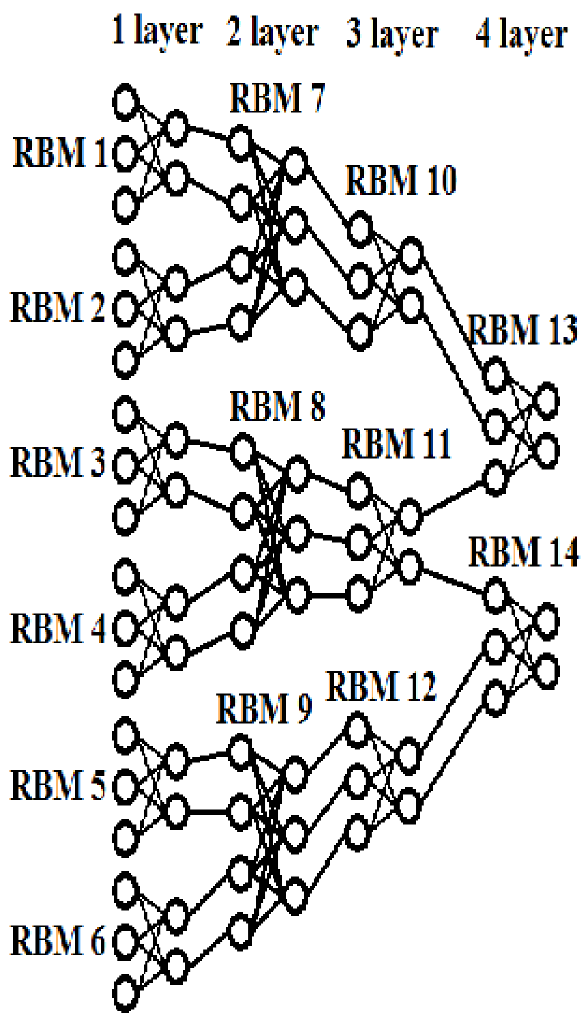


Figure 1. Original deep belief network architecture.

The first layer consists of an ensemble of six restricted Boltzmann machines. The input layer of each machine consists of three neurons, the hidden - of two. The second layer of the network consists of an ensemble of three machines. The size of machine input layer is four neurons, the hidden one is three. The third layer of the network consists of an ensemble of three restricted Boltzmann machines, each of which has three neurons in the input layer and two in the hidden one. The last layer of the deep belief network consists of an ensemble of two machines, each of which has three neurons in the input layer and two in the hidden one.

The proposed architecture has several advantages:

1) decomposition of network layers ensures complete independence of the trained machines-components within the network layer, which allows to parallelize the training process;
2) the architecture can significantly reduce the number of configurable network parameters, which reduces the training dataset size and significantly reduces the computational complexity of the training process;
3) the architecture fully meets the constraints of the problem for an effective solution using heterogeneous computing devices [10].

The trainDeepNN module implements the main function of the framework. He provides training of received deep belief network. The internal composition of this block is presented in the form of the following scheme (see Fig. 2). At the beginning of the module execution, OpenCL is configured on computing devices. Then, a deep belief network training cycle by layers begins. When moving to the next layer, data is preliminarily converted to the architecture of the current layer. After this, cyclic training of the restricted Boltzmann machines that form the current layer is started. The cycle includes initialization of the initial state of the machines, data transfer to computing devices and a training cycle using the original annealing method algorithm.

The following algorithm is proposed that implements the ideology of this method.

At the preliminary stage, initialization (setting the initial values) of the parameters (W, VB, HB, σ), initial temperature T_0 is performed.

The main stage of the training algorithm implements a procedure for sequentially updating the values of the specified parameters using a certain quality functional.

Describe the process to update settings in more detail. For simplicity, consider it on the example of the set of para-meters W. For other sets, this procedure is identical.

To the set of parameters W, we associate a segment L_w of length l. After that, each element of the set W is sequentially placed in the center of the given segment. To determine the direction of change of parameter values, we generate a random variable from 0 to 1. If it is more than 0.5, then the value of the parameter increases, otherwise it decreases.

New parameter values are defined as follows. A random permutation is generated, the number of elements of which is equal to the number of elements of the set W. We order the elements of the set W in accordance with the permutation and change the values of the first W_p elements of the set. The new value of the parameter is determined as a result of the implementation of a uniformly distributed random variable on the segment, the ends of which are the current value of the parameter, and the end of the segment towards which the change is made.

Similarly, actions are sequentially performed for the sets VB, HB, σ.

For newly obtained parameter values, the quality functional is calculated.

As the latter, it is proposed to use the following function:

$$F(W, VB, HB, \sigma) = \frac{1}{NR} \sum_{i=1,N} \sum_{j=1,R} |x_{ij} - f^{-1}(y_{ij})|, \tag{8}$$

where y_{ij} – reconstituted input signal of restricted Boltzmann machine, f^{-1} - inverse function of the preliminary transformation of input data.

Then a decision is made to move to a new state with probability:

$$P(y|x) = \min\{1, \exp(-(F(y) - F(x))/T_i)\}, \tag{9}$$

where x – current state, y – state selected for transition, F – minimized objective function, T_i - temperature of i-th iteration.

— in case of change of state cooling takes place by the rule:

$$T_{i+1} = T_0 / \ln(k+1), \tag{10}$$

where k is the number of completed transitions to a new state.

— otherwise the temperature does not change.

After cooling, the received solution is checked for optimality:

— the solution is optimal if the time allocated for training has expired.

If the received solution is optimal then:

— algorithm stop,

— otherwise move to the next iteration.

We will check the efficiency of using the neural network of the proposed architecture using the example of the problem of compressing color images.

IV. Experiments and Results

The «STL-10» data from the Stanford University repository was used as baseline [11]. The dataset contains one hundred thousand unmarked color images measuring 96x96 pixels. Each image is described by 27648 integer numbers (in the range from 0 to 255) specifying the content of red, green and blue colors [12]. Based on the characteristics obtained (the sample is given approximately $2,8 * 10^9$ numbers, contains descriptions of arbitrary, unmarked objects), we can conclude that the process of compressing images of a given sample with low losses is a rather difficult problem.

For data processing, a standard computer with an 4-core processor and a video card was used: video card: nvidia 1060 3gb; processor: intel i7 4770k 3.5 GHz; RAM: 2x8 Gb 1600 MHz; hard disk: samsung 850 pro 256 Gb; operating system: Lubuntu 16.04.

The compiler gcc was used as software (libraries OpenMP and CUDA version 9.1 [13]) with options:

«gcc superOpenCLFramework.cpp -lstdc++ -D_FORCE_INLINES -O2 -l OpenCL -lgomp -lm -fopenmp». Measurement of operations time was performed using function «gettimeofday».

The following deep belief network architecture was used in the experiments.

The first layer consisted of a combination of 432 restricted Boltzmann machines of Gauss-Bernoulli type. The number of neurons in the input layer and hidden was 64 and 16, respectively, for all machines in the layer. The second layer consisted of a combination of 108 restricted Boltzmann machines of Bernoulli-Bernoulli type. The number of neurons in the input layer and hidden was 64 and 16, respectively, for all machines in the layer. The third layer consisted of a combination of 27 restricted Boltzmann machines of Bernoulli-Bernoulli type. The number of neurons in the input layer and hidden was 64 and 16, respectively, for all machines in the layer.

Images compression ratio was tuned by disabling the last layers of the deep belief network. So 3 layers of the net-work provided 512-fold compression, 2 layers – 128-fold, and the first layer of the network – 32-fold. For the training of the first layer of the network, 2000 images were used, for the second and third – 4000.

In experiments, we will use the number of bits to encode one pixel of the image as a compression ratio. For encoding without compression of one pixel of a color image, 24 bits are required. Reducing the number of bits for encoding a pixel leads to image compression, for example, with 24-fold compression, the number of bits per image pixel will be equal to one.

There are many different functionals for evaluating the quality of image compression [14], however, many of them are not widely used due to the fact that for their calculation it is necessary to fix the parameter values in the functional itself for the correct comparison of the results or require a large amount of calculations.

As a quality functional, to estimate losses during image compression, we will use the most common PSNR (peak signal-to-noise ratio) quality functional that does not require fixing parameters and a large amount of computations. This quality function calculates the ratio of the maximum possible signal to the mean square error for color raster images as follows.

$$PSNR(X, Y) = 10 \log_{10} \frac{3N * 255^2}{\sum_i (x_i - y_i)^2}, \tag{11}$$

where X, Y – compared images, N – the number of pixels in the image X.

The higher the value of the functional, the less compression loss and higher quality. For lossless compression, this quality function is equal to plus infinity. For a high compression ratio (approximately 0.05 bits per

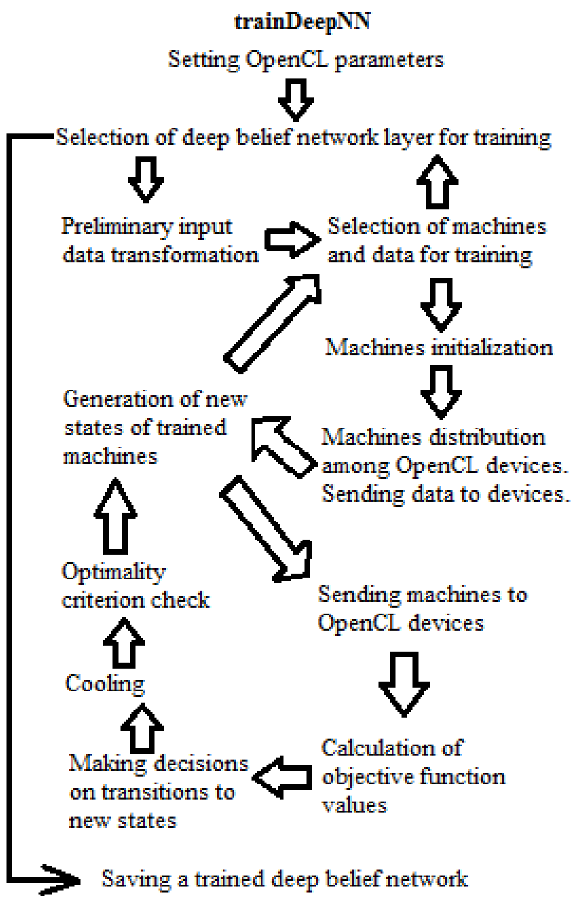


Figure 2. Function trainDeepNN.

pixel or less), values in the range of 8 and higher are considered adequate. For medium compression (1 bit per pixel), normal values are considered to be from 20 above.

Based on the results of the experiments, the following results were obtained (see Table 1):

Table I
DEEP BELIEF NETWORK TRAINING

Framework Efficiency	Compress Ratio (bit/pixel)		
	0,75	*0,1875*	*0,046875*
quality function (PSNR)	19	16,9	14,72
training time (h)	6	10	11

The results show the high efficiency of the deep belief neural network architecture and the training algorithm based on the annealing method. To configure a separate machine consisting of 1168 parameters in the first layer of the net-work, only 2000 images were needed, while for machines of subsequent layers consisting of 1104 parameters, only 4000 images were needed, which indicates a very high efficiency of the annealing method in training neural networks.

The time spent on training the network shows that, with the proper selection of the training algorithm parameters, the annealing method can have a high convergence rate.

The obtained results confirm the assumption that the annealing method can be used to train deep belief networks [15]. Considering that the STL-10 dataset is rather complicated, so in comparison with other results [16] [17], it can be argued that the developed original annealing method algorithm is quite efficient.

V. Conclusion

The paper provides a description of the framework for solving applied problems using a deep belief network. An original network architecture and a set of algorithms implementing training processes and problem solving are proposed.

The effectiveness of the framework functioning is demonstrated on the example of the problem of compression of colored images. The problem was solved on a computer with a multi-core processor and video card.

It is shown that the developed deep belief networks training algorithm lacks many of the disadvantages of gradient methods. Based on the obtained experimental results, we can conclude that the developed framework with appropriate revision (for a wider class of applied problems) has a certain potential.

The proposed framework technology focused on deep belief networks is easily integrated into the methodology of open semantic networks and can be used for further development of this technological area.

References

[1] V.V. Krasnoproshin, V.V. Matskevich Statistical approach to image compression based on a restricted Boltzmann machine // Proceedings of the 12-th International Conference "Computer Data Analysis and Modeling"- CDAM'2019, Minsk, 2019, -p.p. 207-213.

[2] Vadim Matskevich, Victor Krasnoproshin Annealing method in training restricted Boltzmann machine // Proceedings of the 14-th International Conference - PRIP'2019, Minsk, 2019, -p.p. 264-268.

[3] Aicher C., Foti N.J., Fox E.B Adaptively truncating backpropagation through time to control gradient bias [electronic resource]. – link: arxiv.org/abs/1905.07473 – Access date: 04.01.2020.

[4] Xavier Glorot, Yoshua Bengio Understanding the difficulty of training deep feedforward neural networks // Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR Vol.9, 2010, -p.p 249-256.

[5] Treadgold N.K., Gedeon T.D. Simulated annealing and weight decay in adaptive learning: the SARPROP algorithm // IEEE Transactions on Neural Networks vol. 9, Issue: 4 , Jul 1998.

[6] V.A. Golovko, A.A. Kroshenko Using deep belief neural networks to highlight semantically significant attributes // Proceedings of the 5-th International Conference - OSTIS'2015, Minsk, 2019, -p.p. 481-486.

[7] Locatelli M. Convergence properties of simulated annealing for continuous global optimization // Jornal of Applied Probability Vol. 33, Issue 4 December 1996, pp. 1127-1140.

[8] Hajek B. Cooling schedules for optimal annealing // mathematics of operations research vol.13, No 2, May 1988.

[9] Sanguthevar Rajasekaran On the Convergence Time of Simulated Annealing [electronic resource]. – link: repository.upenn.edu/cis_reports/356/ – Access date: 04.01.2020

[10] V.V. Krasnoproshin, V.V. Matskevich Effective Data Processing on Heterogeneous Computing Devices // Bulletin of Brest state technical university. Series - physics, math, computer science. 2018. Vol. 5 (113), -p.p 15-18.
[11] STL-10 dataset [electronic resource]. – link: academictorrents.com/ details/a799a2845ac29a66c07cf74e2a2838b6c5698a6a – Access date: 25.02.2018.
[12] STL-10 dataset description [electronic resource]. – link: stanford.edu/ acoates//stl10/ – Access date: 24.02.2018.
[13] CUDA toolkit [electronic resource]: – link: developer.nvidia.com/cuda-downloads – Access date: 23.02.2018.
[14] Dogancan Temel. Ghassan AlRegib Perceptual Image Quality Assessment through Spectral Analysis of Error Representations // Signal Processing: Image Communication, Vol. 70, 2019, -p.p 37-46, ISSN 0923-5965.
[15] L.M. Rasdi Rere, Mohamad Ivan Fanany, Aniati Murni Arymurthy Simulated Annealing Algorithm for Deep Learning // Procedia Computer Science Vol. 72, 2015 -p.p 137–144.
[16] Toderici G., Vincent D., Johnston N., Sung Jin Hwang, Minnen D., Shor J., Covell M. Full Resolution Image Compression with Recurrent Neural Networks [electronic resource]: – link: arxiv.org/abs/1608.05148 – Access date: 04.01.2020.
[17] Zhou X., Xu L., Liu S., Lin Y., Zhang L., Zhuo C. An Efficient Compressive Convolutional Network for Unified Object Detection and Image Compression // Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, 2019. -p.p 5949-5956.

Программная технология глубокого обучения доверительных нейронных сетей

Краснопрошин В.В., Мацкевич В.В.

В докладе предлагается описание структуры и состава фреймворка для решения прикладных задач с использованием глубоких доверительных сетей. Предложены оригинальная архитектура сети, ориентированная на параллельную обработку данных, набор алгоритмов реализующих процессы обучения на основе метода отжига и решения задач. Эффективность работы описанного фреймворка демонстрируется на примере решения задачи сжатия цветных изображений.

Ключевые слова: Фреймворк, метод отжига, глубокая доверительная сеть, параллельные вычисления, обучение, выборка.

Received 12.02.2020

Application of virtual reality technology based on ontology in education of aviation

Sergey Vlasov, Nikolay Borgest
Institute of Aviation Engineering
Samara National Research University
Samara, Russia
vlasov.ssau@mail.ru, borgest@yandex.ru

***Abstract*—Education system is continuously evolving and to make process of learning easier people may use AR and VR technologies. It has a lot of benefits and can be used either for learning in simulators via VR headset, or for designing something new. Vuforia Engine provides variety of opportunities for getting a new knowledge. To be able to get full experience while using VR and AR technologies, people should get knowledge of ontology.**

***Keywords*—Augmented reality, Virtual reality, 3D model, ontology.**

I. Introduction

In recent years, because of the formation of a powerful production base, the rapid development of science, technology and information technology and their implementation in almost all spheres of human life, new psychological and technical phenomena have emerged. They are called “virtual” or “imaginary” reality in the scientific literature. Thanks to the development and improvement of computers, programming languages, and the technical component, a number of special means of transmitting information to humans, as well as feedback, have been created, which made it possible to achieve a new quality of perception and experiences that people recognize as virtual realities.

Developing virtual environments remains a technical and time-consuming process. Usually, low-level programming languages are used to define the virtual world, the interaction of the user with virtual objects and the objects’ behaviours within the virtual environment. [1].

Virtual reality - this is a world created by technical means and transmitted to a person through his usual sensations for the perception of the material world: vision, hearing, smell and others. In addition, the term “virtual reality” means a reality that can exist both in a potential, possible state, and in an actual, existing state [2].

II. IMPLEMENTATION OF VR AND AR FOR EDUCATION

In practice, virtual reality is implemented in the fields of engineering and design, mining, military technology, construction, marketing and advertising, the entertainment industry, and various simulators and simulators are being created. In virtual reality systems, imitation of tactile sensations, imaging with a stereo effect, multichannel acoustic systems, and control by a certain movement of a person in space are widely used. So, a special helmet was created with displays for each eye, with headphones and sensors that provide information about the position of the head, a power vest (creates effort on the muscles of the trunk and arms of a person, simulating the illusion of interacting with objects in virtual space), gloves and boots (equipped special sensors that provide information about the movement of arms, legs and even individual fingers). Having put on such a “suit”, the observer enters the virtual world. At the same time, you can turn your head, look around, walk, touch objects with your hand or foot, lift them and feel their weight and temperature. That is, a virtual world created by a computer is capable of deceiving the senses of the observer. Another way to immerse yourself completely in the virtual world is to use a special virtual room in which the floor, walls and ceiling are equipped with screens on which images are projected. Motion and sounds are simulated (for example, a car, plane, train, or spaceship). All this is important for creating special simulators for pilots, astronauts, car drivers. Aggregates acting on the human vestibular apparatus have also been created. An example is rotating cabins for training astronauts. Moreover, data can be transmitted directly to nerve endings, and even directly to the brain through brain interfaces. This approach allows you to increase the effectiveness of "immersion" of a person in virtual space. However, such a technology is too expensive for everyday use and does not achieve the quality of data transfer acceptable for virtual reality transmission.

In addition, virtual reality should not be considered as a augmented reality [3]. Their fundamental difference is that the virtual constructs a new artificial world, while the augmented one only introduces individual artificial elements into the perception of the real world [4].

Augmented Reality (AR) allows you to enrich the world with the latest technology, creating a unique combined interactive experience. Although augmented reality is still rarely used in education, more and more teachers, researchers and developers are beginning to move towards more interactive teaching methods. Many of these techniques grow into really interesting and creative projects. Augmented reality, or AR is undoubtedly a huge breakthrough in the way of presenting educational material and in the assimilation of information. AR allows you to enrich the world with the latest technology, creating a unique combined interactive experience. Virtual images that students can see right in the lecture hall make the teaching material vivid, memorable and easier to understand. The effectiveness of its use is confirmed by various tests and experiments that show excellent results. For example, a series of experiments were conducted in which teacher showed to one group of children visual material with AR during the lessons, and showed the usual group posters and diagrams to the second group. It was found that in the group where augmented reality was used, the percentage of information assimilation by children approached 90 percent, the level of discipline increased and it was possible to maintain the attention of about 95 percent of the audience, while in the group with two-dimensional benefits, all indicators were half or three times less. One of the reasons for this influence is that AR creates the effect of presence, very clearly reflects the connection between the real and virtual worlds, which psychologically attracts people and activates their attention and sensitivity to the information component.

III. THE IMPORTANCE OF ONTOLOGY BASED CONCEPT

Virtual Humans are virtual entities with a rich set of functionalities and potential, present in a VE. One of the main instruments used to lay-down the foundations of a knowledge-based system are ontologies. Ontology defines a common vocabulary for domain-users (researchers or experts in a particular area) who need to share information in a particular domain. It includes machine interpretable definitions of basic concepts in the domain and relations among them. The semantics-based representation can be enhanced by means of knowledge-management techniques and tools. One of the main tools used to lay-down the foundations of a knowledge-based system is therefore an ontology. A first one focuses on the adaptive multimodal interfaces system and a second one formalizes the knowledge related to the creation of virtual humans and serves as a basis for a general ontology for Virtual Environments. Many consider W3C'sWeb Ontology Language (OWL) the prospective standard for creating ontologies on the Semantic Web. OWL has three species: OWL Lite, OWL DL and OWL Full, in ascending order according to expressiveness. We can divide the use of ontologies in the domain of Virtual Environments into three uses; the first use: Ontologies in the context of Virtual Environments, the second use: Ontology for interactive Virtual Environments, the third use: Ontology for Virtual Human. The conceptualization step is to identify a body in the knowledge domain and to clarify the conceptual nature (concepts, relations, properties and relations of concepts, rules, constraints, etc.) The ontologization step consists in modeling in a formal language the domain properties, the objective is to obtain a model in which almost all the ambiguities inherent in natural language are lifted. Without ontology based concept people won‘t be able to get full experience of VR and AR In Samara Students study how an airplane works, develop an ontology of a given subject area [5]–[7], and build 3D models of airplane assemblies. The work is aimed at the application of these models. There is a class containing both decommissioned aircraft units and operating earlier aircrafts, figure 1. Such as the Yak-26, MiG-21, Su-15 and others. It helps students to get a better understanding about aircrafts. Students also create drawings of both units and the entire aircraft as a whole, figure 2. Aircraft drawings are taken as a marker for augmented reality.

Figure 1. Aircraft class

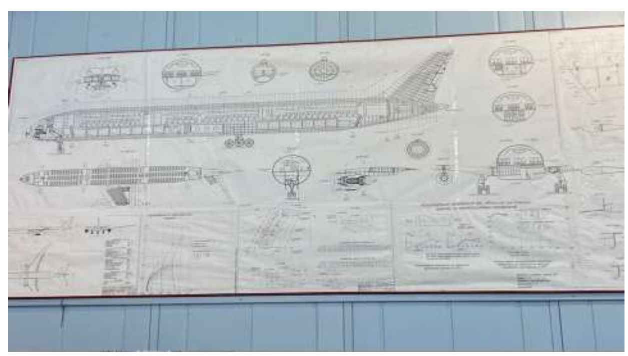


Figure 2. Aircraft drawings made by students

IV. PTC VUFORIA

There are many platforms on the market that will be useful in developing projects useful in training, and one of them is PTC Vuforia.

PTC Vuforia is an augmented reality platform and software development kit (SDK) for mobile devices developed by Qualcomm. The Vuforia project was purchased from Qualcomm in November 2015 and has since been wholly owned by PTC Inc. Vuforia uses computer vision technology, as well as tracking flat images and simple volumetric real objects (for example, cubic) in real time. From version 2.5, Vuforia recognizes text, and from 2.6 it has the ability to recognize cylindrical markers. More than half a million developers have chosen advanced computer vision technology, incredible performance and cross-platform Vuforia Engine. This made the Vuforia project the most popular AR on the planet. Key Benefits of PTC Vuforia are:

A. Unrivaled Accuracy

The advanced computer vision technology in the Vuforia Engine provides incredible AR accuracy in a variety of environments.

B. Creative opportunities

The robust technology offered by Vuforia Engine allows developers to freely create proprietary AR applications for new or existing applications.

C. Maximum reach - multi-platform

Vuforia supports AR devices such as smartphones, tablets and headsets on leading platforms to reach the largest audience.

D. Extended Vision Concept

Vuforia Engine offers dynamic object recognition and can define images, 3D models and environments, providing development flexibility.

Augmented reality is one of the fastest growing technology segments in the corporate environment. However, its distribution is slowed by a lack of tools with which content authors could use existing 3D resources [8]. Vuforia Studio solves this problem by allowing you to develop applications for the creation, operation and maintenance of smart networking products [9]. Vuforia Studio provides all the components necessary for the implementation of augmented reality. To work with them, neither deep knowledge in programming, nor expert experience in augmented reality technologies is required [10].

Using augmented reality technology, a special educational application is being developed at the Department of Aircraft Design and Design at Samara University. The application loaded 3D models of aircraft, units of which are available at the department. Using this application, the user can point the camera at the drawing of the aircraft and receive a parameterized 3D model of the aircraft on his smartphone screen. The drawing serves as a kind of marker by which the program determines the model that needs to be shown to the user. Demonstration of the application is shown in Figure 1.

In the future, it is planned to improve and refine the application, by adding new functions to it, such as: hiding the skin of the aircraft and demonstrating its power elements, breaking it into separate units, into separate 3D objects.

Figure 3. Augmented reality, 3D-model of the Tu-154

CONCLUSION

Being able to use modern technologies correctly can significantly facilitate and diversify the education process. It also allows you to expand knowledge of both the simplest objects and planes. Learning VR and AG technologies will help to achieve significant success in all areas including education, engineering, aircraft etc.

REFERENCES

[1] Lode Vanacken Chris Raymaekers Karin Coninx. Introducing Semantic Information during Conceptual Modelling of Interaction for Virtual Environments Hasselt University, 2012. pp. 17–24.

[2] Sherman, William R.; Alan B. Craig.: Understanding Virtual Reality: Interface, Application,and Design, Morgan Kaufmann, Cambridge, MA, 2018.

[3] Francisco Grimaldo, Fernando Barber, and Miguel Lozano. Augmented Reality in Education: Current Technologies and the Potential for Education An ontology-based approach for Computer Science Department, 2013. University of Valencia

[4] Mezati Messaoud , Foudil Cherif , Cédric Sanza and Véronique Gaildrat An ontology for semantic modelling of virtual world Department of Computer Science, University of Ouargla, Algeria 2018. pp. 65-74.

[5] Walter TERKAJ, Giovanni Paolo VIGANO‘ Semantic GIOVE-VF: an Ontology-based Virtual Factory Tool, Institute of Industrial Technologies and Automation (ITIA-CNR), Milan, Italy

[6] A. Van Dam, A. Forsberg, D. Laidlaw, J. LaViola, and R. Simpson. Immersive. VR for scientific visualization: A progress report. IEEE Computer Graphics and Applications, 20(6): 26-52, 2000

[7] Borgest N.M., Vlasov S.A., Gromov Al.A., Gromov An.A., Korovin M.D., Shustova D.V., 2015. Robot-designer: on the road to reality. Ontology of designing. – No.4(18). – pp. 429-449 DOI:10.18287/2223-9537-2015-5-4-429-449

[8] Borgest N.M., Korovin M.D. Formalization of design procedures in the engineer’s educational strategy. Proceedings of the 2016 conference on Information Technologies in Science, Management, Social Sphere and Medicine Editors: Olga Berestneva, Alexei Tikhomirov, Andrey Trufanov ISBN (on-line): 978-94-6252-196-4 part of the series ACSR, ISSN 2352-538X, volume51. itsmssm-16, part of series: Advances in Computer Science Research. ISBN: 978-94-6252-196-4, ISSN: 2352-538X, pp.524-527.

[9] Jakub Flotyński, 2012 Semantic 4-dimensional Modeling of VR Content in a Heterogeneous Collaborative Environment. - 2015. - No. 10. - Pp. 18-21: ill- ISSN 1560-4640

[10] Ye J, Badiyani S, Raja V, Schlegel T. Applications of virtual reality in product design evaluation. In Human-Computer Interaction. HCI Applications and Services, 2007; pp. 1190-1199. Springer Berlin Heidelberg.

Применение технологии виртуальной реальности на основе онтологии в обучении авиационных студентов

Власов С.А. Боргест Н.М.

В последние годы, в образовании начали часто использовать технологии виртуальной и дополненной реальности. Это позволяет обогащать мир новейшими технологиями, порождая уникальный комбинированный интерактивный опыт. Онтологический подход определения сущности виртуальности связан с рассмотрением последней как универсального естественного феномена, одного из фундаментальных свойств бытия, бытия вообще, а не только социального. Подобный подход позволяет сделать онтологически значимые выводы, определив создаваемую в киберпространстве и мире Интернет виртуальность как один из онтологических подуровней «тонкой структуры» виртуальности вообще.

Используя технологию дополненной реальности, на факультете проектирования и проектирования самолетов Самарского университета разрабатывается специальное учебное приложение. В приложение загружены 3D модели самолетов, части которых доступны на кафедре. Используя это приложение, пользователь может направить камеру на чертеж самолета и получить параметризованную 3D-модель самолета на экране своего смартфона. Чертеж служит своего рода маркером, с помощью которого программа определяет модель, которую необходимо показать пользователю.

Received 03.02.2020

Security patterns based approach to automatically select mitigations in ontology-driven threat modelling

Andrei Brazhuk
Yanka Kupala State University of Grodno
Grodno, Belarus
brazhuk@grsu.by

***Abstract*—Common approach of the threat modelling includes an analysis of computer system architecture on early stages of development process and creation of threat model. Data Flow Diagrams (DFD) are often used to represent the system organization. The main challenge with threat modelling is that there are no formal approaches to describe the computer system architecture and structured knowledge sources of threats and countermeasures.**

To overcome these restrictions we have created ontology-driven threat modelling (OdTM) framework based on base threat model, used to develop various domain-specific threat models. Each domain-specific threat model contains a set of typical components of some architectural domain, threats and countermeasures. A system architect describes its computer system with DFD diagram; then automatic reasoning procedures are used to semantically interpret the diagram and figure out relevant threats and countermeasures for the system.

We approach a conception of context security patterns as countermeasures. Context security pattern contains a precise description of security problem and its solution. Also it has criteria that allow to automatically map it to system component. We propose three ways to integrate context security patterns with domain-specific threat models: with data flow templates, through association with threats, and the use of labels.

All the models, discussed in this work, can be implemented as OWL (Web Ontology Language) ontologies with Description logics (DL) as a mathematical background.

***Keywords*—software security, knowledge management, threat modelling, OWL, DFD**

I. Introduction

Threat modelling is a process of identification of security threats and their countermeasures in order to increase security level of computer system. Common approach of the threat modelling includes two stages. Firstly, an analysis of computer system organization (i.e. its architecture) happens on the early stages of its development process (requirements, design, redesign). Secondly, they build the threat model that represents all security aspects of the system.

Different informal graphical representations of system architecture are used during this process, more often Data Flow Diagrams (DFD) [1]. DFD consists of the objects (stencils) and data flows between these stencils; also a special type of entities exists to group objects and define trust boundaries between them (Figure 1 shows an example of DFD from cloud computing field). Using such diagrams, development team can build an informal threat model of computer system, i.e. figure out possible threats to the system and their countermeasures, through discussions, making notes, and evaluations.

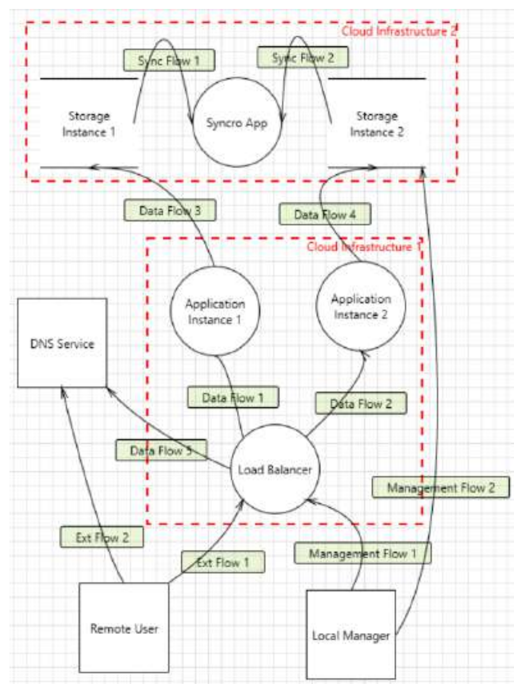


Figure 1. Example of DFD

The main challenge with the threat modelling is that it is too hard to employ formalization and automation there. There is a lack both of formal approaches to describe the computer system architecture, and structured knowledge sources of threats and their countermeasures.

In order to bring a formal approach to this filed [2], we have created *ontology-driven threat modelling (OdTM) framework*. The OdTM framework includes a common approach of the architectural security analysis, method of semantic interpretation of DFD, and automatic reasoning procedures of relevant threats and countermeasures. Our approach is based on the *base threat model* that enables creation of various *domain-specific threat models*. Each domain-specific threat model holds a set of typical components of some architectural domain, threats and countermeasures (security patterns) associated with these components. System architect can describe computer system in terms of a domain-specific model with diagram(s). Then the automatic reasoning procedures can be used to build threat model of the system.

All the models, proposed in this work, can be implemented as OWL (Web Ontology Language) ontologies. OWL has description logics (DL) as a mathematical background. The DL means are able to describe concepts of a domain and relations between them in very formal way and apply automatic reasoning features with relatively low computational complexity.

Another challenge, discussed in this work, refers to employing an approach to choose the mitigations based on security patterns. Security patterns are known as descriptions of security problems that appear in specific contexts and present well proven solution for them [3]. They are created by security experts and represent best security practices for inexpert computer system architects.

We approach a conception of *context security patterns*. Context security pattern is a security pattern that has been placed into a context, i.e. contains more precise interpretation, directly applicable for a system component, rather then generic descriptions of problem and solution. Also, it has criteria that allow to automatically map it to a system component. We propose three ways to integrate context security patterns with domain-specific threat models: with data flow templates, association with threats, and the use of labels.

Also, a problem of formalization of domain specific knowledge is discussed in this work.

II. Ontology-driven threat modelling framework

The Ontology-driven threat modelling (OdTM) framework is aimed to employ formalization and automation to architectural security analysis of computer systems. It consists of the ontology-driven models and methods that enable automatic analysis of DFD diagrams and building particular threat models. Figure 2 shows structure of the OdTM framework.

The base threat model, implemented as OWL ontology, enables the automatic reasoning features and contains basic concepts and individuals, representing components of the diagrams, threats, countermeasures, and their properties.

To involve the threat modelling of a particular type (domain) of computer systems (e.g. Cloud computing,

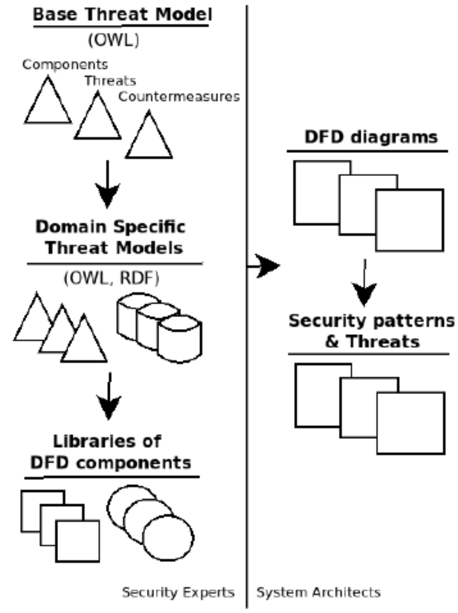


Figure 2. Structure of OdTM framework.

Fog/Edge computing, SaaS, IaaS, or Software Defined Networks), it requires the following steps:

- *Building domain-specific threat model*. To build the model, security experts should extend the base model with specific components, threats and countermeasures, related to the domain. The domain-specific threat model can be considered as a meta model, which depicts the security aspects of this type of computer systems. They are represented as OWL ontologies and can be connected to various external linked data sources (OWL ontologies and RDF data sets).
- *Building domain-specific library of DFD components*. Background procedures automatically extract the component hierarchy from the domain-specific threat model and create the library of DFD stencils, which can be used to draw DFD diagrams.

To create a threat model of a particular computer system, it requires the following steps:

- *Depiction of an architecture of the system as DFD*. A system architect draws a structure of its computer system as the DFD diagram (or the set of diagrams), using the stencils of DFD component library.
- *Semantic interpretation of the DFD*. The background procedures automatically interpret the diagram as a set of semantic instances (components, data flow, boundaries and relations between them) and combine this set with the domain-specific threat model.
- *Automatic reasoning of relevant threats and countermeasures*. Automatic reasoning procedures infer relevant threats and countermeasures from the

semantic interpretation and domain-specific threat model. This allows the background procedures to build lists of the threats and countermeasures for the system.

III. Semantic interpretation of data flow diagrams

Common ontology description with the DL means uses separation of axioms to the TBox (concepts and properties) and ABox (individuals and their relations) parts. Also, it is supposed that automatic reasoning procedures exist, which allow to get extra facts from an ontology (inferred axioms).

The OdTM base threat model, implemented as OWL ontology, enables semantic interpretation of the diagrams and automatic reasoning of threats and countermeasures. To model a DFD diagram we use a set of concepts (classes) and their properties, described by Figure 3 as DL axioms.

```
01   Stencil ⊆ ⊤
02   Target ⊆ Stencil
03   Process ⊆ Target
04   ExternalInteractor ⊆ Target
05   DataStore ⊆ Target
06   DataFlow ⊆ Stencil
07   TrustBoundary ⊆ Stencil
08   TrustLineBoundary ⊆ TrustBoundary
09   TrustBorderBoundary ⊆ TrustBoundary
10   ∃ hasSource.⊤
11   ∃ isSourceOf.⊤
12   isSourceOf ≡ hasSource⁻
13   ∃ hasTarget.⊤
14   ∃ isTargetOf.⊤
15   isTargetOf ≡ hasTarget⁻
16   ∃ crosses.⊤
17   ∃ divides.⊤
18   divides ≡ crosses⁻
19   ∃ includes.⊤
20   ∃ isIncluded.⊤
21   isIncluded ≡ includes⁻
```

Figure 3. Semantic interpretation of DFD diagram (a part of TBox).

Axioms (Ax.) 01-09 in Figure 3 model the hierarchy of the DFD base stencils. Three main concepts are derived from the “Stencil” concept: “Targets”, “Trust-Boundaries”, and “DataFlows”.

An instance of the "DataFlow" concept represents a directional flow from a source "Target" instance to a target "Target" instance. To model this, a data flow should have two properties: “hasSource” (Ax. 10) and “hasTarget” (Ax. 13); for both of them a range is supposed to be Target. The properties “isSourceOf” (Ax. 11) and “isTargetOf” (Ax. 14) are inverse to previous two ones (Ax. 12, 15) and allow to infer that a target is a starting edge or ending edge of some data flow.

A data flow can cross some instance of the “TrustLineBoundary”; to tell this, the “crosses” property is used (Ax. 16). Also, a line boundary might divide some data flow; to describe this, the “divides” property is used (Ax. 17). The last two properties are inverse (Ax. 18).

A target might be included into some “TrustBorderBoundary” instance; to model this, the “isIncluded” property is used (Ax. 20). Also, a TrustBorderBoundary instance contains, or “includes” (Ax. 19) some targets. The “include” and “isIncluded” properties are reverse (Ax. 21).

To semantically model a diagram that describes an architecture of a computer system, we should apply the ABox axioms to the ontology. This includes instances of the base concepts (or derived of them) and their properties: “Targets”, “DataFlows” with “hasSource” and “hasTarget”, “TrustLineBoundaries “with “crosses”, “TrustBorderBoundaries” with “includes”.

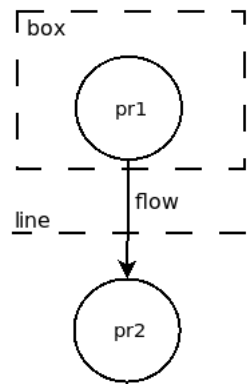


Figure 4. A simple DFD diagram.

Figure 4 depicts a simple example of DFD, and below is shown a possible semantic interpretation of the diagram:

$Process(pr1)$
$Process(pr2)$
$DataFlow(flow)$
$hasSource(flow, pr1)$
$hasTarget(flow, pr2)$
$TrustLineBoundary(line)$
$crosses(flow, line)$
$TrustBorderBoundary(box)$
$includes(box, pr1)$

Note, the automatic reasoning procedures would be able to infer additional facts from the ABox sets like this. We do not need to tell in this example, that “pr1” is a source of “flow”, “pr2” is its target, “line” divides “flow”, “pr1” is included into “box”. These facts would be inferred by the automatic reasoning procedures (see Ax. 12, 15, 18, 21).

IV. Automatic selection of security patterns

In modern computer systems data flows are origins of security issues, because most of the attacks are remote and sourced from the local and remote networks. Usually threats are applied to computer system by data flows. Also, reasons for adding a countermeasure (security pattern) to particular architecture depend on the presence of a data flow.

To model threats and countermeasures we use a set of concepts and their properties, described by Figure 5.

```
22  Threat ⊆ T
23  Countermeasure ⊆ T
24  ContextSecurityPattern ⊆ Countermeasure
25  ∃ affects.T
26  ∃ isAffectedBy.T
27  isAffectedBy ≡ affects⁻
28  ∃ protects.T
29  ∃ isProtectedBy.T
30  isProtectedBy ≡ protects⁻
31  ∃ mitigates.T
32  ∃ isMitigatedBy.T
33  isMitigatedBy ≡ mitigates⁻
34  ∃ labelsSO.T
35  ∃ labelsSTRIDE.T
```

Figure 5. Countermeasures and threats (a part of TBox).

An instance of the "Threat" concept (Ax. 22) "affects" some data flow. Also the inverse property called "isAffectedBy" is used (Ax. 26, 27).

An instance of the "Countermeasure" concept "protects" (Ax. 28) some data flow, and a data flow "isProtected by a countermeasure (Ax. 29, 30). Also, countermeasure "mitigates" (Ax. 31) some threat, and a threat "isMitigated" by some countermeasure (Ax. 32, 33).

Security patterns in the model are extended to the "ContextSecurityPattern" concepts. They are derived concepts of "Countermeasures" (Ax. 24). The base threat model has three ways to enable the automatic reasoning of context security patterns, as well as other kinds of mitigations, through countermeasures: A) data flow templates, B) association with threats, C) the use of labels.

A) *The first option is the use of data flow templates*. A flow template should be defined as a concept with the "hasSource", "hasTarget", "crosses" (and other) properties like:

$$Template1 \equiv DataFlow$$
$$\cap \exists hasSource.Process$$
$$\cap \exists hasTarget.Process$$
$$\cap \exists crosses.TrustLineBoundary$$

To enable automatic reasoning, it requires to create a instance of context security pattern and associate it with a data flow template, like:

$$ContextSecurityPattern(pattern1)$$
$$Template1 \subseteq \exists isProtectedBy.\{pattern1\}$$

Using last three axioms, the "flow" instance (from Figure 4) would be recognized by the automatic reasoning procedures as an instance of the "Template1" concept, so "flow" would be protected by "pattern1", and "pattern1" would protect "flow".

By creation flow templates and descriptions of patterns, like shown above, it is possible to form a model of context security patterns. The same way it can be possible to create a model of threats by mapping the Threat instances to data flow patterns with the "isAffected" property.

B) *The next way to employ the context security patterns is to map them to threats with the "mitigates" property*, like:

$$Threat(threat2)$$
$$Template1 \subseteq \exists isAffectedBy.\{threat2\}$$
$$ContextSecurityPattern(pattern2)$$
$$mitigates(pattern2, threat2)$$

C) *For more precise classification of threats and countermeasures a set of security objectives (SO) and the STRIDE model can be used*. STRIDE stands from Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. A list of the security objectives used here includes: Confidentiality, Integrity, and Availability (the CIA triad) and extra objectives like Authentication, Non-Repudiation, and Authorization.

Using the "labelsSO" and "labelsSTRIDE" properties (Ax. 34-35), it is possible to label threats, countermeasures and context security patterns:

$$labelsSO(pattern2, SO_Availability)$$
$$labelsSTRIDE(threat2, STRIDE_Denial_of_Service)$$

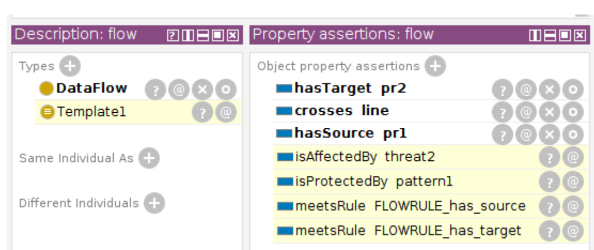


Figure 6. Automatic reasoning example in Protege.

It would be easy to implement the discussed above model (Figures 3 and 5) as OWL ontology and, using a reasoner (like HermiT, Fact++, or Pellet), check the feasibility of proposed ideas.

Our implementation of the OdTM base threat model has freely been published with the GitHub service (https://github.com/nets4geeks/OdTM) as the *OdTM-BaseThreatModel.owl* file.

Figure 6 shows the properties of the "flow" instance from Figure 4 after the automatic reasoning performed with Protege.

V. BUILDING DOMAIN-SPECIFIC THREAT MODELS

Figure 7 shows the process of the domain specific knowledge formalization, used to build domain-specific threat models.

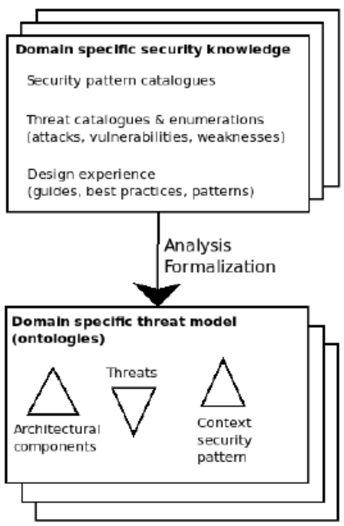


Figure 7. Formalization of domain specific knowledge.

Steps to build a domain-specific threat model are following:

- *Create a sub model of architectural components*. That includes findings all the items (components, relations, boundaries) that a computer system of particular domain can have. This enumeration should be used to extend the component hierarchy of the base threat model with domain specific concepts.
- *Create a sub model of threats*. This includes an enumeration creation of possible domain specific threats, giving proper definitions for the threats, mapping them to data flows and patterns. To make easier the threat analysis, it is possible to build additional semantic models. In particular, we have built the OWL ontology [4] that is based on the attack pattern (CAPEC - Common Attack Pattern Enumeration and Classification) and weakness (CWE - Common Weakness Enumeration) concepts, and it is able to classify security concepts according given criteria. To process raw sources of vulnerability information, like NVD (National Vulnerability Database), it strongly requires implying various NLP (natural language processing) methods [5].
- *Create a sub model of context security patterns*. This includes an enumeration creation of context security patterns and mapping them to data flows and threats. However, well-known security patterns are textual descriptions of security problems, made in some format in technology-independent way, e.g. the POSA (Pattern Oriented Software Architecture) templates. Security experts are able to understand generic descriptions of security patterns and employ them as design decisions to specific computer systems. But applying them as automatically inferred solutions (i.e. putting in a context) requires some extra adaptation steps. It can be argued about two hundred common security patterns exist [6], and there are hardly any sources able to create patterns automatically. So, manual and semi automatic methods of creation of context security patterns are preferred for this kind of job.
- *Label the threats and context security patterns to the security objectives and STRIDE items.*

VI. RELATED WORK

Works [6] and [7] have presented an ontological approach to manage security patterns. The ontologies facilitates mapping between the context aspects and security patterns themselves, and therefore enables automatic pattern selection during the secure system development process.

The most powerful effort for the threat modelling automation has made by Microsoft with the Threat Modelling (TM) software. Microsoft TM consists of a drag-and-drop DFD editor, simple rule-based reasoner, report subsystem, and built-in threat template editor. Microsoft uses a simple rule language to describe associations threats with data flows. The XML format is used to save threat templates.

Some works have used the Microsoft approach for research and creation of security threat models based on DFD [8], [9], [10]. However, there are some issues with the Microsoft implementation. The Microsoft tool only operates with two level hierarchy of objects (stencils and derived stencils) and threats (categories and threat types), however for description of complex computer systems and their threats it usually requires more layers of abstraction. Also there is a lack of full-featured countermeasure hierarchy, which would allow a user to choose a countermeasure to a threat from a relevant list. Our work has gone to find a way to overcome these restrictions with creation of semantic models with well-formed hierarchies of components, threats and countermeasures.

Work [11] has proposed an approach to architectural risk analysis that leverages the threat modelling by introduction of extended Data Flow Diagrams (EDFD), declaring a few improvements to DFD (their knowledge base uses a domain-specific rule language, based on a graph query language), and creation of a visual EDFD viewer. Work [12] has proposed very similar findings to what our research has offered. They have researched a challenge of automatic correction of security issues in declarative deployment models of cloud services based on the ontological approach and security patterns. It can be argued that our OdTM approach conceptually satisfies their topology-based deployment meta model. However, their implementation is directly based on First-order logic (FOL) and the low-level logical programming (Prolog).

An advantage of our approach to compare with other works [11] and [12] is the use of Description logics (DL) through OWL and the automatic reasoning features as an implementation. This allows to employ an object-oriented approach to the knowledge management system design, i.e. provide better representation for users, stricter formalization, and easier ways to implement. Also it is possible to apply (if necessary) various high-level means, like the SWRL (Semantic Web Rule Language) rules, the SPARQL (SPARQL Protocol and RDF Query Language) queries. And an implementation based on OWL enables integration with linked open data (LOD) sources.

VII. Conclusions

The OdTM framework is based on domain-specific threat models with appropriate libraries of DFD components. Each domain-specific threat model is a meta model of threats of particular computer system domain, represented as the DL compatible formalization. It contains axioms that can be considered as a TBox (mixed with the instances of threats and countermeasures). To model a DFD diagram that describes an architecture of a particular computer system, it requires to interpret the diagram as a set of instances (ABox). Using these ABox and mixed TBox, the automatic reasoning procedures can infer relevant threats and countermeasures for the computer system.

Informally, to build domain-specific threat it is necessary to extend the base threat model by creation of a hierarchy of domain typical components, association of threats and context security patterns to data flow templates, mapping threats and patterns to each other, and labeling them by the security objectives and STRIDE items. However, formalization of domain specific security knowledge should be considered as its transformation to sub models of architectural components, threats and context security patterns. In future research we are going to learn these processes in order to create methods and models that enable (semi) automatic building of domain-specific threat models.

A special field of interest there would be transformation of general security patterns to context security patterns. A system architect thinks of a computer system in domain specific terminology and expects that proposed solutions would be described the same way. Also, criteria to map or not a pattern to a system design are quite important.

References

[1] M. Abi-Antoun, D. Wang, P.Torr, "Checking threat modeling data flow diagrams for implementation conformance and security," Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering. ACM, pp. 393-396, 2007.

[2] E.V. Olizarovich, A.I. Brazhuk "Kontseptual'nye osnovy analiza modelei informatsionnoi bezopasnosti oblachnykh sistem klassa «infrastruktura kak usluga»" [Conceptual framework of analysis of information security models of cloud systems of the class «Infrastructure as a Service»], Doklady BGUIR, 2019. vol. 6(124), pp. 12-20.

[3] M. Schumacher, et al. Security Patterns: Integrating security and systems engineering. John Wiley and Sons, 2013.

[4] A. Brazhuk "Semantic model of attacks and vulnerabilities based on CAPEC and CWE dictionaries," International Journal of Open Information Technologies vol. 7, no. 3, pp. 38-41, 2019.

[5] A. Brazhuk "Building annotated semantic model of software products towards integration of DBpedia with NVD vulnerability dataset." IJOIT, vol. 7, no. 7, pp. 35-41, 2019.

[6] A.P. Vale, E. B. Fernández, "An Ontology for Security Patterns". In 2019 38th International Conference of the Chilean Computer Science Society (SCCC), pp. 1-8, 2019.

[7] H. Guan, H. Yang, J Wang, "An ontology-based approach to security pattern selection." International Journal of Automation and Computing 13.2, pp. 168-182, 2016

[8] M. Tasch, et al. "Security analysis of security applications for software defined networks." Proceedings of the AINTEC 2014 on Asian Internet Engineering Conference. ACM, 2014.

[9] M. Abomhara, M. Gerdes, and G. M. Køien. "A stride-based threat model for telehealth systems." Norsk informasjonssikkerhetskonferanse (NISK) 8.1, p. 82-96, 2015.

[10] L. Sion, et al. "Solution-aware data flow diagrams for security threat modeling." Proceedings of the 33rd Annual ACM Symposium on Applied Computing. ACM, 2018.

[11] B.J. Berger, K. Sohr, R. Koschke. "Automatically extracting threats from extended data flow diagrams." International Symposium on Engineering Secure Software and Systems. Springer, Cham, p. 56-71, 2016.

[12] K.Saatkamp, et al. "An Approach to Determine and Apply Solutions to Solve Detected Problems in Restructured Deployment Models using First-order Logic." SICS Software-Intensive Cyber-Physical Systems, vol, 34, no. 2-3, pp.85-97, 2019.

Подход на основе шаблонов безопасности для определения контрмер в онтологическом моделировании угроз

Бражук А.И.

Описан подход к моделированию угроз компьютерных систем на основе предметно-ориентированных моделей угроз, который позволяет автоматически определять угрозы и контрмеры по графическому представлению структуры системы в виде диаграми потоков данных. Описана базовая онтологическая модель угроз. Предложена концепция контекстных шаблонов безопасности для определения контрмер. Для реализации предлагается использовать язык онтологий OWL и функции автоматического логического вывода.

Received 26.11.19

Numerical conditioning of neural network components in time series forecasting

Stanislav Sholtanyuk
the Department of Computer Applications and Systems
Faculty of Applied Mathematics and Computer Science
Belarusian State University
Minsk, Belarus
SSholtanyuk@bsu.by

***Abstract*—In this paper, an issue of stability of fully connected neural network (perceptron), forecasting time series with sliding window method, is considered. A task of finding conditions, providing best results on accuracy and stability, was set. To accomplish it, training of the neural network with various sets of hyperparameters, such as amount of neurons on input and hidden layers, as well as activation function, has been performed. Training was performed on modeled data, which have been perturbed with unbiased normally distributed random variables with various variance. Such training results as number of conditioning of weight matrix between input and hidden layers, number of conditioning of activation functions, as well as statistics on retrospective forecasting absolut error, have been shown. An approach for estimation of condition number for some activation functions has been proposed. Impact of activation function on accuracy and stability of forecasting perceptron has been shown.**

***Keywords*—time series forecasting, neural network, perceptron, numerical conditioning, activation function**

I. Introduction

Artificial neural networks are widely used in various tasks, for example, pattern classification, speech recognition, decision making. Neural networks have also proved themselves as powerful tool for time series forecasting [1], [2]. There are plenty implementations and libraries on various programming languages, which can build neural networks with given architecture.

One of the modern researchers' focus is on stability and robustness of neural networks. There are different approaches for understanding and considering of neural networks stability. For example, dynamical stability (low sensitivity of the result to perturbations of input data) and structural stability (insignificant error of output in case of changes in numerical characteristics of the neural network itself) are distinguished [3]. The most vivid examples of neural networks instability appear in pattern recognition with deep convolutional networks. When training on certain dataset, these networks become unable to classificate objects on noisy images, which are indistinguishable from a human's point of view, but are completely different from a point of view of neural nets [4], [5]. Techniques, using such vulnerabilities of intelligence systems, called adversarial attacks. Connection between existence of adversarial examples and conditioning of weight matrices in neural networks has been revealed and systemically considered in [6]. In [7], it's also argued that stability of neural network training is an analogue of identification whether specified matrix is well- or ill-conditioned, besides, a measure of stability of neural networks training has been proposed for various applications and tasks in biometrics.

The aim of this paper is revealing the factors, which have impact on dynamical stability of neural networks in forecasting time series. For this, an issue of conditioning of neural nets and their components, calculation of condition numbers and affecting these indicators on neural network performance have been considered. Comparative analysis of neural networks with different hyperparameters sets has been performed as well.

II. Architecture of the forecasting perceptron and its training

The object of study of this article is three-layered perceptron with the following architecture:

- input layer contains arbitrary amount of neurons p. p sequential values of standardized time series are given to these p neurons;
- a hidden layer with arbitrary amount of neurons n;
- output layer contains the only neuron, where a forecasted value for the next, $(p+1)$th value, is received;
- input and hidden layers also contain bias neurons;
- the neural network is fully connected, i.e. all neurons on every layer connected with all neurons on the next layer.

Forecasting perceptron works with standardized time series, which is calculated by the following formula:

$$\tilde{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}},\ i = \overline{1, N}, \tag{1}$$

where x_i is i-th value of the initial time series, $N \in \mathbb{N}$ – length of the time series, $x_{\min}$ and $x_{\max}$ – minimal and maximal values of the time series respectively. Values of the standardized time series $\{\tilde{x}_i\}_{i=1}^{N}$ lie between 0 and

1. To obtain final result of forecasting, the output value should be processed with inverse transformation:

$$y = \tilde{y}(x_{\max} - x_{\min}) + x_{\min}. \quad (2)$$

On every neuron of the hidden layer, an activation function is acting. In this article the following activation functions are considered:

- rectifier linear unit (ReLU):

$$\mathrm{ReLU}(x) = \max(0, x); \quad (3)$$

- leaky rectifier linear unit (LeakyReLU, or LReLU):

$$\mathrm{LeakyReLU}(x) = \begin{cases} \alpha x, x < 0, \\ x, x \geq 0, \end{cases} \quad (4)$$

  where $\alpha << 1$ - parameter, preventing from vanishing gradient problem [8];
- softplus, which can also decrease the affect of vanishing gradient problem [9]:

$$\mathrm{soltplus}(x) = \ln(1 + e^x). \quad (5)$$

The perceptron is trained on the dataset, being obtained with sliding window method, i.e. vectors $\vec{x}^j = (\tilde{x}_j, \tilde{x}_{j+1}, ..., \tilde{x}_{j+p-1})$ with their expected outputs $\tilde{x}_{j+p}$, $j = \overline{1, N-p}$ represent the training set. After the training, predicting of future values of considered time series is possible, for example, one can give vector $(\tilde{x}_{N-p+1}, \tilde{x}_{N-p+2}, ..., \tilde{x}_N)$ to the input of the neural net in order to obtain an estimation for $\tilde{x}_{N+1}$.

III. Conditioning of the neural network

The forecasting perceptron can be considered as a vector operator $\mathfrak{N} : \mathbb{R}_{p,1} \to \mathbb{R}$. One of the measures of vector and matrix operators stability is number of conditioning. For a non-degenerate operator A can be calculated as follows:

$$\mathrm{cond}\, A = \|A\| \left\|A^{-1}\right\|, \quad (6)$$

where $\|\cdot\|$ is operator norm, induced by Euclidean vector norm $\|\vec{x}\|_2 = \sqrt{\sum\limits_{i=1}^{p} x_i^2}$, i.e.

$$\|A\| = \sup_{\|\vec{x}\| \neq 0} \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2}. \quad (7)$$

Operator $\mathfrak{N}$ can be decomposed into the following operators $\mathfrak{N}_i$, $i = \overline{1, 7}$:

- standardizing of the initial vector $\vec{a} \in \mathbb{R}_{p,1}$ with formula (1):

$$\begin{gathered} \mathfrak{N}_1 : \mathbb{R}_{p,1} \to \mathbb{R}_{p,1}, \\ \mathfrak{N}_1(\vec{a}) = \frac{1}{x_{\max} - x_{\min}} \vec{a} - \frac{x_{\min}}{x_{\max} - x_{\min}} \vec{e}, \end{gathered} \quad (8)$$

  where $\vec{e} = (\underbrace{1, 1, ..., 1}_{p})^T$;
- left multiplicating received vector by weight matrix $A_1 \in \mathbb{R}_{n,p}$, whose element on i-th row, j-th column is equal to weight of the synapse, connecting i-th neuron on the input layer with j-th neuron on the hidden layer:

$$\mathfrak{N}_2 : \mathbb{R}_{p,1} \to \mathbb{R}_{n,1}, \ \mathfrak{N}_2(\vec{a}) = A_1 \vec{a};$$

- adding received vector to bias vector $\vec{b}^1 \in \mathbb{R}_{n,1}$, whose i-th component is equal to weight of the synapse, connecting i-th bias neuron on the input layer with i-th regular neuron on the hidden layer:

$$\mathfrak{N}_3 : \mathbb{R}_{n,1} \to \mathbb{R}_{n,1}, \ \mathfrak{N}_3(\vec{a}) = \vec{a} + \vec{b}^1;$$

- activation function, acting on every neuron of the hidden layer:

$$\mathfrak{N}_4 : \mathbb{R}_{n,1} \to \mathbb{R}_{n,1}, \ \mathfrak{N}_4(\vec{a}) = (\mathrm{act}(a_1), ..., \mathrm{act}(a_p))^T,$$

  where act – one of the activation functions (3)-(5);
- left multiplicating received vector by weight matrix $A_2 \in \mathbb{R}_{1,n}$, whose element on i-th row is equal to weight of the synapse, connecting i-th neuron on the hidden layer with the neuron on the output layer:

$$\mathfrak{N}_5 : \mathbb{R}_{n,1} \to \mathbb{R}, \ \mathfrak{N}_5(\vec{a}) = A_2 \vec{a};$$

- adding to weight b of the synapse between the bias neuron on hidden layer and the neuron on the output layer:

$$\mathfrak{N}_6 : \mathbb{R} \to \mathbb{R}, \ \mathfrak{N}_6(a) = a + b;$$

- inverse of the standardizing transformation with formula (2):

$$\mathfrak{N}_7 : \mathbb{R} \to \mathbb{R}, \ \mathfrak{N}_7(a) = a(x_{\max} - x_{\min}) + x_{\min}.$$

Number of conditioning of the operator $\mathfrak{N}$ can be estimated from above with multiplication of condition numbers of $\mathfrak{N}_i$, $i = \overline{1, 7}$ according to submultiplicativity of the norm (7):

$$\mathrm{cond}\, \mathfrak{N} \leq \prod_{i=1}^{7} \mathrm{cond}\, \mathfrak{N}_i.$$

Thus, conditioning of the whole neural network depends on conditioning of its components, such as weight matrices, activation functions, as well as on preprocessing data techniques, being used before training. By estimating of condition numbers for these operators, we can judge conditioning of the neural network in general.

Let us estimate condition numbers of the operators $\mathfrak{N}_i$, $i = \overline{1, 7}$. Initially, it can be proved, that $\mathrm{cond}\, \mathfrak{N}_1 = 1$. For proving, the following formula will be used [10]:

$$\mathrm{cond}\, f = \sup_{\|\vec{a}\| \neq 0} \frac{\|J\|}{\|f(\vec{a})\| / \|\vec{a}\|}, \quad (9)$$

where J is Jacobian matrix for function f, and matrix norm $\|J\|$ is induced by the vector norm from the same

Table I
STABILITY INDICATORS OF THE FORECASTING PERCEPTRON IN DEPENDANCE ON ACTIVATION FUNCTION AND STANDARD DEVIATION OF PERTURBATIONS IN INPUT DATA

Standard deviation of perturbations	**Mean estimation of condition number of A_1**			**Mean estimation of condition number of activation function**			**Mean of norm of absolute forecasting error w.r.t. the input series**			**Standard deviation of norm of absolute forecasting error w.r.t. the input series**		
	ReLU	**LReLU**	**sp**	**ReLU**	**LReLU**	**sp**	**ReLU**	**LReLU**	**sp**	**ReLU**	**LReLU**	**sp**
0	1.58	1.70	1.75	2.08	1.56	0.37	588	602	749	237	191	322
1	1.63	1.67	1.71	2.13	1.52	0.37	582	607	786	215	252	474
5	1.65	1.65	1.70	2.00	1.54	0.37	647	643	787	235	239	351
10	1.66	1.67	1.71	1.94	1.50	0.37	692	743	927	202	250	482
50	1.66	1.70	1.70	1.95	1.56	0.37	1975	2018	2398	231	347	733
100	1.65	1.74	1.71	1.94	1.55	0.37	3583	3623	4116	266	302	916
500	1.63	1.79	1.74	1.86	1.51	0.35	16915	17052	18236	623	751	2416

formula. For function from formula (8), matrix J is equal to

$$\begin{pmatrix} \frac{1}{x_{\max}-x_{\min}} & 0 & \cdots & 0 \\ 0 & \frac{1}{x_{\max}-x_{\min}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{x_{\max}-x_{\min}} \end{pmatrix},$$

and its norm is $\frac{1}{x_{\max}-x_{\min}}$. After substitution into (9) we got

$$\operatorname{cond}\mathfrak{N}_1 = \sup_{\|\vec{a}\|\neq 0} \frac{\|J\|}{\|\mathfrak{N}_1(\vec{a})\|/\|\vec{a}\|} = \frac{1}{x_{\max}-x_{\min}} \times \\ \times \sup_{\|\vec{a}\|\neq 0} \frac{\|\vec{a}\|}{\|\mathfrak{N}_1(\vec{a})\|} = \frac{1}{x_{\max}-x_{\min}} \left\|\mathfrak{N}_1^{-1}\right\|. \quad (10)$$

Similary, for operator $\mathfrak{N}_1^{-1}(\vec{a}) = (x_{\max}-x_{\min})\vec{a} + x_{\min}$ we got

$$\operatorname{cond}\mathfrak{N}_1^{-1} = (x_{\max}-x_{\min})\left\|\mathfrak{N}_1\right\|. \quad (11)$$

Multiplication of (10) and (11) with taking into account (6) leads to the identity $\|\mathfrak{N}_1\|\left\|\mathfrak{N}_1^{-1}\right\| = 1$, whence follows that condition number of operator $\mathfrak{N}_1$ is equal to 1. $\operatorname{cond}\mathfrak{N}_i = 1$ for $i = 3, 6, 7$ is proved in the same way.

For operators $\mathfrak{N}_2$, $\mathfrak{N}_5$ condition numbers are equal to condition numbers of matrices A_1 and A_2 respectively, besides, $condA_2 = 1$, as far as the matrix $A_2 \in \mathbb{R}_{1,n}$ has only one singular number, i.e. such ς that $\exists \vec{u}, \vec{v}$: $A_2\vec{v} = \varsigma\vec{u} \wedge A_2^T\vec{u} = \varsigma\vec{v}$. Estimation of condition number for matrix A_1 was performed after forecasting perceptron training with various hyperparameters and training data.

As regards operator $\mathfrak{N}_4$, its Jacobian matrix is diagonal, hence, the matrix norm, induced by Euclidean vector norm, is equal to maximum among absolute values of diagonal elements. For functions ReLU and LeakyReLU $\|J\| = 1$, because $\mathrm{ReLU}'(x) = \mathrm{LeakyReLU}'(x) = 1$ for positive x, and $\mathrm{ReLU}'(x) = 0$, $\mathrm{LeakyReLU}'(x) = \alpha << 1$ for negative x. In case of using softplus function the following estimation is taking place:

$$\|J\| \leq \sup_x \mathrm{softplus}'(x) = \\ = \sup_x \left(\ln(1+e^x)\right)' = \sup_x \frac{e^x}{1+e^x} = 1.$$

By using (9), we got

$$\operatorname{cond}\mathfrak{N}_4 = \sup_{\|\vec{a}\|\neq 0} \frac{\|J\|}{\|\mathfrak{N}_4(\vec{a})\|/\|\vec{a}\|} \leq \\ \leq \sup_{\|\vec{a}\|\neq 0} \frac{\|\vec{a}\|}{\|\mathfrak{N}_4(\vec{a})\|}. \quad (12)$$

IV. COMPUTATIONAL EXPERIMENT AND ITS RESULTS

Testing of the forecasting perceptron was performed on the examples of time series $\vec{x} = (x_1, x_2, ..., x_N)$, modeled with the following formula:

$$x_t = t^\alpha + \beta \sin\gamma t, \ t = \overline{1, N}.$$

Time series $\vec{x}^j = \vec{x} + \vec{\xi}^j$, $j = \overline{1, Q}$, where $\vec{\xi}^j = (\xi_1^j, \xi_2^j, ..., \xi_N^j)$, $\xi_t^j \in N(0, \sigma_j^2)$ are random variables with unbiased normal distribution with variance σ_j^2, were also considered. Training of the forecasting perceptron was carried out with various values of p, n, as well as various activation functions (3)-(5) for 7 epochs. An optimization algorithm, called AdaGrad (adaptive gradient algorithm) [11], which has shown the best results on stability of the neural network [12], was used. Condition numbers of activation functions were estimated according to formula (12), where components of the vector $\vec{a}$ are equal to values on inputs of neurons on the hidden layer, and vector $\mathfrak{N}_4(\vec{a})$ consists of values outputs of the same neurons.

During the experiment neural networks with different architectures have shown approximately equal results on stability and conditioning. Forecasting perceptron performance for $N = 1000$, $\alpha = 1.2$, $\beta = 100$, $\gamma = 0.05$, $Q = 60$ in dependance on standard deviations σ_j and an activation function is shown on the I. As this table shows, when using activation function ReLU, matrix A_1 has lower condition number, then with other considred activation functions. What about LeakyReLU and softplus, the first function gives lower condition number of A_1, when perturbations in input data have standard deviation σ, not exceeding 10, and the second one – when $\sigma \geq 100$. In case $\sigma = 50$ both functions give approximately equal condition numbers of A_1. On the contrary, for the operator $\mathfrak{N}_4$, ReLU gives the highest condition number, and softplus gives the lowest one, moreover, when using ReLU, increasing of standard deviation of perturbations in input data leads to condition number decreasing, which can be explained with the aforementioned issue of vanishing gradient. The best results on accuracy and stability have been obtained with using ReLU function.

V. Conclusions and further research

Results of the forecasting perceptron training evidence that ReLU function is the most suitable as activation function in the neural network. Compared with LeakyReLU and softplus functions, ReLU has given better conditioning of weight matrix, as well as better accuracy and stability of retrospective time series forecasting. These results, however, don't correlate neither matrix A_1 conditioning nor activation functions conditioning. Thus, accuracy and stability of the forecasting perceptron don't directly depend on conditioning of its components.

In further researches, stability of the forecasting perceptron in sense of numerical conditioning will be considered. For exploring this problem, theoretical estimation of condition number of activation functions is necessary, as well as considering of factors, affecting on weight matrices and their conditioning, for example, weight initialization methods.

References

[1] L. S. Maciel, R. Ballini. Design a Neural Network for Time Series Financial Forecasting: Accuracy and Robustness Analysis. *Anales do 9º Encontro Brasileiro de Finanças*, 2008. Availaible at: https://www.cse.unr.edu/~harryt/CS773C/Project/895-1697-1-PB.pdf (accessed 2019, December)

[2] L. Falat, Z. Stanikova, M. Durisova, B. Holkova, T. Potkanova. Application of Neural Network Models in Modelling Economic Time Series with Non-Constant Volatility. *Business Economics and Management Conference*, 2015, pp. 600-607

[3] B. Sengupta, K. J. Friston. How Robust are Deep Neural Networks? *arXiv preprint arXiv:1804.11313*, 2018.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations, arXiv:1312.6199*, 2014

[5] I. J. Goodfellow, J. Shlens, C. Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations, arXiv:1412.6572*, 2015.

[6] M. Singh, A. Sinha, B. Krishnamurthy. Neural Networks in Adversarial Setting and Ill-Conditioned Weight Space. *arXiv preprint arXiv:1801.00905*, 2018.

[7] S. V. Kachalin. Otsenka Ustoichivosti Algoritmov Obucheniya Bol"shikh Iskusstvennykh Neironnykh Setei Biometricheskikh Prilozhenii [Assessment of Stability Learning Algorithms Large Artificial Neural Networks of Biometric Application]. *Vestnik SibGAU* [*SibSAU Bulletin*], 2014, no. 3, pp. 68-72.

[8] A. L. Maas, A. Y. Hannun, A. Y. Ng. Rectifer Nonlinearities Improve Neural Network Acoustic Models. *International Conference on Machine Learning*, 2013.

[9] H. Zheng, Z. Yang, W.-J. Liu, J. Liang, Y. Li. Improving Deep Neural Networks Using Softplus Units. International Joint Conference on Neural Networks, 2015. Available at https://www.researchgate.net/publication/30886346 (accessed 2019, December).

[10] L. N. Trefethen, D. Bau. *Numerical Linear Algebra*, Philadelphia, Society for Industrial and Applied Mathematics, 1997, 390 p.

[11] J. Duchi, E. Hazan, Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2121-2159.

[12] S. Sholtanyuk. Sravnitel'nyi analiz neirosetevoi i regressionnykh modelei prognozirovaniya vremennykh ryadov [Comparative Analysis of Neural Networking and Regression Models for Time Series Forecasting]. *Tsifrovaya transformatsiya* [*Digital Transformation*], 2019, no. 2, pp. 60-68.

Обусловленность компонентов нейронной сети в задачах прогнозирования временных рядов

Шолтанюк С.В.

В данной работе рассмотрен вопрос устойчивости полносвязной нейронной сети (персептрона), прогнозирующей временные ряды методом скользящего окна. Поставлена задача нахождения условий, обеспечивающих лучшие результаты по точности и устойчивости, для чего было проведено обучение нейронной сети при различных наборах гиперпараметров, таких как количество нейронов на входном и скрытом слоях, а также функция активации. Обучение проводилось для смоделированных данных, которые подвергались зашумлению несмещёнными нормально распределёнными случайными величинами с различной дисперсией. Представлены такие результаты обучения нейронной сети, как число обусловленности матрицы весов между входным и скрытым слоями, число обусловленности функций активации, а также статистические характеристики абсолютной погрешности ретроспективного прогноза. Предложен подход к оцениванию числа обусловленности некоторых функций активации, использующихся при решении задач прогнозирования. Показано влияние выбора функции активации на точность и устойчивость прогностического персептрона.

Received 29.12.19

Stock Prices Dynamics Forecasting with Recurrent Neural Networks

1[st]Tatyana Vasyaeva
Automated Control Systems
Donetsk National Technical University
Donetsk, Ukraine
vasyaeva@gmail.com

2[nd] Tatyana Martynenko
Automated Control Systems
Donetsk National Technical University
Donetsk, Ukraine
tatyana.v.martynenko@gmail.com

3[rd] Sergii Khmilovyi
Automated Control Systems
Donetsk National Technical University
Donetsk, Ukraine
hmelevoy.sergey@gmail.com

4[th] Natalia Andrievskaya
Automated Control Systems
Donetsk National Technical University
Donetsk, Ukraine
nataandr@yandex.ru

***Abstract*—The application of deep neural networks was examined in the area of stock prices forecasting of pharmacies chain "36 and 6". The learning sample formation in the time series area was shown and the neural network architecture was proposed. The neural network for exchange trade forecasting using Python's Keras Library was developed and trained. The basic parameters setting of algorithm have been carried out.**

***Keywords*—Machine Learning, Deep Learning, Recurrent Neural Networks, Simple Recurrent Neural Network, Update Gate and Reset Gate, Long short-term memory, Time Series, Stock Prices.**

I. Introduction

Effective actions on the stock exchange are connected with a careful analysis of all that's happening on the market. Good and effective prediction systems for stock market help traders, investors, and analyst by providing supportive information like the future direction of the stock market. In order to make the forecasting as reliable as possible, and the forecasts – well grounded, traders use exchange analysis. Widely used methods of analysis, to which most market participants are accustomed [1], are not always effective. That's why over recent years, financial analysts have become of great interest in such a direction as machine learning [2] and in particular artificial neural networks [3] . In contrast to the classical methods of exchange analysis [1], which involve the implementation of ready-made algorithms, machine learning allows the system to learn how to recognize patterns and make forecasts.

In our work we will consider the problem of predicting the future stock prices of the pharmacies chain "36 and 6" based on the values of their stock prices in the past, as well as additional data of exchange trade for the period under report. Future stock price prediction is probably the best example of time series forecasting [4], [5].

II. Data Representation

The data that we are going to use for this article can be downloaded from servers "FINAM" and their web resource [6]. From the list of securities (instruments) provided by "FINAM" company we will take the data of quotations of the pharmacies chain "36 and 6". The received information provided by Moscow Exchange PJSC has the following format (Table I): date, opening price, maximum price, minimum price, closing price, volume. Usually datasets for working with time series form a "sliding window" [4], [5] with a width equal to the depth of immersion. After the data are prepared in this way they will have such form (see Table II). As a rule of thumb, whenever you use a neural net-work, you should normalize or scale your data. We normalize the obtained data and divide it into parts. A common dataset

Table I
Data representation.

OPEN	HIGH	LOW	CLOSE	VOL
13.4100	13.7500	13.4000	13.5900	255490
13.4800	13.6900	13.3300	13.4200	151960
13.5400	14.6100	13.4400	14.5200	433440
14.6900	14.9100	14.4100	14.7600	178030
…	…	…	…	…
14.7500	15.2300	14.6500	15.1000	151950
14.8800	14.9500	14.6500	14.7900	77570

Table II
Training data.

№	OPEN	HIGH	LOW	CLOSE	VOL
1					
2					
…	…	…	…	…	…

after the transformations shown in Tab. II, with depth of immersion =100 has a size of 5x100x1260. To build the model, we will use a set consisting of 851 simples (of 1260 total), and under the validation set we will allocate 1% of this data. As a result, we get (see Tab. III).

Table III
TRAINING, VALIDATION AND TESTING DATASET.

Training	Validation	Testing
...	...	...
765 Data Points	86 Data Points	309 Data Points

III. MATERIALS AND METHODS

Nowadays deep neural networks become one of the most popular approaches to solving a wide variety of problems [7]–[10]. There is no unambiguous definition of what a deep neural network is. In this work, the term deep neural network will be understood as a neural network that contains more than one hidden layer. To train neural networks, including deep ones, we use the error back-propagation algorithm [3] based on the gradient descent method.

Recurrent Neural Networks (RNN). RNN [11]– are the networks with loops in them, allowing information to persist. There is a chunk of neural network (See Fig. 1). The network takes the input value xi and returns value hi. A loop allows information to be passed from one step of the network to the next. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a subsequent copy. If we unroll the loop, we will get the following (see Fig. 1). So, RNN is a deep neural network. Simple RNN (see Fig. 2) is the

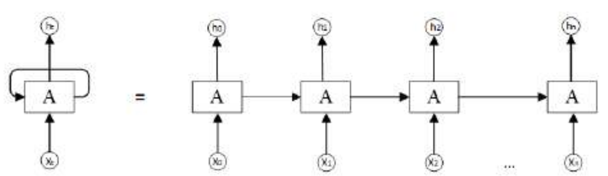


Figure 1. The fragment of recurrent neural network and an unrolled recurrent neural network..

RNN, where output is calculate as simple multiplication of Input (xi) and Previous Output (hi-1). Passed through Tanh activation function. No Gates present. Deep neural

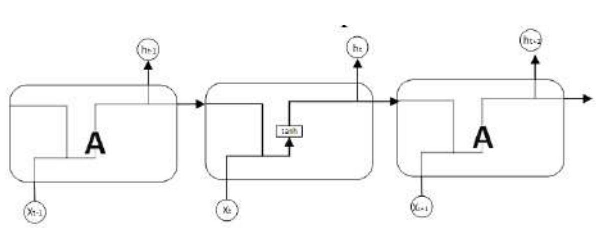


Figure 2. The repeating module in Simple RNN.

networks with many hidden layers are difficult to train because of the vanishing gradient problem. The problem of vanishing gradient can be solved by the architecture of a recurrent neural network called a network of long short–term memory [11], [12] and Update Gate and Reset Gate [13]. Long short-term memory (LSTM). The structure LSTM [11] reminds of same chain as on Fig. 3 but modules look and interact in a different way (see Fig. 3).

The key component of LSTM – is cell state. Cell state participates in several linear transformations. LSTM can erase information from cell state; this process is regulated by structures called gates. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation. LSTM-network gates: "input gate layer", "output gate layer" and "forget gate layer" [12]. "Forget gate

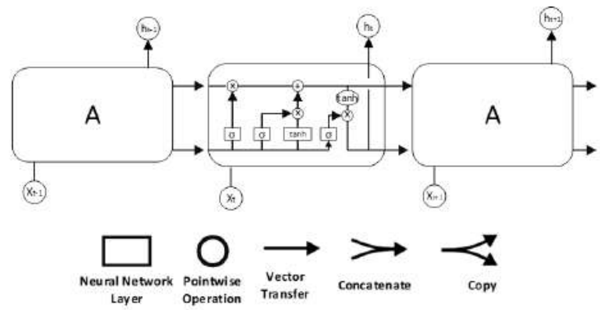


Figure 3. Recurrent LSTM network.

layer". The "forget gate layer" controls erasing or saving the cell state information based on the new input value and the value received from the previous iteration. The sigmoid layer returns numbers from zero to one which indicate what part of information should be passed on the network. Zero in this case means "do not miss anything", one – "skip all". "Input gate layer". "Input gate layer" determines when data should be written to the cell. The sigmoid layer, as before, returns numbers from zero to one based on the new input value and the value from the previous iteration. Now the sigmoid layer indicates what part information will be recorded, and the tanhlayer generates values that can be added to the cell state. As a result combination of new value and previous one will be recorded to the cell. When the data should be saved and to what extent and when they should be replaced with new ones and to what extent the neural network "decides itself" in learning process. "Output gate layer". "Output gate layer" determines what information we receive at the output. The output data will be based on the cell state and some gates will be applied to it. First, a sigmoid layer decides what information from the cell state we will output. The cell state values then pass through the tanhlayer to obtain values from the range -1 to 1 at the output and are multiplied with the output values of the sigmoid layer allowing only the required information to be displayed. Gated Recurrent Unit (GRU). The structure GRU (See Fig. 4) as LSTM has some gates: "Update gate layer" and "forget gate layer", their purpose is the same as one.

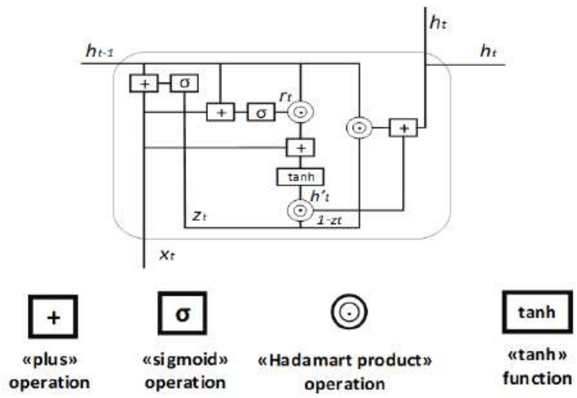


Figure 4. Recurrent GRU network.

IV. Neural network architecture.

The following neural network architecture was developed experimentally (see Fig. 5). The network consists of an input layer, three LSTM layers and one dense output layer with activation function Relu. The input data sets are 5100 in size (depth of immersion = 100). On the first LSTM layer there are 500 units and the other two have 100 units each. On the last layer there is 1 unit as we need only one output (forecast horizon = 1). After every LSTM layer the Dropout [14] is used, i.e. when training every training sample the network is rebuilt in such way that a certain set of units falls out (see Fig. 6). Developed neural network has the following view (see Fig. 7).

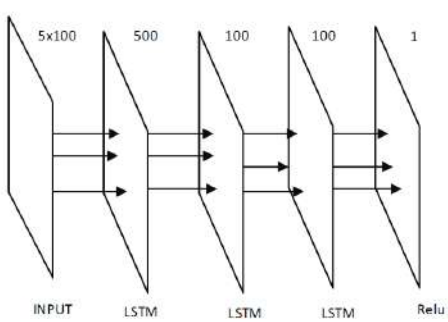


Figure 5. The architecture of proposed model.

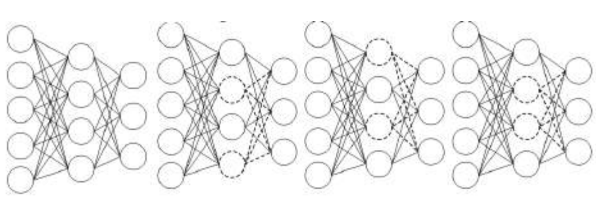


Figure 6. Dropout.

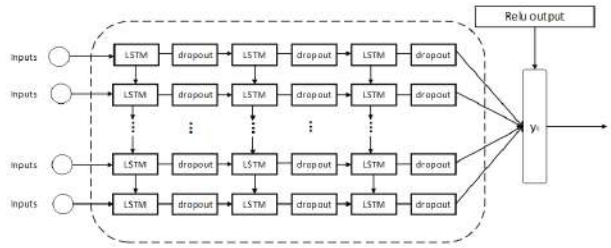


Figure 7. Neural network architecture.

V. Experiment

In the environment of Colab Laboratory [15] there was performed a software implementation of a neural network for future stock price prediction with LSTM using Python's Keras Library. Colab Laboratory is a free environment for Jupyter notebooks which requires no setting up and runs entirely in the cloud. Colab Laboratory allows you to write and execute code, as well as to get access powerful computing resources, which is an important advantage when working with resource-intensive machine learning algorithms. Keras [16] is one of the most powerful and easy-to-use Python's libraries for the development and evaluation of deep learning models covering the effective libraries of numerical computations Theano and TensorFlow. The advantage of this is mainly that it is possible to work with neural networks in a fairly simple way. Let‘s conduct experiments on neural network training (Tables IV- XII). As an optimizer we will use a stochastic gradient descent optimizer (Tables IV- VI), which is traditional for neural networks training. Loss function (function of optimization estimation) – mean squared error. The metric (the function that is used to evaluate the model) - mean absolute error.

Table IV
Experiments on network training for optimizer = «SGD», Simple RNN.

Ep	Training set		Validation set		Test set	
	MSE	***MAE***	***MSE***	***MAE***	***MSE***	***MAE***
50	0.0931	0.2512	0.0505	0.2245	0.0938	0.3038
100	0.00	0.0562	0.0053	0.0711	0.0039	0.0583
150	0.0041	0.0502	9.05e-4	0.0244	0.0004	0.0138
200	0.0026	0.0367	6.54e-4	0.0236	0.0020	0.0420
250	0.0019	0.0321	0.003	0.0537	0.0032	0.0530
300	0.0023	0.0356	4.99e-4	0.0176	0.0004	0.0148
350	0.0016	0.0306	3.31e-4	0.0147	0.0004	0.0162
400	0.0022	0.0351	1.85e-4	0.0116	0.0005	0.0170
450	0.0014	0.0275	6.88e-4	0.0215	0.0014	0.0286
500	0.0014	0.0274	3.06e-4	0.0141	0.0007	0.0223

VI. Results and discussion

In this work the neural network model has been suggested in the area of stock prices forecast for pharmacies chain "36 and 6". The neural network architecture has been developed. Network consists of LSTM layers and a

Table V
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = «SGD», GRU.

Ep	Training set		Validation set		Test set	
	MSE	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*
...	...	...	...	...	...	...
250	0.0019	0.0321	0.0034	0.0537	0.0032	0.0530
275	0.0023	0.0371	9.29e-4	0.0259	0.0004	0.0160
300	0.0023	0.0356	4.99e-4	0.0176	0.0004	0.0148
325	0.0017	0.0301	5.09e-4	0.0196	0.0005	0.0170
350	0.0016	0.0306	3.31e-4	0.0147	0.0004	0.0162
375	0.0015	0.0293	2.68e-4	0.0139	0.0005	0.0168
400	0.0022	0.0351	1.85e-4	0.0116	0.0005	0.0170
425	0.0013	0.0275	5.68e-4	0.0191	0.0012	0.0228
450	0.0014	0.0275	6.88e-4	0.0215	0.0014	0.0286
475	0.0019	0.0322	1.40e-4	0.0079	0.0003	0.0155
500	0.0014	0.0274	3.06e-4	0.0141	0.0007	0.0223

Table VI
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = «SGD», LSTM.

Ep	Training set		Validation set		Test set	
	MSE	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*
150	0.0024	0.0347	1.58e-4	0.0090	0.0002	0.0118
175	0.1967	0.4219	0.0538	0.2313	0.0148	0.1039
200	0.0021	0.0318	1.72e-4	0.0099	0.0002	0.0128
225	0.0019	0.0303	2.24e-4	0.0119	0.0004	0.0182
250	0.0019	0.0297	1.41e-4	0.0084	0.0001	0.0107
275	0.0019	0.0299	1.90e-4	0.0105	0.0002	0.0144
300	0.0019	0.0295	2.33e-4	0.0122	0.0004	0.0189
325	0.0020	0.0310	1.65e-4	0.0094	0.0003	0.0129
350	0.0018	0.0297	2.28e-4	0.0119	0.0005	0.0199
375	0.0018	0.0291	1.31e-4	0.0080	0.0002	0.0108
400	0.0017	0.0279	2.16e-4	0.0118	0.0005	0.0199

It is recommended to use "RMSprop" optimizer for recurrent neural networks according to [11]

Table VII
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = "RMSPROP", SIMPLE RNN.

Ep	Training set		Validation set		Test set	
	MSE	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*
...	...	...	...	...	...	...
200	0.1967	0.4219	0.0538	0.231	0.0148	0.1039
250	0.0210	0.1201	0.0518	0.231	0.1297	0.3545
300	0.1967	0.4219	0.0538	0.231	0.0148	0.1039
350	0.0205	0.1180	0.0251	0.157	0.0852	0.2849
400	0.0094	0.0755	2.87e-4	0.013	0.0215	0.1319
450	0.1967	0.4219	0.0538	0.231	0.0148	0.1039
500	0.1967	0.4219	0.0538	0.231	0.03523	0.1877
600	0.1967	0.4219	0.0538	0.231	0.0148	0.1039
700	0.1967	0.4219	0.0538	0.231	0.0148	0.1039
800	0.0201	0.1163	0.0129	0.112	0.0615	0.2399
900	0.0196	0.1158	0.0299	0.172	0.0938	0.2997
1000	0.1967	0.4219	0.0538	0.231	0.0148	0.1039

Table VIII
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = "RMSPROP", GRU.

Ep	Training set		Validation set		Test set	
	MSE	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*
100	0.0012	0.0280	2.88e-5	0.0047	2.24e-5	0.0042
125	0.0011	0.0258	1.16e-5	0.0028	0.0002	0.0114
150	9.01e-4	0.0233	1.68e-4	0.0127	0.0005	0.0227
175	7.62e-4	0.0217	1.44e-4	0.0117	7.79e-5	0.0082
200	5.24e-4	0.0175	1.12e-5	0.0027	7.71e-5	0.0075
225	6.43e-4	0.0193	8.59e-5	0.0088	0.0004	0.0182
250	4.70e-4	0.0169	6.39e-4	0.0251	0.0018	0.0417
275	4.46e-4	0.0162	1.80e-5	0.0036	7.51e-5	0.0080
300	3.72e-4	0.0151	1.48e-5	0.0032	3.13e-5	0.0049
325	3.62e-4	0.0144	5.22e-5	0.0069	0.0001	0.0107
350	1.85e-4	0.0100	6.24e-6	0.0018	1.18e-5	0.0019
375	3.26e-4	0.0140	3.72e-4	0.0191	0.0007	0.0265
400	2.77e-4	0.0127	1.55e-4	0.0121	0.0006	0.0231

Table IX
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = "RMSPROP", LSTM.

Ep	Training set		Validation set		Test set	
	MSE	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*
100	6.026e-4	0.0185	1.487e-4	0.0119	0.0003	0.0220
125	8.697e-4	0.0228	1.723e-5	0.0034	0.0003	0.0173
150	3.978e-4	0.0146	4.024e-5	0.0057	0.0005	0.0256
175	6.145e-4	0.0189	8.031e-5	0.0086	5.27e-5	0.0066
200	7.348e-4	0.0208	1.654e-4	0.0125	0.0002	0.0103
225	5.222e-4	0.0179	9.030e-5	0.0090	0.0008	0.0259
250	4.442e-4	0.0166	1.405e-5	0.0032	6.09e-5	0.0069
275	4.244e-4	0.0149	1.830e-4	0.0133	0.0003	0.0170
300	3.782e-4	0.0146	0.0013	0.0341	0.0083	0.0861
325	4.915e-4	0.0176	7.736e-5	0.0076	0.0026	0.0423
350	3.653e-4	0.0139	4.351e-4	0.0207	0.0007	0.0256
375	0.1967	0.4219	0.0538	0.2313	0.0148	0.1039
400	0.1967	0.4219	0.0538	0.2313	0.0148	0.1039

Table X- XII shows the results using another common optimizer «adam». A good results are obtained for a LSTM neural network trained in 150 epochs, optimizer «adam». The best results are obtained for a neural network trained in 350 and 400 epochs.

Table X
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = «ADAM», SIMPLE RNN.

Ep	Training set		Validation set		Test set	
	MSE	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*
100	0.0209	0.119	0.0516	0.226	0.1292	0.3538
125	0.0212	0.119	0.0646	0.253	0.1491	0.3809
150	0.1967	0.421	0.0538	0.231	0.0148	0.1039
175	0.0201	0.117	0.0619	0.248	0.1424	0.3723
200	0.0017	0.036	5.32e-5	0.005	0.0012	0.029
250	0.1967	0.421	0.0538	0.231	0.0321	0.1468
300	0.1967	0.421	0.0538	0.231	0.0148	0.1039
325	0.0045	0.053	0.0038	0.057	0.0032	0.0497
350	0.1967	0.421	0.0538	0.231	0.0148	0.1039
400	0.1967	0.421	0.0538	0.231	0.0148	0.1039
450	0.1967	0.421	0.0538	0.231	0.0148	0.1039
475	0.0012	0.026	6.05e-5	0.006	0.0016	0.0331
500	0.1967	0.421	0.0538	0.233	0.0148	0.1039

Table XI
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = «ADAM»,GRU.

Ep	Training set		Validation set		Test set	
	MSE	***MAE***	***MSE***	***MAE***	***MSE***	***MAE***
100	5.11e-4	0.0164	1.42e-5	0.0027	3.84e-5	0.0055
125	4.94e-4	0.0164	4.57e-5	0.0058	8.36e-5	0.0086
150	3.27e-4	0.0136	3.28e-5	0.0053	2.70e-5	0.0047
175	3.51e-4	0.0138	6.15e-5	0.0073	7.88e-5	0.0086
200	2.91e-4	0.0123	2.98e-5	0.0047	0.0148	0.1039
225	3.25e-4	0.0137	8.18e-6	0.0022	1.90e-5	0.0038
250	2.36e-4	0.0111	2.62e-5	0.0045	1.11e-5	0.0027
275	2.31e-4	0.0113	1.20e-5	0.0028	3.66e-5	0.0055
300	1.87e-4	0.0099	9.99e-6	0.0026	6.20e-5	0.0066
325	2.04e-4	0.0106	1.39e-5	0.0029	3.94e-5	0.0057
350	2.08e-4	0.0109	2.64e-5	0.0047	1.52e-5	0.0032
375	1.72e-4	0.0095	5.43e-6	0.0016	9.69e-6	0.0025
400	1.84e-4	0.0099	1.20e-5	0.0028	5.27e-6	0.0017

Table XII
EXPERIMENTS ON NETWORK TRAINING FOR OPTIMIZER = «ADAM», LSTM.

Ep	Training set		Validation set		Test set	
	MSE	***MAE***	***MSE***	***MAE***	***MSE***	***MAE***
100	7.98e-4	0.0215	3.71e-5	0.0041	3.69e-5	0.0041
125	4.73e-4	0.0159	1.19e-4	0.0099	0.0305	0.1746
150	5.43e-4	0.0174	2.58e-5	0.0036	3.44e-5	0.0043
175	4.08e-4	0.0153	2.17e-5	0.0032	4.75e-5	0.0058
200	3.54e-4	0.0141	2.76e-4	0.0162	0.0003	0.0177
225	3.27e-4	0.0134	1.44e-5	0.0028	5.05e-5	0.0061
250	2.56e-4	0.0117	5.37e-5	0.0065	4.58e-5	0.0061
275	0.0019	0.0306	2.21e-4	0.0118	0.0003	0.0165
300	2.81e-4	0.0126	2.51e-5	0.0041	2.93e-5	0.0048
325	2.30e-4	0.0110	9.75e-6	0.0024	2.61e-5	0.0041
350	1.85e-4	0.0096	8.27e-6	0.0022	3.34e-5	0.0049
375	2.03e-4	0.0102	1.61e-4	0.0124	0.0003	0.0166
400	1.75e-4	0.0098	8.63e-6	0.0022	1.66e-5	0.0033

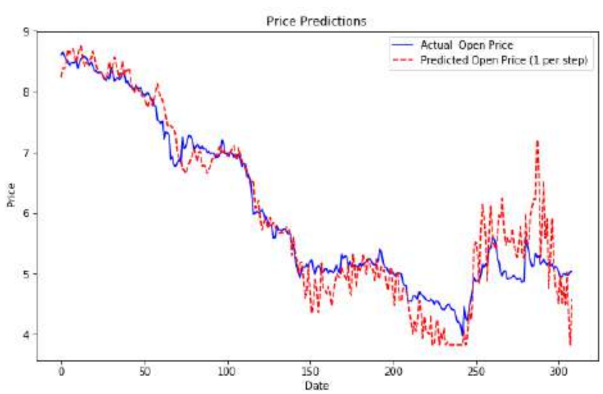


Figure 8. The results of prediction on the test sample (Steps = 150, optimizer = "SGD"). SimpleRNN.

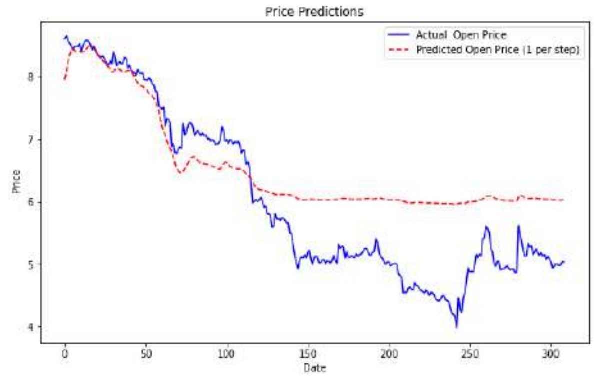


Figure 9. The results of the forecast on the test sample (P = 475, optimizer = "adam"). SimpleRNN.

Figure 10. The results of prediction on the test sample (Steps = 150, optimizer = "RMSprop").GRU.

Figure 11. Forecast results on the test sample (Steps = 250, optimizer = "adam"). GRU.

Figure 12. Forecast results on a sample of ten (Steps = 175, optimizer = "RMSprop"). LSTM

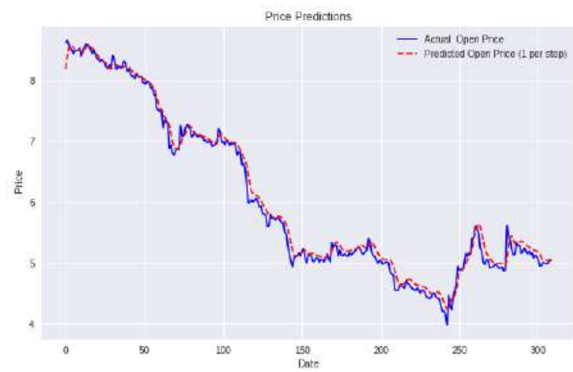


Figure 13. The results of prediction on the test sample (Steps = 150, optimizer = "adam").LSTM.

Dense layer with the Relu activation function. Dropout is used to solve the problem of retraining. Experimental studies have shown that the best results are achieved using the optimizer "adam". Error on a test set is MSE = 1.664e-05, MAE = 0.0033. This error makes it possible to forecast the price dynamics.

References

[1] Iacomin R.: Stock market prediction. In: 19th International Conference on System Theory, Control and Computing (ICSTCC), Cheile Gradistei, pp. 200-205. (2015), DOI: 10.1109/ICSTCC.2015.7321293.

[2] Shai, Shalev-Shwartz, Shai, Ben-David: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA, 416p. (2014)

[3] Schmidhuber J.: Deep Learning in Neural Networks: an Overview. Neural Networks. Vol. 1. pp. 85–117. (2015.) DOI: 10.1016/j.neunet.2014.09.003

[4] Brockwell, Peter J., Davis, Richard A.: Introduction to Time Series and Forecasting. Second Edition. Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA, 449 p. (2002)

[5] Sergii, K., Yurii, S., Tatyana, V., Natalia, A.: Feature Selection for Time-Series Prediction in Case of Undetermined Estimation. In: Samsonovich A., Klimov V., Rybina G. (eds.) Biologically Inspired Cognitive Architectures (BICA) for Young Scientists, Advances in Intelligent Systems and Computing, vol. 449, pp. 85-97. Springer, Cham (2016).

[6] Stock quotes, https://www.finam.ru/, last accessed 2019/03/29.

[7] Vargas, R., Ruiz, L.: Deep learning: previous and present applications. Journal of awareness, 2(Special 3), 11-20 (2018).

[8] Chen Y., Lin Z., Zhao X., Wang G., and Gu Y.: Deep Learning-Based Classification of Hyperspectral Data. Selected Topics in Applied Earth Observations and Remote Sensing. IEEE Journal of. 7, 2094-2107(2014).

[9] Bao, W, Yue, J, Rao, Y.: A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS One 12(7), 1-24 (2017).

[10] Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’05), vol. 1, pp. 539–546. IEEE, (2005).

[11] Manaswi, N.K.: Regression to MLP in Keras. In: Deep Learning with Applications Using Python, pp. 68-89. Apress, Berkeley, CARecurrent Neural Networks (2018).

[12] Vasyaeva T., Martynenko T., Khmilovyi S., Andrievskaya N. (2019) Stock Prices Forecasting with LSTM Networks. In: Kuznetsov S., Panov A. (eds) Artificial Intelligence. RCAI 2019. Communications in Computer and Information Science, vol 1093. Springer, Cham

[13] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. arXiv preprint arXiv:1502.02367, 2015

[14] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958. (2014).

[15] Google Colaboratory, https://colab.research.google.com/notebooks/welcome.ipynb, last accessed 2019/02/11.

[16] Keras: The Python Deep Learning library, https://Keras.io/, last accessed 2019/03/10.

Прогнозирование динамики изменения цен на акции с помощью рекуррентных нейронных сетей.

Васяева Т.А., Мартыненко Т.В.,
Хмелевой С.В., Андриевская Н.К.

В статье рассмотрено применение рекуррентных нейронных сетей для прогнозирования цен акций сети аптек «36и6». Предметом исследования данной работы являются методы анализа фондовых бирж. Показано получение котировок акций, описан формат полученной информации. Данные представляют собой многомерный временной ряд, который содержит 5 каналов: цена открытия, максимальная цена, минимальная цена, цена закрытия, объем торгов. Рассмотрено формирование обучающей выборки «скользящим окном» в задачах машинного обучения применительно к многомерным временным рядам. Рекуррентные нейронные сети трудно обучать из-за проблемы исчезающего градиента. Показано применение GRU и LSTM сетей для решения данной проблемы. Предложена архитектура нейронной сети. Разработана и обучена нейронная сеть для прогнозирования биржевой торговли с использованием языка программирования Python, а также библиотеки Keras. Проведены эксперименты по настройке основных параметров.

Received 14.12.2019

Decision-making process analysis and semantic explanation extraction from trained supervised machine learning models in medical expert systems

1st Aliaksandr Kurachkin
faculty of radiophysics and computer technologies
Belarusian State University
Minsk, Belarus
alex.v.kurochkin@gmail.com

2nd Vasili Sadau
faculty of radiophysics and computer technologies
Belarusian State University
Minsk, Belarus
sadov@bsu.by

3rd Tatiana Kachan
Belarusian State Medical University
Minsk, Belarus
tvk35@yahoo.com

***Abstract*—Supervised machine learning provides a mechanism for establishing an approximation of input-output relationship between arbitary dataset. However, semantic interpretation of an underlying decision-making process of a trained model is very hard, especially considering the probabilistic nature of machine learning. The paper discusses possible ways to semantically explain decision-making process of a trained supervised machine learning model in order to gain insights to the dataset and derive new expert knowledge from such models.**

***Keywords*—supervised machine learning, machine learning explanation, decision support systems, medical expert systems**

I. Introduction

Supervised machine learning has become a staple of data mining techniques in the recent years, solidly establishing itself as a separate field of research with a multitude of approaches and algorithms, suitable for solving a wide array of problems. One of the most important applications of such techniques is creating decision support systems – given a sufficiently large dataset, supervised machine learning algorithms are able to derive a decision path by establishing non-linear dependencies between input features and expected results. However, one of the main problems of supervised machine learning algorithms is a semantic interpretation of acquired results [1], [2].

Although some models, like predictive decision trees, are relatively easy to examine, the models they produce are usually too simple to correctly approximate datasets with non-linear dependencies. Because of this, most of the dependencies established by a more complex supervised machine learning model, like gradient-boosted decision trees, random forests or neural networks, are highly non-linear in nature and are usually very hard to interpret. Moreover, such models usually employ a degree of randomness in order to make the learning process more robust, making the learning itself probabilistic, which means that several retrainings of the same model architecture on the same dataset may yield different models, with their own underlaying decision-making processes.

II. Semantics of decision-making process

In many applications, supervised machine learning models used as part of decision-making process must not only be correct on the dataset provided, but also be able to generalize for new data and provide semantically correct results. While the former is usually solved by introducing a test and cross-validation dataset split, the latter is impossible to establish for a general case. In other words, a trained supervised machine learning model remains a black box that might produce correct decisions on the items provided in dataset, but the decision-making process itself remains obscure, making it harder to "trust" such a model from a semantic interpretation standpoint.

Another important problem is generation of new expert knowledge based on data, or gaining insights into the relationship between a particular data item input and output in a given model. Even if a model demonstrates a high prediction accuracy in a certain dataset, the prediction itself is sometimes not as important as the decision-making process behind making that prediction. This is especially true for medical expert systems and sympthomogenesis – semantic explanation of dependencies between some measurable inputs and the prediction of a certain pathology may become, in essence, the definition of a symptom – a piece of expert knowledge that establishes, based on dataset given, that specific types of inputs may, in fact, serve as indicators that a certain pathology is present. In order to establish such a dependency, the nature of relation must be explained semantically, i.e. in an understandable and human-readable form [1].

One of the main problems with establishing an understandable explanation for a trained model's decision-making process is the neccesity to create a reasonable semantically succint representation that doesn't hide away the complexity of the model itself. For this purpose, it's necessary to distinguish what kind of relationships are generally percieved as semantically understandable by a human being.

One of the more "understandable" decision-making classification techniques is a threshold-based linear binary classification using a single feature, i.e. feature exceeding a certain threshold signifies one class, and feature being lower than this threshold signifies the opposite class. More formally, given a dataset $\mathbf{T}$ of k elements for binary classification with n-dimensional input vectors

$$\mathbf{T} = \{\mathbf{X}, \vec{y}\}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{k \times n}$ – inputs matrix, $\vec{y} \in \mathbb{B}^k$ – expected outputs vector, $\mathbb{R}^{k \times n}$ – a set of rational-valued matrices with k rows and n columns, $\mathbb{B}^k$ – a set of k-dimensional vectors of boolean $\mathbb{B} = \{0, 1\}$ values, a threshold-based linear binary classification using a feature j is established as follows:

$$C_j(\vec{x}, p) = [x_j > p], \tag{2}$$

where x_q denotes a feature with the number q in input vector, and parameter p is some threshold value. The classifier (2) assumes that each individual feature of the input vector is sufficient to perform an entire classification. It is also possible to use binary classifier metrics, like precision, recall and F1-score, in order to establish the best threshold value p with respect to a dataset (1). Threshold value can also be adjusted to perform ROC-analysis and compare individual features for statistical significance using AUC (area under ROC-curve) metric [2], [3].

Such a classifier is quite primitive – it only uses a single feature and assumes a direct linear discrimination for classification. However, it has the advantage of being intuitively easy to explain and extract semantic meaning from. While it is rare for a single feature to serve as a good discriminator for output on the entire dataset, it is quite possible that a simple linear discriminative dependency in a local input neighbourhood would be representative of a more complex classifier behavior, while maintaining the ease of explanation.

III. Feature-dropping as a way to establish the most statistically significant features

Within the problem of deriving decision-making process from a trained supervised machine learning model it is possible to establish a sub-problem of determining the statistical significance of individual features in terms of how much their specific values affect the predicted output. While impossible to derive from a single model, it is possible to use ensembling to evaluate individual feature significance. In order to do this on a dataset (1), the following algorithm is proposed:

- A neural-network based feedforward model M is created for the dataset. Model's hyperparameters can be optimized at this stage using grid search methods in order to find the model that best fits the dataset (1) with respect to its F1-score $F_1(M)$.
- For each feature j of total n input features, the same model architecture (i.e. with the same set of hyperparameters) is used to train the model M_j on the dataset (1) with feature j excluded from the dataset.
- For each model M_j, F1-score $F_1(M_j)$ is calculated.
- Amongst n models M_k and base model M, the best model M^* is found based on its F1-score.
- For every feature, a metric is calculated as follows:

$$Q(M_k) = 1 - (F_1(M^*) - F_1(M_k)) \tag{3}$$

The process can be repeated for any number of combinations of individual features – i.e. for pairs, triples, etc. of input features in the dataset, constrained by computational complexity and combinatorial explision of working with individual features. In such a case, it's possible to aggregate the metric (3) per feature.

The metric (3) can be used as a relative measure of statistical significance amongst the features considered. The idea behind it is that a statistically significant feature would have a large enough impact on the overal model that its removal would lead to a greater decrease in the F1-score whenever this particular feature is not used in the next model.

Features with high enough statistical significance are candidates for being discriminative enough to try and establish a single-feature threshold classifier (2), although without traversing all the possible input feature combinations, it's usually impossible to determine whether a specific feature is representative by itself, or if it relies on relationships (possibly non-linear) with other features in order to make a decision.

IV. Instance explanation using local linear approximations

While simple linear threshold binary classifiers (2) aren't usually sufficient to approximate dependencies in real data, their defining feature of being susceptible to semantic analysis makes them a useful tool for generating approximate explanation that may be used to describe a local behavior of a more complex model [3].

The paper proposes using local linear approximations as a method for explaining a specific instance prediction result for a trained supervised machine learning model. The explanation itself takes a form of a threshold condition for an arbitary classifier (2) that is representative of a model's behavior in proximity to a specific instance data point, while not necesserily being a good approximation of the model's behavior as a whole on the entire input parameter space.

Given a trained supervised machine learning model for binary classification $C_f(\vec{x}, \mathbf{\Theta}) : \mathbb{R}^n \to \mathbb{B}$ with a set of parameters $\mathbf{\Theta}$ that are determined during model's learning on dataset (1), a local linear approximation of this model at data point $\vec{x}^*$ can be derived using the following algorithm:

1) A subset of the input vectors from source dataset (1) are selected based on their proximity to the data point $\vec{x}^*$. For proximity measure, it's possible to use any n-dimensional vector measure, for instance, Euclidean distance:

$$d(\vec{x_1}, \vec{x_2}) = \sqrt{\sum_{i=1}^{n} (x_{1_i} - x_{2_i})^2}. \tag{4}$$

It's possible to select an arbitary number of vectors closest to the data point, or select vectors within a certain threshold distance.
Dataset point selection is used to determine a radius d_{min} around the point $\vec{x}^*$ that is representative of a model's behavior without bias introduced by other points. The simplest selection method is to assume d_{min} to be the distance to the closest point in the dataset.
2) The local proximity $P_{\vec{x}^*} = \{\vec{x}_i\}$ of $\vec{x}^*$ is sampled within d_{min}: $\forall \vec{x}_i | d(\vec{x}^*, \vec{x}_i) < d_{min}$. For generated sample points $\{\vec{x}_i\}$ respective classifier responses are obtained by evaluating the model C_f:

$$\{y_i\} = \{C_f(\vec{x}_i)\} \tag{5}$$

3) An explanatory classifier $C_{exp}^{(\vec{x}^*)}$ with a linear model is defined:

$$C_{exp}^{(\vec{x}^*)}(\vec{x}) = \theta_0^* + \vec{\theta}^* \cdot \vec{x} \tag{6}$$

The linear classifier $C_{exp}^{(\vec{x}^*)}$ represents a hyperplane in feature space that is used to discriminate between two classes in the original classification problem solved by C_f. Hyperplane parameters θ_0^* and $\vec{\theta}^*$ are initialized randomly.
4) Point sets $\{\vec{x}_i\}$ and $\{y_i\}$ are used as a training set for explanatory classifier $C_{exp}^{(\vec{x}^*)}$ in order to obtain specific hyperplane parameters.

The resulting linear classifier $C_{exp}^{(\vec{x}^*)}$ is, essentially, a local linear approximation of a more complex behavior exhibited by the original model C_f. However, this local approximation is guaranteed to be representative of the original classifier within d_{min} radius of the original data point $\vec{x}^*$, with the added benifit of being semantically representative and understandable by a human being. It is possible to analyze projections of the hyperplane defined by classifier model (6) in order to establish specific threshold values for individual features, and use the projection angle as a measure of this particular feature impact. For example, if a local approximation hyperplane is perpendicular to a feature axis k, that means that original classifier C_f in the proximity of an explained data point relies only on feature k to produce a prediction. With the threshold value t_k, it is possible to produce a semantic form of an explanation using the following statement:
For inputs $\vec{x}^*$ classification result is $y^* = C_f(\vec{x}^*)$, because x_k is greater (lesser) than t_k.

For arbitary plane alignment, similar statements can be produced for any input feature, while their impact can be determined by the cosine of the angle between a feature axis and the plane.

In other words, given a trained supervised machine learning model for binary classification $C_f(\vec{x}, \mathbf{\Theta}) : \mathbb{R}^n \to \mathbb{B}$ with a set of parameters $\mathbf{\Theta}$ that are determined during model's learning on dataset (1), a local linear approximation $C_{exp}^{(\vec{x}^*)}$ of this model at data point $\vec{x}^*$ can be used to create a hyperplane, that defines threshold values t_k (axis intersection points) and impact values $w_k = cos\varphi_k$ (where φ_k is the angle between feature axis and the plane) of an arbitary k-th feature. These individual values can be used to produce semantic form of an explanation with the following statements:
For inputs $\vec{x}^*$ and classification result $y^* = C_f(\vec{x}^*)$, the fact that x_k is greater (lesser) than t_k affects the output with impact w_k.

V. Example data analysis in ophtolmological decision support system

One of the domain areas where proposed approaches were used to generate expert knowledge is ophtolmological disease diagnosis, specifically optical nerve disorders like multiple sclerosis associated optic neuritis and glaucoma. The most common diagnostic tool in this area is optical coherent tomography (OCT) and scanning laser polarimetry (SLP), producing optical nerve and retina images and allowing to produce structural features. On the other hand, clinical practice also uses certain functional features. One of the unsolved problems is discriminating feature selection – some papers suggest that specific OCT or SLP features are the most impactful, while the others indicate that functional features should be used instead [5].

Based on historical data analysis, a simple feedforward binary classifier was created that uses all structural and functional features measured during the research for confirmed cases pathology and control group of healthy people. Separate classifiers were implemented for multiple sclerosis and glaucoma. The initial research demonstrated that the classifiers were performing better than linear thresholding using any single feature, with F1 scores of 0.75 versus 0.68 for multiple sclerosis, and 0.93 versus 0.81 for glaucoma [5].

The neural network architecture used in proposed classifiers comprised 26 input features, 3 hidden layers of 30 neurons each with ReLU activation function and dropout, and single-neuron output layer with variable classification threshold.

Black-box nature of proposed classifiers made it hard to justify their use in clinical practice, because semantic correctness of the underlaying decision-making process was impossible to determine.

The statistical significance analysis based on the proposed metric (3) allowed to confirm that certain sturctural and functional features were more impactful than the others. The general domain knowledge about the impact of individual features coincided with acquired impact predictions, confirming their correctness [5].

When examining individual samples using proposed local linear approximations method, it was possible to determine a clinical explanation for many of the inconclusive samples. Established input-output dependencies explained in an understandable way were used in order to propose new insights into clinical pathologenesis and confirm certain forms and characteristics of researched pathologies.

Conclusion

Explanation of decision-making process in trained supervised machine larning models provides an effective way to evaluate datasets used for training the model. The methods proposed in this paper include individual feature statistical significance evaluation based on the reduction of F1-score when this particular feature is excluded from training, as well as local approximations method that allows to explain the local behavior of a trained model within the proxmity of an individual data point.

Proposed algorithms can be used in order to gain insights into the dataset used to train the data, evaluate decision-making processes established within the models to derivate new expert knowledge, and semantically validate individual predictions. In certain applciations, such analysis allows to mitigate the issue of black-box behavior of supervised machine learning algorithms and enable their applications to domain fields where transparency of a decision-making process is required, like complex control systems or medical decision support systems.

The evaluation of proposed feature impact analysis and semantic instance explanation were used to evaluate a medical decision-support system for diagnosing optical disorders. Feature impact analysis based on the reduction of F1-score on the same model showed results that are similar to the general domain knowledge, while smenatic interpretation of the decision-making process allowed to produce new domain knowledge and found a better understanding of specific dosorder's pathologenesis.

References

[1] А. В. Курочкин, В. С. Садов, "Методы анализа и семантической интерпретации процессов принятия решения в классификационных нейросетевых моделях машинного обучения с учителем," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, 2019.

[2] M. A. Nielsen, "Neural Networks and Deep Learning," Determination Press, 216 p., 2019

[3] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference and Prediction," Springer Series in Statistics, Springer-Verlag New York, 2nd ed., 745 p., 2009.

[4] D. M. W. Powers, "From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, iss. 1, BioInfo Publications, pp. 37–63, 2011.

[5] Т. В. Качан, А. В. Курочкин, Е. А. Головатая и др., "Роль искусственных нейронных сетей в выявлении ранней гибели ганглионарных клеток сетчатки у пациентов с дегенеративными оптиконейропатиями," Офтальмология. Восточная Европа. – 2019. – Т. 9, № 4. – С. 445–458.

Анализ процессов принятия решений и извлечение семантического описания в обученных моделях машинного обучения с учителем в медицинских экспертных системах

Курочкин А.В., Садов В.С., Качан Т.В.

Модели машинного обучения с учителем предоставляют механизм установления аппроксимации взаимодействия между входными и выходными значениями произвольных наборов данных. Тем не менее, семантическая интерпретация лежащего в основе этих моделей процесса принятия решения является сложной задачей, особенно в контексте вероятностного характера некоторых методов машинного обучения с учителем. В статье рассматриваются методы семантического объяснения процесса принятия решения обученной модели машинного обучения с учителем, что позволяет выделить сложные зависимости из наборов данных и вывести с их помощью новые экспертные знания.

Received 30.12.19

Classifier for Confocal Thyroid Imaging on Single Board Computers

1st Konstantin Kurochka
Sukhoi State Technical University of Gomel
Gomel, Belarus
kurochka@gstu.by

2nd Alexander Demidov
Sukhoi State Technical University of Gomel
Gomel, Belarus
demidovmainmailbox@gmail.com

***Abstract*—The paper proposes the architecture of a convolutional neural network for classification of digital images of the thyroid gland obtained using a laser confocal microscope. The possibility of using single-board computers as a basis for building a full-fledged device for classifying images with high mobility and versatility is being considered. The process of preparing data for training and comparing existing single-board computers available.**

I. Introduction

Recently, interest in building mobile energy-efficient and resource-intensive system, with using micro-platforms has increased in many areas of industry, and medicine is no exception. This area has always been at the forefront of technology, while the capabilities of personal computers were relatively small, and distributed systems [1] were used to process digital images in view of the required capacities.

However, the development of technology has allowed the same amount of calculations, but on more compact devices.

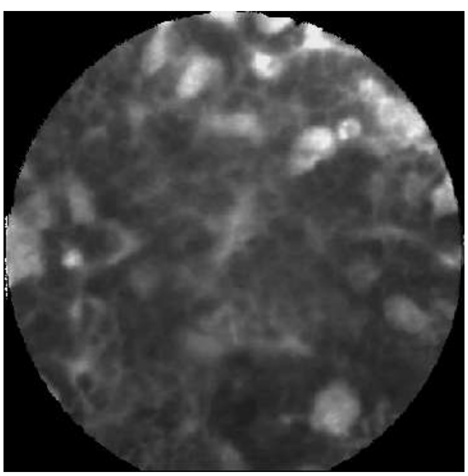


Figure 1. An example of a confocal image of the thyroid gland with a pronounced adenoma.

The development of machine learning has enabled the use of neural networks to recognize and localize objects in medical images [2] [3]. One of these areas has become microscopy. Data from various electron microscopes should be analyzed quickly enough to establish a diagnosis. One such data source can be considered a confocal laser microscope. Confocal microscopy helps microbiologists solve a number of specific problems.

This method is used to study the structure of cells, their nucleus and mitochondria, chromosomes and genes. With its help, one can observe many dynamic processes: features of cellular transport of various compounds, redistribution of ions within a living cell. Confocal microscopy helps to model three-dimensional images. Using photos of optical sections recorded in the computer's memory, you can recreate any object in maximum detail.

On the fig. 1 shows the image of the thyroid gland during surgery obtained using a confocal laser microscope. Due to the fact that the microscope is a closed device, the introduction of new functionality is impossible or is an extremely rare and expensive procedure.

It is proposed to use a convolutional neural network and a third-party device based on a single-board computer to classify modified tissues using confocal images of the thyroid gland. This approach will minimize the cost of adding new features to a confocal microscope.

This solution will also significantly speed up the time and number of patients under consideration due to the fact that ordinary cases with a pronounced presence or absence of thyroid modifications will be detected automatically, this will allow patients to be sorted for various types of tests or various techniques. It should be noted right away that when using single-board computers and placing a neural network on them, the problem of limited resource capabilities of devices arises.

The selection conditions for testing devices were their technical characteristics, price and availability of platforms on the Belarusian market. A comparative analysis of single-circuit devices in solving classification problems using the proposed convolutional neural network is carried out. The corresponding software was developed using tools in the Python language and the operating systems Armbian and Ubuntu, adapted for use on single-board computers with limited resources.

II. Convolutional Network Classifier

When analyzing data obtained from a confocal microscope, two main parameters are important - this is the quality and speed of the network. In the course of solving the problem of classification of images of the thyroid gland, a comparative analysis of various devices for the implementation of the image classifier was carried out.

The solution to the problem of recognition of confocal images of the thyroid gland consists of several stages:

- Building a network model.
- Data preprocessing.
- Network quality training and validation.

The training data-set consists of 250 images of expressed thyroid adenoma, 250 images of a expressed normal state and 400 images of boundaries and poorly defined areas. Based on the limited sampling, it was decided to artificially increase the dimension of the data-set. For this, the following operations were carried out:

- Collection of graphic data. Getting confocal images
- Filtering - checking images for a number of requirements: a sufficient level of illumination of objects on them, the presence of the necessary object. All noisy and low-quality images were used to create a negative class.
- Preparation of tools for marking. Selection of the necessary image generation algorithms.
- Create new images. Apply filters and distortion.

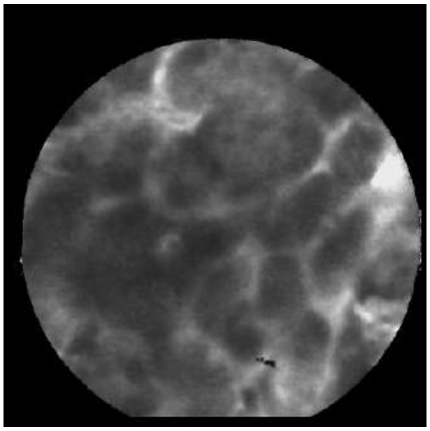


Figure 2. An example of a confocal image of a healthy thyroid gland.

After carrying out these operations, the dimension of the data-set was increased by 1.5 times. This will increase the flexibility of the network in solving the problem. The data-set was artificially expanded due to augmentation of data to 1350. The source data is divided in the proportions of 75% of educational to 25% of validation. On the fig. 2 illustrates confocal images of healthy thyroid.

Neural network written in high level Python using TensorFlow machine learning library [4] and with framework Keras [5]. On the fig. 3 illustrates the network architecture that was used in training the network.

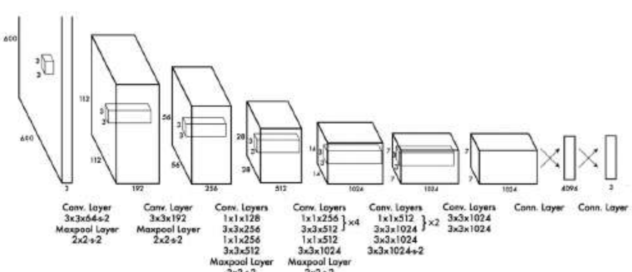


Figure 3. Model of neural network.

The network consists of convolutional and fully connected layers. The first convolutional layer contains 128 convolution cores 3 * 3, after each convolutional layer there is an activation layer ReLU:

$$f(x) = max(0, x) \tag{1}$$

Several fully connected layers follow convolutional layers; the last layer contains the Softmax function. Adam was used as an optimization function.

A three-channel image with a size of 600px * 600px is fed to the network input, after which the data passes through a set of convolutional and ReLU layers, which transmit a set of features to fully connected layers. The output of the network is a vector of dimension equal to the number of classes of objects.

To train the network, a personal computer with a video card was used NVIDIA GTX 1080TI. On the fig. 4 and 5 shows the results of the quality of accuracy after training the network in 50 epoch, for the training samples is 83% and 79% for the validating samples.

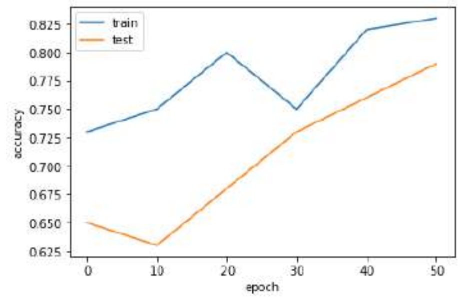


Figure 4. Model accuracy.

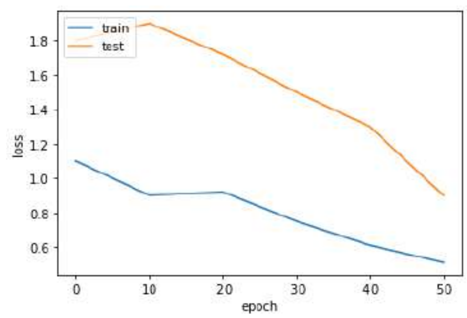


Figure 5. Model loss.

Testing was conducted under the control of the Armbian 4.19 operating system, the following software was used: Python version 3.5.4, version Keras and TensorFlow 2.2.3 and 1.12, respectively.

III. Single Board Computer Test

The following single-board computers were used in this test

- Banana PI M1.
- Orange PI One.
- Khadas VIM.
- Asus Tinker Board.

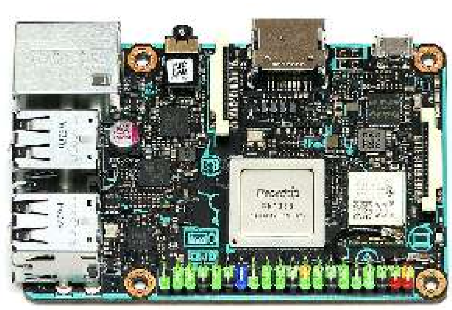


Figure 6. Example of Single Board Computer.

The Banana Pi BPI-M1 and Orange PI One are a business card-sized and low-power single-board computers. They are an open platform devices, that provide possibility for community build decision. Khadas VIM is more powerfully then previous two sample and it is actually a TV box. Khadas Vim is completely open source and all the resources available on GitHub too. Asus Tinker Board is a SBC in an ultra-small form factor that offers class-leading performance while leveraging outstanding mechanical compatibility. The Tinker Board offers makers, IoT enthusiasts, hobbyists, PC DIY enthusiasts and others a reliable and extremely capable platform for building and tinkering their ideas into reality. Prices for devices range from 23$ to 100$ and directly affect the level of their capabilities.

Table I
SINGLE BOARD COMPUTER PROPERTIES

Feature list	Devices under consideration	
	Banana PI M1 [6]	*Orange PI One [7]*
Release date	09.02.2015	29.01.2016
Price	25	23
Length Width	92 mm 60 mm	69 48
CPU	A20	H3
Frequency	1GHz	1.2GHz
Number of Cores	2	4
RAM	1 Gb	512 Mb
GPU	Mali-400 MP2	Mali400 MP2

As follows from the considered characteristics of the devices, the main and most productive solutions are Khadas VIM and Asus Tinker Board, these models can even handle 4K video playback, in addition, these devices have a built-in WI-FI and Bluetooth. It should be noted that all solutions offer an almost identical set of input / output devices, all devices can be connected to an Ethernet network, and also have from 2 to 4 USB ports, there is also a universal input / output interface (GPIO), through this interface you can listen and give commands to any external device, this interface significantly increases the area of use of single-board computers.

Table II
SINGLE BOARD COMPUTER PROPERTIES

Feature list	Devices under consideration	
	Khadas VIM [8]	*Tinker Board [9]*
Release date	12.06.2017	04.04.2017
Price	69	60
Length Width	82 mm 57 mm	79 mm 60 mm
CPU	Amlogic S905X	Rockchip RK3288
Frequency	1.5GHz	1.8GHz
Number of Cores	4	4
RAM	2	2
GPU	Mali-T450MP5	Mali-T764

IV. DEVICE TESTING

After a series of experiments with the corresponding devices, a graph was compiled containing the results of the recognition quality, as well as data on the cost of the device on which the recognition was carried out. ASUS Tinker Board shows the highest speed, beating competitors by 33%. If we consider only the lower segment of solutions, then here is the advantage for Orange PI One.

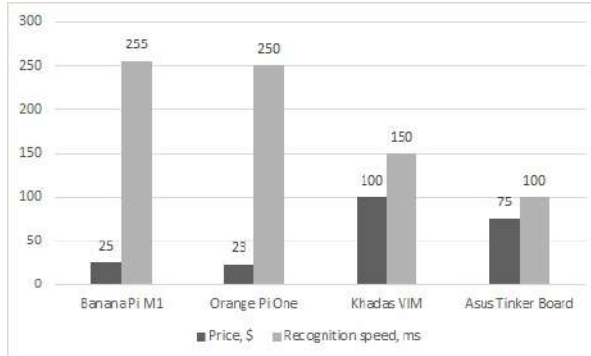


Figure 7. The result of testing the speed of the neural network on the devices.

As a result of testing, the following strengths and weaknesses of these devices can be distinguished.

Banana PI M1

This device is cheaper than all submitted samples, which allows its use in solutions with a limited budget. It has one of the largest communities because, it is used by many enthusiasts to create systems. But the device relatively few additional sources of input / output data, which creates certain problems when connecting a large number of peripheral devices. Have problem with weak equipment for graphical interface.

Orange PI One

The device have a large number of different operating systems, supports various builds of Linux and Android. As well as the previous example, there are relatively few additional sources of input / output data, which creates certain problems when connecting a large number of peripheral devices.

Khadas VIM

Khadas is have a built-in Wi-Fi module, which significantly increases the mobility of the device. High-performance equipment capable of playing 4K video. The presence of built-in flash memory. Main weak is quite a small community around this device, due to the low popularity of this device.

Asus Tinker Board

This device have a large number of I/O sources allows to connect a large number of peripheral devices to it. Availability of complete radiators for additional heat removal from the main elements of the board. High overall performance solutions in many tasks, suitable for use in systems with a graphical interface. wide range of operating systems for this platform, from Linux and Android builds to custom solutions - TinkerOS. It have high heat transfer, which requires additional solutions from the heat sink system.

The paper reviewed the neural network architecture for binary classification of digital images of the thyroid gland. The results of a comparative analysis of four single-board computers in image recognition suggest that the Asus Tinker Board model is most suitable for solving the problem of confocal image recognition. This model has the best rate of image recognition, a sufficient number of additional input / output sources, as well as optimal dimensions.

V. CONCLUSION

In this work, algorithm for constructing a classifier for images of the thyroid gland obtained using confocal microscopy were developed and tested. A comparative analysis of the speed of the developed classifier using several representatives of SBC.

The results of a comparative analysis of five single-board computers allow us to conclude that the Asus Tinker Board model is the most productive model and is suitable for solving the problem of confocal image recognition. This model has the best indicator of image recognition speed, a sufficient number of additional input / output sources, but it is worth noting that this solution is the most expensive. If review each device, can highlight the main features of these solutions: relatively low cost, which allows the use of devices to create home systems or prototypes of more complex solutions; the size of these devices practically does not exceed the size of a plastic card; device performance is sufficient for use as a small server; relatively poorly developed information input / output capabilities, in addition to the hosted GPIO interface, impose some inconvenience when working with them, since not every device has a built-in Wi-Fi or Bluetooth module.

It should be noted that currently there are many devices for various needs, from basic solutions for home use to modern hardware acceleration systems for the operation of trained models of neural networks [10]. Therefore, when developing various systems, an analysis of devices on the market is required. to find the best price / performance ratio. According to the results of testing a model of our own design on a test sample of 14 images, the accuracy of the network can be 83%.

ACKNOWLEDGMENT

The authors are grateful to all those who simply do not intervene.

REFERENCES

[1] Konstantin S. Kurochka, Igor V. Safonau Medical Images Indexing and Retrieval in a Distributed Computing Environment, Journal of Automation and information Science, 2010, vol. 42, no 5, pp 61-71
[2] Zhenghao Shi, Lifeng He, Kenji Suzuki Survey on Neural Networks Used for Medical Image Processing, Int J Comput Sci, 2009, vol 3, no 1, pp 86-100
[3] Konstantin S. Kurochka, Konstantin A. Panarin Algorithm of Definition of Mutual Arrangement of L1–L5 Vertebrae on X-ray Images, Optical Memory and Neural Networks, 2017, vol. 27, no 3, pp 161-169
[4] Bharath Ramsundar, Reza Bosagh Zadex TensorFlow for Deep Learning: From Linear Regression to Reinforcement Learning, Boston, O'Reilly Media, 2018,256 p.
[5] Antonio Gulli, Sujit Pal Deep Learning with Keras: Implementing deep learning models and neural networks with the power of Python, Birmingham, Packt Publishing, 2017, 318 p.
[6] BPI Home. Available at: http://www.banana-pi.org/m1.html (accessed 2018, Dec).
[7] Orange PI. Available at: http://www.orangepi.org/orangepione/ (accessed 2018, Dec).
[8] Khadas.com. Available at: https://www.khadas.com/vim (accessed 2018, Dec)
[9] Asus. Available at: https://www.asus.com/ru/Single-Board-Computer/Tinker-Board/ (accessed 2018, Dec).
[10] N+1. Available at: https://nplus1.ru/news/2018/07/26/edge-tpu (accessed 2018, Nov).

Классификатор конфокальных изображений щитовидной железы на одноплатных компьютерах

Курочка К.С., Демидов А.И.

В статье рассматривается архитектура сверточной нейронной сети для классификации цифровых изображений щитовидной железы, полученных с помощью лазерного конфокального микроскопа. Предлагается способ увеличения возможностей использования лазерного конфокального микроскопа с помощью одноплатных компьютеров для классификации модифицированных тканей. Разработано соответствующее программное обеспечение, адаптированное для использования на одноплатных компьютерах, с ограниченными ресурсными возможностями. Проведен сравнительный анализ одноплатных устройств при решении задач бинарной классификации с использованием рассмотренной сверточной нейронной сети. Факторами отбора устройст для испытаний были выделены их технические характеристики, невысокая цена и доступность платформы на рынке Беларуси. Описаны сильные и слабые стороны при использовании данных устройств. Кратко приведены их характеристики. Предполагается, что приведенные устройства или их аналоги могут быть использованы для построения мобильных и энергоемких систем обработки информации при ограниченных возможностях, как пример обработка конфокальных изображений щитовидной железы.

Received 15.01.2020

Construction of individual geometric 3D model of the lumbar spine of a person based on the analysis of medical images

Tatyana Semenchenia
Sukhoi State Technical University of Gomel
Gomel, Belarus
levts@gstu.by

Konstantin Kurochka
Sukhoi State Technical University of Gomel
Gomel, Belarus
kurochka@gstu.by

***Abstract*—In this article, we will consider the algorithm on the basis of which an application was developed to build a three-dimensional model of the lumbar spine using images DICOM. The model is different from the existing ones so far as it has an intervertebral disk. The created application has been successfully tested.**

***Keywords*—computed tomography, DICOM, STL, 3D model, lumbar spine**

I. Introduction

The task of modern medicine is the study of pathogenic factors, as well as the identification of means to eliminate or weaken pathological processes. However, it seems obvious that in order to foresee the development and consequences of pathological processes, as well as to formulate medical recommendations, it is necessary to fully study both normal and impaired functioning of organs.

The numerical study of physiological and unnatural processes occurring in the human body is currently one of the most relevant and promising areas in scientific research.

Modeling and three-dimensional reconstruction are increasingly used as applied tools in various fields. Such areas include, in particular, medicine. Today, with the emergence of new diseases, it is increasingly difficult for doctors to make the correct diagnoses in the examination of patients, and when conducting high-precision surgery, it is necessary to have as much as possible a detailed idea of the state of internal organs. For example, it is often necessary to accurately construct a three – dimensional model of the human vertebra for the subsequent selection of the optimal implant design [1-3].

In many cases, to establish the diagnosis, the doctor visually examines the images of individual sections of the object obtained during the tomographic examination. However, for some clinical tasks, such as surgical planning, this is not enough [1].

The model provides much more information about the spine than can be obtained by modern measuring instruments.

Studies of the spine required for:

- analysis of the state of the spinal column, under unusual loads and pathological changes [4 – 6];
- choosing a reasonable method of correction [1];
- development of new correction methods [7];
- development of implant designs.

Pathological diseases and injuries are among the most common diseases of the human lumbar spine [8]. The number of diseases of the spine elements in humans is very large, and each case is individual, corresponding to a particular type of pathological changes or injuries [6].

Most of the pathologies of the spinal column are associated with osteochondrosis. It is this disease that most often leads to the progression of other disorders:

- intervertebral hernia;
- disc protrusion;
- sciatica;
- spinal cord compression;
- loss of sensitivity;
- spinal cord infarction.

Surgical treatment of the spine is performed to prevent compression of the spinal cord and the development of complications in osteochondrosis and intervertebral hernias.

Surgical treatment of diseases and injuries of the spine is carried out with surgical treatment:

- of osteochondrosis is used in the presence of the following complications: hernias and protrusions of the intervertebral discs, spondylolistea, secondary spinal stenosis.
- for traumatic lesions of the spine at all levels, including the consequences of spinal injuries and comprehensive treatment after “unsuccessful” previous operations.
- for cancer with damage to the spinal column, spinal cord, spinal roots.
- of a number of congenital anomalies of the spine.

Currently, the creation of three-dimensional models describing the lumbar spine of a person, and the study of these models will justify the methods of treatment of diseases or injuries, which determines the relevance of research.

Thus, the automation of the process of building a 3D model of the lumbar spine of a person is relevant. The significance is due to the fact that the development of materials and technologies for the production of prosthesis structures, implants, orthoses and other medical special technical means, as well as the combination of different methods of treatment requires research to confirm or refute this combination [9].

II. The main stages of building a three-dimensional model of the lumbar spine of a person

Currently, the treatment of diseases of the musculoskeletal system is impossible without the use and analysis of relevant biomedical images, a special place among which is occupied by the results of computed tomography, and, in particular, computed tomography of the human spine [10 – 13].

The result of computed tomography is a three-dimensional matrix of numbers representing the density of different parts of the object. The values of the matrix elements depend on the type of tissue under study and lie in a certain range, allowing to obtain three-dimensional images of the internal structures of the object under study [14].

Three-dimensional modeling allows to measure angles, lengths and diameters of different anatomical structures in the initial state and spatial transformations. This is essential in assessing the feasibility of reconstructive operations, in determining the most correct method of operation.

At design of a program complex it is possible to allocate several key stages of work (figure 1).

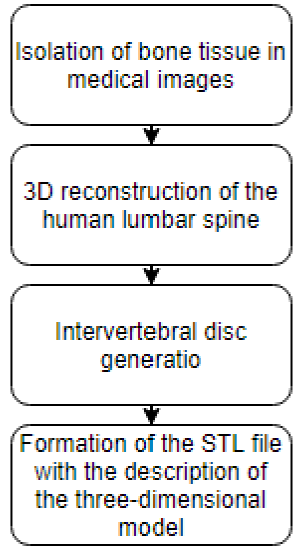


Figure 1. Key stages in the construction of a three-dimensional model of the lumbar spine of a person using computed tomography images.

The first step is the selection of bone tissue in medical images. This step will allow you to leave only significant information on the images.

The second stage is a three-dimensional reconstruction of the lumbar spine, as well as the generation of intervertebral discs.

The final stage is the formation of an STL file with a description of a three-dimensional model for further visualization and strength studies in CAD systems.

To perform the first stage of work, images were segmented [15], since the image elements corresponding to bone tissue in computed tomography images contrast sharply in comparison with other anatomical objects (figure 2), which is connected with the peculiarities of color representation in computed tomography image, which stores the value of the x-ray density of the captured object [16].

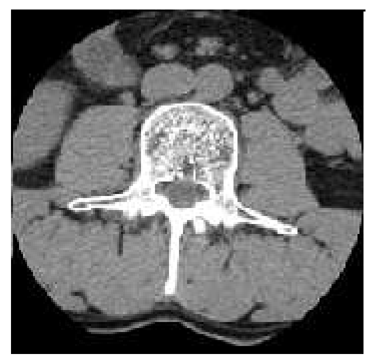


Figure 2. Sample image computed tomography.

To highlight the necessary areas in the images (bone tissue) for further processing and to eliminate noise in the form of a background, threshold transformations are used, such as:

$$T = T(x, y, p(x, y), f) \tag{1}$$

where f is an image; $p(x, y)$ is a local characteristic of the point (x, y) of the image.

The image $g(x, y)$ obtained as a result of the threshold transformation is determined by the formula:

$$\hat{g}(x, y) = \begin{cases} 1, & \text{if } f(x, y) \geq T; \\ 0, & \text{if } f(x, y) < T \end{cases} \tag{2}$$

where T is a threshold value; $f(x, y)$ is a pixel value; $g(x, y)$ is a value of the function that determines the belonging of the pixel to the object.

Thus, the pixels assigned the value 1 correspond to objects, and the pixels with value 0 correspond to the background [17].

Figure 3 shows the original images of spinal slices on computed tomography.

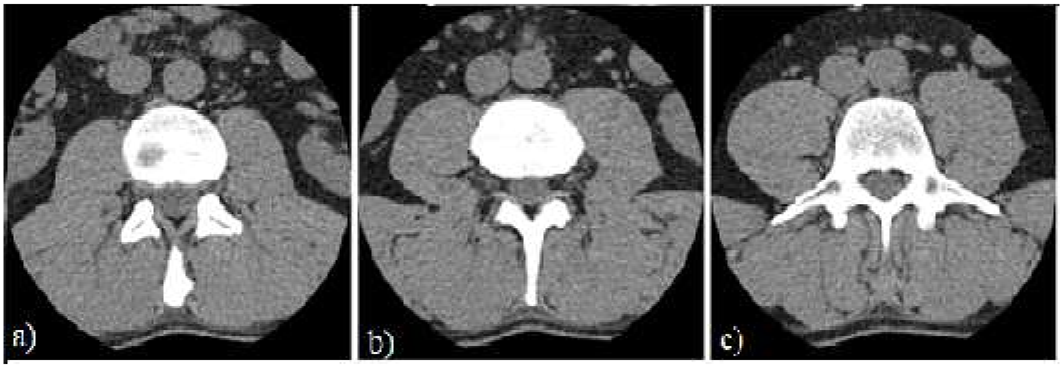


Figure 3. Image to threshold processing method.

After the threshold transformation of the image (figure 3 a, b, c) look as shown in figure 4 a, b and c, respectively.

Transformations allow the detection of bone tissue by selecting the optimal density (brightness) value, which is used as a threshold and allows you to distinguish the required objects from the background.

The image may be damaged by noise and interference of various origins. Averaging filters were used to suppress noise in the images. The idea of using averaging filters is to replace the original values of the image elements with the average values of the filter mask. This reduces the sharp transitions of brightness levels.

The use of averaging filters is to suppress "nonexistent" details in the image. By “nonexistent” we mean a set of pixels that are small compared to the size of the filter mask.

For visualization of the intervertebral disc, which is absent in the images of computed tomography, it was completed. At the first stage, images were found that lack a vertebral body (figure 5), and then images of the previous and subsequent vertebra with the body. The second stage was the reconstruction of the intervertebral disc.

An algorithm was written to restore the intervertebral disc, which consists of the following steps.

To obtain a clear contour of the objects in the processed image, the contour segmentation algorithm, the Laplace operator, is used. Using the Laplace operator [21] on each image, we find the contours of the spine, which in the vast majority of cases are closed curves. Figure 6 shows an example of the operation of the implemented algorithm.

The developed algorithm consists of the following steps:

a) the selected borders are viewed from top to bottom in layers (the layer in this case is the points that have the same y coordinate value);

b) if there is a distance between adjacent boundaries, then all layers already covered form a region and are recorded separately; at the same stage, after selecting the area, this area is divided, passing from left to right:

- the selected area is passed from left to right; if there is a gap between adjacent boundaries, then all layers that have already been covered will form a region and are recorded separately; if there is no clearance, then scanning continues until the region ends;
- if more than one contour was found in the new area being viewed, then they replace it;

c) if there is no clearance, then viewing continues until the layers being viewed are over.

The first stage is the recognition of the contour of the vertebral body in images that were detected earlier (the previous and subsequent vertebra with the body), as well as the formation of a list of points on the contours that are the coordinates of the triangle vertices.

The second step is to determine the contour with the least number of vertices. Figure 7 shows an example of two contours (upper contour and lower contour), where the upper contour is the required contour.

Next we connect the first point of the upper contour with the closest point of the lower contour. The remaining points are connected, as shown in figure 8.

As a result, you may get several cases:

- when the number of points on the contours coincides, then connect the vertices, as shown in figure 9, a.
- when the number of points on the contours differs by one, then connect the vertices, as shown in figure 9, b.
- when the number of points on the contours differs by two or more, then connect the vertices, as shown in figure 9, c.

For the three-dimensional reconstruction of the lumbar spine of a person the algorithm of triangulation of the surface «Marching cubes» was chosen as a basis [18]. The algorithm produces a partitioning of a region of space containing the surface of the cubic cell and approximarely the intersection of surface and every cubic cell. As a result, on the basis of the coordinates of

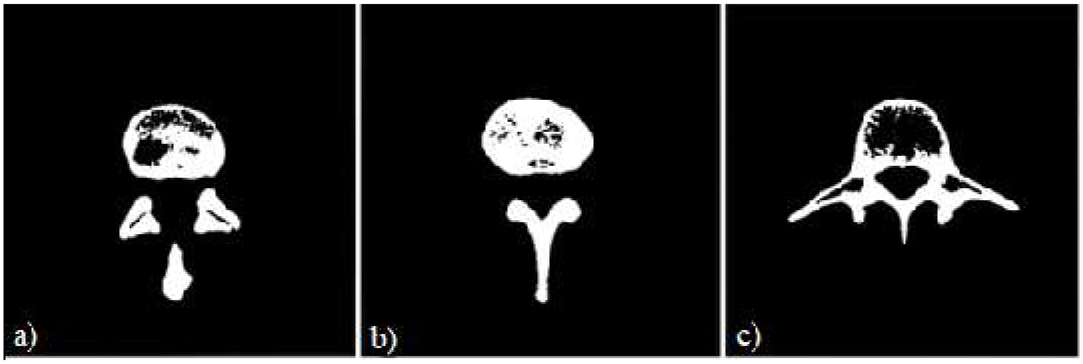


Figure 4. Image after threshold processing method.

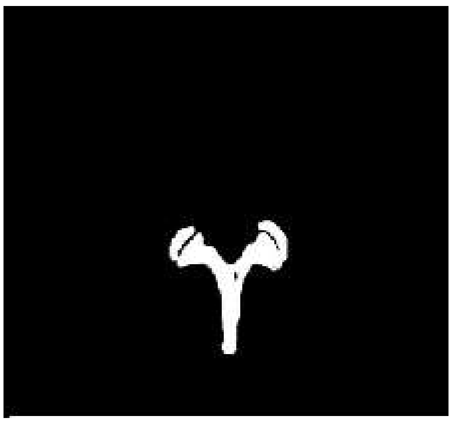


Figure 5. The image on which the vertebral body is missing.

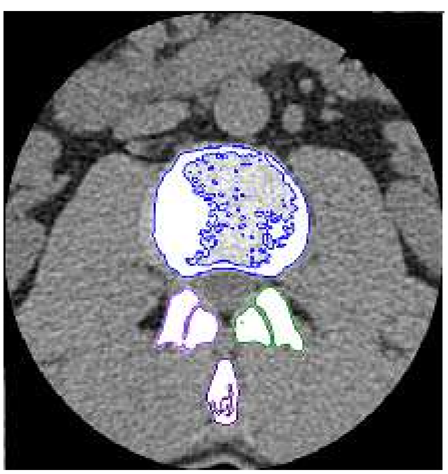


Figure 6. Isolation of individual sections of the vertebra, using the developed algorithm.

the points of bone tissue in the images of computed tomography, a list of the coordinates of the vertices of triangles that describe the three-dimensional surface, and the unit normals to them was formed.

Based on the obtained three-dimensional model of the lumbar spine of a person, an STL file is generated. Each

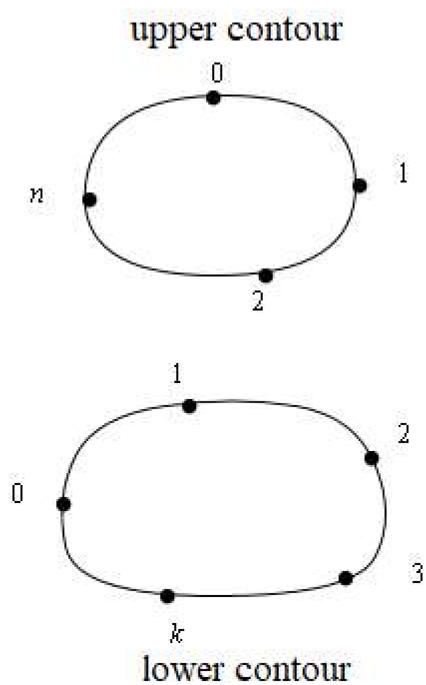


Figure 7. Contours of the vertebral bodies.

resulting triangle is described by twelve 32 bit floating point numbers: three numbers for the normal and three numbers for each of the three vertices for the X/Y/Z coordinates.

III. Result

During the study, an algorithm was developed for constructing a 3D model of the lumbar spine of a person, which includes vertebrae and intervertebral discs, based on the analysis of DICOM images.

Figure 10 shows the result of visualization of a geometric 3D model of the human lumbar spine. The constructed computer models can be the basis for the development of the method of preoperative forecasting of

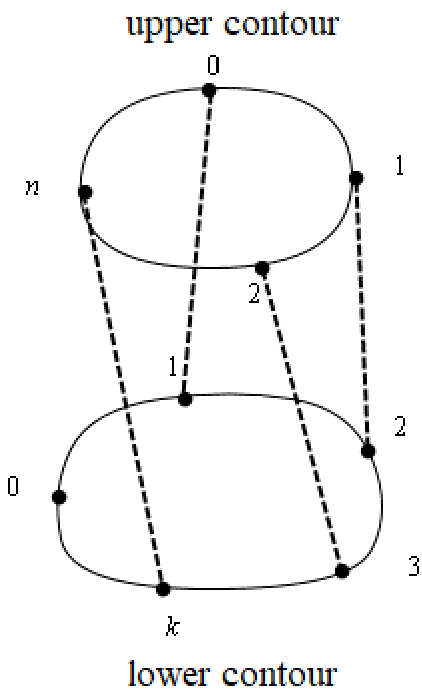


Figure 8. Connecting contour points.

the spine in various pathological formations, correction and prosthetics.

The developed software package for the construction of an individual geometric 3D model of the lumbar spine of a person based on the analysis of medical images differs from the existing possibility of building a three-dimensional model, which includes the vertebrae and intervertebral discs.

Also, the software package allows you to export models in STL format for further numerical experiments in computer-aided design systems aimed at obtaining information about the behavior of the lumbar spine of a person under the influence of various loads, as well as the impact on its stress-strain state of implants and other supporting devices.

Individual geometric 3D models of lumbar calving of the human spine can be used:

- to improve the assessment of spatial relationships of organs and structures;
- as a good example in the educational process (use of the model in the educational process for students and interns of neurosurgeons [1]);
- in scientific research (for example, studies to determine the risk of diseases of the musculoskeletal system [19]);
- to select the optimal fixing structure, options for attaching it to the vertebra;
- predict the effects of implantable elements (materials and structures that can replace the damaged item or some segment of the spine and to perform the functions assigned to the replaced element/segment) n the vertebral part of the patient;
- planning of some treatments (virtual simulation of radiation therapy, surgical navigation);
- assess the development of various diseases and pathologies [20] (for example, such as osteoporosis);
- to assess the normal and pathological anatomy of a particular patient, to predict surgical intervention, possible complications, and their prevention [2].

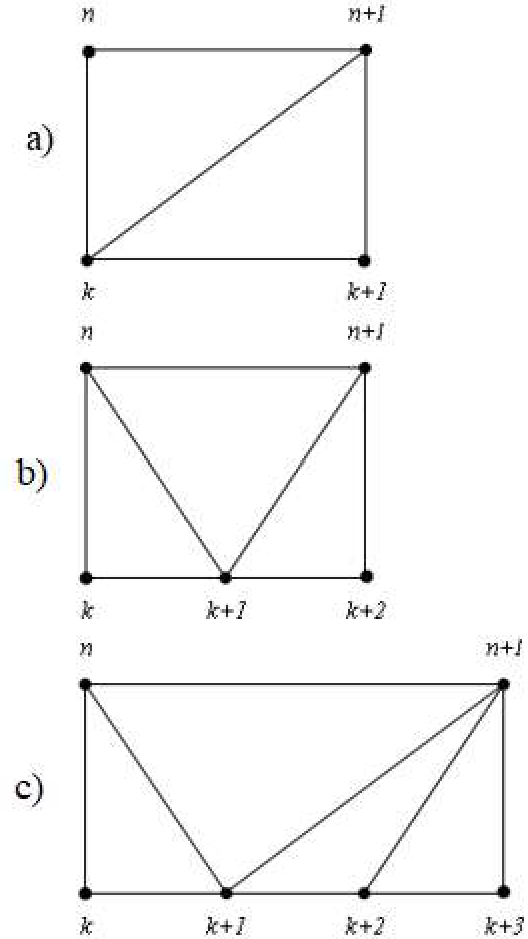


Figure 9. Vertex connection example.

IV. Conclusion

As a result of the research, the program complex «Vertebra 3D» was developed, which has the following advantages:

1. possibility to build an individual geometric 3D model of the human lumbar spine, which includes vertebrae and intervertebral discs;

2. export of a three-dimensional model to the STL file format, which allows to directly materialize the

Figure 10. 3D model visualization.

reconstructed objects in the software complexes of three-dimensional modeling or complexes of finite element analysis, which allow to simulate the functional properties of the object and the implant, necessary to predict the impact of implantable elements or planning treatments;

3. the presence of means of visualization of the three-dimensional model;

4. cross-platform operation of the software package.

The use of the developed software package for constructing an individual geometric 3D model of the lumbar spine of a person contributes to the optimization of the work of medical personnel, reducing the load on medical workers and reducing the impact of the human factor on the accuracy of the results obtained in the analysis of the spine, anatomical studies of vertebral segments or the development of implant design.

References

[1] M. V. Olizarovich, Computer modeling of the lumbar vertebrae in the preoperative period, Health and ecology problems, 2013, no. 12, p. 125-129.
[2] O. N. Yamshchikov, Computer modeling in traumatology and orthopedics, TSU Bulletin, 2014, no. 6, p. 1974-1979.
[3] M. P. Burykh and R. S. Voroshchuk, Voxel anatomical modeling of human internal organs, Clinical Anatomy and Operative Surgery, 2006, no. 5, p. 115-118.
[4] V. P. Balandin and V. S. Tuktamyshev, Computer simulation of the effect of intra-abdominal pressure on the loading of the spinal column, Master's journal, 2015, no. 2, p. 229-237.
[5] N. Kh. Tsurova, Biomechanical research and modeling of the lumbar spine segment, News SPEBGTU "Lati", 2008, no. 3, p. 73-78.
[6] N. Kh. Zinnatova, Biomechanical method for diagnosing the state of the spine in health and pathology, Proceedings of the Southern Federal University. Technical science, 2009, p. 108-113.
[7] S. V. Orlov, R. L. Sedov and N. D. Bobarykin, Mathematical modeling of spinal instability and stabilization methods, Russian Journal of Biomechanics, 2010, no. 3, p. 36-46.
[8] P. V. Netsvetov, A. T. Khudyaev and G. V. Dyachkova, X-ray characterization of fractures of the thoracic and lumbar spine, according to computed tomography, at various stages of treatment with the method of transpedicular fixation, Genius Orthopedics, 2007, no. 1, p. 69-75.
[9] O. V. Veretelnik, N. A. Tkachuk and I. B. Timchenko, Mathematical and numerical study of various designs of orthoses for cervical spinal spondylode, Bulletin of the National Technical University "KPI", 2017, no. 29, p. 27-37.
[10] A. N. Chuyko, An approximate analysis of the anatomy, mechanical characteristics and stress-strain state of the human spine, Injury, 2014, vol. 15, no. 6, p. 100-109.
[11] A. M. Zharnov and O. A. Zharnova, Biomechanical processes in the vertebral motor segment of the cervical spine during its movement, Russian Journal of Biomechanics, 2014, vol. 18, no. 1, p. 105-118.
[12] K. S. Kurachka and I. M. Tsalka, Vertebrae detection in X-Ray images based on deep convolutional neural networks, INFORMATICS 2017: Proceeding of the 2017 IEEE 14th International Scientific Conference on Informatics, Slovakia, 2017, pp. 194-196.
[13] K. S. Kurachka, V. V. Komrakov and N. N. Masalitina, The automated classification system for lumbar spine anatomic elements, Nonlinear Dynamics and Applications, 2017, vol. 23, p. 127-134.
[14] M. Ya. Marusina and A. O. Kaznacheyeva, Modern types of tomography. Tutorial, SPb: SPbSU ITMO, 2006.
[15] Yu. A. Ogluzdina, Image segmentation algorithms, Youth Scientific and Technical Gazette, 2014, no. 8, p. 11-17.
[16] E. V. Doronicheva and S. Z. Savin, Medical image recognition methods for computer-aided diagnostic tasks, Modern problems of science and education, 2014, no. 4, p. 30-38.
[17] R. Gosales and R. Woods, Digital image processing, Moscow: Technosphere, 2005.
[18] W. E. Lorensen and H. E. Cline, Marching Cubes: A high resolution 3D surface construction algorithm, SIGGRAPH, 1987, vol. 21, no. 4, p. 163-169.
[19] E. L. Tsitko, K. S.Kurochka, N. N. Masalitina, I. N. Tsalko, I. N. Komrakov and E. V. Tsitko, The radiometric estimation of kinematics of the lumbosacral spine segments of osteochondrosis with the help of the software «Volot», Health and ecology problems, 2017, no. 4, p. 35-41.
[20] A. Yu. Mushkin, E. V. Ulrich and I. V. Zuev, Spine biomechanics in health and pathological conditions: the main aspects of research, Spinal surgery, 2009, no. 4, p. 53-61.
[21] W. Pratt, Digital image processing, M.: Mir, 1982.

Построение индивидуальной геометрической 3D модели поясничного отдела позвоночника человека на основе анализа КТ-изображений

Семенченя Т.С., Курочка К.С.

Предложен алгоритм, позволяющий по результатам компьютерной томографии поясничного отдела позвоночника человека, представленной в виде DICOM-файлов, сформировать трёхмерную модель, отличающуюся от существующих сгенерированным межпозвоночным диском. На основе алгоритма создано соответствующее программное обеспечение и проведена его аппробация.

Received 15.01.2020

The Problem of Understanding the Content of a Complex Question by BERT Model Construction

Vladimir Kopeyko
Belarusian State University
Minsk, Belarus
volodyakopeyko@gmail.com

Alexander Nedzved
United Institute of Informatics Problems
Minsk, Belarus
Anedzved@bsu.by

***Abstract*—This paper describes the application of the BERT natural language processing model to solve the problem of understanding the content of a complex question. In this paper there are description of the model itself, basic concepts of solution and model for training for classification. Common task are represented by example of the using of this model for practical analysis of the data. Quality results is based on preprocessing such data.**

***Keywords*—natural language processing, automated answer, language analysis, BERT model**

I. Introduction

The task of automated answering of the question is one of the important tasks for natural language analysis and processing. Modern methods for its solution allows to take good results for this task, but they are quite accurate only for limited sets of clear answers. There are problems with answers to questions that involve the expression of a subjective opinion or recommendation. The structure of the questions also has some difficulty for automated understanding. In real life, question can be represented by different forms. It can consist of only a few words or a complex structure consisting of several sentences. The understanding of such questions in natural language processing tools is possible to solve by methods of machine learning[1]. One of the basic trends in natural language processing is the use of models that are trained by simple problems. In this case, machine learning methods can be applied to other tasks through a little refinement. In addition, such models are trained on large datasets. It allow to the full extended sets, to use of an extensive context sets for solving question problems in natural language processing[2]. One of the most popular models of this type at the moment is the BERT model.

II. Creation of BERT model description

In the field of computer vision, the value of transfer training has been repeatedly demonstrated by a pre-trained neural network model. This model has good results for solving a well-known problem. For example the ImageNet with using fine-tuning can serve as the basis for a new model that solves a completely different problem. In recent years, there are many publications that describe a similar methods. It can be useful in many natural language processing problems. One such pre-trained model is the BERT model[3]. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model for solving the problems of processing natural language with open source code, developed by researchers at Google in 2018. Most of the models for solving the problems of natural language processing considered the text during the training either from left to right, or from right to left. This unidirectional approach works well in the task of generating sentences. Usually, it can predict the next word. Then add this word to the sequence and predict the next word until the sentence is not complete. But the BERT is trained by two ways (Fig. 1). It means a deeper understanding of the language context that compared to unidirectional language models. With a change in the approach to text processing, the tasks of the model trains also change. The BERT no predict the next word in a sequence. It simultaneously solves two problems: Masked LM and Next Sentence Prediction [4]. The masked LM (MLM) consist of next basis: 1) sequences of words are used as input, 2) words are replaced by the [MASK] token in each 15 procents. The model tries to predict the initial meaning of the masked words based on the context provided by the other unmasked words in the sequence. Masking of word is meaning the model looking in both directions. After it the full context of the sentence interpreted as left and right surroundings for prediction of the masking word[5][6]. An example is shown in Table 1. The word "Wimbledon" was taken as a masked word. As a result, MLM predicts "Wimbledon" as a mask with a probability of 33

The Next Sentence Prediction (NSP) is the second task of word prediction. The main task of it is prediction of the second sentence in a pair as a continuation of the first. During BERT training, the model receives pairs of sentences as input and learns to predict whether

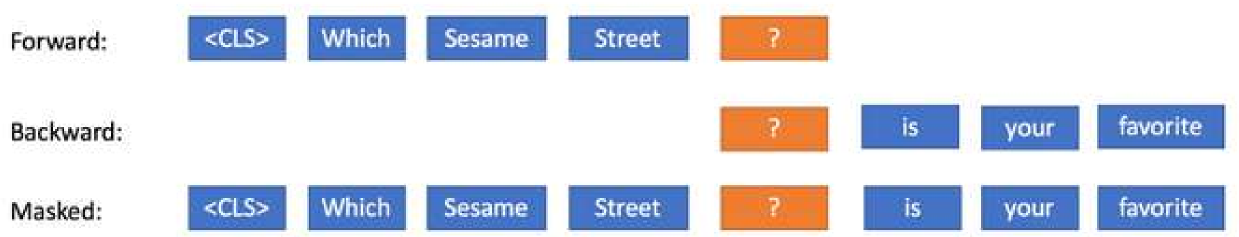


Figure 1. Example of bidirectional processing of sentence by BERT like pipeline.

Table I
EXAMPLE OF MLM ANALYSIS

Sentence	Result
Over 303,000 glasses of Pimm's are served during the [MASK] tennis championships?	32.9 Wimbledon 8.0 table 7.7 world 3.0 professional 2.3 Table

the second sentence in a pair as the next sentence in the source document. During training, 50 Thus, during

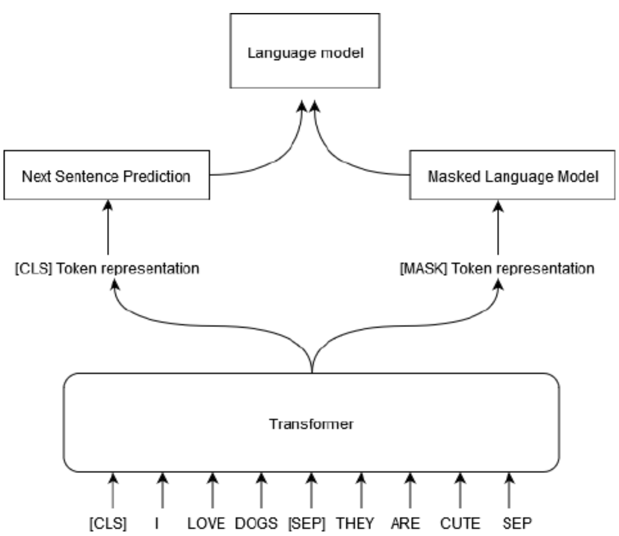


Figure 2. BERT model training scheme

training, the BERT, MLM, and NSP models are trained simultaneously to minimize the combined loss function for two tasks.

III. PREDICTION IN SENTENCE BY THE BERT MODEL

To solve the problem of improving understanding of complex question, the use of the BERT model is effective. An example of such a task can be: each question-answer pair has 30 labels of mask. Some words describe the question itself, and the other part describe the answer. Each label can have a continuous value in the range [0, 1]. At the input there is a labeled training dataset, consisting of many thousand question-answer pairs and the values of all labels for each pair. A description of pair of question and answer:

- qa id;
- question title;
- question body;
- question user name;
- question user page;
- answer
- answer user name
- answer user page
- url
- category
- host

It is necessary to determine the same labels for all question-answer pairs from the test dataset. [1] An example of a description of some tags:

- "question expect short answer" - degree of expectation of a short answer to a question;
- "question interestingness others" - degree of interest of the question for others;
- "answer helpful" - the usefulness of the answer;
- "answer level of information" - informative response;
- "answer relevance" - the relevance of the answer;
- "answer satisfaction" - satisfaction with the answer;
- "answer type instructions" - is the answer an instruction;
- "answer well written" - how well the answer is written.

There are independence of labels among themselves. Figure 3 shows the correlation matrix, where the darker the color of the cell correspond of the more closely relation of labels. It mean that those cells depend on each other. The pronounced diagonal elements that determine the dependence of the label itself. It can be seen that the greatest correlation is between the "question interestingness others and "question interestingness self" labels. Thus, the values of labels determine the degree of interest of the question for the most asking question and the degree of interest of the other question. They are most correlated in this dataset.

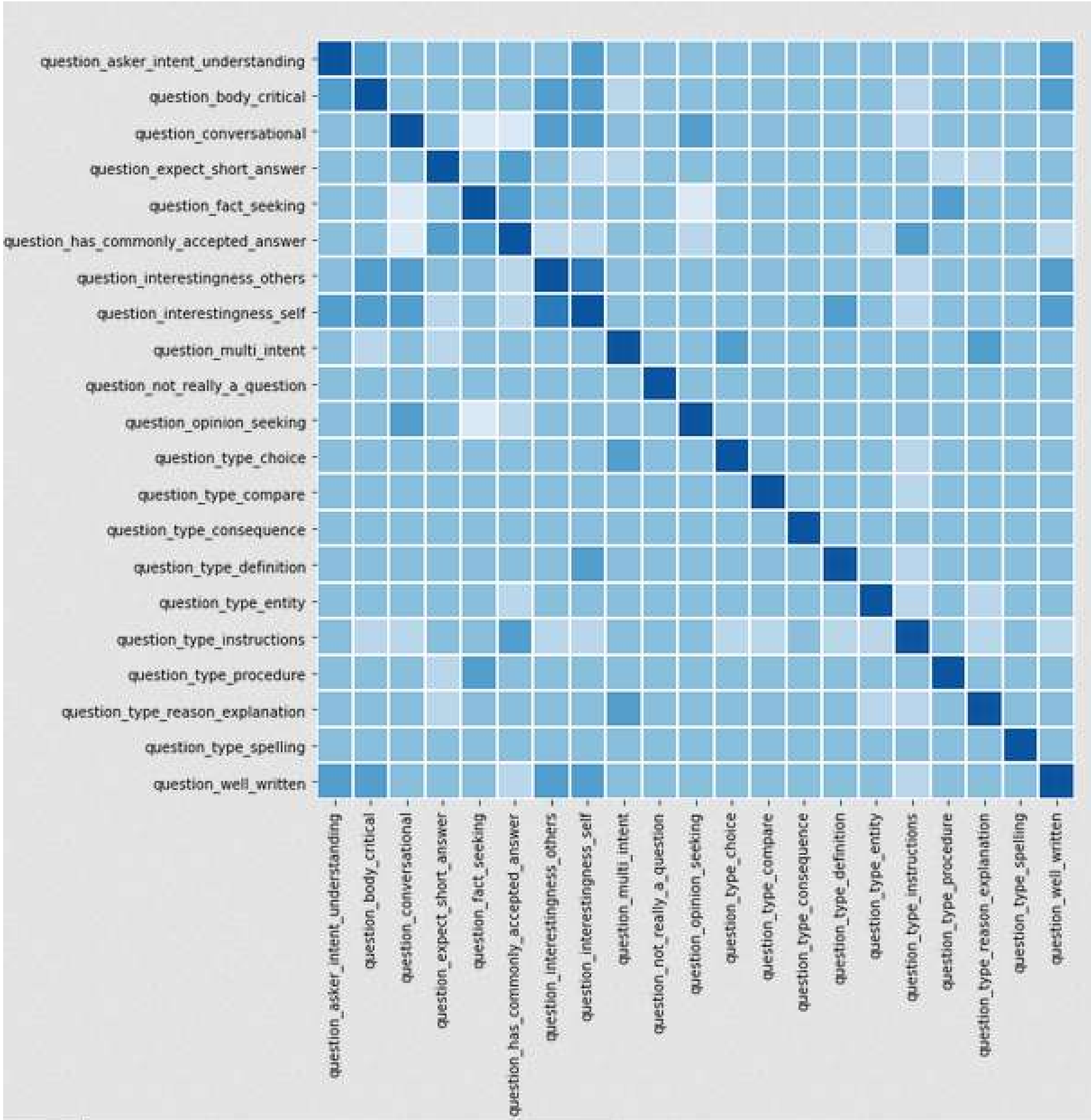


Figure 3. Label correlation matrix for test dataset.

Thus, during training, the BERT, MLM, and NSP models are trained simultaneously to minimize the combined loss function for two tasks. The solution to this problem consists of two stages: 1) training the BERT neural network, 2) its refinement (settings) by using of additional layers that needed to solve a specific task. For the first stage, a pre-trained model is used. However, it is necessary to prepare data accordingly texts of questions and answers in order of using it. The BERT is training on the NSP solution. The sentences are divided into tokens by this context. In this case, tokens are words. Then each sentence is framed by the [CLS] and [SEP] tokens, where [CLS] indicates the beginning of the sentence, [SEP] - its end. Obtained data after preliminary processing is going into the BERT model. Output data are represented as vector of size 728 elements (in the case of the usual implementation of the model) or 1024 elements (for the case of the extended version). Addition an additional layer is used as an ordinary fully connected layer. Every output vector has size 30 elements. It corresponds to a certain label, that must be predicted.

Table 2 shows examples of input pairs of questions and answers. Table 3 shows the result of using the BERT model to identify labels that explain questions

Table II
EXAMPLES OF INPUT PAIRS OF QUESTIONS AND ANSWERS

qa id	question title	question body	answer
39	Can I get back my passport after applying for a Schengen visa?	I need to travel, but the processing time takes up to 15 days according to the website.	Out of first hand experience in two different EU consulates (Italian and Greek), they actually return the passport to you and on the visa issuance date they will ask for it and they will post it there...
46	Why was Ryuk tied to Light's Death Note?	In How to Use: XIII, in the second point it says The god of Death always remains with the owner of the Death Note. indicating that a Shinigami will remain with the owner of their Death Note	According to How to use II: The owner of the note can recognize the image and voice of the original owner, i.e. a god of death. Therefore the only explanation for Light being able to see Ryuk

Table III
IDENTIFY LABELS EXPLAIN QUESTIONS AND ANSWERS IN BERT

qa id	question asker intent understanding	question body critical	question conversational	question except short answer	question fact seeking	question has commonly accepted answer	question interestingness others
39	0.927	0.563	0.192	0.53	0.687	0.556	0.658
46	0.899	0.58	0.006	0.77	0.788	0.926	0.569

and answers in the task of automatically understanding complex questions.

IV. CONCLUSIONS

The BERT model is the best solution for solving natural language processing problems by machine learning algorithms. Its characterised by availability and quick setup. It allows the use of this model for a wide range of practical applications. In addition, the pre-trained model include about 2.5 billion words. It is also an important advantage of using this model for practical purposes. The obtained results indicate applicabilition of this model for solving of problem of complex question construction.

REFERENCES

[1] Z. Chen, H. Zhang, X. Zhang, and L. Zhao Quora question pairs. 2018, Re-trieved from http://static.hongbozhang.me/doc/Quora.pdf
[2] Minh-Thang Luong, Christopher D. Manning, Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. 2016, Re-trieved from https://arxiv.org/pdf/1604.00788.pdf
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019, Re-trieved from https://arxiv.org/pdf/1810.04805.pdf
[4] Kevin Clark, Urvashi Khandelwal, Omer Levy Manning What Does BERT Look At? An Analysis of BERT's Attention. 2019, Re-trieved from https://arxiv.org/pdf/1906.04341.pdf
[5] Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, Songlin Hu, Mask and Infill : Applying Masked Language Model to Sentiment Transfer. 2019, Re-trieved from https://www.ijcai.org/proceedings/2019/0732.pdf
[6] Noah A. Smith Contextual Word Representations: A Contextual Introduction. 2019, Re-trieved from https://arxiv.org/pdf/1902.06006.pdf
[7] Yifu Sun, Haoming Jiang, Contextual Text Denoising with Masked Language Models. 2019, Re-trieved from https://arxiv.org/pdf/1910.14080.pdf
[8] Wei Shi, Vera Demberg Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. 2019, Re-trieved from https://www.aclweb.org/anthology/D19-1586.pdf
[9] Felipe Alfaro, Marta R. Costa-jussà, José A. R. Fonollosa BERT Masked Language Modeling for Co-reference Resolution. 2019, Re-trieved from https://www.aclweb.org/anthology/W19-3811.pdf
[10] Jan Christian Blaise Cruz Evaluating Language Model Finetuning Techniques for Low-resource Languages. 2019, Re-trieved from https://arxiv.org/pdf/1907.00409.pdf
[11] Sneha Chaudhari, Gungor Polatkan An Attentive Survey of Attention Models. 2019, Re-trieved from https://arxiv.org/pdf/1904.02874.pdf
[12] Ashish Vaswan, Noam Shazeer, Niki Parmar, Jakob Uszkoreit Attention Is All You Need. 2017, Re-trieved from https://arxiv.org/pdf/1706.03762.pdf

Использование BERT модели в задаче понимания содержания сложного вопроса

Копейко В.С., Недзьведь А.М.

Данная работа посвящена описанию применения модели обработки естественного языка BERT для решения задачи понимания содержания сложного вопроса. Сначала вводится описание самой модели, ее основные концепции и задачи, на которых модель изначально тренируется. Затем рассматривается пример использования этой модели в практических целях, анализ исходных данных и описание процесс их предобработки.

Received 14.01.2020

An approach combining general and highly specialized semantic analysis in DLP systems

1st Dmitry Purtov
Volga State Technological University
Yoshkar-Ola, Russia
idmitry.purtov@gmail.com

2nd Irina Sidorkina
Volga State Technological University
Yoshkar-Ola, Russia
igs592000@mail.ru

***Abstract*—The paper proposes an approach to combining general and highly specialized semantic analysis. The analysis of the main problems of the semantic analysis of the text is carried out, as well as an approach is proposed in which the semantic analysis in DLP systems immediately spreads to the entire protected system. The presented approach will allow to gradually accelerate the work of the DLP system, as well as analyze the result of the work of semantic analysis to evaluate and maintain semantic analysis in an up-to-date state**

***Keywords*—semantic analysis, DLP system, morphological reference book, visualization of semantic analysis.**

The development of modern computer technology has allowed the creation of systems for the analysis of not only structured data, but also data presented in natural language. Typically, automatic text analysis systems use the following analysis steps: morphological, syntactic, and semantic. There are many different tasks where automatic text analysis is needed, such as in DLP systems.

DLP (Data Loss Prevention) — It is a system designed to prevent confidential information from leaking outside the corporate network. This system is built on the analysis of data streams that go beyond the corporate network. In the event of a certain event characterizing the transfer of confidential information, the system either blocks such a transfer or sends notifications to the operator. A typical operation scheme of DLP-systems is presented in Figure 1.

- Means of intercepting information transmitted via external channels (outside the protected automated system). This category includes drivers for controlling the printing of information, drivers for controlling connected devices, firewalls that control network traffic, etc. [1].
- The categorizer that makes up the core of the DLP system. His work is to analyze the transmitted information, as a result of which the category is uniquely determined (degree of information confidentiality) [1].
- Means of response and registration. Based on the degree of confidentiality determined by the categorizer, the DLP system responds in accordance with the system settings and blocks the transfer of confidential information, or the security administrator is alerted (signaling) about unauthorized transmission (leakage) of information [1].

In such systems, semantic analysis is used in the stream analysis of text data and is part of the categorizer. Semantic analysis helps to determine whether there is confidential information in the text, relying directly on the contents of the text, its meaning, and not on specialized labels (vultures) that may or may not be deleted at all. The meaning of the text refers to the text from the point of view of a person, that is, obtaining facts that are clearly present in the text and revealing the hidden meaning from the text if it is present [2].

Semantic analysis is closely related to structural analysis. Both there and there are dependencies of words, analyzes the connection, their strength. But if structural analysis takes into account only the language rules for constructing sentences, word dependencies through parts of speech, then semantic analysis takes into account the meaning of words to all this. Without semantic analysis, an analysis of the text can make a mistake, since the structural representation of the text can have several representations and, in addition, may not convey meaning even if it is correctly constructed from the point of view of language rules. It is worth noting the fact that since the main task of semantic analysis is understanding the meaning, we can conclude that it cannot exist without the morphology of the language. Morphology here refers to various morphological representations of the word, the so-called word forms. To store word forms, various reference books are used. Reference books store signs and words by which the context is identified with subsequent understanding of the meaning of the text.

In semantic analysis, many features arise that cause problems in the implementation of semantic analysis. Here are some problems that need to be addressed in semantic analysis:

Knowledge of context. Words, or rather their meanings can vary greatly depending on the context. The same words, sentences can have completely different meanings, depending on the context of their use.

Various sentence structures. If we consider the natural language in a broad sense, we can see that Latin, Cyrillic

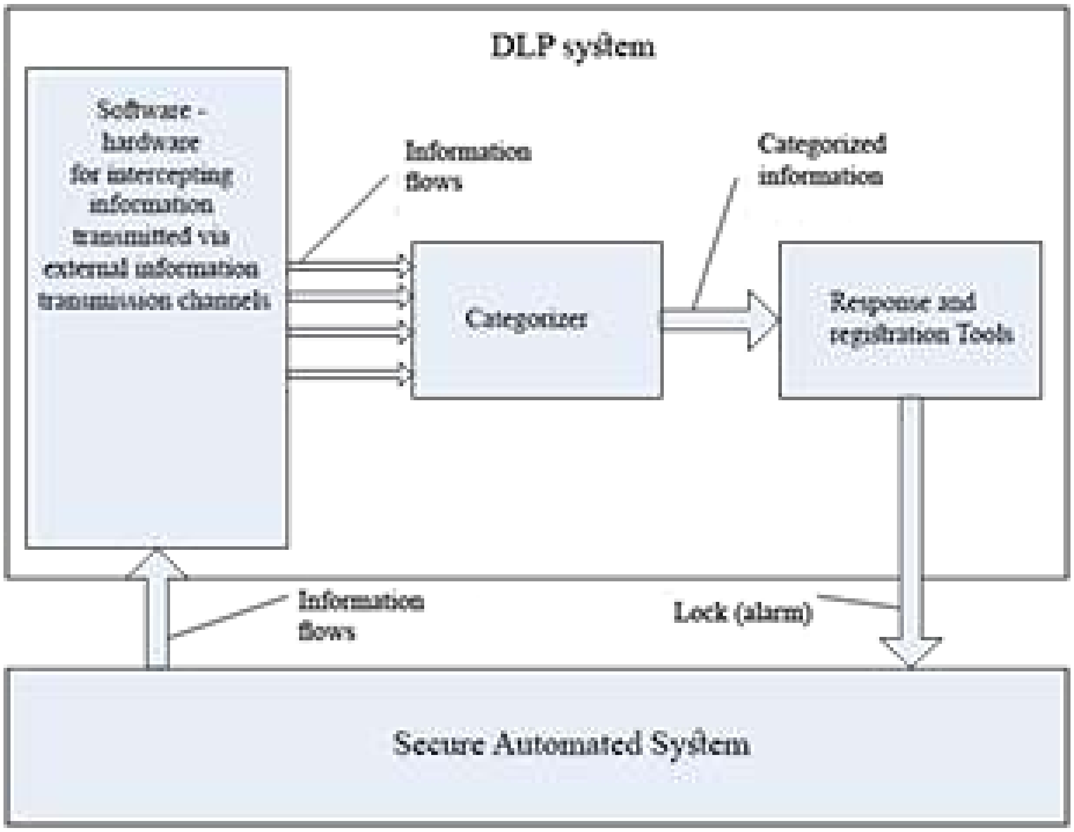


Figure 1. DLP system operation diagram.

languages have their own structure. Data fact does not allow relies on the structure to understand the meaning of the proposal. Yes, and the same sentences can be rewritten so that their structures are different, but the meanings have not changed.

Interpretation problem. One and the same subject can be described in completely different words. In this case, even when using other words, the general meaning has not changed.

The emergence of new words. Over time, new words appear in circulation that can describe existing things or completely new ones.

Contradictions. In a number of languages, it is possible to use seemingly unrelated words to convey a certain meaning. Basically, this phenomenon is observed in works of art when compared, etc.

Ellipses. These are sentences in which words are omitted as their presence is implied in relation to the context.

Semantic analysis has many problems that are quite difficult to solve. Suppose that there are already some solutions to problems, but all of them are not ideal, and there will not be an ideal solution. Many decisions are based on formal grammars. The main representatives of this approach can be called Melchuk I., Chomsky N. and others [3]. At the heart of formal grammar is an attempt to create a new universal language based on mathematical representation [3].

When using semantic analysis in DLP systems, the speed and quality of analysis, or rather the balance between them, is very important, since the analysis takes place in a stream form and these two criteria are mutually exclusive. If we consider semantic analysis in general, we can see that there are systems that use semantic analysis to understand any text in a natural language, such systems are more universal, but slow, as well as highly specialized analysis, which is quite fast, but effective in only one context . DLP systems usually use large digging, where there are many subdivisions, departments, etc. Each of these links usually works strictly with a certain type of data. For example, the economic department with financial documents, and the legal with legal, etc. To increase the speed of analysis of documents from such units, it is more efficient to use highly specialized analysis since the specifics and the context of the information that needs to be checked are already known. But do not forget about units where there is no clear context of information. They can get any kind of information. In such cases, a combination of analysis methods is better. Combination gives a better result, although it complicates the whole system than choosing one analysis method. When combining, the main thing is to find a balance between quality and speed, correctly choose the methods that analyze the text.

When deploying a DLP system with semantic analysis of texts, the following method can be proposed, which can provide maximum efficiency with minimal delay, as well as reduce the costs of its operation. As an example, take a medical organization, where there are many subdivisions, departments. For all units at the beginning, it is necessary to use universal semantic analysis. One of the representatives of universal analysis can be called UNL (Universal Networking Language). UNL is an artificial language designed to store data [4]. UNL in this case is a kind of intermediary, a universal way of storing data, not tied to any context, that is, when analyzing the text, the text is translated into the UNL representation and then it is analyzed. Upon completion of the analysis, the result, also presented in the UNL representation, is translated into natural language. The main problem of a universal system when using semantic analysis is the need to store a reference based on which the system could identify the context, the meaning of words, etc. In view of the fact that this approach uses large directories and many text transformations, the speed of such an analysis will be very slow. But such a solution will secure the company's infrastructure at the initial stage. To speed up the analysis of the text, it is worth introducing a highly specialized analysis in those units where it is possible, such as extracting information about injuries in medicine [5]. But here it is worth noting one feature, namely, in general, the essence of a highly specialized analysis is almost the same; only the context differs. The context here refers to the area of use of words, rules for building texts, that is, the input and output data vary depending on the unit. Thus, it turns out that a choice arises between a long but universal analysis and a fast but highly specialized analysis. Although there is now a choice between the two approaches, it is still worth using both in parallel.

Let us consider two situations when the text passes through a DLP system, and the DLP system in one case has highly specialized analysis and is universal, and in the other cases there is no highly specialized analysis. When there is no highly specialized analysis, more time is spent on analysis, but at the same time we take into account all options for leakage of confidential information. But when we have a highly specialized analysis, we can save and analyze quickly with a small probability that we will miss the case of information leakage if the text contains information that is not in the context of a highly specialized analysis. Therefore, to reduce risks, it is worth additionally sending a text for a general analysis, but if a highly specialized analysis has passed, you should not wait for the result of a general analysis. In this case, a general analysis is only an additional, indirect check, the result of which we can receive belatedly. In addition, do not forget about the maintenance of the system. Directories need to be updated over time, and incidents should be analyzed for their correctness. To do this, it is worth visualizing the result of the analysis. This is necessary for the timely updating of directories, which are used both in general analysis and in highly specialized ones. Here, visualization is understood as a conclusion to the screen for a detailed analysis of word relationships in the text with the possibility of quick updating of the reference [6]. The main task of visualization is to help and partial automation of the data update process. This will allow, with minimal human resources, updating directories even in role-time mode. For the assessment, it is worth using the number of leaks detected. Moreover, you need to consider the total number, missed and false information leaks. This will help identify bottlenecks in the analysis for further improvement by changing the algorithm or updating directories.

The proposed approach with combined general and highly specialized semantic analysis will secure the corporate network. An increase in security will occur due to streaming semantic analysis of all textual information. This completely eliminates the direct leak of textual information. Thanks to the combined method of text analysis, the speed of information processing increases. The increase in speed can be measured, compared with the difference in the system before the introduction of highly specialized analysis and after. The difference in speed will be different. In addition, the results of the analysis will allow you to quickly update the directories. In reference books there are words that identify the context and meaning, as, for example, in our case, the reference may contain medical illnesses. Handbooks of this kind should be updated with the advent of new diseases. Visualization of the analysis in such cases will allow This is necessary in order to get rid of the need to search for the necessary reference. This takes time so that the administrator can find specialized software.

References

[1] Kumunzhiev K.V., Zverev I.N. Method for increasing the efficiency of the dlp-system in the semantic analysis and categorization of information. Modern problems of science and education, 2014, No 5.

[2] Pospelov D. A. Ten hot spots in research on artificial intelligence. Intelligent Systems (Moscow State University), 1996, Vol. 1, No 1-4, pp. 47–56.

[3] Batura T.V. Methods and systems of semantic analysis of texts. Software products, systems and algorithms, 2016, No 4, pp. 45-57.

[4] What is UNL? L. Kreidlin, 2001. Available at: https://old.computerra.ru/2001/390/197284/.

[5] Golovko A.P. Automatic recognition of the meaning of the text on the example of reports on industrial accidents. VESTNIK OF Kostroma State University, 2016, No 3, pp. 106–112.

[6] Purtov D. N. Sidorkina D. N. The problem of training a neural network when extracting key information. Intelligent systems and information technologies, 2019.

Подход с комбинированием общего и узкоспециализированного семантического анализа в DLP-системах

Пуртов Д. Н., Сидоркина И. Г.

Развитие современных компьютерных технологий позволило создавать системы для анализа не только структурированных данных, но также данных, представленных на естественном языке. Как правило системы автоматического анализа текста используют следующие этапы анализа: морфологический, синтаксический и семантический. Подобного рода системы использоваться в DLP-системах. DLP (Data Loss Prevention) -— это система, созданная для предотвращения утечек конфиденциальной информации за пределы корпоративной сети. Наиболее интересной частью анализа текста является семантический анализ. Именно при семантическом анализе выясняется смысл текста. В данной работе были рассмотрены основные проблеме семантического анализа, а также предложен способ построения семантического анализа, который может обеспечить максимальную эффективность при минимальных временных затратах и уменьшит издержки при эксплуатации. Особенность подхода заключается в особом комбинировании общего семантического анализа с специализированным, постепенном увеличении скорости анализа в DLP-системе, визуализации анализа с возможностью улучшения работы семантического анализа. Под общим анализом тут понимается возможность анализировать текст любого рода, а специализированный тексты только определенной тематики. Постепенный плавный переход от общего к специализированному анализу позволяет увеличить скорость работы системы в тех тематических областях. Визуализация дает возможность более быстро анализировать и оценивать результат работы семантического анализа, позволяет сразу актуализировать справочники слов, необходимых для понимания смысла текста. Все эти действия позволяют DLP-системе быть в актуальном состоянии для предотвращения утечки информации.

Received 15.01.2020

Analysis of relation types in semantic network used for text classification

Victor Potaraev
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
vic229@rambler.ru

***Abstract*—Contemporary information systems contain lots of textual information. One of important kinds of textual information processing is text classification. Semantic network is a model which can be used to resolve different tasks including text classification. There are other models which can resolve the same task, and some of them show relatively good results, but using semantic networks has such advantages as human readability and analyzing actual semantic relations between words. Semantic networks have different set of relation types depending on the network purpose. This article is devoted to looking for a set of relation types to be leveraged in semantic network created for text classification. The analysis is performed by generating semantic networks for Russian-language texts leveraging different sets of relation types. Generated networks are used for text classification. Texts were taken from books on several technical disciplines. Proposed algorithms can be used to perform text classification when performing such tasks as dividing electronic messages on categories, spam filtering, text topic recognition and other.**

***Keywords*—semantic network, text classification, text categorization, natural language processing, semantic analysis, machine learning, data analysis.**

I. Introduction

During several previous decades humankind created a large amount of text documents. As a result, it is very important to have ability to perform automatic text classification and natural language processing. Detailed research in the area of automated text classification has started quite long time ago. One of the first researches was performed in 1961 [1] and was based on statistical method of documents indexing.

Primary purpose of text classification is to divide an unstructured set of documents into groups according to their content. Text classification can be used in the following areas:

1) Divide electronic messages on categories.
2) Spam filtering.
3) Text topic recognition.
4) Other.

There are two primary approaches to text classification and topic analysis: frequency analysis and semantic analysis. The first one is based on calculating frequency of words in text and the second is based on meaning of words (more precise). There are also methods mixing characteristics of these two approaches, such as frequency and context analysis [2].

Semantic network is an oriented graph which reflects concepts and relations between them [3]. Semantic network describes knowledge using networking structure [4]. For example, in case of 3 concepts related to each other the network will look like shown on Fig. 1.

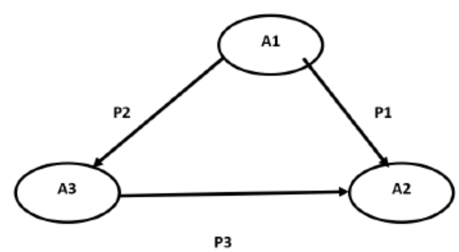


Figure 1. Example of semantic network.

Each relation in semantic network can have a kind, and kinds of relations used in network are usually selected based on specific problem being resolved [5].

There is a lot of different approaches to text classification. One of them is Naive Bayes classifier which shows good results even comparing to more complicated approaches. Other popular approaches are neural networks, support vector machines, regression methods and other [6]. Many of used models are based on working with numbers rather than actual semantic relations between words. Semantic network is a model which has such advantages as human readability and reflecting actual semantic relations.

This article is devoted to looking for a set of relation types to be leveraged in semantic network created for text classification.

II. Relation Types in Semantic Network

Information systems are often used as a tool to find an answer in response to a query. Let's consider the task of answering to a question in natural language. There are five primary kinds of question in Russian language. Open question is a kind of question which requires clarification [7], for example "who?", "where?", "when?", "how much?". A system able to answer different kinds of question allows user to make less effort for formulating queries to the system.

Let's imagine that there is a text "Bag stays near the bed", and user asks question "where does the bag stay?". Semantic network used to answer this question [8] may contain the following kinds of relation: "subject-predicate", "place" (Fig. 2).

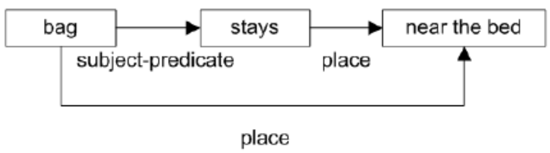


Figure 2. Semantic network used to answer question.

In case of another question other relation types such as "attribute" may be needed. Also, storing synonyms and word forms in network will improve ability to answer questions using the network [8]. Relation types being considered are the following:

1) Word form.
2) Synonym.
3) Subject-predicate.
4) Place.
5) Attribute.

Using these relation types, for the "Black bag stays near the bed" sentence we will get the following more complex semantic network (Fig. 3).

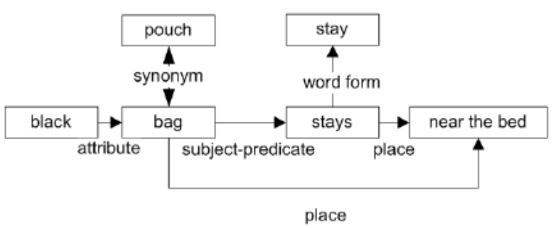


Figure 3. Complex semantic network.

Let's consider using this structure (based on the listed relation types) for text classification. Semantic network is a model which can be used for text classification [9]. Need to note that one of advantages of using this model is that semantic networks correspond to contemporary vision of long-term human memory organization [10].

III. Semantic network generation

Semantic network generation algorithm for a text can be represented as specified below. If a word was already added to the network it is not added the second time.

1) For each meaningful word (e.g. noun, verb, etc.) in each sentence add them to the network.
2) For each meaningful word add stemmed version of the word to the network with "word form" relation type. Porter stemming algorithm can be used to get stemmed version.
3) Load dictionary of synonyms, and for each word in the network find corresponding synonyms. Add found words to the network, connecting them with "synonym" relation type.
4) Find subject(s) and predicate(s) in each sentence. Add them to the network and connect with "subject-predicate" relation.
5) Find words meaning places in the text. Add them to the network and connect with "place" relation.
6) Find words meaning attributes in the text (e.g. adjectives). Add them to the network and connect with "attribute" relation.

This algorithm requires prepared dictionary of synonyms. Subject, predicate, place, attribute words can be found by checking word's part of speech, word form in Russian language (or words order in English language) for simplicity. Also, place can be determined as combination of place pronoun and a noun.

IV. Text Classification

Let's consider a machine learning approach to text classification. To perform text classification, we need to determine a similarity measure for two semantic networks. Similarity S of "network1" to "network2" can be calculated as the following:

$$S = \frac{\sum_{i=1}^{N} Ri}{N} \tag{1}$$

where N is the number of all edges in network1, Ri - similarity of edge number i to network2 (let's call this edge Ei), S - resulting similarity of networks.

The edge Ei has source and target words in network1. If the same words exist in network2 then they may have a $path$ between them in network2. Let's define length of the shortest $path$ as $L2i$.

Ri is calculated as following:

$$Ri = \begin{cases} \frac{1}{L2i}, & \text{if } path \text{ exists,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The idea behind these formulas is that once networks contain the same concepts but connected a bit differently then there is similarity but not 100 percent.

The value of S belongs to range [0;1]. In case of comparing a network with itself S is equal to 1.

Text classification algorithm using semantic network can be represented as following:

1) Determine list of rubrics (i.e. text classes) and texts which will be used for learning. These are texts for which rubric is already known.
2) Define a set of relation types used and a threshold value of similarity required to categorize a text as belonging to a rubric (decision threshold).

3) Concatenate all learning texts for each rubric into single text per rubric. Generate semantic network based on each of texts created. As a result, we have a separate network for each rubric.
4) Create semantic network for classified text and calculate its similarity with networks generated for each rubric.
5) If calculated similarity for one of rubrics is larger then threshold value then the text belongs to that rubric. If threshold is not met for any rubric, or if it is met for several rubrics, then the class is not determined.
6) If there are other texts to classify, go to step 4.

Similarity calculation includes searching for the shortest path between rubric network nodes. This is performed a lot of times for each relation of network created for classified text. So all paths between all nodes of rubric network should be calculated by Floyd–Warshall algorithm. We can make an assumption that words connected with path longer than 2 relations are really not very related and ignore such long chains of relations. Then the algorithm can be limited to look for path with length not longer than 2 which drastically increases performance.

V. Relation Type Sets Comparison

Classification result depends very much on parameters of classification algorithm, in particular on decision threshold amount and relation types used. Contrariwise, algorithms on graphs work relatively slow and count of nodes and edges in the network matters too. So these are primary characteristics being investigated.

Classification algorithm was researched on a set of Russian-language educational texts related to the following 5 disciplines (rubrics): Geometry, Physics, Informatics, Probability Theory, Philosophy, with total size around 2 million characters. The algorithm for calculating similarity includes finding path between all nodes in graph (to improve performance, only path not longer than two edges is considered). The path finding algorithm contains nested loops on nodes and edges so time complexity is $O(n \cdot k^2)$, where n is the number of unique words in text, k - maximal number of relations per word. Average speed of text processing is around 1 Megabyte per 5 minutes.

Based on preliminary check it was decided that it makes sense to use only threshold amounts less than 0.4 because it's very rare that even two texts related to the same rubric would be so much similar. As for relation types, each combination of the 5 relation types was checked:

1) Word form.
2) Synonym.
3) Subject-predicate.
4) Place.
5) Attribute.

The correctness of classification was measured as ratio of correctly classified texts to all texts. In total, 160 experiments were performed. The best results and worst results configurations are displayed in the Table 1. The best result is achieved in case when 3 types of relations are used: Word Form, Subject-Predicate, Attribute. Synonym and Place are not used. It may be related to specific of texts classified: they are mostly technical and synonyms are rarely used in such kind of literature. Also, looks like places written in technical texts are not so different between rubrics as attributes.

On another hand, algorithm based on semantic network containing only Attribute relation shown relatively good results while requesting much less resources. Threshold value in this case is small which means that texts from the same rubric contained small amount of the same attributes, but texts from other rubrics had even less in common.

Semantic network with all five relation types has shown average results, while taking most resources.

Table I
Relation Type Sets Comparison results

Form	Syn.	Subj. -Pr.	Pla ce	At tr.	Correct %	Total Edges	Thres hold
+	-	+	-	+	65	24737	0,14
+	-	+	-	+	57	24737	0,21
-	-	+	+	-	52	17905	0,07
-	-	+	-	+	51	15558	0,07
-	-	-	-	+	46	5366	0,07
-	-	+	-	-	46	10192	0,07
+	-	+	-	+	46	24737	0,28
-	-	-	+	-	42	7745	0,07
+	+	-	+	+	38	30364	0,35
+	-	+	+	+	38	32450	0,14
-	-	-	+	+	37	13111	0,07
-	-	+	-	+	35	15558	0,14
+	-	+	-	-	35	19371	0,21
-	-	+	+	+	34	23271	0,07
-	-	+	+	-	33	17905	0,14
+	+	-	-	+	33	22619	0,35
+	-	+	-	-	33	19371	0,28
...	...	...	...	...	...	...	...
+	+	+	-	+	22	32811	0,21
+	+	+	+	+	22	40524	0,35
-	-	-	-	+	20	5366	0,28
...	...	...	...	...	...	...	...
-	+	+	-	-	1	18266	0,07
-	+	+	-	+	1	23632	0,07
-	+	+	+	-	1	25979	0,07
-	+	+	+	+	1	31345	0,07
+	+	-	-	-	1	17253	0,07
+	+	-	-	-	1	17253	0,14
-	+	-	-	+	0	13440	0,07
-	+	-	+	+	0	21185	0,07
+	+	-	+	-	0	24998	0,07
+	+	-	+	+	0	30364	0,07
+	+	+	-	+	0	32811	0,07
+	+	+	+	-	0	35158	0,07

Need to note that based on all results, this algorithm selects wrong rubric on very rare occasions. The algorithm more likely will not make a decision than make wrong decision. Based on the formula 1, it makes sense because it's unlikely that common relations would exist in texts related to different topics. Only in such case this algorithm based on semantic network could make a wrong decision.

VI. Conclusion

Semantic networks can be used for different purposes, and relation types used depend on the way in which that network will be used. If the purpose of the network is to answer a question then it is better to have more relations in it. But once another problem is being resolved then more relations doesn't mean better. Semantic network used for text classification have been investigated. The correctness of classification was measured as ratio of correctly classified texts to all texts. Text classification involves several (or more) semantic networks and it may become harmful to have too much relations in each network: they require more processing power, and after including "every possible" relation they become too similar to each other.

Comparison of different sets of relation types has been performed and it shows that combination of 3 relation types (Word Form, Subject-Predicate, Attribute) has maximal percentage of correct results.

Also the investigation has shown that semantic network makes wrong decisions in rare cases so if we use another relation types set and it gives some result, that result is most likely correct.

Using semantic networks for text classification has a drawback: it takes time to analyze each relation type and to find ways in graph. Performance of algorithm analysing each relation in a big semantic network is not very fast. It is possible to combine two or more kinds of semantic networks to improve classification speed. Other possible approaches for improving performance may include caching, hash-tables, selecting most valuable relations based on frequency, simplifying algorithm of finding path on graph. During work on this article, Floyd–Warshall algorithm was modified to find ways not longer than 2 to improve performance.

In general, selecting set of relation types for specific semantic network depends on specific task. There are more and less useful relation sets. It makes sense to perform some preliminary learning and testing of network before using it on real data. Proposed algorithms can be used to perform text classification when performing such tasks as dividing electronic messages on categories, spam filtering, text topic recognition and other.

References

[1] M. Maron. Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, 1961, No. 3, pp. 404 – 417.

[2] Model' strukturnogo predstavleniya tekstovoi informatsii i metod ee tematicheskogo analiza na osnove chastotno-kontekstnoi klassifikatsii. Available at: http://www.dissercat.com/content/model-strukturnogo-predstavleniya-tekstovoi-informatsii-i-metod-ee-tematicheskogo-analiza-na (accessed 2020, Jan).

[3] T.A. Gavrilova and V.F. Khoroshevskii. *Bazy znanii intellektual'nykh sistem* [*Knowledge bases of intelligent systems*], Saint-Petersburg, Piter, 2000. 384 p.

[4] Bazy znanii ekspertnykh sistem. Metody predstavleniya znanii: logicheskie modeli, produktsionnye pravila, freimy, semanticheskie seti. Available at: http://daxnow.narod.ru/index/0-18 (accessed 2020, Jan).

[5] D. R. Rakhimova. Postroenie semanticheskikh otnoshenii v mashinnom perevode [Building semantic relations in machine translation]. *Vestnik KazNU im. al'-Farabi. Seriya «Matematika, mekhanika i informatika».* [*KazNU messenger named after al'-Farabi. "Mathematics, mechanics and informatics" series.*], 2014, no. 1, pp. 90–101.

[6] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, vol.34, No. 1, 2002, pp. 1–47.

[7] N. Surin. *Ritorika: Uchebnoe posobie* [*Rhetoric: tutorial*], Moskow, MGIU, 2007. 246 p.

[8] V.V. Potaraev. Algoritm primeneniya semanticheskoi seti dlya poiska otveta na vopros [Algorithm of leveraging semantic network to search for answer to a question]. *Komp'yuternye sistemy i seti: Materialy 54-oi nauchnoi konferentsii aspirantov, magistrantov i studentov BGUIR.* [*Computer systems and networks: Materials of 54th scientific conference of postgraduates, undergraduates and students of BSUIR*], 2018, pp. 103–105.

[9] L.V. Serebryanaya, V.V. Potaraev. Metody klassifikatsii tekstovoi informatsii na osnove iskusstvennoi neironnoi i semanticheskoi setei [Methods of textual information classification based on artificial neural and semantic networks]. *Informatika* [*Informatics*], 2016, no. 4, pp. 95–103.

[10] Osnovy iskusstvennogo intellekta. Available at: http://search.rsl.ru/ru/record/01007574162 (accessed 2020, Jan).

Анализ типов отношений в семантической сети, используемой для классификации текста

Потараев В.В.

Современные информационные системы содержат большое количество текстовой информации. Одним из важных способов обработки текстовой информации является классификация текстов. Семантическая сеть является моделью, которая может быть использована для решения различных задач, в том числе для классификации текстовой информации. Существуют другие модели, способные решать эту задачу, и некоторые из них показывают довольно хорошие результаты, но использование семантических сетей имеет такие преимущества, как читабельность и анализ явных семантических отношений между словами. Семантические сети могут иметь различные наборы связей в зависимости от целей, для которых они создаются. В данной работе производится поиск набора типов связей, который можно использовать в семантической сети, создаваемой для классификакации текстов. Осуществляется анализ структуры сетей путём генерации семантических сетей для русскоязычных текстов с использованием различных типов связей. Сгенерированные сети используются для классификации текстов. Классифицируемые тексты были взяты из книг по нескольким техническим дисциплинам. Предложенные алгоритмы могут быть использованы для классификации текста при решении таких задач как разделение электронных сообщений по категориям, фильтрация спама, определение темы текста и при решении других задач.

Received 15.01.2020

AutoEncoders for Denoising and Classification Applications

Boris A. Assanovich
Yanka Kupala State University of Grodno
Grodno, Republic of Belarus
bas@grsu.by

***Abstract*—Several structures of autoencoders used for the efficient data coding with unsupervised learning and applied to solving the tasks of classification and removing the internal noise from data used in problems of biometric and emotional recognition have been analyzed in this paper. Smile type recognition and biometric identification experiments using the transformed features from UvA-NEMO Smile Database and Caltech Faces datasets showed the possibility of improving the classification accuracy by 10%**

***Keywords*—stacked autoencoder, denoising, classification accuracy, unsupervised learning, smile type recognition, biometric identification**

I. INTRODUCTION

Learning of data representations with little or no supervision is a key challenge in machine learning (ML). Last time, there has been an increasing interest in the use of autoencoders to solve many theoretical and applied data processing tasks based on principals of ML, especially considering how to learn a mapping from high-dimensional observations to a lower-dimensional representation space [1]. However, recent advances in the use of auto-encoders are largely based on the seminal paper [2], which served as the beginning of the development of new algorithms for data processing in ML.

Autoencoder is a special architecture of artificial neural networks that allows the use of unsupervised learning using the method with back propagation of error. Generally speaking, an autoencoder is a direct distribution network without feedback, most similar to a perceptron and containing an input layer, one or several intermediate layers and an output layer. The goal of an autoencoder is to learn a representation (encoding) for a dataset, typically for dimensionality reduction, by training the network and ignore the "noise" signal. Autoencoders are used to solve many applied problems, from face recognition to obtaining the semantic meaning of image and language structures.

In this paper, we consider the use of autoencoders for solving applied problems of classification and removing the internal noise from data used in biometric and emotional recognition. Further, in the first section, the formalization of the construction of autoencoders types are considered, in the second section, the results of experiments using video and image processing datasets from the UvA-NEMO Smile Database and Caltech Faces database are presented. In conclusion, the evaluation of experiments is given, further aspects for the use of autoencoders are determined, which completes the paper.

II. AUTOENCODER STRUCTURE AND TYPES

There are several types of autoencoders. *Sparse Autoencoder* has a dimension of the hidden layer that is greater than the input. It consists of two parts: coder (encoder) G and decoder F as depicted in Figure 1.

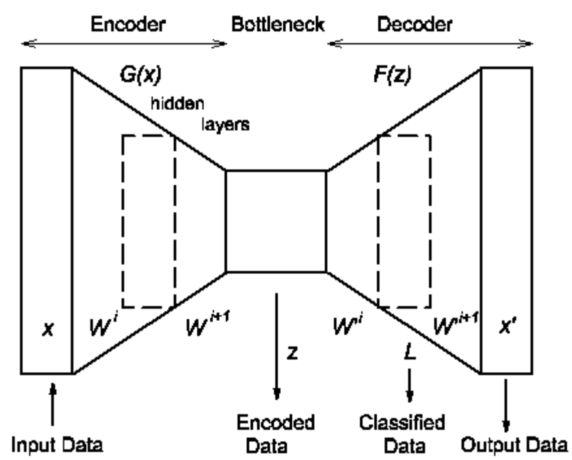


Figure 1. Autoencoder Structure

The encoder translates the input signal into its representation (code): $z = G(x)$, and the decoder restores the signal by its code: $x' = F(z)$. Moreover, the transformation functions F and G contain activation function, weights and biases of trained artificial neural networks [2]. By changing the mappings F, G, the autoencoder tends to learn the identity function $x = F(G(x))$, minimizing kind of error based on some functional $L(x, F(G(x')))$.

Let consider that a vector $x \in R$ connected to the input of an autoencoder. Then the encoder maps the vector x to another vector $z \in R$ as follows $z = h^i(W^i x + b^i)$, where the superscript i indicates the i-th layer. Then h^i is a transfer function for the encoder, $W^i \in R$ is a weight matrix, and $b^i \in R$ is a bias vector. Hence, the decoder maps the encoded representation z back into an estimate of the original input vector x, as follows: $x' = h^{i+1}(W^{i+1}z + b^{i+1})$, where the superscript $i+1$ represents the $i+1$ layer. Then a transfer function h^{i+1} for the decoder has a factor $W^{i+1} \in R$ that is a weight matrix, and $b^{i+1} \in R$ is a bias vector correspondently. Hence, if the encoder has only 2 layers then the expression for the transfer function can be represented as $x' = h^2(W^2 z + b^2)$.

Usually autoencoders are limited in the dimension of the code (it is smaller than the dimension of the signal). The input signal is restored with errors due to coding losses, but in order to minimize them, the network is forced to learn to select the most important features.

To exclude the process of overfitting in ML the sparse autoencoders are used. It is done usually by adding a regularizer to the cost function. This regularizer is a function of the average output activation value of a neuron and encourages the autoencoder to learn a representation, where each neuron in the hidden layer "fires" to a small number of training examples. That makes to respond of neurons to some feature that is only present in a small subset of training examples. Then, for a sparse autoencoder a cost function consists of 3 terms: mean squared error term, regularization term and sparsity regularization term [3]. These parameters are usually used to optimize its unsupervised training.

Another type of autoencoder is a denoising autoencoder (DAE). It gets trained to use a hidden layer to reconstruct a particular model based on its inputs. DAE take a partially corrupted input and is trained to recover the original *undistorted input by minimizing* the loss function between the output node and the damaged input.

Recently, such structure as stacked autoencoder (SAE) has become popular. It is a neural network that includes several layers of sparse autoencoders where output of each hidden layer is connected to the input of the successive hidden layer. In this case the hidden layers are trained by an unsupervised algorithm and then fine-tuned by a supervised method.

A stacked denoising autoencoder (SDA) can be simply represented as several DAE that perform denoising. A key function of SDAs is unsupervised layer by layer pre-training. When a network is being trained, it generates a model, and measures the distance between that model and the benchmark through a loss function. SDA attempts to minimize the loss function involve resampling the damaged inputs and re-reconstructing the data, until it finds those inputs which bring its model closest to what it has been told is true.

Last time, of particular interest is the variational autoencoder, in which instead of mapping an input to fixed vector, input is mapped to a distribution. However, its description is beyond the scope of this study.

III. Application of SDA for Noise Reduction and Classification

In our opinion, the conversion of input signals and the analysis of transformed features received by an autoencoder is more effective approach for data processing, as it allows to reduce the data size in latent space and complexity of the entire overall system, as well as use a more accurate data separation.

We developed the FaceAnalyzer Platform (FAP) [4] based on OpenFace tool [5] to carry out the study of human characteristics and perform the analysis, recognition and verification of human biometric data, elements of his emotions by capturing images and video, features extraction and data processing as shown in Figure 2.

In [6], we have made an analysis of two smiles types temporal characteristics with the use of lip corner displacement, and perform their classification exploiting a well-known k-nearest neighbors (k-NN) algorithm based on the intensity of a person's face Action Units (AU), which describe the physiological dynamics of a human smile. Generally, smiles can be composed of three non-overlapping phases; the onset (neutral to expressive), apex, and offset (expressive to neutral), respectively. We made an attempt to use the detection of these three phases to perform the classification between genuine and posed smiles.

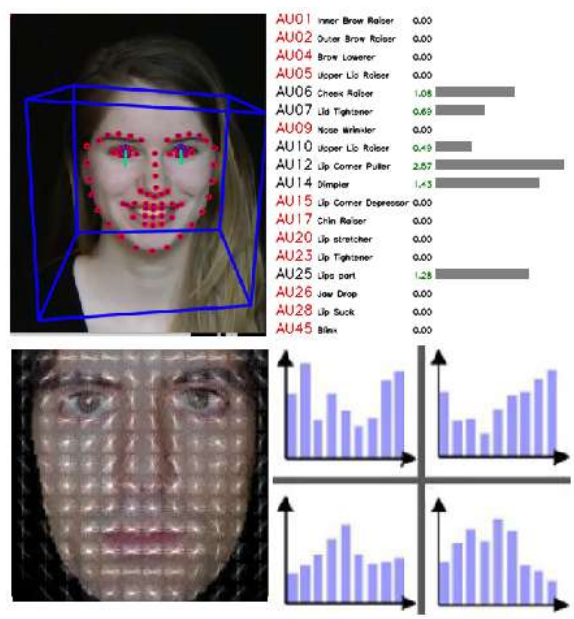


Figure 2. AUs and HOG-data from FAP

The analysis has been performed with FAP that perform 68 Facial Landmark detection and tracking with the help of pre-trained models based on Conditional Local

Neural Fields [5] (CLNF). To perform a classification, the timing smile signals were segmented into three phases, describing the state of a smiling person's face, as it was described in the work [6]. In each frame of analyzed smile, the intensities of the AU measurements were made, averaged over all frames, and a vector with nonzero intensity values $x = [x_1 \dots x_n]$ was formed for learning and classification purposes. The k-NN classification algorithm for the different values of k has been applied in the experimental setup. The results with a vector length of $n = 17$ elements showed that the classification accuracy is 60%. Next, we increased the dimension of vector by separating and averaging the AU intensity values over all 3 phases (Onset Apex and Offset) and used the increased feature vectors with 3x17=51 elements. The classification results with the use of k-NN algorithm have been slightly improved with weakly dependence on k number.

In further studies, we have designed SAE that consisted of 2 sparse autoencoders as depicted in Figure 3. Their convolutional layers were used to apply data de-noising, reduce features dimension and perform classification. The dimension of input vector (hiddenSize) for the first autoencoder was 51 elements which was pre-trained by Matlab with parameters: L2WeightRegularization', 0.001, 'SparsityRegularization', 4, 'SparsityProportion', 0.05, 'DecoderTransferFunction','purelin'. For the second autoencoder we have used similar values of parameters and dimension of 17 elements. The output layer of the second autoencoder had 8 elements, to which the soft-max decision function has been attached to perform a binary classification (upper part of Figure 3). The classification accuracy with the use of SAE achieved 70%.

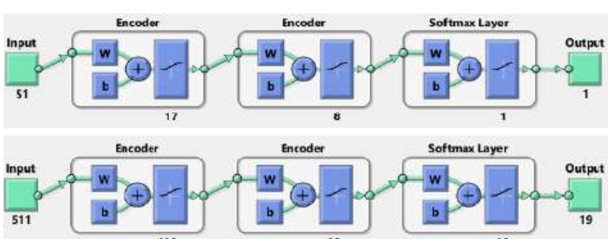


Figure 3. Stacked Autoencoders

The next experiment has been conducted with the use of SAE (lower part of Figure 3) for denoising of the HOG-dataset biometric data (Figure 2) obtained using FAP from the Caltech Faces database. HOG-dataset represented 19 faces of different people with 70 copies for each subject. The designed SAE consisted of 2 sparse autoencoders, that were trained and tuned with similar parameters and in such way as in previous experiment. The layer element dimensions of SAE scheme was chosen to be 511-100-50 ending with a Soft-Max classifier for 19 subjects (lower part of Figure 3). Unsupervised training and classification was carried out with transformed and masked HOG-samples [4] from Caltech Faces database with sizes in ratio 4:1. In these simulations SAE was used both for removing the "biometric noise" and for authentication of a person's face by its biometric HOG-vector with a length of 511 real numbers. The results obtained demonstrated that FRR (False Rejection Rate) takes the value 3.2% without application of error correcting codes as it was realized in Fuzzy Commitment scheme [4], where the application of BCH codes (511,58,91) and (511,28,111) allowed to achieve FRR at almost the same level of 3%.

Conclusion

In this work, an analysis of various types of autoencoders and the principles of their construction has been performed. A method for solving applied problems of processing image data with the use of extraction of all available features of images (video), and then reducing their dimensionality and removing noise with the use of autoencoders was proposed. The experiments based on of UvaNemo and Caltech datasets performed have shown the improvement in the accuracy of classification genuine smile from posed one by 10%, as well as a reduction in the complexity of biometric templates design. In addition, the hypothesis of biometric image data denoiseing from internal noise using SAE was confirmed. The prospects of research how to use the latent space of autoencoders has been determined.

Acknowledgment

This work was supported in part by the COST Action CA16101 "MULTI-modal Imaging of FOREnsic SciEnce Evidence tools for Forensic Science".

References

[1] M. Tschannen, O. Bachem, and M. Lucici, "Recent Advances in Autoencoder-Based Representation Learning," Proc. W. Bayesian DL. NeurIPS, 2018. arXiv preprint: https://arxiv.org/abs/1812.05069 (accessed 2019, Dec).

[2] P. Vincent, H. Larochelle, B. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," Proc. of the 25th ACM Int. conf. ML, 2008. pp. 1096–1103.

[3] B. A. Olshausen and D. J. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1," Vision Research, Vol.37, 1997, pp.3311–3325.

[4] B. Assanovich and Yu. Veretilo, "Biometric database based on HOG structures and BCH codes," Proc. ITS, Minsk, 2017, pp. 286-287.

[5] T. Baltrusaitis, L.-P. Morency, and P. Robinson, "OpenFace: An open source facial behavior analysis toolkit," WACV, 2016, pp. 1-10.

[6] B. Assanovich, Yu. Veretilo, N. Bich, A. Pushkina, and V. Khilmanovich, "Recognition of Genuine Smile as a Factor of Happiness and its Application to Measure the Quality of Customer Retail Services," PRIP 2019, pp.84-89.

[7] H. Guo, X. H. Zhang, J. Liang, and W. J. Yan, "The Dynamic Features of Lip Corners in Genuine and Posed Smiles," J. Front.Psychol. Feb., 2018.

Автоэнкодеры для приложений шумоподавления и классификации

Ассанович Б. А.

Проанализировано несколько структур автоэнкодеров, используемых для эффективного кодирования данных с обучением без учителя и применяемых для решения задач классификации и удаления внутреннего шума из данных, используемых в задачах биометрического и эмоционального распознавания. Эксперименты по распознаванию типов улыбок и биометрической идентификации с использованием преобразованных характеристик данных из видео базы улыбок Smile UvA-NEMO и наборов данных Caltech Faces показали возможность повышения точности классификации на 10%.

Received 19.01.2020

Analysis of Finite Fluctuations as a Basis of Defining a Set of Neural Network Model Inputs

Roman Sheglevatych
Lipetsk Region Health Department
Lipetsk, Russia
schegl111@mail.ru

Anton Sysoev
Department of Applied Mathematics
Lipetsk State Technical University
Lipetsk, Russia
sysoev_as@stu.lipetsk.ru

***Abstract*—The paper describes an approach on the defining a set of neural network model inputs analyzing their influence on the output. The mathematical basis of such procedure is Analysis of Finite Fluctuations based on applying Lagrange mean value theorem. The applied problem under consideration in finding outliers in data from healthcare digital system records.**

***Keywords*—Analysis of Finite Fluctuations, neural networks, Sensitivity Analysis**

I. Introduction. Outliers Detection Problem in Medical Healthcare Datasets

The very important problem which facing Russian healthcare system and consequently insurance companies is the qualitative way of organizing and functioning of healthcare single digital system [1]. It was announced that in two years every Russian will have a digital copy of his medical care story. In accordance, the approaches based on artificial intelligence will be widely implemented to help physicians on making decisions.

A part of the described challenge which is a peace of interest firstly of insurance companies and is a way to increase the quality of medical healthcare in general is the outliers detection in datasets obtained from medical healthcare digital systems.

Outliers (or anomalies) detection refers to the problem of finding data are not corresponding to some expected behavior of the process or indicator of the system [2]. When anomalies are detected, it is sometimes difficult to determine the threshold of the normality. Values which are close to threshold can be both normal and anomaly. That is why it is very important to verify each approach to outliers detection via expert estimates.

In the field of anomalies detection exist many fundamental and applied studies containing different approaches to solve the basic problem. They may be classified according the scheme used in the approach as supervised (cf. [3]), unsupervised and semi-supervised anomalies detection (cf. [4]). And also there is a classification which is based on the mathematical approach underlying used scheme.

This study is partially supported by Russian Foundation for Basic Research (RFBR) and Lipetsk regional administration. Grand No 19-47-480003 r_a.

II. Problem to Define the Most Influenced Factors in Neural Network Model

Constructing a mathematical model of a technical, economic, social system or technological process, the question of choosing the most significant inputs affecting the response (output) of the studied structure is a relevant problem in information processing. Sensitivity Analysis allows to estimate the influence of the model output on its inputs, as well as to assess the importance of each factor. Being a universal mechanism to describe different complicated systems and processes, artificial neural networks are widely used in many areas of theoretical fields and practical applications.

Let $X = (x_1, ..., x_n)$ be system inputs (factors, variables), $G(\cdots)$ is an operator of the system and Y is the system output (indicator), then the system can be represented as $Y = G(X)$.

In this applied study neural network was chosen as a structure of classifier. The most prominent approach to assess the sensitivity of neural network models is the Garson algorithm. This method is based on the study of weights constructed neural network model. It is believed that the variation of the studied coefficients can explain the characteristics of the "black box" neural network. According to the study [5], for three-layer neural network with a classical structure, factor sensitivity coefficients can be found as

$$S_k^p(i) = \frac{\sum\limits_{j=1}^{n} \left(w_{ij} \cdot v_{jk} / \sum\limits_{i=1}^{n} w_{ij} \right)}{\sum\limits_{i=1}^{n} \left(\sum\limits_{j=1}^{n} \left(w_{ij} \cdot v_{jk} / \sum\limits_{i=1}^{n} w_{ij} \right) \right)},$$

where i, j and k are indexes for weights of input, hidden, and output layer weights respectively.

A well known approach which could be used in different cases (and does not depend on model type) is applying techniques in composite indicators [6], [7]. The idea behind these family of methods is to construct a composite indicator aggregating several factors with some weights, where weights define the degree of importance for each indicator. The most prominent approach in this family is Sobol sensitivity coefficients.

Table I
INDICATORS ON THE MEDICAL CARE RECORDS

Type of the indicator	Code of the indicator	Explanation
Indicators belonging to the patient	CEL_OBSL	The purpose of the patient's appeal to the medical organization
	RSLT	The result of the medical care
	ISHOD	The patient's vitals
	voz	The patient's age
Indicators belonging to the medical organization providing medical care	lpu_p	The name of the medical organization to which the patient is assigned
	LPU	The name of the medical organization where the patient was treated
	KOD_TP	The code of the medical organization department where the patient was treated
	PODR	The name of the medical organization department where the patient was treated
	PROFIL	The profile of the provided medical care
	NAZ_PMP	The profile of the medical care to which the referral was given based on the results of the medical examination for patients of the 3rd group of health
	NAZ_PK	The profile of a round-the-clock or daily hospital place for which a referral for the hospitalization was given based on the results of medical examination for patients of the 3rd health group
	PROFIL_K	The profile of the place in a round-the-clock or day hospital where medical care was provided
Indicators belonging to the disease	IDCASE	The case unique identifier
	DS0	The primary diagnosis
	DS1	The basic diagnosis
	DS2	The concomitant disease
	POVTOR	The sign of a treatment repeated case for a single disease
	TYPE_MN	The nature of the basic disease
	VIDTR	The sort of an injury
	KOD_KSG	The code of the clinical and statistical group of the disease in the conditions of a daily or round-the-clock hospital
Indicators belonging to the doctor	PRVS	The doctor's specialization
	SPEC_END	The regional localization of the doctor's specialization
	NAZ_SP	The specialization of the doctor to whom the appointment was made based on the results of medical examination for patients of the 3rd group of health
Indicators belonging to the particular case of the medical care	USL_OK	Conditions for the provided medical care
	VIDPOM	The type of medical care
	FOR_POM	The form how the medical care was provided
	DATE_1	Start date of treatment *
	DATE_2	End date of treatment *
	POL_VIS	The number of visits to a medical organization in a case of the medical care
	HOM_VIS	The number of home visits by the doctor to the patient in the case of the medical care
	ITAP	The stage of the medical examination or the preventive examination
	DISP	The type of the medical examination or the preventive examination
	RSLT_D	The result of the medical examination or the preventive examination
	OBR	The indicator characterizing the method of payment for medical care in case of outpatient treatment
	TIMEV	The time of the call to an ambulance *
	TIMEP	The time of the arrival of an emergency medical services *
	P_PER	The sign of the transfer to the daily department or to the round-the-clock department
	NAZR	The appointment of the doctor based on the results of medical examination for patients of the 3rd group of health
	NAZ_V	The type of examination assigned based on the results of the medical examination for patients of the 3rd group of health
	DN	The type of the dispensary observation

Indicators marked with * are not analyzed directly, but used in some combinations.

A. Analysis of Finite Fluctuations

Analysis of Finite Fluctuations (AFF) could be defined as an approach to analyze complicated system with the goal to build a dependency connecting finite fluctuations of a function and finite fluctuations of its arguments [8].

In a practical applications normally the argument x could be measured, and its measurement $\mu(x)$ could have different forms, but the most used one to define the transition from initial value $x^{(0)} = a$ to the value $x^{(1)} = b$ is the absolute increment $\mu(x) = \Delta x = b - a$.

The main problem of AFF is formulated as follows. Let us have a model

$$y = f(X) = f(x_1, ..., x_n) \tag{1}$$

describing the connection between the response (model output) y and its arguments (factors) x_i, $i = 1, ..., n$. It is necessary to transform Model (1) into the form

$$\mu y = \phi(\mu(x_1), ..., \mu(x_n)), \tag{2}$$

which shows the connection between the fluctuation of the output $\mu(y)$ and the fluctuations $\mu(x_i)$ of its factors x_i, $i = 1, ..., n$.

Mathematical Analysis gives in case of finite increments the model, which allows to represent (1) to an exact connection between the finite fluctuation of the model output and its factors fluctuations (2). This is Lagrange mean value theorem (the formula of finite increments, intermediate value theorem of Differential Calculus) for multivariable functions, defined and continuous in a closed domain and having continuous partial derivatives inside this domain:

$$\Delta y = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \left(X^{(m)} \right) \cdot \Delta x_i, \tag{3}$$

where $X^{(m)} = \left(..., x_i^{(m)} = x_i^{(0)} + \alpha \cdot \Delta x_i, ... \right)$, $0 < \alpha < 1$. Here the mean (or intermediate) values of factors $x_i^{(m)}$ are defined by the value of parameter α.

B. Sensitivity Analysis based on AFF

Let us have a neural network model with k hidden layers, which describes studied technical, economic or social system or technological process

$$Y^{(k)} = \Phi^{(k)} \Phi^{(k-1)} ... \Phi^{(1)} X,$$

where $X = (x_1, ..., x_n)^T$.

In a current moment of time the initial vector of inputs is $X^{(0)} = \left(x_1^{(0)}, ..., x_n^{(0)} \right)^T$ and the value of system indicator is

$$Y_0^{(k)} = \Phi^{(k)} \Phi^{(k-1)} ... \Phi^{(1)} \left(x_1^{(0)}, ..., x_n^{(0)} \right)^T.$$

After a while the inputs vector changes its value to $X^{(1)} = \left(x_1^{(1)}, ..., x_n^{(1)} \right)^T$ and system indicator is respectively

$$Y_1^{(k)} = \Phi^{(k)} \Phi^{(k-1)} ... \Phi^{(1)} \left(x_1^{(1)}, ..., x_n^{(1)} \right)^T.$$

Thus, the difference (finite fluctuation) of the system indicator is

$$\Delta Y^{(k)} = Y_1^{(k)} - Y_0^{(k)} \tag{4}$$

and according to Lagrange mean value theorem (3) the same difference could be estimated as

$$\Delta Y^{(k)} = \sum_{i=1}^{n} \frac{\partial Y^{(k)}}{\partial x_i} \left(..., x_i^{(0)} + \alpha \cdot \Delta x_i \right) \cdot \Delta x_i. \tag{5}$$

Equating (4) and (5) and solving the resulting equation, it is possible to find the value α and then to estimate so called factor loads A_{x_i}:

$$\Delta Y^{(k)} = \sum_{i=1}^{n} \frac{\partial Y^{(k)}}{\partial x_i} \left(..., x_i^{(0)} + \alpha \cdot \Delta x_i, ... \right) \cdot \Delta x_i =$$

$$= A_{x_1} \Delta x_1 + ... + A_{x_n} \Delta x_n.$$

For the approach based on AFF there are obtained $N-1$ estimates (because of existing $N-1$ finite fluctuations for N input values data set) and their median values are taken as a sensitivity measure.

The described approach was tested on 2000 cases from the data set "neuraldat" from NeuralNetTools R package comparing with the other classically used approaches (Sobol sensitivity coefficients and Garson algorithm). All compared strategies gave similar results with a slight variation, which proves, that Analysis of Finite Fluctuations is not contradictory and could be applied in such kind of problems. But proposed approach has an undeniable advantage [9]. In contrast to Sobol sensitivity coefficients it does not use an approximation procedure to model statistical parameters of the studied structure and in contrast to Garson strategy, it operates with both parameters and factors of the studied model.

III. Numerical Experiments

A. Scope of Experiment. Data set Description

The data set on medical healthcare in Lipetsk region includes many indicators defined in Table I. Data were collected from February to May 2019 and show how medical care was provided in more than 1 billion cases. It should be noted, that one patient could be linked with many records (according to his visits to the physician). All the indicators could be divided into consolidated groups presented in Table I.

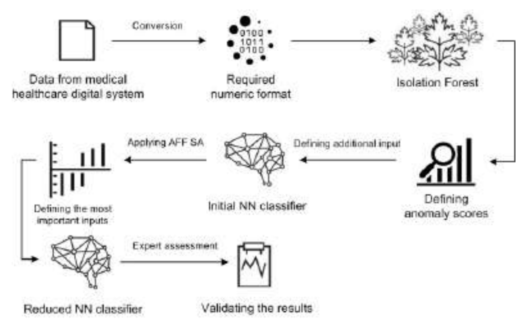


Figure 1. Work flow of the proposed solution

B. Neural Network Classifier to Predict Outliers

As the neural network model has to be reduced according to its inputs, to improve the accuracy of the model it is proposed to add the results (anomaly score) of Isolation Forest algorithm applied to the inputs as an additional input. This step has positively established itself and is a basis of combined approach to detect anomalies in health care data sets described in [10]. The work flow of the proposed approach is given on Figure 1.

The initial neural network model with 34 input factors, 1 hidden layer with 3 neurons, and 1 output was investigated; the logistic activation functions were used, when the reduced one has 15 the most important inputs presented in Table III-A.

The data set used in the numerical experiment was divided into two subsets with 80% in training set and 20% in testing set. The final reduced model has demonstrated the general accuracy of 78% with 24% type I error of classification (false positive) and 16% type II error of classification.

Table II
INDICATORS WITH THE HIGHEST IMPACT ON THE NEURAL NETWORK OUTPUT

Name of the indicator	Explanation	Degree of influence, %
lPU_P	The name of the medical organization to which the patient is assigned	4,11
USL_OK	Conditions for the provided medical care	8,25
SROK_LECH	The length of the treatment or of the hospitalization	6,29
CEL_OBSL	The purpose of the patient's appeal to the medical organization	3,92
PRVS	The doctor's specialization	4,08
SPEC_END	The regional localization of the doctor's specialization	4,82
POVTOR	The sign of a treatment repeated case for a single disease	5,71
TYPE_MN	The nature of the basic disease	4,55
ITAP	The stage of the medical examination or the preventive examination	4,24
RSLT_D	The result of the medical examination or the preventive examination	4,77
OBR	The indicator characterizing the method of payment for medical care in case of outpatient treatment	3,15
RAZN_SKOR	The time between calling to the ambulance and the arrival of medical service	5,75
VIDTR	The sort of an injury	4,39
NAZ_PK	The profile of a round-the-clock or daily hospital place for which a referral for the hospitalization was given based on the results of medical examination for patients of the 3rd health group	5,33
ANOMALY_SCORE	The anomaly score obtained by Isolation Forest algorithm applying	2,85

IV. RESULTS AND DISCUSSION

Analyzing results of the identified by the neural network model outliers it was found that in many cases detected results have deviations from ordinary records. Many of detected outliers could be assigned to one of the following groups. Firstly, in many founded records there were technical errors during filling the record like detecting that « dental consultation» was marked as «chronic disease». Secondly, in several cases there were too many visits to doctor with the minor ailments like visiting dentist 5 times to treat an enamel degradation.

ACKNOWLEDGMENT

Authors are grateful to Lipetsk territorial compulsory medical insurance fund for the applied problem statement and access to the non-personal health care digital system records.

REFERENCES

[1] The Federal law on 29.07.2017 no 242 "On amendments to certain legislative acts of the Russian Federation on the application of information technologies in the field of health protection."
[2] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," in Procedia Computer Science, 2015.
[3] Abe Naoki, Bianca Zadrozny, and John Langford. "Outlier detection by active learning," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006.
[4] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. "Distance-based outliers: algorithms and applications," The VLDB Journal 8.3-4, 2000.
[5] Maozhun Sun, and Liu Ji. "Improved Garson algorithm based on neural network model," 2017 29th Chinese Control And Decision Conference (CCDC). IEEE, 2017.
[6] Bandura Romina. "Composite indicators and rankings: Inventory 2011," New York: United Nations Development Programme, Office of Development Studies (UNDP/ODS Working Paper), 2011.
[7] A. Saltelli, S. Tarantola, and F. Campolongo. '"Sensitivity anaysis as an ingredient of modeling," Statistical Science 15.4, 2000.
[8] S.L. Blyumin, G.S. Borovkova, K.V. Serova and A.S. Sysoev "Analysis of Finite Fluctuations for Solving Big Data Management Problems," 9th Conference on Application of Information and Communication Technologies, (AICT). Rostov-on-Don, 2015, pp. 48-51.
[9] A. Sysoev, A. Ciurlia, R. Sheglevatych, and S. Blyumin, "Sensitivity Analysis of Neural Network Models: Applying Methods of Analysis of Finite Fluctuations", Periodica Polytechnica Electrical Engineering and Computer Science, vol. 63, no. 4, pp. 306-311, 2019.
[10] A. Sysoev, and R. Scheglevatych. "Combined Approach to Detect Anomalies in Health Care Datasets," 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA). IEEE, 2019.

Анализ конечных изменений как основа формирования входов нейросетевой модели

Щеглеватых Р.В., Сысоев А.С.

В статье рассматривается подход к формированию набора входных переменных для нейросетевой модели на основе анализа их влияния на выход. Математической основой такой процедуры выступает анализ конечных изменений, основанный на применении теоремы Лагранжа о промежуточной точке. Прикладная проблема исследования — выявление аномалий в наборах данных, полученных из информационной системы фиксации результатов оказания медицинской помощи.

Received 30.01.20

Forming Production Rules in Intelligent Transportation System to Control Traffic Flow

Elena Khabibullina
Department of Applied Mathematics
Lipetsk State Technical University
Lipetsk, Russia
e.shterenberg@gmail.com

Anton Sysoev
Department of Applied Mathematics
Lipetsk State Technical University
Lipetsk, Russia
sysoev_as@stu.lipetsk.ru

***Abstract*—The paper describes an approach to remodel the capacity of high-speed transportation corridors segments. Tho proposed remodelling scheme is based on applying neural network model combining results of stochastic capacity estimation during congested and free-flow regimes. Based on the obtained model it is proposed to analyze the impact of factors affecting on the estimated capacity with the methods of Analysis of Finite Fluctuations. There is also presented scheme how to form production rules into intelligent transportation system based on the presented approaches.**

***Keywords*—semantic rules, intelligent transportation system, traffic flow, neural network, expert system**

I. Introduction

Due to the increasing number of personal vehicles and the geographical location of many Russian regions the idea to build the one unified intelligent transportation and logistic system delivering the minimum travel time within it, is now the leading one. Such system could be decomposed into similar items with the identical structural scheme but taking into account specific features of the region.

The organizations developing intelligent transportation systems actively implement projects to forecast traffic volumes and flow-control all over the world. They implement systems in Japan, America, European Union, Australia, Brazil, China, Canada, Chile, Korea, Malaysia, New Zealand, Singapore, Taiwan, the UK. In India, Thailand and some countries of South Africa such scientific schools and organizations are just beginning to develop the concept of smart roads ([1], [2]).

Nowadays, the most advanced technologies in the field of intelligent transportation control are designed in Japan, the USA, Singapore and South Korea. The main directions of developing intelligent systems in these countries are connected vehicle technologies, connected corridors, well-managed and resilient traffic flows, Smart Roads and integration these technologies into Smart City Systems and Internet of Things.

The reported study was supported by the Russian Science Foundation within the project 18-71-10034.

According to the inner policy of the Russian Federation the transportation infrastructure and consequently intelligent transportation systems have to be the priority projects. In 2018 there were determined 7 global economic markets supposed to be the leading world projects in the next 10-15 years [3]. These markets are focused on the person as the main subject of public relations and provide all needs of the person. One of the most important market providing the perspectives of transportation mobility is AutoNet with its road map including the most relevant problems to be solved to reach the main goal, which is to connect Europe and Asia with the high-speed transportation corridors for manned and unmanned vehicles.

The purpose of the presented study is to construct the scheme of forming production rules to control traffic flow parameters in intelligent transportation system.

II. Regional Intelligent Transportation System: Information Infrastructure and Used Algorithms

Long-term studies conducted in Lipetsk State Technical University have formed the basis of Lipetsk Regional Centre to Control Traffic Flow being the intelligent transportation system of Lipetsk region. The main task conducted by this centre is to control the traffic situation of manned and unmanned vehicles within the region (cf. Figure 1).

As it could be seen on Figure 1 one of the main tasks of the described centre is to store data on traffic flow and to produce control influences controlling parameters of intelligent transportation system.

A. *Methods to Estimate and Analyze Capacity*

It is reasonable to model high-speed transportation corridor as a set of freeway segments, each with its capacity. There are two types of ways to estimate freeway capacity. The first strategy is connected with taking it as a constant value defined on empirical results or on simulation series. It is typically used way in many national guidelines like Highway Capacity Manual (HCM) [4], HBS (German HSM) [5] or appropriate Russian guidelines [6]. The

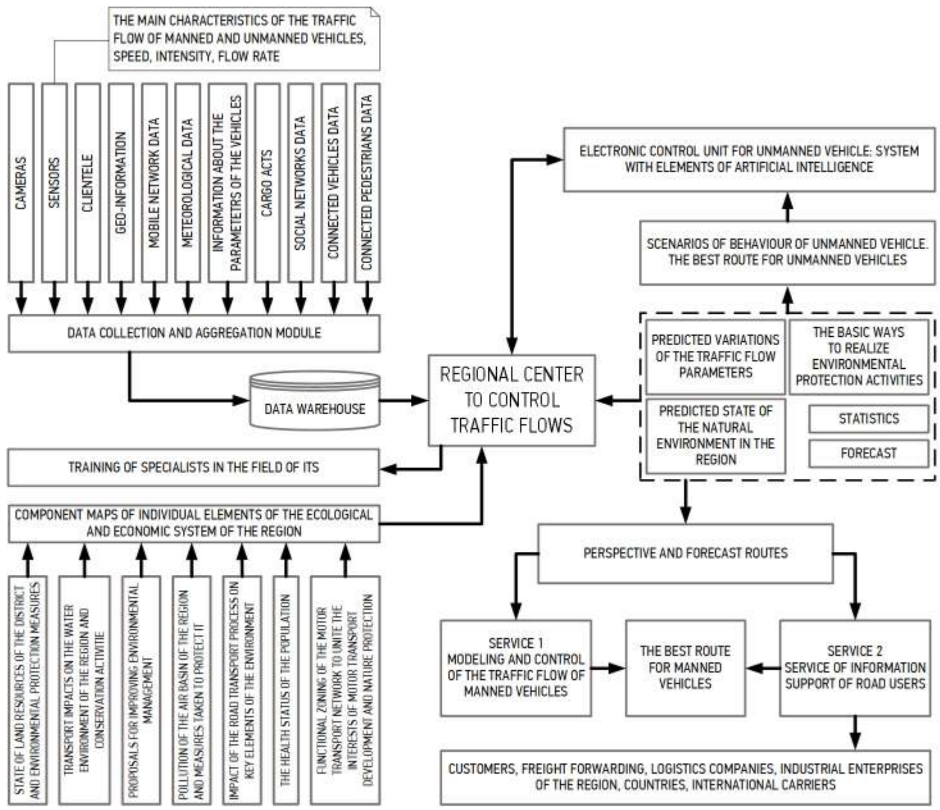


Figure 1. Conceptual scheme of regional intelligent transportation system module

main lack here is that not only each segment has its unique geometrical and physical features, but also each moment of time could have different characteristics, which is explained for example by drivers' behaviour and weather conditions. Such idea has formed the strategy to take freeway segment capacity as a random value specificaly distributed with unique parameters [7]. This strategy is more realistic and could be used in controlling traffic flow via intelligent transportation system.

Based on Mathematical Remodeling approach, involving the substitution of one model by another in complicated system or even mixing models [8], there was build a neural network model to estimate freeway capacity taking into account a set of parameters effecting it [9], [10] (namely they are given in Table II-C). Further it is described the way of constructing such kind of model. Being a stochastic parameter of a transportation system, capacity could be measured only in a time interval prior to the congestion and is precisely equal to the number of vehicles within the survey section during the congestion.

Using the available congested data and statistical assumptions, one can form input-output array dividing all time intervals into two subsets:

- O ("observed"): congested intervals, section capacity is precisely equal to the observed traffic volume;
- E ("estimated"): fluid intervals, section capacity is estimated using the stochastic capacity approach.

These two subsets form an array of initial data to train the chosen structure for further prediction.

B. Algorithm to Remodel Capacity

Step 1. To divide the whole available data into subsets of "observed" and "estimated" intervals. The criteria of this separation must be predefined at this step. Normally it is an average speed threshold, pointing the transition of the traffic flow into a congested regime.

To estimate capacity within congested intervals it has to be applied the Weibull distribution model [11]:

$$F_c(q) = 1 - \exp\left(-\left(\frac{q}{b}\right)^a\right), \qquad (1)$$

where $F_c(q)$ is the distribution function of the capacity rate, q is the traffic volume of the vehicles (veh / h), a, b are Weibull distribution parameters, responsible for the capacity rate variation and for the systematic average value of the capacity rate caused by such constant factors as the number of lanes, the slope, the number of drivers, respectively.

Step 2. Using the data sample obtained on Step 1, to train neural network model with the predefined structure.

A general form of neural network model to be applied on this step is

$$c = \phi^{(k)}(w_0^{(k)} + W_1^{(k)}\phi^{(k-1)}(...(w_0^{(2)} + \\ + W_1^{(2)}\phi^{(1)}(w_0^{(1)} + W_1^{(1)}x))...)), \quad (2)$$

where $c \in \mathbb{R}$ is the output scalar describing capacity, $x \in \mathbb{R}^n$ is the input vector, $\phi^{(i)}$, $i = 1, ..., k$ are vector functions of vector arguments, activation functions, $W_1^{(i)}$ are matrices of weights from layer $(i \breve{} 1)$ to i, $w_0^{(i)}$, $i = 1, ..., k$ are bias weights.

On this step the analysis of the model accuracy must be done and the adjustment (in case of unsatisfied accuracy) should be applied.

Step 3. Using the model obtained on Step 2, to estimate capacity rate within the new data set.

There were conducted numerical experiments data obtained from loop and radar detectors describing the capacity and factors affecting it (cf. paper [10] and study [9]). As an output there was taken nominal capacity for the determination of standardized capacity values in one-hour intervals obtained by applying Kaplan-Meier approach to fit Weibull distribution function. Basis neural network model applied to fit the model of the capacity was:

$$c = \phi(w_0 + W_1 x), \quad (3)$$

where c is the capacity values (veh/h), $x \in \mathbb{R}^8$ is inputs vector (according to study [Nina]), $\phi(net) = \tanh(net)$ is the hyperbolic tangent activation function, w_0 and W_1 are estimated weights for bias and input factors respectively.

Model (3) was fitted in RStudio free software with "nnet" package. Initial data were firstly standardized to obtain weights and then unstandardized to calculate predicted capacity values. It should be noted, that the level of remodeling approximation error was 5.58%, which is acceptable for the described problem.

C. Sensitivity Analysis as a Way of Defining the Most Significant Factors Affecting Capacity

In many applied problems it is very important to find, which input factors are the most significant in order, for example, to control the process or system, ets. Commonly the answer to such kind of question could be obtained after applying approaches of Sensitivity Analysis [12], which is based on statistical and probabilistic techniques and allows to estimate the influence of each model input value (independent variable, argument of the function, factor of the system, etc.) on the output value (dependent variable, function value, index of the system, etc.). Considering the existing mathematical model

$$y = f(X), \quad (4)$$

where $X \in \mathbb{R}^n$ are inputs, $y \in \mathbb{R}$ is output and $f(\cdot)$ can be a function, a system of differential equations, etc., even a program code, the procedure of Sensitivity Analysis is made through the individuation of some indicators, called impact factors, determining quantitatively the influence, that each input has on the output, and consequently, allowing to understand which of them have to be changed the least possible, so that the output of the model does not change too much [12]. Many well-known techniques of Sensitivity Analysis have drawbacks (like stochastic nature or high computational costs, etc.). In contrast to them it is possible to use the approach based on applying Lagrange mean value theorem and called Analysis of Finite Fluctuations [13], [14].

According to the idea of Analysis of Finite Fluctuations let us take the initial instant of time t_0, where the input factors vector is

$$X^{(t_0)} = (x_1^{(t_0)}, ..., x_n^{(t_0)})$$

and respectively the output is

$$y^{(t_0)} = f(X^{(t_0)}) = f(x_1^{(t_0)}), ..., x_n^{(t_0)}).$$

In the next time instance t_1 we have final value of inputs

$$X^{(t_1)} = (x_1^{(t_1)}, ..., x_n^{(t_1)})$$

and output

$$y^{(t_1)} = f(X^{(t_1)}) = f(x_1^{(t_1)}), ..., x_n^{(t_1)}),$$

here $x_i^{(t_1)} = x_i^{(t_0)} + \Delta x_i$, $i = 1, ..., n$.

It follows, that

$$\Delta y = y^{(t_1)} - y^{(t_0)} = f(X^{(t_1)}) - f(X^{(t_0)}) = \\ = f(..., x_i^{(t_0)} + \Delta x_i, ...) - f(..., x_i^{(t_0)}, ...). \quad (5)$$

But according to the Lagrange mean value theorem the same function increment could be estimated as

$$\Delta y = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \left(..., x_i^{(t_0)} + \alpha \cdot \Delta x_i, ... \right) \cdot \Delta x_i \right), \quad (6)$$

and, applying notations $A_i = \dfrac{\partial f}{\partial x_i}(..., x_i^{(t_0)} + \alpha \cdot \Delta x_i, ...)$, the formula becomes

$$\Delta y = \sum_{i=1}^{n} (A_i \cdot \Delta x_i) = \\ = A_1 \Delta x_1 + A_2 \Delta x_2 + ... + A_n \Delta x_n. \quad (7)$$

Equating (5) and (6) and obtaining (7), it is resolved the resultant equation according to the unknown parameter α, finding which and, respectively, estimating impact indexes A_i, determining the influence, that each input factor fluctuation has on the output fluctuation.

Sensitivity Analysis based on applying Analysis of Finite Fluctuations was implemented to Model (3) to gain importance estimates of factors affecting capacity. It should be mentioned, that for this approach there were obtained $N - 1$ estimates (cf. Table II-C) (because of

Table I
COMPARING RESULTS OF SENSITIVITY ANALYSIS FOR MODEL (3) OBTAINED BY DIFFERENT APPROACHES

Affecting factors	**Importance measure, %**	
	Analysis of Finite Fluctuations	***Garson Algorithm***
Speed limit	2,8	1,1
Work zone layout	4,3	2,2
Area	11,1	7,4
Lane widths	5,0	7,7
% of heavy vehicles	9,9	8,1
Grade	11,6	8,5
Lane reduction	12,0	8,5
Number of lanes	43,3	56,5

existing $N-1$ finite fluctuations for N input values realizations in the dataset) and their median values were taken as a sensitivity measure. It is also important, that signs of obtained measures were ignored, so the only degree of sensitivity without its direction was considered. To prove the correctness of the proposed approach the obtained results were compared with Garson algorithm which is common way to estimate the importance of inputs of neural network model.

Analyzing obtained results (cf. Table II-C) we could see, that the number of lanes and the percentage of heavy vehicles are factors which are firstly have the highest impact on the capacity and secondly they are could be controlled to deliver better quality of transportation system functioning. The first parameter (number of lanes) could be varied within the corridor for example by using the hard shoulder as an additional lane or by using reversing traffic lane); the possible way to control the second parameter (the percentage of heavy vehicles) is using ramp metering system to limit access to the corridor from entry-ramps (cf. study [15]).

III. CONCEPTION MODEL OF EXPERT TRAFFIC FLOW CONTROL SYSTEM WITHIN HIGH-SPEED TRANSPORTATION CORRIDORS

According to the results of the study given above, the following factors describing the current and predicted state of the transportation corridor are used in the expert system as input parameters: the estimated capacity (which can be calculated by the proposed neural network model (3)), the average vehicles speed in lanes, the high-speed transportation corridor location, the grade of the highway, the percentage of heavy vehicles, the repair zone layout, the number of lanes in one direction, the lane widths, the lane reduction, the speed limit. The percentage of heavy vehicles, the number of lanes in one direction (shoulders or reverse lanes can be used as additional lanes), the speed limit (using variable traffic signs) are used as the output parameters of the expert system.

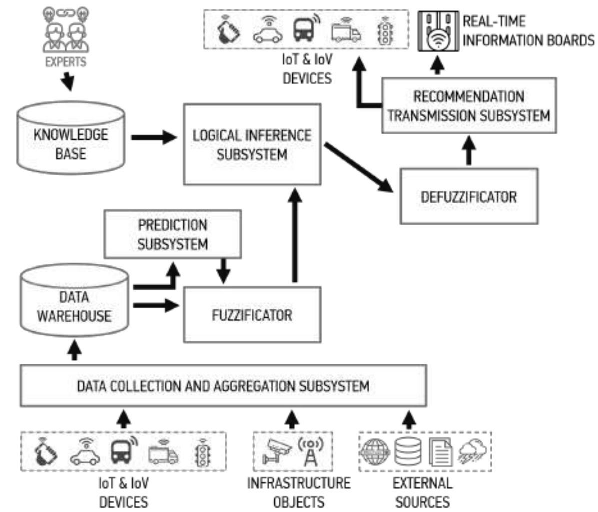


Figure 2. Conceptual model of the Expert Traffic Flow Control System

Task-specific knowledge of experts can be presented as a set of IF-THEN production. Each production rule can include information about one or several input factors in the condition part and in the conclusion part. These parameters can be involved in the production rules in the following forms: linguistic variable, quantitative variable, string variable.

The Figure 2 shows a conceptual model of the Expert Traffic Flow Control System. The values of the input system parameters are calculated and aggregated into Data collecting and aggregating subsystem. Then data is transferred to the data warehouse. In the Prediction subsystem, the estimated capacity is determined using a neural network model. Data coming from the database as a crisp values and corresponding linguistic variables are converted into fuzzy values in the Fuzziffication subsystem. Then, the values of the variable parameters (the percentage of heavy vehicles, the speed limit, the number of lanes) are determined into the Logical inference subsystem using the Mamdani algorithm. In the Defuzziffication subsystem fuzzy values of the output factors are converted to the crisp values. The Recommendation transmission subsystem send recommendations to the devices of road users and road infrastructure objects via wireless interfaces and the Internet.

Due to fuzzy logic algorithms and model describing in a natural language, expert systems have a number of advantages when they are used to solve problems in which information about the system, its parameters, as well as about the inputs, outputs, and system states is unreliable and poorly formalized. However, there is a significant drawback for such systems: the fuzzy rules set is formulated by experts and may be incomplete or contradictory. Therefore, the task of automatically knowledge base construction based on the observable data is urgent.

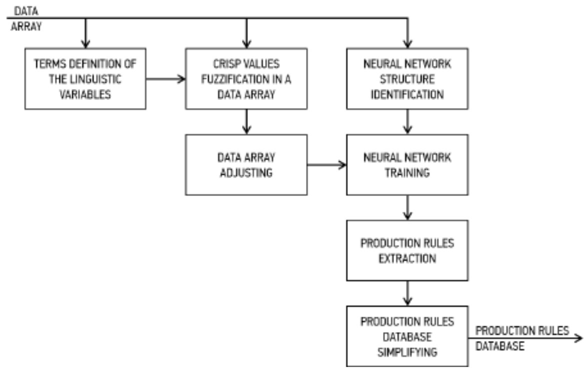


Figure 3. Conceptual scheme of Forming Production Rules in Intelligent Transportation System algorithm

IV. Scheme of Forming Production Rules in Intelligent Transportation System Based on Applying Neural Networks Models

On the 3 is shown a conceptual scheme of the automated knowledge base building process.

Step 1. Terms definition of the linguistic variables based on an experimental data set and sensitivity analysis results. The sensitivity analysis results determine the number of fuzzy values that each linguistic variable can initially accept. Also, the fuzzy membership functions are defined on this step.

At this step, we get the term sets of input parameters $I_i, i = 1, \ldots, N$, where N is the number of input parameters, and output parameters $O_l, l = 1, \ldots, L$, where L is the number of output parameters.

Each input parameter I_i and output parameter O_l can take fuzzy values ("small", "large", etc.) from their term sets (N_i and M_l are the number of corresponding fuzzy values for input and output linguistic variables).

Step 2. Fuzzification of crisp values in an experimental data set.

Step 3. The experimental array pre-processing: normalizing input values that are included in the model as quantitative variables; encoding possible input values that are included in the model as string variables; rows aggregation in the data set which has the same values of the all input parameters and the different values of the output parameters.

Step 4. Determining the structure of a neural network. To solve the task, a neural network can be used without hidden layers [16] or with hidden layers.

To solve the task of expert system knowledge base building for the intelligent transportation system it is considered to use the neural network with one hidden layer (cf. 4).

Input neurons correspond to every possible fuzzy value of input parameters:

$$x_{ij} = \begin{cases} 1 & \text{if } I_i \text{ is } j\text{-fuzzy value from } I_i \text{ term set} \\ 0 & \text{if } I_i \text{ is not } j\text{-fuzzy value from } I_i \text{ term set} \end{cases}$$

where j = 1, ..., N_i. If the value of the parameter I1 is "small", then it cannot be any different fuzzy value at the same time. In other words, only one of $x_{11}, x_{12}, \ldots, x_{1N_i}$ can take the value 1 at a given time.

Similarly, output neurons correspond to every possible fuzzy value of output parameters:

$$y_{ls} = \begin{cases} 1 & \text{if } O_l \text{ is } s\text{-fuzzy value from } O_l \text{ term set} \\ 0 & \text{if } O_l \text{ is not } s\text{-fuzzy value from } O_l \text{ term set} \end{cases}$$

where s = 1, ..., M_l.

Hidden layer neurons are designated h_k, $k = 1, \ldots, K$, $K \geq L$ – number of hidden layer neurons which shouldn't be less then output neurons.

Step 5. Neural network training.

At this stage, the neural network weights will be calculated:

- V_{ijk} - weights of the trained neural network (input – hidden layer)
- W_{kls} - weights of the trained neural network (hidden layer - output)

Step 6. Production rules extraction from the results of parametric identification of a neural network. The number of rules matches the output neurons number. Extracting production rules can be reduced to selecting the most dominant rule for each neuron.

the hidden layer neuron is selected for every output layer neuron as follows

$$h_s = \max_s (h_k * w_{ksl}) \tag{8}$$

Then, for this hidden layer neuron a combination of input neurons is selected (one for each input parameter) as follows:

$$\forall I_i : \qquad \max_j (x_{ij} * v_{ijk} - b_k), \tag{9}$$

where k is the index of the hidden neuron (8) which is chosen for every output layer neuron.

Step 7. Simplifying the knowledge base: removing duplicate rules, merging rules, etc.

V. Conclusion

The article presents an algorithm and results of sensitivity analysis for the model of estimate capacity, describes an expert system which is designed to control traffic flows in high-speed transportation corridors. The scheme of the algorithm for automated knowledge base building based on a neural network is also proposed. A distinctive feature of the algorithm is the neural network structure with a single hidden layer. The proposed extracting production rules algorithm provides the most dominant rule selecting for each output neuron: selecting a neuron from a hidden layer, and then a combination of fuzzy values selecting for each input parameter included in the rule, based on the analysis of weights obtained as a neural network training result.

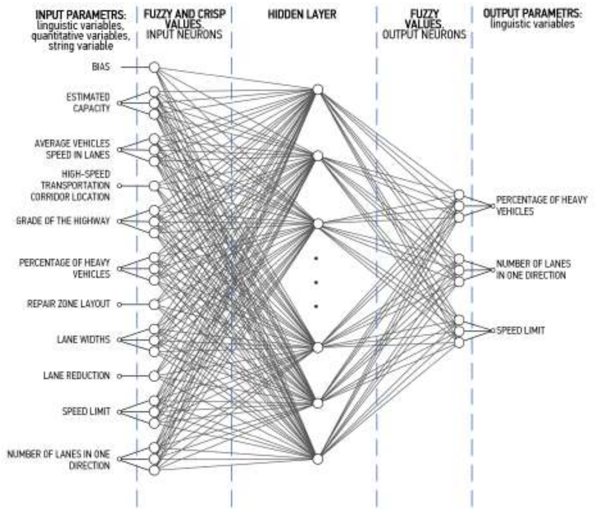


Figure 4. Example of the Neural Network Structure for Forming Semantic Rules Task

In the future, it is planned to study the required neurons number on the hidden layer and build a knowledge base for transport corridors in the research region.

REFERENCES

[1] M. Chowdhury, A. Apon, and K. Dey, “Data Analytics for Intelligent Transportation Systems”. 2017.
[2] ITS-America, “ITS-America. The official website.”, URL: https://www.itsa.org.
[3] “National Technological Initiative”, URL:http://www.nti2035.ru.
[4] TRB, “Highway Capacity Manual — HCM 2010”, 2010.
[5] FGSV, “German Highway Capacity Manual, HBS — 2015” (Handbuch fuer die Bemessung von Strassenverkehrsanlagen), 2015.
[6] Rosavtodor, “Russian National Guidelines to detect the capacity of motor roads ODM 218.2.020–2012”. 2012.
[7] J. A. C. van Toorenburg, “Praktijkwaarden voor de Capaciteit” (Practical Values for Capacity). Transport Research Centre (DKV), Ministry of Transport, Public Works and Water Management, Rotterdam, 1986.
[8] P. V. Saraev, S. L. Blyumin, A. V. Galkin, and A. S. Sysoev, “Neural remodelling of objects with variable structures,” Proceedings of International Conference on Intelligent Information Technologies for Industry, 2017, pp. 141–149.
[9] N. von der Heiden and J. Geistefeldt, “Capacity of freeway work zones in Germany,” Transportation research procedia, No. 15, 2016, pp. 233–244.
[10] A. Sysoev and N. Voronin, "Approach to Sensitivity Analysis of Stochastic Freeway Capacity Model Based on Applying Analysis of Finite Fluctuations," 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russia, 2019, pp. 621-626.
[11] W. Brilon, J. Geistefeldt, and M. Regler, “Reliability of freeway traffic flow: a stochastic concept of capacity,” Proceedings of the 16th International symposium on transportation and traffic theory. Vol. 125143. Maryland: College Park. 2005.
[12] A. Saltelli, “Sensitivity analysis for importance assessment,” Risk analysis, vol. 22.3, 2002, pp. 579–590.
[13] S.L. Blyumin, G.S. Borovkova, K. V. Serova, and A.S. Sysoev, “Analysis of finite fluctuations for solving big data management problems,” Proceedings of the 9th IEEE International Conference on Application of Information and Communication Technologies (AICT), 2015, pp. 48–51.
[14] A. Sysoev, A. Ciurlia, R. Sheglevatych, and S. Blyumin, “Sensitivity Analysis of Neural Network Models: Applying Methods of Analysis of Finite Fluctuations”, Periodica Polytechnica Electrical Engineering and Computer Science, vol. 63, no. 4, pp. 306-311, 2019.
[15] A. Sysoev, S. Hohmann, and J. Geistefeldt, “Differential Evolution Approach To Calculate Optimal Ramp Metering Rates,” Int. J. Traffic Transp. Eng., 2017.
[16] S.H. Huang, H. Xing, "Extract intelligible and concise fuzzy rules from neural networks", Fuzzy Sets and Systems, vol. 132, no. 2, pp. 233–243, 2002.

Построение семантических правил для управления транспортным потоком в интеллектуальной транспортной системе

Хабибуллина Е.Л., Сысоев А.С.

В статье приводится подход к математическому ремоделированию пропускной способности участков высокоскоростных транспортных коридоров. Предложенный метод основан на применении нейросетевой модели, сочетающий результаты оценки стохастической пропускной способности в период транспортного затора и свободного движения транспортных средств. Основываясь на полученной модели предлагается проводить анализ важности факторов, оказывающих влияние на оцененную пропускную способность с использованием методов анализа конечных изменений. Также на основании предлагаемых подходов представлена схема формирования продукционных правил в интеллектуальной транспортной системе.

Received 30.01.20

Information Advising System for Safe Fleet Operations

Dariya A. Karaeva
Federal Research Center computer Sciences and Control,
Russian Academy of Sciences
Moscow, Russia
dariandr95@gmail.com

***Abstract*—Safe fleet operations requirest the implementation of a large number of documented rules. When preparing a vessel for a voyage or directly during operation, there is often a need of actual information concerning safe operation regulations in effect, which volume can be quite considerable. Help for the users can be provided through automation of search and extraction processes.**

The article deals with an approach to the design of information advising system on safe fleet operations, based on semantic technologies. System users and their main informational needs are described. The problems, solved by the system are determined. Functional blocks of the system are highlightened. The requirements for the informational and documentational supply are set. Such a system will fasten and simplify relevant information acquisition.

***Index Terms*—information advising systems, semantic technologies, precedents search, safe fleet operations**

I. Introduction

Maritime transportation is a kind of human activity, involving a great number of risks, both for people working on vessels and environment. There are a lot of rules sets, conventions and other kind of orders, regulating any aspects of this activity to provide safe fleet operations. A vessel can be operated, in particular allowed to enter an international port, only after the fulfillment of all the regulations and rules has been checked. To search for the particular data and verify all orders and constraints to be fulfilled, one should analyze large volumes of information, quite a difficult, and what's more important, time-consuming process. One possible solution for this kind of problems, related to the automatic search over document corpus, is based on the use of information advising intellectual systems and semantic technologies.

A vessel represents a complex autonomous engineering system, operated by the minimum required for navigation number of crew members. For the cargo handling operations to be done in ports the vessel should come through the vetting procedure, i. e. get a permission from a certified specialist. Vetting is carried out by independent experts according to strictly determined parameters, established by international rules and conventions.

Safe fleet operations problem includes the following main aspects:

- Route working out including national border crossing, domestic waters of foreign countries, traffic through neutral waters and regions with regulation restrictions.
- Vessel operation during interrepair period: life boat drill schedule, equipment inspection, health and safety regulation compliance on the board, environmental discharge norms, etc.
- Crew training: crew members certification, medical examination and training preparation.
- The berth take up in host country.

The above aspects are strictly regulated and described in various normative documents. In order to work effectively in these cases, it is necessary to search for the necessary information, which may take a long time depending on the complexity of the issue or the number of regulating documents. For example, when navigating a complex route through a large number of regions with different regulations in action, to verify that the ship's equipment meets all regulations becomes quite a time-consuming process. This problem's solution is based on precedents search and identification of similar situations. To increase efficiency of relevant documents search automation of this activity is required.

II. System concept

A. Users

Let's define systems users. Broadly speaking, main user is represented by the shipping company, interested in the possibility of vessel operation with commercial or any other purposes, and thus interested in respecting rules, regulating such an operation. As mentioned above, norms compliance check is carried out during vetting process, which results unified register forms companies activity assessment, its reputation and expertise level. Nevertheless, systems interface focuses on terminal users – persons, implementing ship operation. Let's identify users, according to the described typical situations, requiring informational support.

While route working out, two main system users can be distinguished – logistic specialists and navigators. The former accomplish preliminary route working out according to existing rules for a specific vessel. Often situations arise (precedents) when ships do not come up to ecological standards, taking place in some regions, which is why such ships should choose alternative routes. Logisticians have to conduct complex analysis of all the norms in action so that to find an optimal route, with possible additional ship preparation or if it is not possible due to design or any other features, a route with norms that are already come up by the vessel.

In addition, due to weather or other conditions it may not be possible to follow the planned path. The navigator choosing a detour route needs information support in such situations. In this case, it is sufficient to analyze only those rules that apply in a particular area or several areas.

As the ship is operated by its crew, different crew members may use the system for information support according to their areas of responsibility. In contrast to the first situation, the questions listed are governed by fewer strictly defined rules, and in this situation it is not so much the analysis of all sources that is important but the responses to specific requests. In addition, the system is useful to the master at this stage. Since the master is responsible for the crew actions, he needs to check the integrity of the work even in areas where his competence is limited. The advising system effectively performs such checks by searching only for important information without the necessity for the mater to analyse a large amount of data.

While the preparation and certification of crew members, the system provides substantial assistance both to crew members checking or deepening their knowledge in certain areas and to certifying persons to check crew knowledge. In addition, at this stage the user can be also represented by a shipping company that monitors what competencies the crew should have in accordance with current international and local standards.

Finally, in order to moor a ship, it is necessary to prepare it according to the rules of the host country. For this purpose, the system is used by both crew members and representatives of the shipping company responsible for preparing the vessel for the voyage.

B. Goal and tasks

The main purpose of the system is to provide information that responds to a user request, which contains a description of a certain problem situation or specific question. To achieve this goal, it is necessary to solve the tasks of analyzing the request, formalizing the situation corresponding to the request, searching for a precedent for such a situation and relevant regulatory documents. If there is no such precedent, it is necessary to refer to the entire body of the regulatory documents.

Formally, advice generation task can be posed as follows:

$$D = \langle F, O, X, G \rangle \tag{1}$$

Here F is a task formulation including a description of the problem, definition of the goal and the result form. The task is represented by a user question related to the operation of the vessel and somehow reflected in the regulatory documents. The question is formulated in a limited natural language and the problem is formalized as some semantic structure. In general, the aim of the system will be to find the most consistent answer based on the analysis of the situation, search for a precedent and, if necessary, search in the corpus of all regulatory documents on ship operation. Presentation of the result is most preferable in the form of a coherent answer to a question in natural language with references to the documents on the basis of which this answer is drawn up. In this work we will limit ourselves to present quotations from sources containing an answer to a question with references to documents.

O - a set of possible options from which the choice is made in the absence of precedents, or if there are several similar, but not fully corresponding to the initial situation. Answering the given question the system will carry out search both in set of formalized situations - precedents, and in the whole set of available documents. Therefore, in the broad sense, a set of options will constitute the entire body of texts or the entire knowledge base, in some way formed from such a body.

X - a formalized description of situations - precedents. Since the solution is to search for some textual information, case studies will include semantic structures of texts describing situations as well as structures of documents related to these situations. Semantic design technologies are used in this subsystem.

G - conditions that restrict the search area. An example of such a restriction is the restriction of information sources. In particular, when entering territorial waters, local rules will dominate international rules, which may be temporarily omitted or treated with less priority.

Problematic situations related to the safe operation can widely range both in the extent and content. The main feature distinguishing the problem being solved is the possibility to formulate the question in natural language, which answer is somehow presented in explicit or implicit form in the convention texts. Nevertheless, the question can be both very generic (e. g. “Which scope of competence should the master possess?”), or very concrete and contain a direct answer in the text (e. g. “What length should mooring lap have?”).

C. System functional blocks

Fig. 1 shows the conceptual scheme of the system identifying main functional blocks that implement the described process of developing the council.

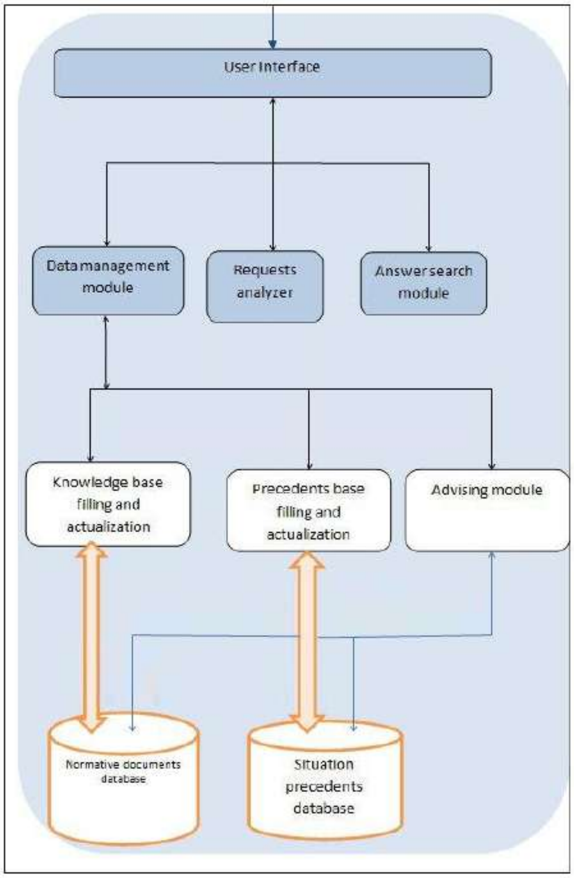


Figure 1. Conceptual scheme of automated system

As the given system co-operates with the user, as well as any similar system, it contains the user interface. The interface problem includes development of communication between the user and the system. The system contains data management module, requests analyzer and answers search module. Each of these modules provides both functioning and support of corresponding knowledge bases, and search and construction of the answer responding to the users request. The requests analyzer aims to perform semantic analysis of a user request presented in a limited natural language, to build request's semantic structure and to search for a precedent corresponding to this request.

Precedents data base and normative documents on ship operation upload is carried put through data management module. It is necessary to store initial documents in the formalized form as some knowledge base on which effective search will be carried out. Besides, as documents are periodically updated, it is necessary to have a possibility of documents verification and loading in a knowledge base. This module converts documents into knowledge, solves the tasks of storing documents, converting them into knowledge and storing transformed data. One of the most important functions of the given subsystem is filling, actualization and support of precedents database according to precedents model.

III. Conclusion

Safe fleet operations is an important task from the point of view of both environmental and human life protection. The importance of the task is confirmed by the abundance of rules and documents regulating this activity. For practical application such volume of data is redundant and brings with it certain difficulties related to analysis of all available information. The application of intellectual systems based on semantic technologies will make it possible to cope with this problem.

The field of application of this consulting system is quite wide. It will allow shipping companies to significantly reduce preparation time for vetting, as well as quickly and easily check compliance with the norms of the host country. Such system may also help in resolving disputes arising during inspections. Finally, the intelligent system can be used in crew training.

References

[1] M. Ju. Popov, S. A. Fomenkov, and A. V. Zaboleeva-Zotova. Principy raboty i osobennosti realizacii universal'nogo semantiko-sintaksicheskogo analizatora russkogo jazyka. *Vestnik komp'juternyh i informacionnyh tehnologij*, (7(13)):37–42, 2005.

[2] A. B. Petrovsky. Group verbal decision analysis. In F.Adam and P.Humphreys., editors, *Encyclopedia of Decision Making and Decision Support Technologies. Vol.1.*, pages 418–425. Information Science Reference, Hershey, New York, 2008.

[3] A. B. Petrovsky. *Teorija prinjatija reshenij*. Izdatel'skij centr "Akademija", Moscow, 2009.

[4] A. B. Petrovskij, A. N. Romanov, G. I. Shepelev, and N. V. Ogurcova. Sistema podderzhki prinyatiya reshenij v nestrukturirovannyh situaciyah. Zayavka # 2012151196/08 ot 29.11.2012.

[5] A. B. Petrovsky. *Gruppovoj verbal'nyj analiz reshenij*. Nauka, Moscow, 2019.

[6] V. N. Volkova. *Teorija informacionnyh processov i sistem*. Jurajt, Moscow, 2014.

[7] V. V. Persianov and E. I. Logvinova. *Informacionnye sistemy*. Direct-Media, Moscow-Berlin, 2016.

[8] P.R. Varshavskiy and A.P. Eremeev. Modelirovanie rassuzhdenij na osnove precedentov v intellektual'nyh sistemah podderzhki prinjatija reshenij. *Iskustvenniy intellekt y prinyatiye resheniy*, (2):45–57, 2009.

[9] C.C. Aggarwal and C.X. Zhai. *Mining Text Data*. Springer Science and Business Media, New York, 2012.

[10] W. Maass and T. Kowatsch. *Semantic Technologies in Content Management Systems: Trends, Applications and Evaluations*. SpringerLink : Bücher. Springer Berlin Heidelberg, 2012.

[11] Averchenkov V.I., Zaboleeva-Zotova A.V., Kazakov YU.M., Leonov E.A., and Roshchin S.M. *SISTEMA FORMIROVANIYA ZNANIJ V SREDE INTERNET*. Moskva, (3-e izdanie, stereotipnoe), 2016.

[12] Dolbin A.V., Rozaliev V.L., Orlova YU.A., and Zaboleeva-Zotova A.V. Semanticheskij analiz teksta dlya raspoznavaniya elementov vneshnego vida cheloveka. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nyh sistem.*, (6):401–404, 2016.

[13] Dmitriev A.S., Solov'ev I.S., and Zaboleeva-Zotova A.V. Izvlechenie vzaimosvyazej mezhdu ob'ektami i terminami v tekstah na ekonomicheskuyu tematiku. *Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta.*, (13(177)):55–60, 2015.

[14] Zaboleeva-Zotova A.V., Orlova Y.A., Rozaliev V.L., Fomenkov S.A., and Petrovsky A.B. Formalization of inital stage of designing multi-component software. In *Proceedings of the IADIS International Conference Intelligent Systems and Agents 2013, ISA 2013, Proceedings of the IADIS European Conference on Data Mining 2013, ECDM*, pages 107–111, 2013.

[15] Orlova YU.A., Rozaliev V.L., and Zaboleeva-Zotova A.V. Annotirovanie i vizual'noe predstavlenie tekstovoj informacii v graficheskom vide. *Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta.*, (13(177)):74–85, 2015.

[16] Zaboleeva-Zotova A.V. Orlova YU.A., Fomenkov S.A. *OSNOVY SISTEMNOGO ANALIZA. Uchebnoe posobie(grif). Dop. UMO vuzov po universitetskomu politekhnicheskomu obrazovaniyu / Volgogradskij gosudarstvennyj tekhnicheskij universitet.* Volgograd, (2-e izdanie, pererabotannoe i dopolnennoe), 2013.

[17] Zaboleeva-Zotova A.V. Berdnik V.L. Model' 'semanticheskoe pyatno' v slozhnoformalizuemyh zadachah intellektual'noj obrabotki informacii. *Izvestiya YUFU. Tekhnicheskie nauki.*, (1(126)):116–121, 2012.

[18] Polovinkin A.I. Zaboleeva-Zotova A. Logiko-lingvisticheskoe modelirovanie pri analize i sinteze intellektual'nyh sistem : Inform. *Analiticheskij bankovskij zhurnal.*, (3):213, 2011.

[19] Kamaev V.A. Zaboleeva-Zotova A.V. *LINGVISTICHESKOE OBESPECHENIE AVTOMATIZIROVANNYH SISTEM. Uchebnoe posobie dlya studentov vysshih uchebnyh zavedenij, obuchayushchihsya po napravleniyu "Informatika i vychislitel'naya tekhnika", special'nosti "Programmnoe obespechenie vychislitel'noj tekhniki i avtomatizirovannyh sistem".* Moskva, 2008.

[20] Zaboleeva-Zotova A.V. Formalizaciya semantiki teksta pri avtomatizacii slabostrukturiruemyh procedur v processe sinteza tekhnicheskih sistem. *Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta.*, (4(21)):36–43, 2006.

[21] Horoshevskij V. F. Preobrazhenskij A. B., Rybina G. V. Generaciya mnogocelevyh intellektual'nyh voprosno-otvetnyh sistem. *Izv. AN SSSR. Tekhn. Kibernetika.*, (6):142–151, 1979.

[22] Cinman L. L. Yazyk dlya zapisi lingvisticheskoj informacii v sisteme avtomaticheskogo perevoda etap. *Semiotika i infomatika*, (27):82–120, 1986.

[23] Zaboleeva-Zotova A.V. Popov M.YU., Fomenkov S.A. Principy raboty i osobennosti realizacii universal'nogo semantiko-sintaksicheskogo analizatora russkogo yazyka. *Vestnik komp'yuternyh i informacionnyh tekhnologij.*, (7(13)):37–42, 2005.

[24] Popov E. V. *Obshchenie s EVM na estestvennom yazyke.* M.: Nauka, 1982.

[25] Pospelov D. A. *Logiko-lingvisticheskie modeli v sistemah upravleniya.* M.: Energoizdat, 1981.

[26] Cejtin G. S. Programmirovanie na associativnyh setyah. *EVM v proektirovanii i proizvodstve*, (2):16–48, 1985.

[27] SHenk R. *Obrabotka konceptual'noj informacii.* M.: Energiya, 1980.

[28] Boran-Keshish'yan A.L. Astrein V.V., Kondrat'ev S.I. Predstavlenie precedentov v sudovoj sppr bezopasnosti sudovozhdeniya. *Morskie intellektual'nye tekhnologii*, 3(4(46)):147–152, 2019.

[29] YArushkina N.G. Moshkin V.S. Gibridizaciya algoritmov izvlecheniya znanij iz tekstov i mekhanizma precedentov v processe rasshireniya ontologii. In *Pyatnadcataya nacional'naya konferenciya po iskusstvennomu intellektu s mezhdunarodnym uchastiem Trudy konferencii: v 3 tomah.*, pages 80–87, 2016.

[30] Pleshkova O.A. Nikolaenko D.V., Nikolaenko V.L. Precedenty sistemy vneshnego upravleniya na dorozhnom primitive. *AKTUAL'NYE PROBLEMY AVTOTRANSPORTNOGO KOMPLEKSA Mezhvuzovskij sbornik nauchnyh statej (s mezhdunarodnym uchastiem)*, pages 162–165, 2019.

[31] Fastovickij E.I. Nikolaenko V.L. Analiz precedentov sistemy monitoringa meteo uslovij na dorozhnom primitive. *Nauchno-tekhnicheskie aspekty razvitiya avtotransportnogo kompleksa Materialy V mezhdunarodnoj nauchno-prakticheskoj konferencii*, pages 373–376, 2019.

[32] YUrin A.YU. Metody gruppovogo vybora dlya adaptacii reshenij, poluchennyh v rezul'tate rassuzhdenij na osnove precedentov. *Iskusstvennyj intellekt i prinyatie reshenij.*, (3):78–85, 2013.

[33] ZHukova M.N. Zolotarev A.V. O zadache formirovaniya bazy znanij s ispol'zovaniem precedentov. *Aktual'nye problemy aviacii i kosmonavtiki.*, 7(7):420–421, 2011.

[34] Alekhin R.V. Metod izvlecheniya precedentov na osnove teorii strukturnogo otobrazheniya. In *RADIOELEKTRONIKA, ELEKTROTEKHNIKA I ENERGETIKA tezisy dokladov Dvadcat' vtoroj Mezhdunarodnoj nauchno-tekhnicheskoj konferencii studentov i aspirantov: v 3-h tomah*, page 215, 2016.

[35] Nikolajchuk O.A. Maltugueva G.S. Primenenie nechetkih mul'timnozhestv dlya ocenki blizosti precedentov. In *Pyatnadcataya nacional'naya konferenciya po iskusstvennomu intellektu s mezhdunarodnym uchastiem Trudy konferencii: v 3 tomah.*, pages 115–122, 2016.

[36] Troshina G.V. Aleksandrov A.M. Ispol'zovanie diagrammy precedentov pri razrabotke bazy dannyh na primere obrabotki informacii o muzejnyh eksponatah. *Sbornik nauchnyh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta.*, (3(89)):78–91, 2017.

[37] Antsiperov V.E. Identification of the point process intensity shape with the precedents maximum likelihood distributions. *Jurnal radioelectroniki.*, (12):11, 2017.

[38] Myznikov P.V. Izvlechenie informacii iz novostnyh tekstov s pomoshch'yu modelirovaniya rassuzhdenij na osnove precedentov. In *MOLODYOZH' I SOVREMENNYE INFORMACIONNYE TEKHNOLOGII sbornik trudov XV Mezhdunarodnoj nauchno-prakticheskoj konferencii studentov, aspirantov i molodyh uchyonyh. Nacional'nyj issledovatel'skij Tomskij politekhnicheskij universitet.*, pages 51–52, 2018.

Информационно-советующая система для безопасной эксплуатации судов

Караева Д.А.

Безопасная эксплуатация судов требует выполнения многих правил, регламентированных утвержденными документами. При подготовке судна к рейсу или его эксплуатации необходима актуальная информация о действующих регламентах по безопасной эксплуатации, объем которой может быть весьма значительным. Существенную помощь пользователям может оказать автоматизация процедур поиска и выдачи необходимых данных и документов.

В статье описан подход к созданию информационно-советующей системы по безопасной эксплуатации судов, основанной на семантических технологиях. Описаны пользователи системы и их основные информационные потребности. Определены задачи, решаемые системой. Приведена концептуальная схема системы. Наличие подобной системы позволит ускорить и упростить получение релевантной информации.

Received 30.01.2020

Agent-space transfer learning for sign-based world model

Maksim Rovbo
National Research Center «Kurchatov Institute»
Moscow, Russia
rovboma@gmail.com

***Abstract*—Several algorithms were studied for transfer learning problem for a foraging task. These algorithms are suitable for a semiotic control system to facilitate experience transfer between two learning agents with different agent-space descriptions of the state in a predicate form. They require that the target agent's description of the task is a subset of the source agent's description.**

The algorithms are based on the Q-learning algorithm and uses a custom transfer function to initialize the target agent's state-action table, which matches the corresponding predicates. The test problem used to test the algorithms is a foraging task, where an agent must gather randomly generated food in a grid world.

The results show that they provide improvement in the learning curve for the target agent after transferring experience from an agent with more input predicates. The improvement is inconsistent and sometimes does not bring noticeable difference in performance, but the target agent performs at least as good as an agent with no prior experience.

***Keywords*—transfer learning, reinforcement learning, semiotic control, robotics**

I. Introduction

Robotic systems, both single- and multi-agent, need to have adaptive properties to robustly achieve various goals in real, uncontrolled environments, which is often the case for mobile robots. One of the prominent research areas that can achieve that is reinforcement learning, which views robots as agents with a predefined input space and discrete or continuous actions that act in an environment and receive rewards for useful actions. This allows robots to learn and optimize goal-directed behavior. On the other hand, semiotic control systems allow robots to have a structured and more interpretable world models based on rules and predicates. To use reinforcement learning methods requires some adaptation of the corresponding models and algorithms since semiotic control systems use signs and connections between them as a basis for description of the world, while reinforcement learning mostly uses a vectorized description.

Sign has four parts, including the name, the percept, the functional meaning and the personal meaning [6]. The percept is a set of predicates used to describe the concept encoded by the sign. It can be used to connect the sensors of the agent and its logical control part of the control system by implementing some of the predicates as algorithmic functions of the sensor data. Different agents can have different predicates available. To use reinforcement learning in a multi-agent system for such agents requires a way of using different predicate descriptions when transferring experience between agents.

Transfer learning in reinforcement learning deals with the application of the knowledge about solving a problem by an agent to another, somewhat similar problem, which can also be used to transfer experience between different agents. A distinction can be made between action-space and problem-space descriptions of the task, which is important to consider, since it can lead to more efficient learning algorithms [3]. Problem-space descriptions are usually full descriptions of the state of the environment and are thus impractical for mobile robots, so the focus in this work is on agent-space descriptions. Transfer learning can be done in different ways, which also depend on the problem specifics such as whether the input representation is the same space, are there common useful sequences of actions between tasks, etc. Transfer learning methods can be based on modifying action selection, for example, by adding a heuristic function that contains the relevant knowledge [1], deep learning [7], transfer of samples from one task to another [5]. Manual mapping between tasks and transferring data between approximators [10] is the most similar to the problem considered here, since it can be used for agents with different input spaces and also uses averaging.

The idea of the current paper is to use descriptions of the world inherent in a semiotic control system to facilitate experience transfer between two learning agents with different agent-space descriptions of the state. In terms of the taxonomy given in [4], the problem considered here is the transfer across tasks with different domains (from the agent-space point of view) or transfer across tasks with fixed domain (from the problem-space point of view), while the algorithm uses a special case of parameter transfer that accounts for the difference in state representations of the agents.

A foraging problem is considered as a test environment for the algorithms since it is common for multi-agent robotic research, was used for reinforcement learning evaluation and can serve as a model for some application such as resource gathering or energy collection [2], [11].

II. Methods and algorithms

The objective of the method is to improve the performance of the learning agent by transferring experience to it from a source agent with a possibly different input space. To connect them, a description is proposed based on the notion of a sign [6] as used in some semiotic control systems [8].

Formally, a sign is defined as

$$S = < n, p, m, a >, \tag{1}$$

where n is the name of the sign, p is the percept that describes it (in this case, a corresponding sensor description), m is the knowledge about the sign encoded in rules (which is not important for task in this paper) and a is the encoded personal experience of the agent, such as actions associated with the sign. The name can be used to match different agents' descriptions of the input even when the sensors used for detection of the objects are different and the personal experience can be used to store a combination of an agent's state sign and the corresponding action and results for reinforcement learning.

A grid foraging problem is used as a test environment in the following sections, but it is useful to introduce some of the elements here. An agent moves on a grid and receives the local state of the environment as input. Two types of agents' inputs are considered. One of them will be called 'type A', which has sensors that allow it to determine, what is located in the 3x3 square around it. The other one, which will be called 'type B', does not have diagonal sensors, so it only has 5 squares as the input (a cross with itself in the center).

The developed transfer algorithm is based on the idea of grouping state-action values using the common parts of predicate descriptions or names of the corresponding semiotic state descriptions. In the relatively simple case considered in this work, they allow matching between different state descriptions of agents by determining which subset of states of one agent corresponds to a single state description of the other agent. Let's consider a world description for an agent of type A shown in "Fig. 1". For simplicity, statements about sensor data such as $"agent1"has"S1";"S1"has("0")$ will be shortened to $S1("0")$.

If another agent ("agent 2" of type A) had a similar description, but with more sensors (with names "S1", "S2", ..., "S9"), so that "agent 1" has a subset of sensors of "agent 2", then any state of "agent 2" of type A is a part of the subset of a state of "agent 1", which can be matched by names of the sensors. More complex descriptions of states in an agent based on semiotic descriptions can involve multiple signs with intersecting predicate descriptions, but those cases are not considered in this work and are a subject of further research.

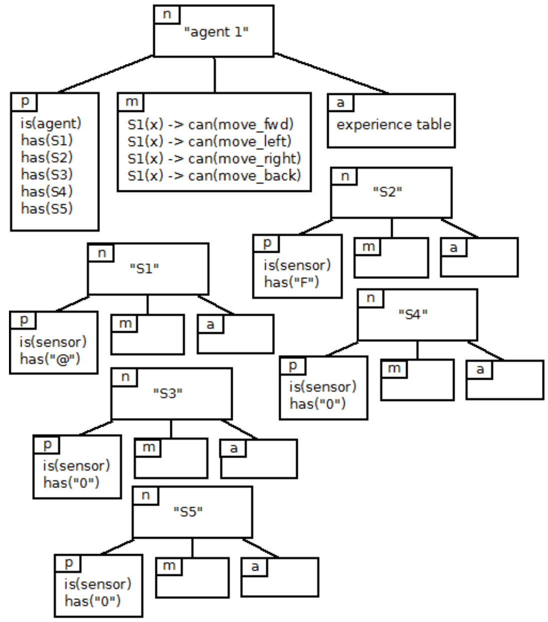


Figure 1: Sign-based description of a world model for an agent with 5 sensors that can each show an object denoted by "@" (name of a food item), "0" (name of an empty cell), etc. The agent sign connects an agent with the name "agent 1" to its sensor objects in *p*, expresses rules that determine when a certain action can be performed in *m* and can store its experience as transition rules in *a*.

The ϵ-greedy Q-learning algorithm [9] is used for individual learning of the agents, while the transfer function is custom as described further, so the individual learning step updates the state-action value function according to the formula

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + Q(s',a') - Q(s,a)), \tag{2}$$

Action selection for the Q-learning algorithm in this work is done in two different ways, continuous using the softmax function

$$P(a') = \frac{e^{Q(s',a')}}{\sum_{i=0}^{num\_actions} e^{Q(s',a_i)}}, \tag{3}$$

and discontinuous using the argmax function

$$a' = argmax_{a_i} Q(s', a_i), \tag{4}$$

After a mapping between different agent state spaces has been established, a way to transfer the state-action value function needs to be defined. In this case a simple averaging function is used for the corresponding subset.

So, the state-action value function of the source agent $Q_s(s_s, a)$ is transferred to the initial state-action value function of the target agent by the formula:

$$Q_t(s_t, a) = \frac{1}{n_t} \sum_{s_s \subseteq s_t} Q_s(s_s, a), \tag{5}$$

where n_t is the number of nonzero $Q_s(s_s, a)$ such that $s_s \subseteq s_t$.

However, it does not take into account the probability of the agent to end up in the corresponding state. A useful estimate may be to track the number of visits $c(s, a)$ to the corresponding states by the source agent and use a weighted average for the target agent with the rate of visits as the weight. In this case, the initialization is performed as:

$$Q_t(s_t, a) = \frac{\sum\limits_{s_s \subseteq s_t} c(s_s, a) Q_s(s_s, a)}{\sum\limits_{s_s \subseteq s_t} c(s_s, a)}. \tag{6}$$

III. Experimental setup

Foraging problem was chosen to evaluate the algorithms. The environment is a grid in which an agent can move left, right, up and down (or remain in the same place). The steps and actions are discrete. Each cell of the grid can contain a number of objects or be empty. The objects are: the agent, a food item, an obstacle. A reward (1 point) is received by the agent when it picks up a food item. Food items are automatically picked up when the agent is in the same cell as a food item and it only gathers one food item at a time. Food items are initially placed randomly in cells without obstacles and are instantly created in a random place after being picked up. The environment is reset and generated randomly each experiment. The size of the grid world is 15 by 15 cells with 100 units of food (“Fig. 2”).

The agent can either have a problem-space input or an agent-space input for the problem. The problem-space input is the coordinates of the agent in the grid. This description of the state is a full observation of the state of the environment. The agent-space descriptions, on the other hand, are partial observations and consist of the descriptions of the local surroundings of the agent. As mentioned in algorithm descriptions, two agent-space descriptions were used. Type A agent receives a description the nearby 9 cells (its own included) with a list of objects in each of them (“Fig. 3a”). Type B agent only receives predicates about the up, down, left, right and its own cell (“Fig. 3b”).

The transfer task is defined as a source agent of type A and the target agent of type B acting in this environment. First, the source agent learns for a fixed amount of steps. Then, its experience is transferred to the target agent of type B by transforming the state-action value table according to the function defined in

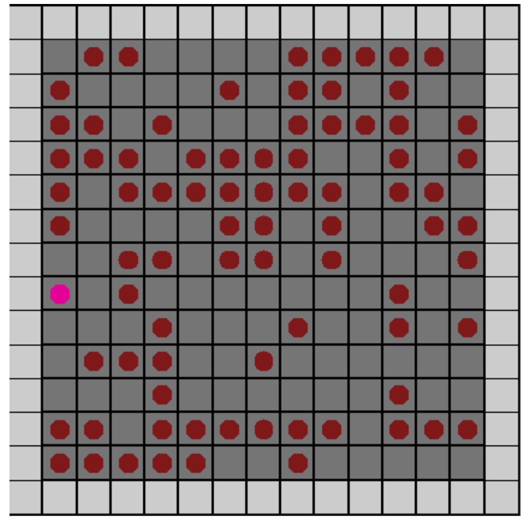


Figure 2: Simulation of the foraging problem environment. The light squares around the field are the obstacles, the bright circle is the agent, other circles are food items.

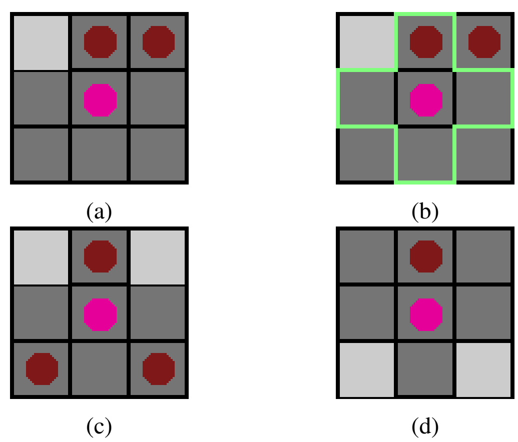


Figure 3: Examples of agent-space states, where dark red circles are food items, bright pink circle in the center is the agent, light grey squares are cells with obstacles and dark gray are empty cells. (a) shows an example of an agent-space state for an agent of type A. This corresponds to a tuple of length 9 $[@, F, 0, 0, 0, \#, F, 0, 0]$ (b) has the 5 elements of type B agent’s state highlighted, which are the center, left, right, up and down cells $[@, F, 0, 0, 0]$. (a), (c) and (d) are examples of agent-space states of and agent of type A that are considered elements of the subset of states generated by the agent-space state of (b).

the algorithms section and initializing the table of the agent of type B with it. That agent then acts and learns for the same amount of steps. The results were averaged over a number of experiments (10 or more) and grouped into 'episodes' (the task itself is continuous, not episodic) consisting of several steps (1000 unless noted otherwise) for a more readable plot.

To define the problem formally, the corresponding spaces and variables must be written out. The state of the agent in agent-space is a tuple of objects seen by the corresponding sensors, for example, $[@, F, 0, 0, 0]$ for type B agent means that there is only the agent itself on the central cell, a food item in front of it and nothing on all other sides. Since only the 'topmost' object is recorded by the sensor and there are a limited amount of objects, this state *s* is an element of a finite set of possible states S_B. Similarly, for the agent-space of the agent if type A, the tuple looks like $[@, F, 0, 0, 0, \#, F, 0, 0]$, where $@$ denotes the agent, F is a food item, $\#$ is an obstacle, 0 is an empty cell, and defines a finite set S_A. It can be noted that a state $s_b \in S_B$ forms a template for a set of states $s_a \in S_A$, where the corresponding elements of the tuple are the same. When a state $s_a \in S_A$ is from the corresponding set generated by s_b, it is denoted in this work as $s_b \subseteq s_a$ for simplicity. For example, the state of an agent of type B in "Fig. 3b" generates a set of states for an agent of type A, where the corner elements can have any object in them, like in "Fig. 3c" or "Fig. 3d".

The possible actions are always the same for all types of agents, available in any state (moving into an obstacle just does not change the current state) and simply correspond to a number in the set $A = 0, 1, 2, 3, 4$. Thus, the state-action pairs sets $S_A \times A$ and $S_B \times A$ are also finite and their value estimates can be expressed as a state-action table $\{Q(s, a) | s \in S, a \in A\}$ for the Q-learning algorithm.

The reward function $R : S \times A \times S \mapsto \mathbb{R}$ is deterministic both in agent-space and in problem-space and does not need to include a probability distribution. The reward $r(s, a, s')$, where s is the previous state, a is the chosen action, s' is the state after the transition, equals 1 when s' includes a food item in the agent's cell (first element of the state tuple) and 0 otherwise.

According to the goal of maximizing the agent's food-gathering efficiency, the goal function to maximize is the discounted return from the current time *t*:

$$G_t = \sum_{k=0}^{\inf} \gamma^k R_{t+k+1} \tag{7}$$

The learning algorithm used for the agent's individual learning is ϵ-greedy Q-learning with the following parameters:

- random action chance $\epsilon = 0.05$;
- learning rate $\alpha = 0.1$;
- reward discounting factor $\gamma = 0.9$;
- initial state-action table is generated as a uniform distribution U(A) for each state (in target algorithm a transfer function is used to set it).

The corresponding policies that govern the agent's choice of actions are the ϵ-greedy and softmax Q-learning alternatives respectively:

$$\pi(s, a) = \frac{\epsilon}{num\_actions} + (1 - \epsilon)\mathbb{1}_{argmax_{a_i} Q(s', a_i)} \tag{8}$$

$$\pi(s, a) = \frac{e^{Q(s', a')}}{\sum_{i=0}^{num\_actions} e^{Q(s', a_i)}} \tag{9}$$

For comparison, during the evaluation of performance of the algorithms, the performance of the agent of type B without transfer is also plotted so that the effects of experience transfer can be compared directly to the same type of agent.

Computational experiments were carried out in a custom Python simulation system.

IV. Results and discussion

The transfer algorithm based on averaging of state-action values grouped by common parts of predicate descriptions for the source agent of type A and the target agent of type B had the learning curve in "Fig. 4" for argmax action selection and "Fig. 5" for softmax action selection.

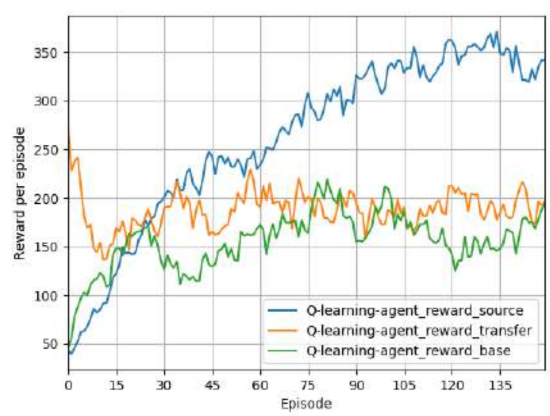


Figure 4: The asymptotically best performing algorithm of these three is the source algorithm which is in full agent-space (type A), its experience is transferred to the target algorithm with limited agent-space (type B) which shows a better starting efficiency and the worst performing algorithm is a type B agent without transferred experience that is a baseline for comparison. The source and target algorithms use argmax for action selection.

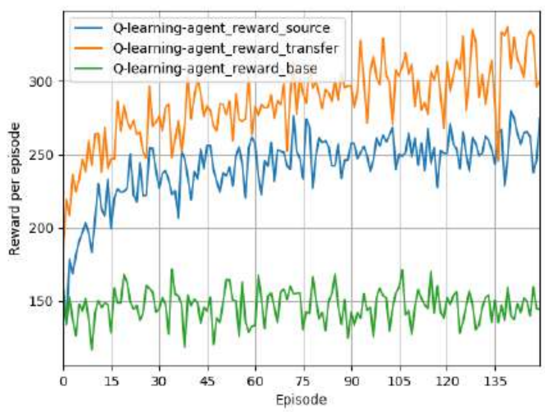


Figure 5: The full agent-space agent (type A) acts as a source and has its experience is transferred to the target algorithm with limited agent-space (type B) which shows a better starting efficiency and the worst performing algorithm is the baseline agent. The source and target algorithms use softmax for action selection.

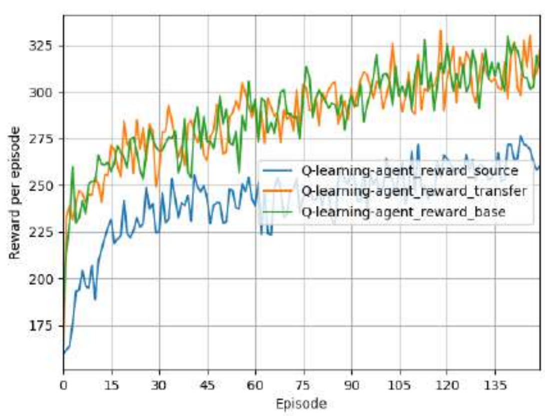


Figure 6: The weighted averaging algorithm alternative. The source agent with full agent-space (type A) here did not learn its optimal performance and the subsequent transfer to the agent of type B shows no improvement over the base non-transferred agent. The source and target algorithms use softmax for action selection.

The weighted average alternative did not show any improvement over the base algorithm "Fig. 6"

The transferred experience immediately improves the efficiency of the agent in some cases (in the argmax experiment) while the learning speed is mostly improved in the softmax experiment. These results may indicate that continuous action selection benefits more from useful hints for better learning and exploration, while non-continuous argmax selection can get stuck on the initial strategy for some time before adjusting the estimations enough and thus needs a good initial strategy. For the foraging environment with agent-space input the full state is only partially observable, which favors probabilistic (softmax) strategies over deterministic (argmax). Both the target agent and the untrained type B agent eventually learn policies with similar efficiency. The alternative weighted algorithm may have overestimated common but not beneficial situations and thus not given any improvement over an untrained agent.

V. Conclusion

Several algorithms were studied for transfer learning problem for a foraging task. These algorithms can be used for special cases of agents with a semiotic control description of world models. The results show that they provide improvement in the learning curve for the target agent after transferring experience from an agent with more sensors (in the form of input predicates), but this improvement is inconsistent and sometimes does not bring noticeable difference in performance.

The most noticeable limitation of the method is the requirement of the target agent's description of the task to be a subset of the source agent's description. This constraint may be relaxed by making better use of the connections between signs within the agents' world models, and applying logical inference, which is a subject of further research.

Acknowledgment

References

[1] Bianchi, R. A. C., Celiberto, L. A., Santos, P. E., Matsuura, J. P., & Lopez De Mantaras, R. "Transferring knowledge as heuristics in reinforcement learning: A case-based approach," Artificial Intelligence, vol. 226, pp. 102—121. 2015. DOI: 10.1016/j.artint.2015.05.008.

[2] Kernbach, S., Nepomnyashchikh, V. a., Kancheva, T. "Specialization and generalization of robot behaviour in swarm energy foraging," Mathematical and Computer Modelling of Dynamical Systems, vol. 18, no. 1, pp. 131—152. 2012. DOI: 10.1080/13873954.2011.601421.

[3] Konidaris, G., & Barto, A. "Building portable options: Skill transfer in reinforcement learning," IJCAI International Joint Conference on Artificial Intelligence, pp. 895—900. 2007.

[4] Lazaric, A. "Transfer in Reinforcement Learning: A Framework and a Survey," Reinforcement Learning. Adaptation, Learning, and Optimization, vol. 12, pp. 143—173. 2012. DOI: 10.1007/978-3-642-27645-3_5.

[5] Lazaric, A., Restelli, M., & Bonarini, A. "ransfer of samples in batch reinforcement learning," Proceedings of the 25th International Conference on Machine Learning, pp. 544—551. 2008. DOI: 10.1145/1390156.1390225.

[6] Osipov, G. S., Panov, A. I., Chudova, N. V., & Kuznecova, J. M. "Znakovaja kartina mira sub'ekta povedenija (Semiotic view of world for behavior subject)," Artificial Intelligence, vol. 226, p. 261. 2017.

[7] Parisotto, E., Ba, J., & Salakhutdinov, R. "Actor-mimic deep multitask and transfer reinforcement learning," 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, pp. 1—16. 2016.

[8] Rovbo, M. A., & Sorokoumov, P. S. "Control system architecture of an intelligent agent based on a semiotic network," Open Education, vol. 22, no. 5, pp. 84—93. 2018. DOI: 10.21686/1818-4243-2018-5-84-93.

[9] Sutton, R. S., & Barto, A. G. “Reinforcement learning: an introduction,” UCL,Computer Science Department, Reinforcement Learning Lectures. Cambridge, MA: The MIT Press. p. 552. 2018. DOI: 10.1109/TNN.1998.712192.

[10] Taylor, M. E., Stone, P., & Liu, Y. “Transfer learning via inter-task mappings for temporal difference learning,” Journal of Machine Learning Research, vol. 8, pp. 2125—2167. 2007.

[11] Yogeswaran, M., & Ponnambalam, S. G. “Reinforcement learning: Exploration-exploitation dilemma in multi-agent foraging task,” Opsearch, vol. 49, no. 3, pp. 223—236. 2012. DOI: 10.1007/s12597-012-0077-2.

Передача опыта в задаче обучения в пространстве агента для знаковых моделей мира

Ровбо М.А.

Исследовано несколько алгоритмов для задачи передачи опыта при обучении на примере задачи фуражировки. Эти алгоритмы подходят для семиотической системы управления, чтобы облегчить передачу опыта между двумя обучающимися агентами с различными описаниями состояния в пространстве агента в форме предикатов. Они требуют, чтобы описание задачи целевого агента было подмножеством описания исходного агента.

Алгоритмы основаны на алгоритме Q-обучения и используют специальную функцию передачи опыта, инициализирующую таблицу состояний-действий целевого агента, которая сопоставляет соответствующие предикаты. Тестовая задача, используемая для оценки алгоритмов, является задачей фуражировки, когда агент должен собирать случайно сгенерированную пищу в мире-сетке.

Результаты показывают, что они обеспечивают улучшение кривой обучения для целевого агента после передачи опыта от агента с большим количеством входных предикатов. Улучшение непостоянно и иногда не приносит заметной разницы в производительности, но целевой агент работает по крайней мере так же хорошо, как агент без предшествующего опыта.

Received 31.01.2020

Classification approach to determining emotional state of person based on analysis of video information*

1st Ivan A. Gurin
"Computer Sciences and Control" of Russian Academy of Sciences
Moscow, Russia
covielle12@gmail.com

2nd Alexey B. Petrovski
"Computer Sciences and Control" of Russian Academy of Sciences
Moscow, Russia
a.b.petrovsky@gmail.com

3rd Alla V. Zaboleeva-Zotova
Volgograd State Technical University
Volgograd, Russia
zabzot@gmail.com

***Abstract*—This work considers recognition of person's emotional state with body movements and postures as a problem of group classification of multi-attribute objects. The analysis of existing approaches to the recognition of the human emotional state based on video information was made. An approach proposed includes using markerless recognition of body movements based on video information, identification of poses and movements with a model of human motor activity, characteristics description of postures and movements in a limited natural language and the group classification method of poses and movements having verbal description for emotional state identification.**

***Keywords*—human emotional state, analysis of the video information, movement and postures recognition, group classification**

I. Introduction

Modern information technologies are becoming more and more entrenched in all spheres of human activity. One of the actual and demanded themes is recognition of human emotional state based on the automated analysis of video information of his movements and poses, as human emotions in many respects define his actions, acts and behavior. Physiologists have established that emotional excitation of a person is accompanied by coordinated reductions and relaxation of facial and body muscles, changes in heart, breathing rhythm, stomach and intestinal motility, etc. Tracking and analyzing reactions to emotional excitation, it is possible to define human emotions. There are many works devoted to the development of emotion detection and recognition systems (EDRS) [14], [15]. A rather large number of publications refer to the recognition of emotions by facial mimicry [16], [17]. Essentially less works in which the emotional state is determined by human body movements and poses [2], [19]. However, the problem of reliable automated recognition of human emotions is far from being solved. In the given work the approach to define the human emotional state as solving the problem of group classification of multi-attribute objects which are the poses described in the limited natural language is offered. Verbal characteristics of poses are formed based on the analysis of body movements, which are recognized by video information and identified using a vector model of the human.

*The work was partially supported by RFBR (projects 17-29-07021, 18-07-00132, 19-29-01047).

II. Human posture identification

The first stage of automated recognition of a human emotional state is identification of poses and movements based on the analysis of video information. This task is not trivial and can be solved in different ways. In this paper we will limit ourselves to determining the emotional state of the person only by positions. First, we need to recognize the pose on the input video and present the input data in some digitized form. For this purpose, we use certain model, which contains information about the position of the human body. The models can be very different and depend primarily on the purpose of the analysis. For example, if the purpose is to determine the relative position of objects, the model can be quite simple. If the purpose is to read hand gestures, then a model capable of describing the position of fingers is needed. In our case, the model should allow setting any posture expressing some emotional state. The vector model of a human being described in works [18], [19] is well suited for automated recognition of the emotional state of a person. The specified model of a human being consists of 22 nodes and 19 body elements - on each of 22 points of bending in a human body, in which the change of angles between body elements is fixed. The position of each node is set by three angles of rotation relative to the axes X, Y, Z (Figure 1). The vector model of the person is presented by file format bvh (BioVision Hierarchy) which contains hierarchical structure of a skeleton and parts of a body, the description of position of nodes and the information on dynamics of movement, i.e. changes of position of nodes and duration of these changes. Durations are measured in frames. It is worth noting that in a reality one defines a pose or body movement not by positions of bending points on someone's body, but by positions of particular body parts. Besides, a huge number of poses differ from each other by position of only a small number of nodes. And motions frequently are position changing not of concrete node, but set of the nodes modeling a part of a body. Besides, psychologists tell an emotional state not so much by a posture as a whole, but by noticing specific behavior of separate parts of the body (such as compressed fists, crossed hands, the lowered head, etc.). All this leads to the idea that it makes sense to present a posture as a set of positions of different bodyparts. Each part of the body is defined by a certain number of nodes, the position of which, in turn, set the angles of rotation relative to the axes X, Y, Z. It means that each part of a body can also be set as a vector. It should also be noted that for motions this approach is also suitable, because, to say it rough, motions are just position changes of one or more parts of the body. The difference lies only in the vectors that define the body parts: for motions, the parameters responsible for the duration must be carefully set and accounted in emotional state identification.

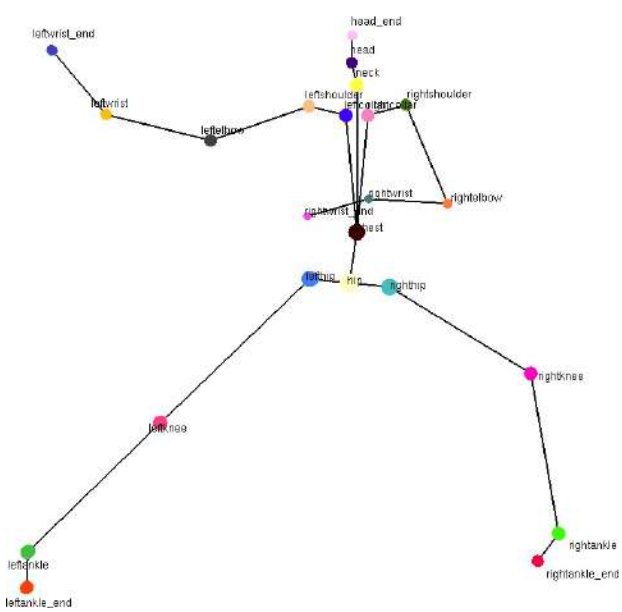


Figure 1. Vector model of the human

Using the specified model, each pose can be represented as a set of vectors describing positions of different parts of a body. For example, the hand position is defined by four nodes <n1, n2, n3, n4>, and each node is defined by three parameters - rotation angles. Thus, it is possible to divide the whole pose into separate parts of a body, each of which is set by some vector describing position of the nodes included into a certain part of a body.

III. Description of human posture

At the following stage it is necessary to transform model representation of each posture which is specified as a set of vectors describing positions of different parts of a body into the description of a posture in the limited natural language habitual for expert psychologists. It can be done, for example, in the following way. Based on the expert's knowledge, the range of values of all elements of the vector representing the posture is divided into intervals, which can be used to match with some verbal characteristics of the posture in the natural language. Obviously partition of the area of values of parameters into intervals cannot be universal as from the physiological point of view people's bodies differ. However, it is possible to divide people into groups (for example, by sex or age - from "children" to "old people") and to build a partition for these groups. Thus, for each image of the pose we get a set of expressions consisting of features that describe this pose.

Figure 2. Example posture

For example, figure 2 shows posture which can be described by the following set of expressions: back is straight; body standing position; head is not bent or overturned; hands are bent in elbows and crossed at chest level, one hand lies on the other; palm of one hand embraces forearm of the other; fingers are not bent, pressed against each other. The description of a pose in natural language, as a matter of fact, represents the position of various parts of a body, thus the certain position of any part of a body is a feature of the given pose. In view of the fact that position of each part of a body is described by some quantity of nodes which position in turn is set by three corners of rotation, we receive that features of a posture is a set of certain elements from the vector model describing a posture.

IV. Determining human emotional state

The third stage is the definition of the emotional state of the person by posture. As shown in works in the field of psychology, there are 61 characteristic poses of the person and 8 basic emotional conditions. However there is no unequivocal conformity between a pose and emotional conditions. We shall consider definition of human emotional state as the decision of a problem of group classification of multi-attribute objects with teachers where objects are poses described in the limited natural language. Let's use for this purpose method of Multi-Dimensional Harmonized Classification Alternatives [7], [8], [9]. The initial array of multi-signed objects is a set of text descriptions of pose images obtained after analysis and identification of motions by video images. Each image of a pose is evaluated by number of experts (t, teachers) and belongs to one of the specified characteristic poses. Then each pose is a multi-signed object O_i, i = 1,…,m, that is present in several versions (copies, instances) $O_i^{<s>}$, s = 1,…,t, which are the essence of the pose image, differing in the values of the attributes $K_1, \ldots, K_n$. Features $K_1, \ldots, K_n$ describe positions of body parts in a limited natural language, each of which is defined by the scale $X_l = x_l^1, \ldots, x_l^{h_l}$, l = 1,…,n, where the gradation $x_l^{e_l}, e_l = 1, \ldots, h_l$ – is a verbal position characteristic. The basic emotional states are classes $D_1, \ldots, D_g$. The expert estimation of the pose image is given by the set of sorting features $R = r_1, \ldots, r_g$, where each of rf shows membership of class D_f, f = 1,…,g. Let's represent each version of $O_i^{<s>}$, i = 1,…,m, s = 1,…,t as a multiset:

$$\begin{gathered} A_i^{<s>} = \{k_{A_i}^{<s>}(x_1^1) \circ x_1^1, \ldots, k_{A_i}^{<s>}(x_1^{h1}) \circ x_1^{h1}; \\ \ldots \\ k_{A_i}^{<s>}(x_n^1) \circ x_n^1, \ldots, k_{A_i}^{<s>}(x_n^{h_n}) \circ x_n^{h_n}; \\ k_{A_i}^{<s>}(r_1) \circ r_1, \ldots, k_{A_i}^{<s>}(r_g) \circ r_g\} \end{gathered} \tag{1}$$

above the set X={x_1^1,…,$x_1^{h_1}$;…; x_n^1,…,$x_n^{h_n}$; r_1,…,r_g} the gradation of the feature scale. Here the value of the multiplicity function $k_{A_i^{<s>}}(x_l^{e_l})$ shows how many times the value $x_l^{e_l} \in X_l, e_l = 1, \ldots, h_l$ of the attribute K_l is present in the description of the object version $O_i^{<s>}$ by the expert s. The value of the multiplicity function $k_{A_i}^{<s>}(r_f) = 1$, if the expert s has referred the version of the object $O_i^{<s>}$ to the class D_f, and $k_{A_i^{<s>}}(r_f) = 0$ – in the opposite case. Let's assume that all images of the same pose are equivalent. Let's form an expert estimation of each O_i as a multiset of features, which is a sum of multiplicity $A_i^{<s>}$, representing images $O_i^{<s>}$ of this posture:

$$\begin{gathered} A_i = A_i^{<1>} + \ldots + A_i^{<t>} = \\ = \{k_{A_i}(x_1^1) \circ x_1^1, \ldots, k_{A_i}(x_1^{h_1}) \circ x_1^{h_1}; \\ \ldots \\ k_{A_i}(x_{n_1}) \circ x_{n_1}, \ldots, k_{A_i}(x_n^{h_n}) \circ x_n^{h_n}; \\ k_{A_i}(r_1) \circ r_1, \ldots, k_{A_i}(r_g) \circ r_g\}. \end{gathered} \tag{2}$$

where the multiplicity functions are calculated by the rules $k_{A_i}(x_l^{e_l}) = \sum_s k_{A_i^{<s>}}(x_l^{e_l})$, $k_{A_i}(r_f) = \sum_s k_{A_i^{<s>}}(r_f)$. As a result, an array of text descriptions of 61 poses will be formed with expert assessments of their possible belonging to 8 basic emotional states. By means of the Multi-Dimensional Harmonized Classification Alternatives method there are generalized decisive rules of group classification of human postures by their characteristics, which are best coordinated with individual rules of expert sorting. The obtained classification rules allow to determine a person's emotional state by video image of his body in an automated mode.

Conclusion

At present, determining a human emotional state by body movements and poses is an urgent task in the areas of health care, public safety and others. However, this area is still poorly developed. So far scientists-psychologists have not offered the reasonable theory about connection between human emotional state and postures and movements. Most of the work in the field of automated recognition of emotions is aimed at determining the emotional state by mimicry. The works on estimation of emotional condition by pantomimics are few, they often have only research purpose. In this work, recognition of emotional state by video image of body movements is considered as a task of group classification of postures. In the future, this approach is also intended for group classification of movements.

References

[1] M.I. Abramov, "Poligoniometricheskie issledovania podvizhnosti osnovnikh otdelov oporno-dvigatelnogo apparata vokrug prodolnoi osi," in Teoria I praktika fizkulturi 1971, pp 34-36.

[2] A.V. Zaboleeva-Zotova, V.L. Rozaliev, A.S. Bobkov, O.S. Fedorov, Trudi kongressa po intellektualnim sistemam i informacionnim tekhnologiam, 2010, pp 398-405.

[3] N.A. Belaya, "Lechebnaya fizkultura i massazh,", 2001 p 272.

[4] A.S. Bobkov, "Interpretiruushaya model poiska dvizheni cheloveka po slovesnomy opisaniu na ogranichennom estestvennom yazike," 2012, pp 53-57.

[5] C. Darvin, "Proishozhdenie cheloveka i polovoi otbor. Virazhenie emocii u cheloveka i zhivotnih," 1940, 320 p.

[6] K. Izard, "Psihologia emocii,", 2003, pp. 464.

[7] A.B. Petrovski, "Metodi gruppovoi klassifikacii mnogopriznakovih objectov (chast 2),", 2009, pp. 3-14.

[8] A.B. Petrovski, "Metod "MASKA" dlya gruppovoi expertnoi klassifikacii mnogopriznakovih objectov,", 2010, pp. 27-31.

[9] A.B. Petrovski, "Gruppovoi verbalniy analiz resheni,", 2019, p. 287.

[10] M. Coulson, "Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. Journal of Nonverbal Behavior, 28(2), 117–139.", 2004, pp. 117-139.

[11] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y., "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2017.
[12] Saini, S., Zakaria, N., Rambli, D. R. A., and Sulaiman, S., "Markerless Human Motion Tracking Using Hierarchical Multi-Swarm Cooperative Particle Swarm Optimization", 2015.
[13] Castillo, S., Legde, K., and Cunningham, D. W., "The semantic space for motion-captured facial expressions. Computer Animation and Virtual Worlds", 2018.
[14] Martin Magdin, Ľubomír Benko, Štefan Koprda, "A Case Study of Facial Emotion Classification Using Affdex", 2019.
[15] D. YangAbeer AlsadoonP.W.C. PrasadA. K. SinghA. Elchouemi, "An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment", 2017.
[16] R.I. Ahmetshin, A.P. Kirpichnikov, M.P.Shleymovich, "Raspoznanie emocii cheloveka na izobrazheniyah", 2015.
[17] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, Remigiusz J. Rak, "Emotion recognition using facial expressions", 2017.
[18] Hao Guo1, Zhenjiang Miao1, Feiyue Zhu2, Gang Zhang2, Song Li , "Automatic Labanotation Generation Based on Human Motion Capture Data", 2014, pp. 3-14.
[19] F. Noroozi, C.A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, G. Anbarjafar , "Survey on Emotional Body Gesture Recognition", 2017.

Классификационный подход к определению эмоционального состояния человека на основе анализа видеоинформации

Гурин И.А., Петровский А.Б, Заболеева-Зотова А.В.

В работе распознавание эмоционального состояния человека по телодвижениям и позам рассматривается как проблема групповой классификации многопризнаковых объектов. Проведён анализ существующих подходов к распознаванию эмоционального состояния человека на основе анализа видеоинформации. Предложен подход, включающий безмаркерное распознание телодвижений по видеоинформации, идентификацию поз и движений с использованием модели двигательной активности человека, описание характеристик поз и движений на ограниченном естественном языке, выявление эмоциональных состояний с помощью метода групповой классификации поз и движений, имеющих вербальные описания.

Received 31.01.2020

Automation of the translation of Russian sign language into text

Vladimir Bondar
National Research Center «Kurchatov Institute»
Moscow, 123182, Russia
vovabond328@Rambler.ru

***Abstract*—This article is devoted to setting the task of automation of recognition of gestures of Russian language of the deaf. The author designed a prototype of gesture recognition system using a camera with depth sensor.**

***Keywords*—sign language, gesture recognition, depth sensor camera, Kinect**

I. Introduction

In Russia, more than a million people suffer from hearing loss. They are forced to communicate among themselves and with normal hearing people through special sign language. People with no hearing problems usually do not know sign language. This language barrier prevents people with different physical abilities from communicating freely. The only connecting and stable link between hearing and non-hearing people is sign language interpreters. However, their number is not increasing, but only decreasing.

Since most people now have access to a computer, it is logical to use emerging computer technologies to establish communication between hearing impaired people and people with hearing disabilities. [8] In recent years, technical devices and ways of recognizing video images have been improved - there are cameras with depth sensors that can produce three-dimensional images.

Microsoft has developed cameras with the depth sensor Kinect, which allows you to use new features for gesture recognition. Depth sensors have existed for a relatively long time, but Kinect has a number of significant advantages in front of them: large distribution, relatively low cost and availability of RGB camera. The depth map obtained from Kinect is invariant to lighting conditions and background, as it is based on infrared radiation. Only strong fog and some other weather conditions can be an obstacle for this type of recorders.

Thus, there are technical possibilities to develop an automatic translator of Russian sign language into text. As for any natural language processing system, here, in addition to the direct recognition of video information, there are problems of constructing the semantic structure of a sentence, solving problems of polysemy, etc.[1] These questions are beyond the scope of this article. We will focus on gesture recognition and text identification by a recognized gesture.

Microsoft Kinect camera was used as a video input device for gesture information.

II. Concept of Gestural Speech

So, what's a gesture speech? "Gesture speech is a way of interpersonal communication between hearing-impaired people through gestures, characterized by peculiar lexical and grammatical patterns. Gesture speech is used as an additional tool (along with basic verbal speech) in the education and upbringing of children with hearing loss. For official communication, calculating gesture speech is used: the consistent use of gestures to reproduce words." [2] In this case, a gesture (from Latin Gestus - body movement) is understood as "some action or movement of the human body or its part having a certain meaning or meaning, i.e. being a sign or symbol." [3]

The sign language program for people with hearing loss has a complex structure that includes two types of language: spoken sign language, which is used during conversational communication, and calculating sign language, which is used in official communication.

The natural national language is single, the official and ordinary forms differ only in vocabulary, which cannot be said about the language of the deaf, or rather languages, because they are two and use a different set of gestures and different grammar.

Gesture speech is divided into two types: - spoken gesture speech - calculating gesture speech Fingerprinting - alphabetic reproduction of words by using finger configurations. It is used to show the names of own and specific terms. Because it is easy enough to learn, fingerprinting can be a bridge between the deaf and the hearing while the latter are trying to learn spoken sign language. By remembering only a few movements, you can start communicating with the deaf, although on a primitive level.

All gestures in terms of number of movements can be divided into two groups: static and dynamic.

In order to show a static gesture, a simple photo is enough, because it is a fixed position of the body in a certain position. A dynamic gesture (movement) will require animation because it is a sequence of static gestures over a certain period of time.

III. Analysis of gesture recognition tools

A. Microsoft Kinect

Kinect (formerly Project Natal) is a "contactless game controller originally introduced for the Xbox 360 console and much later for Xbox One and Windows PCs. It was developed by Microsoft".

The greatest interest in this work is the 3D depth sensor. The sensor consists of a projector that emits an infrared field and a sensor that reads the field and together they create a depth map. Kinect connects the RGB image and the depth map.

A depth map is an image that stores for each pixel its distance from the camera (instead of color). You can segment the image and find the necessary areas where the gesture is shown using the distance information from the camera.

B. Leap motion

Two monochrome infrared cameras and three infrared radiators are used in Leap Motion to detect the movement of the user's hands. The cameras "scan" the space above the table surface at up to 300 frames per second and transmit the data to a computer where it is processed by proprietary software. Despite their seeming similarity to Microsoft Kinect, these devices are still different.

Leap Motion should be placed under the screen, more precisely - instead of the keyboard (in fact, since it is impossible to do without the traditional control, the controller will have to be placed either behind the keyboard or in front of it as it will be more convenient for the user). As a result, Leap Motion has a rather small scanning area, which, in fact, gets only a small space above the keyboard.

C. 3D camera Intel RealSense

The RealSense camera is an embedded solution for a variety of portable and stationary devices. The camera allows you to get a picture in HD format and provides all the necessary functions of interaction with the person: face and gesture recognition" tracking emotions, highlighting the background, and more.

Simple color camera transmits the picture in the plane. For example, if your left hand is closer to the camera than your right hand, the machine can 'understand' that the hands are at different distances only using special algorithms for pattern recognition. A depth camera solves this problem much more easily, and works like a bat or echo sounder, sending an IR beam and measuring how long it has returned. Knowing the speed of light, it is easy to calculate the distance from the camera to the object.

Thus, gesture recognition tools were considered. The analysis of the tools is shown in Figure 1. For our system was selected device Microsoft Kinect, because it has the necessary characteristics for gesture recognition and has a well-designed software development kit.

Evaluation criterion	Kinect	Leap motion	Intel RealSense	2 cameras
Visibility of key points	Detects up to six people in space and 25 joints of each of them.	Tracks the movement of fingers and hands. Supports 27 degrees of freedom for each hand.	Allows to track up to 76 key points of the face, tracking gestures and position of hands and fingers in the range from 0.2 to 1.2 m. The number of vision points increased up to 22.	none
Range of conditional visibility	With lens support from 0.5 m, without - from 1 m to 5 m.	Up to 1.5 m	Up to 1.2 m	Depends on the cameras' resolution
Sensitivity to light	low	low	low	Depends on the cameras' resolution and quality
Need for additional software	No	No	No	Yes
SDK quality	Good realisation.	Qualitatively realized only for fine motor skills.	It has some defects in operation.	none

Figure 1. Analysis of gesture recognition tools.

IV. Information model for gesture recognition

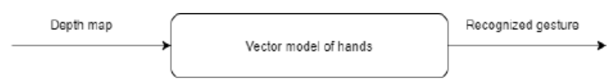


Figure 2. Informational model.

Information from the Kinect camera comes as a depth map. Then, after some processing of the received image, the area of interest is highlighted on it - the human hand [3]. Based on the depth map of the selected part of the image the parameters of the vector model of the hand are calculated.

$$IM = < FV, C, I, O > \quad (1)$$

where IM – Information model;

FV – a vector of hand features. Initial information about the image (color and distance) is analyzed: the position of the hands and the vector of their characteristic features are determined;

C - gesture classifier. According to the found vector of characteristic features of the wrists, it is determined which gesture was shown;

I - set of input values - depth map - distance to each point in the image;

O - set of output values - gesture recognized in the image

V. Gesture recognition system design

The system under design consists of the following modules, which are interconnected with each other:

- image segmentation module;

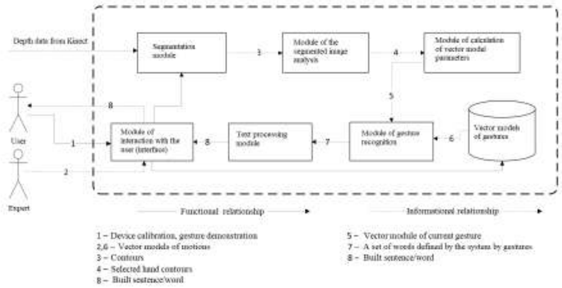


Figure 3. Gesture recognition system design.

- module of the segmented image analysis;
- module of calculation of vector model parameters;
- module of gesture recognition;
- text processing module;
- module of interaction with the user (interface).

VI. General algorithm of gesture recognition system operation

Several works have already been translated from Russian into text, but they focus on extracting information from the hand position. In this work it is suggested to get information about gestures also from facial expressions, as face recognition technologies have stepped very far in the last few years.

The general algorithm of the program consists of the following steps: detect and recognize hands and face on the screen, recognize the gesture and display the text on the screen. The designed system includes two main features: recognize the gesture and display text on the screen.

A. User interface module

The purpose of the module is to provide the user with a convenient interface of working with the program. This includes: displaying the processed image on the screen, the ability to change the settings of the program, displaying the recognized gesture on the screen.

B. Text processing module

The purpose of this module is to form text from recognized gestures, display the formed text on the screen. The recognition of gestures will take place in real time, so the program should recognize the gestures shown, write them to the buffer and display them one after another on the screen. Sentences in the resulting text can be separated by a gesture, which is indicated as a dot.

C. Module of image processing and calculation of vector model parameters

The purpose of image processing modules is to segment the depth map obtained with the Kinect camera, to process the image segmented at the previous stage and to calculate parameters of the vector model of the processed image. Every 1/30 second information from the sensor is supplied to the input, the modules process the input data and calculate the parameters of the vector model of the hand.

Segmentation of the obtained image from the Kinect sensor is necessary for finding the hand or both hands on it. In such cases, as face detection on the image, appearance is a good sign, as eyes, nose and mouth will always be approximately in identical proportions. Therefore, the Haar Cascade method, based on the appearance characteristics of the object, is well suited for facial recognition. In the case of hand recognition, the situation is more complicated: a reliable recognition method can be implemented based mainly on the colour characteristics. Since the colour of the hands can vary depending on the person and context, it seems reasonable to first find the person's face in the image and get information about the colour of the hands based on the colour of the face. The introduced restriction on the presence of a person's face in an image is in any case mandatory, since recognition of a hard language without facial recognition would be unreliable.

Having information about the colour of an object, you should detect it in the image. The task was performed using the Camshift algorithm, the reliability of which has been proved in [Hai et al., 2011]. The model of this algorithm is based on histograms and is trained in the recognition process. Naturally, this algorithm will find all objects of a given color in an image. To prevent it, the information about the distance to the objects in the image is used, i.e. depth map from Kinect sensor.

So, after finding the position (x,y) and dimensions (w,h) of the face in the image using the Haar Cascades method, you can find the average distance to the face using the depth map D:

$$d_f = 1/wh \sum_{i=x}^{x+w} \sum_{i=y}^{y+w} D(i,j) \tag{2}$$

All objects that are closer to the camera than the human face itself can be found using the threshold:

$$D(i,j) > d_f + t_h \tag{3}$$

where t_h– a parameter that determines how close your hands should be to the camera so that the gestures shown can be recognized by the system.

In the second stage of recognition on the segmented image are the contours of human hands with the Canny edge detector.

D. Module of gesture recognition

The purpose of the module is to attribute data obtained from the modules of image processing and calculation of vector model parameters to one of the gestures of the Russian language, embedded in the program. Further it transfers the recognized gesture to the text processing module. The gesture is recognized, as it was already described earlier, by means of the vector model of a hand. The process is concluded in the following. Originally we receive a vector model from the above described modules. According to this data there are matches, which are stored in the database (matches should not be ideally similar, it is possible to find the most suitable features that will signal the similarity), then formed sets of words obtained from this data.

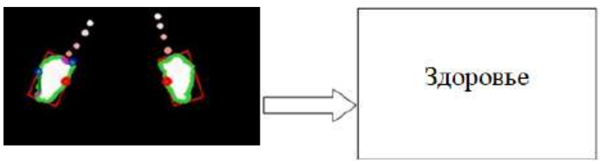


Figure 4. Recognition of "health" gesture.

Thus, the system of translation of Russian sign language into text was designed. This system will be able to translate both static and dynamic gestures.

CONCLUSION

In the course of this study, a gesture recognition system for Russian gestures was designed to recognize gestures by obtaining information about the position of the hands and facial expressions.

As noted, the sign language consists of a number of channels for the transmission of information, so face recognition and hand gestures do not solve the problem of recognition of Russian sign language completely, but is an important element of the future full system. The subject of further research is the implementation of this project as a software product.

In the process of image analysis and text acquisition in natural language semantic technologies of intellectual systems will be used.

REFERENCES

[1] Popov M.YU., Fomenkov S.A., Zaboleeva-Zotova A.V. Principy raboty i osobennosti realizacii universal'nogo semantiko-sintaksicheskogo analizatora russkogo yazyka[Principles of work and features of implementation of the universal semantic-syntactic analyzer of the Russian language]. Vestnik komp'iuternykh i informacionnykh tekhnologii [Herald of computer and information technologies], 2005, No 7 (13), pp. 37-42

[2] Zaboleeva-Zotova A.V., Orlova YU.A., Rozaliev V.L., Fedorov O.S. Predstavlenie harakternyh emocional'nyh zhestov i dvizhenij cheloveka v vide nechyotkih posledovatel'nyh temporal'nyh vyskazyvanij [Presentation of characteristic emotional gestures and movements of a person in the form of vague consecutive temporal expressions],Obozrenie prikladnoj i promyshlennoj matematiki [Overview of Applied and Industrial Mathematics]. 2011, vol 18, no 4, pp. 537-544

[3] Dorofeev N.S., Rozaliev V.L., Zaboleeva-Zotova A.V. Sistema raspoznavaniya zhestov russkogo yazyka gluhih [The system of recognition of gestures of Russian language of deaf people], Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2013, pp. 399-402.

[4] Petrovskij A.B., Zaboleeva-Zotova A.V., Bobkov A.S. Formalizovannoe opisanie dvizhenij cheloveka na osnove nechetkih temporal'nyh vyskazyvanij [Formalized description of human movements based on vague temporal expressions] Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte VII-ya Mezhdunarodnaya nauchno-prakticheskaya konferenciya, k 80-letiyu Dmitriya Aleksandrovicha Pospelova: sbornik nauchnyh trudov v 3 tomah. [Integrated models and soft computations in artificial intelligence]. 2013. pp. 488-495.

[5] Rozaliev V.L., Zaboleeva-Zotova A.V. Methods and models for identifying human emotions by recognition gestures and motion Journal of Politics. 2013, vol. 2. p. 67.

[6] Konstantinov V.M., Rozaliev V.L., Orlova Y.A., Zaboleeva-Zotova A.V. Development of 3D human body model Advances in Intelligent Systems and Computing. 2016. vol. 451. pp. 143-152.

[7] Vybornyi A.I., Rozaliev V.L., Orlova Yu.A., Zaboleeva-Zotova A.V., Petrovsky A.B. Controlling the correctness of physical exercises using microsoft kinect Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2017, Iss. 1, pp. 403-406.

[8] Orlova Yu.A. Modeli i metody informacionnoj podderzhki kommunikacii lyudej s ogranichennymi vozmozhnostyami [Models and methods of information support for communication of people with disabilities.]. – 2016

[9] Sean Kean, Jonathan Hall, Phoenix Perry, Meet the Kinect: An Introduction to Programming Natural User Interfaces (Technology in Action). Apress, 2011. – 220 c.

[10] H. Hai, L. Bin, H. BenXiong and C. Yi, "Interaction System of Treadmill Games based on depth maps and CAM-Shift", IEEE 3rd International Communication Software and Networks (2011), pp.219-222

Автоматизация перевода русского жестового языка в текст

Бондарь В.В.

Данная статья посвящена постановке задачи автоматизации распознавания жестов русского языка глухих. Автором спроектирован прототип системы распознавания жестов с использованием камеры с сенсором глубины.

An approach to sheet music information retrieval

Yulia Korukhova
Computational Mathematics and Cybernetics faculty
Lomonosov Moscow State University
Moscow, Russia
yulia@cs.msu.ru

Aiken Tolybayeva
Computational Mathematics and Cybernetics faculty
Lomonosov Moscow State University
Moscow, Russia
aikesha28@gmail.com

***Abstract*—The paper presents a sheet music retrieval system. A search query is written in MusicXML format. MusicXML is a very popular format for sheet music, supported by various music notation software. With the appearance of large electronic collections the problem of music retrieval becomes particularly important. Known information retrieval methods are not directly applicable for this particular domain: the task differs from the structured retrieval in texts because, for example, two sheet music files with totally different symbolic content may represent the same composition written in different keys or arranged for different instruments.**

Our search starts from a fragment containing a melody or a part of a composition and the goal is to find the whole composition in a digital collection. The collection should be preliminary indexed and the index is stored as database. For MusicXML files we propose a software tool that performs indexing automatically. For other formats the indexing process can be performed manually and database is extended by means of SQL–queries. To perform a fuzzy search we mark the longest common subsequence in the query and index item and then to remove their differences using automated reasoning method called rippling. The appropriate rewriting rules and differences annotation techniques for musical notation are implemented in prototype sheet music retrieval system.

***Keywords*—music information retrieval, automated reasoning, rippling**

I. Introduction

Nowadays with wide spread of networks, computers and electronic devices, a lot of data, information and knowledge is stored in electronic form for almost every domain. That includes sheet music: many of that is now presented in electronic collections[1]. Such collections are often quite large and information retrieval tools are required to work with them. In addition to traditional search tools that are able to find an object by it's name and/or it's author the retrieval of a piece that contains a given fragment is required. This task has been solved for audio fragments in Shazam application[2], for humming queries and melody fragments in Musipedia [1]. However the search of given sheet music corresponding to a fragment is still a challenging task. From one point of view it is similar to information retrieval for texts, but it has several specific issues. First of all, there is no strict analogy with words in texts, that can be determined syntactically. Also the same music piece can have different symbolic representations, being transposed or arranged for another instrument. In this paper we are presenting a context-based approach to sheet music retrieval.

The paper is organized as follows: first we describe a task of music scores information retrieval, next will discuss an appropriate internal representation, that consider the particular features of the domain, then will pass to index organization. Next we describe our search algorithm and give a short example of it's work. Finally the conclusions are drawn.

II. Music scores retrieval task

As an input data we take a music score fragment, written in MusicXML format [2]. This format has been chosen because it is widely supported, it can be used in more than 150 applications [2] including music notation software like MuseScore[3], Finale[4], Guitar Pro[5] etc. We search for the given fragment in indexed music library. The digital music score library can contain sheet music of any kind and in any form, however it has to be preliminary indexed. As an example in this work we take a test library of MusicXML files and present a software tool that performs indexing automatically. As a search result we expect a list of music scores, containing the given fragment or similar to it. In the case the search query is not presented in our index, we propose to change it by a set of rewriting rules, that can help to correct automatically some possible misprinting. A special measure is introduced to estimate the similarity

[1]MuseScore sheet music library and editor https://musescore.com (accessed 2019, Dec), Petrucci Music Library, https://imslp.org (accessed 2019, Dec) , НотоМания, http://www.notomania.ru (accessed 2019, Dec)

[2]Shazam acoustic fingerprint music search system https://www.shazam.com/ru (accessed 2019, Dec)

[3]MuseScore sheet music library and editor https://musescore.com (accessed 2019, Dec)

[4]Finale, https://www.finalemusic.com (accessed 2019, Dec)

[5]Guitar Pro, https://www.guitar-pro.com (accessed 2019, Dec)

between the search query and a fragment from index. It takes into account the number of differences between fragments and the number of rules, that have been applied to remove differences between the search query and a fragment from index.

III. Internal representation

To solve a problem with different symbolic representation of transposed music fragments or played in different tempos we use normalized representation. Each note is presented as a pair of two numbers (relative pitch, relative duration), the same approach is used in [5], [6]. Relative pitch is computed as a distance in tones from the previous note in the melody. Relative duration for the notes are calculated as relation to the previous note. The particular problem of this representation that it is not valid for chords. We propose to select one note of chord. The relative duration for pauses are calculated as relation to the previous pause duration or to the previous note duration if there is no previous pause.

IV. Indexing of Sheet Music Collection

The indexing of music files is a challenging task, because it is difficult to formalize the process of splitting a composition into melodies for building a dictionary. The automatic discovering logically finished part of melody is a process that is difficult to automate. Usually the search query is much shorter than a full music piece that's why we put in our index short fragments corresponding to a period (8 bars) and its parts (4, 2 bars) taken from the score. The obtained fragments are translated into normalized representation and duplicates are not added to the index. For each of the fragments we have a list of compositions in which it occurs.

We propose to organize index as a database implemented on PosgreSQL [3]. That allows us to use the benefits of this database management system: to perform effective search, and also that give other potential users a possibility to add new items to index by writing SQL–queries either manually or by some external software. In this work the special software tool is developed for automatic indexing of MusicXML files. The program was written on Python and stored as an auxiliary module. However indexing can be performed for other formats manually or using external software and new content is added to PosgreSQL database – index by means of SQL-queries.

V. Music Retrieval Algorithm

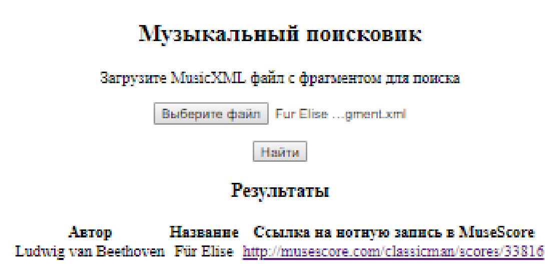
In the our sheet music retrieval system we have to perform fuzzy search. We use an approach based on rippling [5], and continue and extend the ideas proposed in [6], [12]. In the information retrieval we are faced with the situation than we have two similar (but not identic) fragments: a search query and a candidate to be an answer. We need to find if they represent the same composition. This task is quite similar to a mathematical one: while performing mathematical induction proof we have a hypothesis and a goal that are syntactically similar, and we need to rewrite the goal to prove that it is implied by the given. To solve the problem we can use rippling, previously proposed as an heuristic for automation of mathematical induction method [5]. The precondition for using the method is the following: we are rewriting one sentence to another and the sentences have to be similar. The differences are annotated as a wave front and the common parts form a skeleton. We do rewriting of the goal to reduce differences and the skeleton has to be preserved. To estimate differences reduction a measure is proposed. It has to become smaller with every rewriting step and it equals 0 when all the differences are removed. We extend the application of this approach from mathematical formulas to a new domain – a sheet music fragments. For this we have to introduce:

- a set of wave rules for musical domain;
- a measure to estimate differences between musical fragments;
- a proof that each of wave rules reduces the number of differences according to the proposed measure.

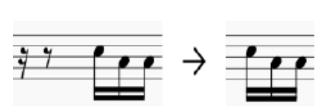
In this work we introduce a set of wave rules (extended version of [6], [12]. Differences are estimated as number of notes in wave front and each wave rule rewrites fragments reducing differences.

The work of our search engine consist of several steps:

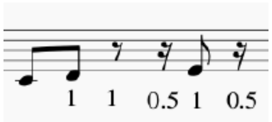
1) Get a search query in MusicXML format.



2) Translate a query into internal normalized representation
   a) Remove pauses in the begining of the query



   b) Relative duration of other pauses is calculated based on previous note. The first note after pauses is divided on the length of the previous note.



3) Comparing the internal representation of the search query with fragments from the index. Since we are performing fuzzy search, for each pair (query and database fragments) we perform the following steps:
   a) Marking the longest common sequence (LCS) using the dynamic programming method [7]. If it is shorter than a half of the shortest fragments, not considering them similar. Otherwise mark all common notes as a skeleton and others as a wave-front.
   b) Deleting wave fronts in the beginning (prefix) and in the end (suffix) of sequence.

   The longest common subsequence is marked after deleting prefix and suffix
   c) Remove pauses after the final note in our fragment.
   d) Compare the fragments as they are build from the same note.
   e) Estimating the similarity using the formula

$$Similiriaty(F_1, F_2) = \frac{2 \cdot S}{L_1 + L_2 + \alpha \cdot N}, \quad (1)$$

   where F_1, F_2 - compared fragments, S - length of the longest common sequence, L_1 - the length of fragment F_1, L_2 - the length of F_2,
   α - heuristic coefficient, taken now as 0.5 after experiments with the system, N - number of applied rules.
   f) If the similarity is greater or equal 0.9, the fragments are considered to be similar and the resulted composition from the database should be mentioned in the list of search results.

Otherwise we are trying to remove differences using wave rules from [6] and the set of rules is extended by new ones. We use rules of three categories:

- Error correction rules.
  The aim of the rule is correction of possible errors made by user in the query, because it is written as he had heard it. For example an error in intervals like fifth, sixth and bigger user can make a mistake writing a note half tone higher or lower. Such differences in the border notes of the wave front can be considered as an error made by user in a query and they are proposed to be corrected by rewriting.
- Alternative notations.
  The same musical fragment can be written and performed in different ways, however the melody can be considered the same. This happens in variations and in different arrangements. The rules of this category assume different placement of pauses and elimination of non-chord notes.
- Elimination rules.
  After applying the previously mentioned rules the longest common sequence becomes longer. However the rippling method does not allow to change skeleton. That is why the newly constructed common parts are removed from the sequence. Only the notes that are at the border with the skeleton can be removed by this rule.

Since we have many rules in our database we have to avoid cycles in their applications. We propose to use a measure, that should become smaller with every rewriting step. We take a number of notes in the wave front as such a measure. The rewriting process stops in two situations; either the measure becomes equal 0, or there are no applicable rules. So at every step of rewriting we use only the rules that decreases measure of the sentence and we avoid infinite loops. An example of rule is presented here. A note (x,y), where x – pitch and y – duration and a pause (y) directly after it can be replaced by the same note with longer duration (x,2*y):

[...](x,y)(y)[...] ==> [...](x,2*y)[...]

After applying rules we calculate the similarity again and make a decision about correspondence of the fragment to the search query.

VI. Example of Work

A short example of music retrieval performed by our prototype system is presented in the Table I. The search query is presented in line 9, among the library fragments we have 9 similar ones. Their differences are reduced by wave rules and the similarity is calculated using formula (1) and shown in the right column.

Table I
RETRIEVAL RESULTS WITH ESTIMATION OF SIMILARITY

Query	*Result*	*Similiriaty*
9	1	0.81
	2	0.75
	3	0.83
	4	0.83
	5	0.63
	6	0.97
	7	0.97
	8	1
	9	1

CONCLUSIONS

In the paper we propose an approach to music information retrieval, where the search query is a MusicXML sequence. Being considered as a structured documents retrieval, however it is a challenging task, because specific knowledge about the domain needed to be taken into account. We perform retrieval based on rippling and our index is stored as a PosgreSQL database, that is constructed automatically for MusicXML documents or manually for other formats. The work is still in progress: we analyze the possibility to improve retrieval efficiency, however the approach is already implemented in a prototype system, that allows to perform fuzzy search for sheet music.

REFERENCES

[1] Musipedia, http://www.musipedia.org (accessed 2019, Dec)

[2] MusicXML format, https://www.musicxml.com (accessed 2019, Dec)

[3] PostgreSQL, https://www.postgresql.org (accessed 2019, Dec)

[4] C. D. Manning, P. Raghavan and H. Schütze Introduction to Information Retrieval, Cambridge University Press. 2008.

[5] A. Bundy, D. Basin, D. Hutter, A. Ireland Rippling: Meta-level Guidance for Mathematical Reasoning. Cambridge University Press, 2005.

[6] Korukhova Y.S., Mytrova M.V. Sheet Music Retrieval / Open Systems. DBMS — 2013. — № 7. — p. 57–58, http://www.osp.ru/os/2013/07/13037358/ (accessed 2019, Dec)

[7] Dan Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press. 1997.

[8] Viro V., Peachnote: Music Score Search and Analysis Platform/// Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011. pp. 359–362.

[9] Typke R. Music Retrieval Based on Melodic Similarity. Ph.D. Thesis, University of Utrecht, 2007

[10] Typke R., Giannopoulos P., Veltkamp R.C., Wiering F., van Oostrum R. Using Transportation Distances for Measuring Melodic Similarity // Proceedings of the Fourth International Conference on Music Information Retrieval. 2003. pp. 107–114.

[11] Yang X., Chen Q., Wang X. A Novel Approach Based on Fault Tolerance and Recursive Segmentation to Query by Humming // Advances in Computer Science and Information Technology: AST/UCMA/ISA/ACN. Conferences Joint Proceedings, Springer–Verlag Berlin, Heidelberg, 2010. pp. 544–557.

[12] Korukhova Y.S., Mytrova M.V. Information retrieval for music scores // Preprint of Keldysh Institute of Applied Mathematics — 2013. — № 48. — p. 1–16, https://keldysh.ru/papers/2013/prep2013_48.pdf (accessed 2019, Dec)

Об одном подходе к музыкальному информационному поиску нотных записей Корухова Ю.С., Толыбаева А.

В работе рассматривается задача поиска нотных записей музыкальных произведений по заданному в виде последовательности нот фрагменту мелодии. В связи с появлением больших электронных нотных библиотек, такая задача становится особенно актуальной.

Поисковый запрос формулируется в виде файла формата MusicXML, который поддерживается большинством программ – нотных редакторов наряду с их собственными форматами. Поиск ведется в электронной нотной библиотеке, которая должна быть предварительно проиндексирована. Индекс хранится в виде базы данных PostgreSQL. Для автоматического построения индекса по MusicXML файлам создан специальный программный инструмент. Файлы других форматов могут быть проиндексированы вручную либо с использованием дополнительных программ. Взаимодействие с индексом – базой данных возможно с помощью SQL – запросов.

При реализации поиска возникает задача сравнения фрагмента – запроса с элементом индекса, причем представляет интерес не только поиск точно совпадающих, но и похожих на запрос произведений. Реализация нечеткого поиска выполнена на основе волновых правил, традиционно применяемых в автоматическом доказательстве теорем методом математической индукции. При построении такого доказательства гипотеза и заключение индукции являются синтаксически похожими, и стоит задача устранения символьных различий между ними. Для ее решения применяются специальные правила переписывания, которые по построению уменьшают количество различий после их применения. В работе составлены волновые правила для устранения различий похожих музыкальных фрагментов и возможности корректировки неточности в поисковом запросе.

Предложенный в работе подход реализован в прототипной программной системе поиска нотных записей.

Received 15.12.2019

Development of web service for generating musical composition from an image

1st Nikita Nikitin
Software Engineering Department
Volgograd State Technical University
Volgograd, Russia
set.enter@mail.ru

2nd Vladimir Rozaliev
Software Engineering Department
Volgograd State Technical University
Volgograd, Russia
vladimir.rozaliev@gmail.com

3rd Yulia Orlova
Software Engineering Department
Volgograd State Technical University
Volgograd, Russia
yulia.orlova@gmail.com

***Abstract*—This paper describes the process of designing and developing a web service for the automated generation of sounds from an image. Also describes the main method for synthesizing music from an image, which is based on the joint use of neural networks and light-music theory. It also contains testing of developed service.**

***Keywords*—recurrent neural network, light-music theory, automated music generation, schemes for correlating colors and notes, Keras, Flask.**

I. Introduction

Nowadays, more and more papers are being published aimed at automating the process of creating musical compositions, however, this process is creative, depends on many factors, starting from the experience and mood of the composer, ending with the area of residence and other external factors, so music cannot be created in automatically mode, therefore, the role of the user-composer is very high and we can only talk about the automation of this process. The emotionality that is conveyed by music and paintings is difficult to recognize [1]. Although the process of creating music is based on clearly defined musical rules, it cannot be completely formalized. To reduce the role of the user in the process of choosing the characteristics of a musical composition, as well as to take into account the emotional component (for example, the emotional state of the user-composer), in this work it is supposed to obtain the characteristics of the composition from an image.

In the framework of this work, automation of the process of creating music by means of automated generation of sounds from an image is assumed. In other words, the generation of sounds from an image is the process of converting an image into one or more sequences of notes, with a certain fundamental tone and duration [2].

II. Basic method

The first step in developing a service is to determine the main method of the program - the method of generating musical material from an image. This method consists of two component:

- an algorithm for correlating color and musical characteristic;
- algorithm for generating melodic parts using neural networks.

A. Algorithm for correlating color and musical characteristic

The main parameters of the resulting musical composition is tonality and tempo. These parameters describe the emotional component of the composition, and should be determined by analyzing the color of the image. For this, first of all, it is necessary to determine the ratio of color and musical characteristics [3]:

- the color hue correlates with the height of the note;
- the color group with the musical mood;
- the brightness with the octave of note;
- the saturation with the duration.

Then, it is necessary to determine the correlation scheme between the color and the pitch. At the moment, there are a large number of such schemes, however, the schemes of I. Newton, Louis-Bertrand Castel, A. Wallace Rimmington, A. Eppley and L. J. Belmont were implemented in this work [4].

The algorithm for determining tonality is based on image analysis and consists of 4 steps:

- The first step is to convert the input image from the RGB color space to HSV. This step allows to convert the image to a more convenient form, since the HSV space already contains the necessary characteristics - the color name (determined by the hue parameter), saturation and brightness parameters [5];
- The second step - analyzing the whole image to determine the predominant color;
- The third step is to determine the hue and color group of the primary color;
- The fourth step - according to the selected scheme of correlation between colors and notes, as well as the results obtained in the previous steps, it is necessary to determine the tonality of the composition. To determine the tempo, it is necessary to obtain the brightness and saturation of the primary color, and calculate the tempo according to these parameters;

B. algorithm for generating melodic parts using neural networks

In this paper, the following algorithm is proposed for obtaining a composition from an image (algorithm for generating the melodic part using neural networks):

- according to the method of correlating color and musical characteristics, the tonality of the work and the sequence of the first 20 percents of the notes read from the image are obtained [6];
- then, according to the obtained sequence of notes, the continuation of the work using the trained model and neural network is predicted;
- according to the final sequence of notes and tonality, and also according to the method of correlation of color and musical characteristics, the harmonic part of the work is built.

III. Developing of the web service

To develop the web site for sound generation based on image color spectrum, the following architecture was proposed ("Fig. 1"):

- the main subsystem to work with user input (also includes the image analysis module);
- subsystem to work with neural networks;
- the subsystem for sound synthesising.

A. The main subsystem

This subsystem provides functionality to work with user input, analysis the given image and then pass the results to other modules. It consists of the 5 modules:

- playback control module - provides functionality to save, play and stop music;
- composition parameter definition module - this module provides functionality to set the characteristics of the result music composition, such as musical instrument (piano or guitar) and correlation scheme (correlation between colors and pitches);
- image uploader - provides ability to choose and load the image from the user's PC;
- image analysis module - this module analysis the image and determines the characteristics of the result musical composition from an image, such as predominant color, color sequence in suitable format and other image characteristics;
- module for generating xml text - this module generates xml text based on image characteristics provided by previous module.

B. The neural network subsystem

This subsystem provides functionality to work with neural network. The aim of this subsystem is to predict the melodic part of the result musical composition and determine the harmony part. It consists of the following parts:

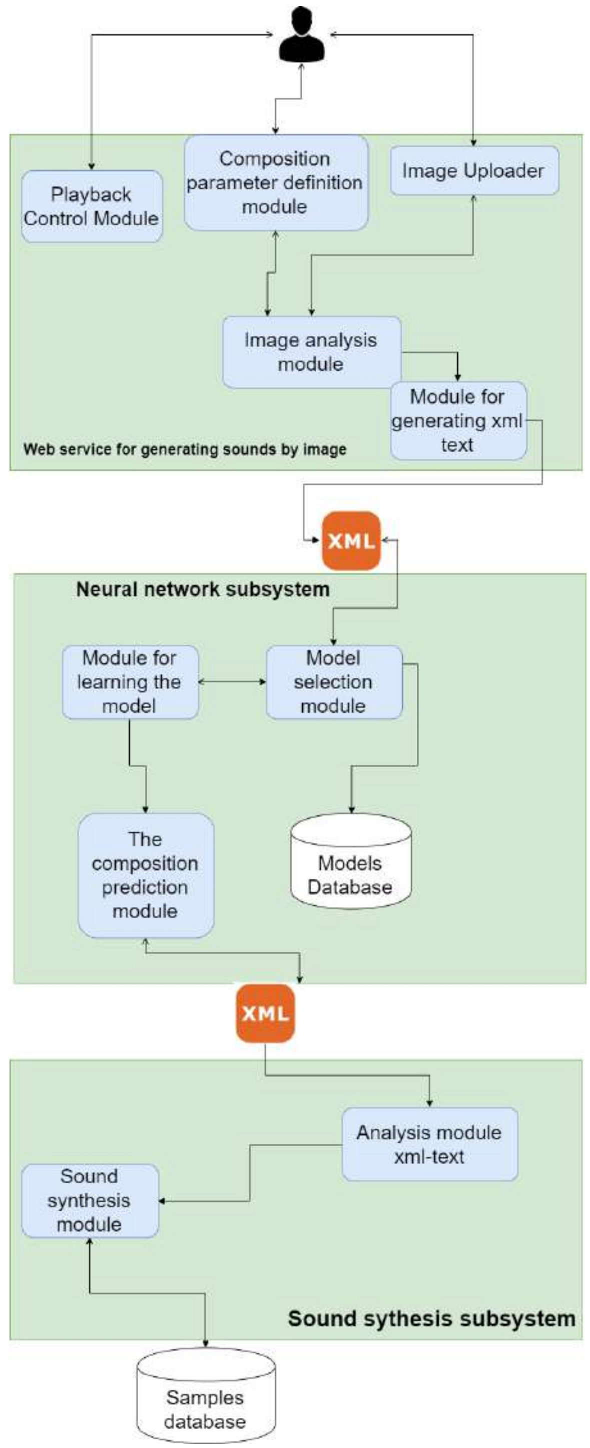


Figure 1. Architecture of the web site.

- learn module - this module is the helper part, this means that the user does not directly interact with this module. It provides the ability to learn the model based on given corpus of musical compositions in midi format, After learning, the module saves model in HDF5 file and saves in the database

(Models database) [7];

- model selection module - this modules determines suitable model from the database (with coordination with the previous module), based on given image characteristics and loads it into memory to subsequent prediction;
- the composition prediction module - this module predicts the melodic part and determines the harmonic part based on certain musical rules [8]. It uses the model has been loaded by the previous module, and also checks the image characteristics and sequence of colors. Also, it forms the xml text to send the composition in internal notation for the subsequent sound synthesis.

C. The sound synthesis subsystem

This subsystem generates the sound in .mp3 format based on given xml text with the result musical composition. It consists of two main modules and one database:

- xml-text analysing module - this module analysis the given xml text from the previous subsystem and translates this text to suitable format for subsequent synthesis;
- sound synthesis module - it combines different pieces of music (samples) [9] from the database with overall mp3 sound file based on the result from the previous module. As the result, this module saves the mp3 file to the server's internal folders and passes the path to the playback control module.

D. Description of the structure of xml text

For the interaction of subsystems, a proprietary xml file structure was developed. It contains the following tags:

- tonality - tag indicating the tonality of the composition;
- tempo - tag indicating the tempo of the composition;
- harmony - tag containing a description of harmony. Inside this tag is the *chord* tag;
- chord - the tag inside which *type*, *chord_name* and *mode* tags are located. The property of the tag is *duration*, indicating the duration of the chord;
- type - the tag, an important part of which is the *value* property, which indicates the type of chord - *standard* or *own*;
- chord_name - the tag denoting the name of the key note (C, D, E, F, G, A, H, etc.);
- mode - tag describing the musical mode of the chord (major or minor);
- if the chord type is *own*, then instead of the *type*, *chord_name* and *mode* tags, the *notes* tag is written inside the *chord* tag, which contains the value properties for *notes* and *durations* of the chord to create. Notes are separated by commas, the pairs are separated by a semicolon;
- after closing the *chord* tag, the *melody* tag follows, which indicates the melodic part of the composition. This tag, as well as the *chord* tag, requires explicit closing with the *</chord>* and *</melody>* tags respectively;
- inside the *melody* tag the *note*, *note_name* and *octave* tags are located. The first contains the *duration* property, which indicates the duration of the note, the second contains the *value* property, which indicates the name of the pitch (C, D, E, F, G, A, H etc.). Finally, the third *octave* tag has a *value* property that represents the octave of the note;

Example of xml text with harmonic and melodic parts is represented at "Fig. 2"

```
<doc>
    <tonality>d_minor</tonality>
    <tempo value="60"></tempo>
    <harmony>
        <chord duration="4">
            <type value="standard"/>
            <chord_name value="d"/>
            <mode value="minor"/>
        </chord>
        <chord duration="4">
            <type value="standard"/>
            <chord_name value="d"/>
            <mode value="major"/>
        </chord>
        ..........................
        <chord duration="4">
            <type value="own"/>
            <notes value="f,2;c,3;f,3;g,3;c,4"/>
        </chord>
    </harmony>
    <melody>
        <note duration="2">
            <note_name value="silence"/>
        </note>
        <note duration="8">
            <note_name value="f"/>
            <octave value="4"/>
        </note>
        <note duration="8">
            <note_name value="g"/>
            <octave value="4"/>
        </note>
        <note duration="4">
            <note_name value="c"/>
            <octave value="5"/>
        </note>
        ............
        <note duration="8">
            <note_name value="f"/>
            <octave value="4"/>
        </note>
    </melody>
</doc>
```

Figure 2. Example of xml text with harmonic and melodic parts.

E. Screenshots of the web service

Screenshots of a web service for generating a musical sequence from an image are shown in "Fig. 3" and "Fig. 4".

Figure 3. Image sound generation page.

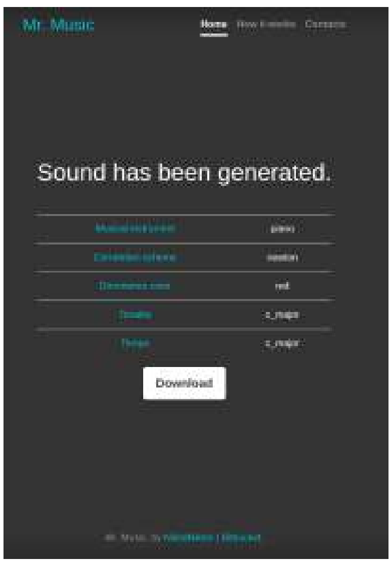


Figure 4. Generated sounds download page.

IV. Testing of neural network parameters

A. Number of LSTM layers

The size of LSTM layers has a big impact on the quality of the resulting musical composition. The larger the size, the faster the network converges, however, this leads to retraining. This may be due to the fact that a larger number of neurons allows the network to store more training set data in the scales. It was found that the most optimal configuration for the proposed model for generating musical compositions is the presence of 4 LSTM levels, each of which contains 512 cells [10].

B. Batch size for network training

The training data should be divided into small parts, according to which the neural network is trained. After each such batch, the state of the LSTM network is reset and the weights are updated. For the testing, various batch sizes were tested and it was found that smaller batch sizes make the model more efficient. Nevertheless, the presence of small batch sizes makes training very long. It was found that a batch size of 512 gives the best result while maintaining relatively fast training [10].

V. Conslusion

In this work, the main stages of developing a web service for generating musical compositions from an image was described: the main method of the program, then the application architecture, and finally, the testing process.

Acknowledgment

The work is partially supported by the Russian Foundation for Basic Research and the administration of the Volgograd region (projects 18-07-00220, 19-47-340003, 19-47-343001, 19-47-340009 project).

References

[1] V. Rozaliev Methods and Models for Identifying Human Emotions by Recognition Gestures and Motion. The 2013 2nd International Symposium on Computer, Communication, Control and Automation 3CA 2013, 2013, pp. 67–71.
[2] Wu.Xiaoying A study on image-based music generation. Master's thesis, Simon Fraser University, 2008.
[3] D. Cope Computer Models of Musical Creativity, Cambridge, MIT Press, Cambridge Mass., 2005.
[4] D. Chernyshev Tsveta i noty. Available at: http://mi3ch.livejournal.com/2506477.html (accessed 2019, Dec)
[5] R. Szeliski Computer Vision: Algorithms and Applications, New York, Springer Science & Business Media, 2011. 979 p.
[6] J. Fernández, F. Vico AI Methods in Algorithmic Composition: A Comprehensive Survey. Journal of Artificial Intelligence Research, 2013, no 48, pp. 513-582
[7] J. Schmidhuber Deep learning in neural networks: An overview, Neural Networks, 2015, vol 61, pp 85-117
[8] V. Vakhromeev, Elementarnaya teoriya muzyki [Elementary music theory], Moskow, Gosudarstvennoe muzykal'noe izdatel'stvo, 1962. 248 p.
[9] M. Russ Sound Synthesis and Sampling, London, Taylor & Francis Group, 2012. 568 p.
[10] N. Hewahi, S. AlSaigal and S. AlJanahi Generation of music pieces using machine learning: long short-term memory neural networks approach. Arab Journal of Basic and Applied Sciences, vol. 26, no. 1, 2019, pp. 397-413

Разработка веб сервиса для генерации музыкальной композиции по изображению

Никитин Н.А., Розалиев В.Л., Орлова Ю.А.

В данной статье описывается процесс проектирования и разработки веб-сервиса для автоматической генерации звуков (музыкальной композиции) по изображению. Также описывается основной метод синтеза музыки из изображения, который основан на совместном использовании нейронных сетей и светомузыкальной теории. В первом разделе кратко описано развитие данной области и описана постановка задачи. Второй раздел посвящён определению основного метода работы программы - преобразования худодественных характеристик в музыкальные. Третий раздел содержит детальное описание процесса разработки веб-сервиса. Последний раздел описывает процесс тестирования.

Received 30.12.2019

The Smart Urban Transport System

1st Alena Shviatsova
dept. Information intelligent technology
Brest State Technical University
Brest, Belarus
helengood@gmail.com

2nd Vasily Shuts
dept. Information intelligent technology
Brest State Technical University
Brest, Belarus
lucking@mail.ru

***Abstract*—In the paper principles of functioning an intelligent urban transport system based on unmanned electric vehicles which capable of adapting to passenger traffic's changes in real time are considered. The algorithm of drawing up a plan of passenger transportation by means of such transport system and cassette-conveyor method of passenger delivery are described. The upper limit of the necessary number of transport units for the implementation of the distribution plan is given.**

***Keywords*—Smart Urban Mobility, Smart Urban Transport Vehicle System, Infobus, Unmanned Vehicles**

I. Introduction

Continued urbanization and the increasing density of population and traffic flows in cities are creating the discomfort of living and using transport systems that built on the solutions of past years. At the same time, the development of information technologies and technology platforms makes it possible to address emerging social and environmental issues and to balance different types of traffic flows (pedestrian, bicycle, new-mobility, motor vehicles). One of these approaches is called Smart Urban Mobility (SUM, Smart Urban Mobility), Fig.1.

Figure 1. Example of a figure caption.

Within the paradigm SUM the concept of public transport system, developing from the point of view of priorities of passengers and pedestrians interests - PRT (Personal Rapid Transit) is received recognition. PRT - transport systems use small unmanned vehicles which moving on a dedicated line, which carry out transportation mainly in the mode of "origin - destination", i.e. movement "on demand" of the passenger from the initial stop to the final one without intermediate stops. Also distinguishing features of PRT-transport are minimization of waiting time and small volume of the cabin, which provides privacy of the trip, comparable to the conditions of private transport. At present, PRT transport systems are successfully used in the West in highly connected cluster logistics terminals: airports, railway stations, university and medical infrastructures located in different parts of the metropolis. As a concrete example, the use of PRT-transport system in Heathrow airport, England.

II. Smart urban transport passenger system

A. Description of smart transport system

The proposed article describes a PRT-like transport system designed for mass passenger transportation based on the use of unmanned electric vehicles and describes the principles of its operation. This intelligent transport system carries all the features of PRT, but can also act as an alternative to traditional public transport, offering a new method for the implementation of passenger transportation: cassette-conveyor transportation, the essence of which is a continuous (conveyor), with small intervals, sending to the route of electric cars connected to virtual couplings in cassettes, consisting of the necessary number of vehicles, as in road trains [1]. The passenger capacity of such cassettes provides with a small excess of the volume of requests for service from the passenger trains. The cassette picks up all passengers travelling to several neighboring stops on the route, thus combining public transportation and PRT: the cassette can carry quite a large number of passengers, comparable to transport with increased passenger capacity, and at the same time, passengers travel with a minimum number of stops during the journey. Thus, the proposed intelligent transport system consists of:

- a fleet of small-capacity unmanned electric cars, called infobuses, which are remotely controlled by a single information center (server) and are sent to the route depending on the intensity of passenger traffic so as to slightly overlap it, Fig. 2a;
- a system of terminals for payment and collection of requests for delivery from passengers that are placed at the route stops, as shown in Fig. 2b;
- a two-way route consisting of k stops, as shown in Fig. 3;

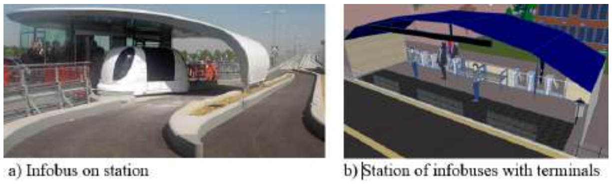


Figure 2. Infuses and its stations

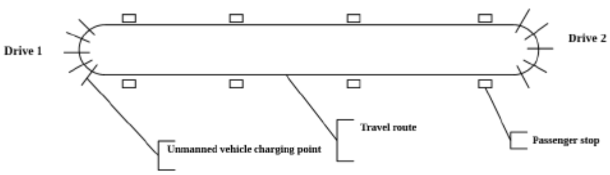


Figure 3. The Route of infuses

B. Functioning of smart transport system

Infobuses move along a dedicated line on the carriageway one after the other without overtaking, which is risky road maneuver, thus significantly improving the safety of the journey, if necessary, united by virtual connections into cassettes. When paying at a stop through the terminal, the passenger also indicates the stop to which he wishes to go. Requests from stop points are received by the information transport system server and it forms a special correspondence matrix M_z, *Z=1,2, ...*, recording each arriving passenger at the stop. Each element m_{ij} of matrix determines the number of passengers going from stop *i* to stop *j*, $i = \overline{1, k-1}, j = \overline{2, k}$,where *k* - the number of stops in one direction of the route.

$$M_z = \begin{pmatrix} 0 & m_{1,2} & \cdots & m_{1,k-1} & m_{1,k} \\ 0 & 0 & \cdots & m_{2,k-1} & m_{2,k} \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & 0 & m_{k-1,k} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \quad (1)$$

The matrix of correspondence is fixed and the server begins to form the delivery plan of passengers on it , when one of its elements will satisfy a condition:

$$m_{ij} = a * V, a \in [0.5, 1), i = \overline{1, k-1}, j = \overline{2, k} \quad (2)$$

where a is the *elasticity coefficient* that provides places for passengers, who will arrive to a stop between the moment of beginning of drawing up of plan and the moment of arrival of an infobus on stop, V - volume of an infobus.

Thus, of drawing up a distribution plan, it is determined:

- the number of infobuses involved
- a serial number $n_i \in N$ for each used infobus, where *i* - indicates the initial stop from which the infobus will take the passengers
- a set of delivery stops for all infobuses of a delivery plan $\bigcup_{i=1}^{k-1} \bigcup J_{n_i}$, where the content of the set J_{n_i} indicates the sequential numbers of stops on which passengers will leave the infobus.

During the drawing up the delivery plan each string *i* of M_z , which contains information about the requests of passengers travelling from stop *i* to next stops of route, is processed. For insurance non - conflict motion of infobuses during transportation of passengers from origin stop *i*, the infobuses will be sent first to distant stops, then to the nearest: $j = k, k-1, ..., i+1$. Thus, each infobus receives its own sequential number n_i. That means, the first infobus receives a number 1_1,when it will transport passengers from the first stop (when processing the first line of M_z matrix) . It will deliver passengers to the last stop *k* and, possibly, to some neighboring stops $k-1, k-2, ...$. The total number of passengers , which travel to these stops, can not exceed the *V* volume of the infobus.

Each infobus with an sequential number n_i has its own set of available stops. In the algorithm it will be called a *potential set of stops* and will be marked J_{n_iP} . It includes all stops on the route behind the original infobus stop, except for those stops to which previous infobuses from the same stop have already delivered passengers. Such infobus will deliver the passengers to stopping points that constitute a *real set of stops* J_{n_i} , and which is a subset of a *potential set of stops* $J_{n_i} \in J_{n_iP}$.

The exact upper boundary (the smallest upper boundary) of numerical set *M* in mathematics is called *supreme* and is signed *sup M* [2].

The stop with the greatest ordinal number of a *potential set of stops* J_{n_iP} of an infobus n_i will be an exact upper boundary of set J_{n_iP} and is signed as $supJ_{n_iP}$ (supreme J_{n_iP}), and always will enter into *real set of stops* J_{n_i} of an infobus n_i. Whether this set will include other stops depends on the volume of the infobus and the number of passengers coming to them.

To determine the *real set of stops* J_{n_i} of the infobus n_i , the algorithm uses a value δ_{n_i} that represents the number of stops that entered the *real set of stops* J_{n_i} of the infobus without stopping $supJ_{n_iP}$, or $\delta_{n_i} = |J_{n_i}| - 1$. Thus, for an infobus n_i , the *potential set of stops* J_{n_iP}, the value δ_{n_i} and the *real set of stops* J_{n_i} are determined from the following conditions [3-12]:

$$\begin{cases} J_{n_iP} = \{i+1, ..., k\} \setminus \bigcup J_{n_i-1}, \\ J_0 = \emptyset \, n_i, \delta_{n_i} \in N_0, \\ \sum\limits_{j=SupJ_{n_iP}-\delta_{n_i}}^{SupJ_{n_iP}} m_{ij} \leq V, \\ \sum\limits_{j=SupJ_{n_iP}-\delta_{n_i}-1}^{SupJ_{n_iP}} m_{ij} > V \\ J_{n_i} = \{j | j \in N_0, \\ SupJ_{n_iP} - \delta_{n_i} \leq j \leq SupJ_{n_iP}\} \end{cases} \tag{3}$$

Also, the proposed smart transport system is able to determine stopping points from real set of stops delivery J_{n_i} of infobus n_i, where vehicle can after disembarking also pick up additional passengers who are going to go to his next stops of the set delivery J_{n_i} . In another words, all passengers (who going to the next stopping points of real stop set J_{n_i}) are taken in the vehicle from the disembarking stop of set J_{n_i}, if their total number does not exceed the current amount of available seats in the infobus.

As an example, let's consider the fourth line of some correspondence matrix 10×10 :

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 2 & 4 & 6 & 3 & 7 & 15 \end{pmatrix}$$

Let at this stop after disembarking of passengers the amount of available seats in some infobus with the number $n_i, i < 4$ is equal to 11, and real set of delivery stations is $J_{n_i} = \{4, 5, 6, 7, 8, 9\}$. In this case passengers are travelling from the fourth stop to the stops {5,6,8} can enter the infobus: $2 + 4 + 3 < 11$, or to {6,7} : 4+6<11, or {6,9}: 4+7=11 and so on. In another words, the set of such stops is variable and determines the number of infobuses that are involved in the transportation plan and the volume of used passenger capacity of the vehicle's cabin .

With the help of the developed software for drawing up delivery plan the simulation of the model's work was carried out using three algorithms of forming such sets stopping points: sequential selection of stops, greedy algorithm and dynamic method of the task of filling the knapsack , using such indicators as the *number of infobuses involved* and *coefficient of use passenger capacity of vehicle* K_{pc}, which is determined by the formula:

$$K_{pc} = \frac{\sum_{i=1}^{n_{span}} (1 - V_{fp_i})}{V \times n_{span}} \tag{4}$$

where V - infobus passenger capacity, n_{span} - the total number of all spans (intervals between neighbouring stops) on which the infobuses of delivery plan carried passengers, V_{fp_i} - is the amount of free places in the infobus on the span.

On the basis of 1000 tests data, that represent the dependence of the average *number of infobuses involved* and the average value of *the coefficient of use of passenger capacity* K_{pc} as a function of the value of the *elasticity coefficient* ***a*** on the above three algorithms of choice of stops for additional boarding of passengers, are shown in Fig. 4 and Fig.5 .

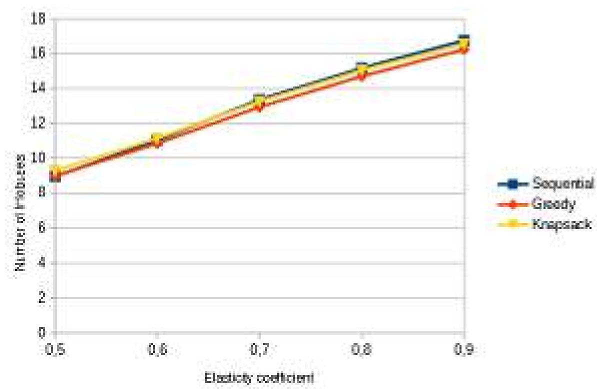


Figure 4. Dependence of the average infobus number on the elasticity coefficient.

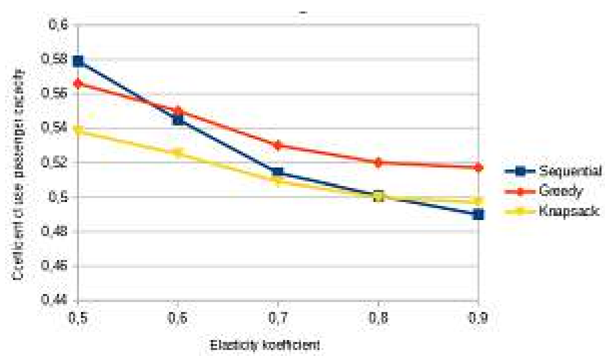


Figure 5. Dependence of the average coefficient of use passenger capacity on the elasticity coefficient.

These tests have demonstrated the greatest effectiveness of the greedy method of selecting additional boarding stops over the entire determination area.

C. Estimation of the upper limit the number of infobuses used

The upper limit of the required number of infobuses can be estimated with an odd number of stops as well as $N_{ul} = \frac{k^2-1}{4}$, with an even number of stops $N_{ul} = \frac{k^2}{4}$. The correctness of calculation of the upper limit has been confirmed experimentally. Fig. 6 shows a graphic of the dependence the number of used infobuses on the number of stops $k = 7$ ($N_{ul} = 12$) in one direction of the route for three value of elasticity coefficient .

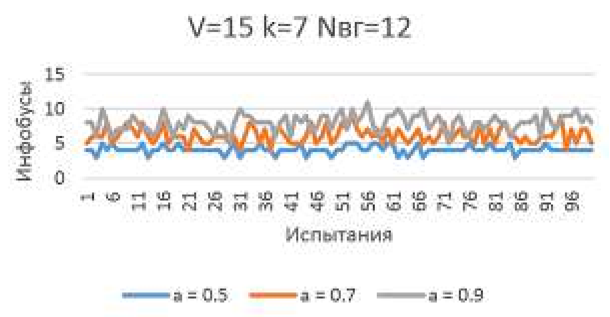


Figure 6. Dependence of the number of used infobus on the number of stations.

Conclusions

The paper describes the principles of functioning a smart urban passenger transport system based on the use of remotely operated by server unmanned vehicles called infobuses, which could become a new type of public transport. This type of transport is capable of operating without interference from other vehicles in a busy street and road environment and carrying a number of passengers comparable to the subway, with not only economic but also environmental benefits. And also the algorithm of drawing up of the plan of transportation of passengers by means of the given transport system with use of a conveyor-cassette mode of transportation which allows functioning it independently without participation or with the minimum participation of the person is described. The work is relevant, as the proposed transport system is able to show adaptability to the dynamics of changes in road and transport conditions.

References

[1] SARTRE.Available at: http://en.wikipedia.org/wiki/Safe Road Trains for the Environment (accessed:09.01.2020).

[2] Rudin, W. 1976. Chapter 1. The Real and Complex Number Systems. Principles of Mathematical Analysis. McGraw-Hill:p.4

[3] A.N. Zhogal, V.N. Shut', E.V. Shvetsova. Avtomaticheskii gorodskoi intellektual'nyi passazhirskii transport[Automatic urban intelligent passenger transport].Transport i innovatsii: vyzovy budushchego: materialy Mezhdunarodnoi nauchnoi konferentsii [Transport and innovation: challenges for the future: proceedings of the International Scientific Conference],2019, Minsk: Natsional'naya biblioteka Belarusi,pp.23-33

[4] E.V. Shvetsova, V.N. Shut'. Algoritm sostavleniya plana perevozok na gorodskikh liniyakh v intellektual'noi sisteme upravleniya bespilotnymi transportnymi sredstvami[Algorithm of drawing up a plan of transportation on urban lines in the intellectual system of management of unmanned vehicles].Vestnik Khersonskogo natsional'nogo tekhnicheskogo universiteta[Herald of Kherson National Technical University],2019,vol.2(69),no3,pp.222-230

[5] E.V. Shvetsova, V.N. Shut'. Algoritm sostavleniya plana perevozok na gorodskikh liniyakh v intellektual'noi sisteme upravleniya bespilotnymi transportnymi sredstvami [Algorithm of drawing up a plan of transportation on urban lines in the intellectual system of management of unmanned vehicles]. Materialy XX mezhdunarodnoi konferentsii po matematicheskomu modelirovaniyu: sbornik materialov konferentsii [Proceedings of the XX International Conference on Mathematical Modelling: a collection of conference proceedings],2019, p.115

[6] V. Shuts, A. Shviatsova. System of urban unmanned passenger vehicle transport. ICCPT 2019: Current Problems of Transport: Proceedings of the 1st International Scientific Conference, 2019,pp. 172-184

[7] V.N. Shut', E.V. Shvetsova. Algoritm organizatsii gorodskikh passazhirskikh perevozok posredstvom rel'sovogo bespilotnogo transporta "Infobus"[Algorithm of organization of city passenger transportation by means of rail unmanned transport "Infobus"].ACTUAL PROBLEMS OF FUNDAMENTAL SCIENCE: third international conference,2019,pp.222-226

[8] V. Shuts, A. Shviatsova. Cassette robotized urban transport system of mass conveying passenger based on the unmanned electric cars. Science. Innovation. Production. Proceedings of the 6th Belarus-Korea Science and Technology Forum, 2019,pp.81-83

[9] V. Shuts, A. Shviatsova. Intelligent system of urban unmanned passenger vehicle transport. Abstracts of the 16th European Automotive Congress (EAEC 2019) hosted jointly the Academic Automotive Association (Belarus), the European Automobile Engineers Cooperation (EAEC) and the Federation Internationale des Societes d'Ingenieurs des Techniques de l'Automobile (FISITA),2019,pp.18

[10] E.V.Shvetsova, V.N.Shut'. Bezopasnost' bespilotnykh transportnykh sistem[Safety of Unmanned Transport Systems]. Materialy IX mezhdunarodnoi nauchno-prakticheskoi konferentsii:V 2-kh ch [Materials of IX International Scientific and Practical Conference: In 2 parts],2019,v.1,pp.90-91

[11] V.N. Shut', E.E. Prolisko, E.V. Shvetsova. Avtomatizirovannaya sistema upravleniya razdelyayushchimsya passazhirskim transportom[Automated management system for split passenger transport]. Doklady XVIII Mezhdunarodnoi konferentsii «Razvitie informatizatsii i gosudarstvennoi sistemy nauchno-tekhnicheskoi informatsii»[Reports of the XVIII International Conference "Development of Informatization and State System of Scientific and Technical Information"],2019,pp.176-180

[12] A.A. Burov, D.O. Boiko, E.V. Shvetsova. Algoritmy funktsionirovaniya bespilotnykh gorodskikh passazhirskikh transportnykh sistem[Unmanned urban passenger transport systems operation algorithms]. Sbornik konkursnykh nauchnykh rabot studentov i magistrantov[Collection of competitive scientific works of students and master students],2019,pp.62-65

Умная городская транспортная система

Швецова Е.В., Шуть В.Н.

Данная статья посвящена описанию интеллектуальной городской транспортной пассажирской системы на основе беспилотных электрокаров, называемых инфобусами, и принципам ее фукнционирования, а именно конвейерно-кассетному способу развозки пассажиров. Во введении приведены предпосылки возникновения новых видов транспортных систем в городах и в частности PRT-транспорта. Второй раздел посвящен описанию самой транспортной системы на основе беспилотных электрокаров и алгоритмам составления плана развозки пассажиров посредством конвейерно-кассетного способа. В третьем разделе приведена оценка верхней границы использованных в плане развозки инфобусов.

Received 08.01.2020

AUTHOR INDEX

АВТОРСКИЙ УКАЗАТЕЛЬ

Научное издание

Открытые семантические технологии проектирования интеллектуальных систем

Open Semantic Technologies for Intelligent Systems

Сборник научных трудов

Основан в 2017 году

Выпуск 4

В авторской редакции

Ответственный за выпуск *В. В. Голенков*
Компьютерная вёрстка *Н. В. Гракова*

Подписано в печать 30.01.2020. Формат 60*84 1/8.
Бумага офсетная. Гарнитура «Таймс». Ризография.
. Усл. печ. л. 42,55. Уч.-изд. л. 76,28.
Тираж 140 экз. Заказ 21.

Полиграфическое исполнение
УП «Бестпринт». Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий №1/160 от 27.01.2014.
Ул. Филатова, д. 9, к. 1, 220026, г. Минск

OSTIS-2021

11th international scientific and technical conference «Open Semantic Technologies for Intelligent Systems»

Open Semantic Technologies for Intelligent Systems

February 24-27, 2021 Minsk. Republic of Belarus

C A L L F O R P A P E R S

We invite you to take part in X International Scientific and Technical Conference “Open Semantic Technologies for Intelligent Systems” (OSTIS-2021), which will focus on areas of use of the semantic technologies.

Conference will take place from ***February, 24th to February, 27nd, 2021*** at the Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus.

Conference proceedings language: English

Working languages of the conference: Russian, Belarusian, English

MAIN ORGANIZERS OF THE CONFERENCE

- Ministry of Education
- Ministry of Communications and Informatization
- State Institution “Administration of High Technologies Park” (Republic of Belarus)
- Educational-scientific association in the direction of "Artificial Intelligence" (ESA-AI)
- Belarusian State University of Informatics and Radioelectronics (BSUIR)
- Brest State Technical University (BrSTU)
- The State Scientific Institution «The United Institute of Informatics Problems of the National Academy of Sciences of Belarus» (UIIP NASB)
- Russian Association of Artificial Intelligence (RAAI)
- Belarusian Public Association of Artificial Intelligence Specialists (BPA of Artificial Intelligence Specialists)

CONFERENCE TOPICS:

- *Underlying principles of semantics-based knowledge representation, and their unification.*
  Types of knowledge and peculiarities of the semantics-based representation of various knowledge and metaknowledge types.
  Links between knowledge; relations, that are defined on the knowledge.
  Semantic structure of a global knowledge base, that integrates various accumulated knowledge.
- *Parallel-oriented programming languages for processing of the semantics-based representation of knowledge bases.*
- *Models for problem solving, that are based on knowledge processing, which occurs directly at the semantics-based representation level of knowledge being processed. Semantic models of information retrieval, knowledge integration, correctness and quality analysis of knowledge bases, garbage collection, knowledge base optimization, deductive and inductive inference in knowledge bases, plausible reasoning, pattern recognition, intelligent control. Integration of various models for problem solving*
- *Semantic models of environment information perception and its translation into the knowledge base.*
- *Semantic models of multimodal user interfaces of intelligent systems, based on the semantic representation of knowledge used by them, and unification of such models.*
- *Semantic models of natural language user interfaces of intelligent systems. The structure of semantic representation of linguistic knowledge bases, which describe natural languages and facilitate solution of natural language text and speech interpretation problems, and of natural language texts and speech messages synthesis, that are semantically equal to certain knowledge base fragments.*
- *Integrated logic-semantic models of intelligent systems, based on semantic knowledge representation, and their unification*
- *Various technical platforms and implementation variants of unified logic-semantic models of intelligent systems, based on semantic knowledge representation*
- *Models and means, that are based on the semantic representation of knowledge and that are oriented to the design of various typical components of intelligent systems (knowledge bases, programs, problem solvers, user interfaces).*
- *Models and means, that are based on semantic representation of knowledge and that are oriented to the complex design of various classes of intelligent systems (intelligent reference systems, intelligent learning systems, intelligent control systems, intelligent robotics systems, intelligent systems for design support etc.)*
- *Applied intelligent systems, that are based on the semantic representation of knowledge used by them*

CONFERENCE GOALS AND FORMAT

The goal of the conference is to discuss problems of creation of the **Open Complex Semantic Technology for Intelligent Systems Design**. This determines the Conference format, which involves wide discussion of various questions of creating of such technology and poster sessions.

During the **poster sessions** every participant of the conference will have an opportunity to demonstrate his results. Conference format assumes exact start time of each report, and exact time of its exhibition presentation.

One of the major objectives of the conference is to attract not only scientists and postgraduate students, but also students who are interested in artificial intelligence, as well as commercial organizations willing to collaborate with research groups working on the development of modern technologies for intelligent systems design.

PARTICIPATION TERMS AND CONDITIONS

All those interested in artificial intelligence problems, as well as commercial organizations willing to collaborate with research groups working on the development of modern technologies for intelligent systems design are invited to take part in the Conference.

To participate in the OSTIS-2021 conference, it is necessary to register in the CMT system before December 15, 2020, find conference page, and from there:

- submit a **participation form** for the OSTIS-2021 conference. Each participation form field is required, including indication of the reporter. By filling in the registration form, you agree that your personal data will be processed by the Organizing Committee of the Conference, and that the paper and information about the authors will be published in printed and electronic format. Participation form should contain information on all of the authors. If author(s) are participating with a report, participation form should have their **color photo(s)** attached (they are needed for the Conference Program);
- upload an **article** for publication in the OSTIS-2021 Conference Proceedings. Papers should be formatted according to the provided template (see https://proc.ostis.net/for-authors/). Four full pages is a minimum size of a paper.
- send the signed **scan of the letter of consent**

If a report is submitted to participate in one of the contests, this intention should be clearly indicated in the participation form.

The selection of papers for publication in the Conference Proceedings and participation in the Conference is performed by a number of reviewers from among the members of the Conference Program Committee.

Incompliant applications and papers will be rejected.

Authors, whose articles were included in the Conference Program, will receive the invitations for participating in the Conference before February 10th, 2021.

Conference participation does not require any fees.

PAPERS SUBMISSION PROCEDURE

Papers (only on topics mentioned above) should be submitted ready for publication (http://proc.ostis.net/eng/main.html -> For authors). The text should be logically complete and contain new scientific and practical results. Each author is allowed to submit two reports maximum.

After the article was submitted, it is sent for review. Review results will become available to the paper author(s) on the CMT website before January 25th.

The Organizing Committee reserves the right to reject any paper, if it does not meet the formatting requirements and the Conference topics, as well as if there was no participation form submitted for the paper.

YOUNG SCIENTIST REPORTS CONTEST

Authors of the report submitted to the contest may include scientists with scientific degrees, but the report should be made by those without a degree and under 35 years old.

To take part in the young scientists report contest, it is necessary to:

1) fill in the participation form, where your participation in the contest is clearly indicated;
2) write an article and upload it to the CMT website;
3) fill in, sign, scan and send letter of consent via the email.
4) make a report at the conference (in person);

YOUNG SCIENTIST PROJECTS CONTEST

Projects of applied intelligent systems and systems aimed at supporting the design of intelligent systems are allowed to take part in the contest; they have to be presented by a scientist without a degree and under 35 years old.

To take part in the young scientist projects contest, it is necessary to:

1) fill in the participation form, where your participation in the contest is clearly indicated;
2) write an article and upload it to the CMT website;
3) make a report at the conference (in person);
4) make an exhibition presentation of the software

STUDENT INTELLIGENT SYSTEM PROJECTS CONTEST

To participate in the contest, a project must meet the following criteria: (a) it was developed by students and/or undergraduates of the higher eduction institutions, and (b) project consultants and advisors must hold a scientific degree and title. To participate in this contest, it is necessary to:

1) familiarize yourself with contest's terms and conditions (http://www.conf.ostis.net);
2) fill in the participation form for the contest (http://www.conf.ostis.net);
3) prepare a summary of the project (http://www.conf.ostis.net).
4) submit the participation form and project summary to the student projects' email address: ostis.stud@gmail.com.

CONFERENCE PROCEEDINGS PUBLICATION

The Conference Organizing Committee plans to publish the papers selected by the Program Committee based on the results of their review, in the Conference Proceedings, on the official Conference website http://conf.ostis.net and on the Conference Proceedings website http://proc.ostis.net**.**

Upon successful review author sends a letter of consent to the Organizational Commitee. Author therefore agrees that his paper can be made freely available in electronic form at other resources at the Editorial Board's discretion.

KEY DATES OF THE CONFERENCE

October 1st, 2020	paper submission opens
December 15th, 2020 ***January 25th, 2021***	paper submission deadline paper review deadline
January 30th, 2021	final decision on paper publication; sending out invitations and notifications on inclusion of a paper in the OSTIS-2021 Conference Proceedings
February 20th, 2021	Draft Conference Program publication on the conference website http://conf.ostis.net
February 22th, 2021	Conference Proceedings and Conference program publication on the conference website http://proc.ostis.net
February 24th, 2021	Participant registration and OSTIS-2021 conference opening
February 24th to 27th, 2020	OSTIS-2021 conference
Marth 1th, 2021	photoreport and conference report publication on the conference website: http://conf.ostis.net
March 15th, 2021	conference proceedings will be uploaded to the Russian Science Citation Index database

CONFERENCE PROGRAM FORMATION

Conference program is formed by the Program Commitee according to the paper review results; author(s)' confirmation of participation is required as well.

CONTACTS

All the necessary information about the forthcoming and previous OSTIS Conferences can be found on the conference website http://conf.ostis.net and http://proc.ostis.net.

For questions regarding conference participation and dispute resolution please contact: ostisconf@gmail.com.

Methodological and advisory support to the conference participants shall be provided through the conference e-mail only.

The conference venue is the 5th academic building of the Belarusian State University of Informatics and Radioelectronics (39, Platonov str., Minsk, Republic of Belarus).

OSTIS-2021

XI международная научно-техническая конференция «Открытые семантические технологии проектирования интеллектуальных систем»

Open Semantic Technologies for Intelligent Systems

24 – 27 февраля 2021 г. Минск. Республика Беларусь

И Н Ф О Р М А Ц И О Н Н О Е   П И С Ь М О

Приглашаем принять участие в XI Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS-2021), которая будет посвящена вопросам области применения семантических технологий.

Конференция пройдет в период с ***24 по 27 февраля 2021*** года в Белорусском государственном университете информатики и радиоэлектроники, г. Минск, Республика Беларусь.

Язык статей сборника конференции: английский

Рабочие языки конференции: русский, белорусский, английский.

ОСНОВНЫЕ ОРГАНИЗАТОРЫ КОНФЕРЕНЦИИ

- Министерство связи и информатизации Республики Беларусь
- Министерство образования Республики Беларусь
- Государственное учреждение «Администрация Парка высоких технологий» (Республика Беларусь)
- Учебно-научное объединение по направлению «Искусственный интеллект» (УНО-ИИ)
- Белорусский государственный университет информатики и радиоэлектроники (БГУИР)
- Брестский государственный технический университет (БрГТУ)
- Государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси» (ОИПИ НАН Беларуси)
- Российская ассоциация искусственного интеллекта (РАИИ)
- Белорусское общественное объединение специалистов в области искусственного интеллекта (БОИИ)

НАПРАВЛЕНИЯ РАБОТЫ КОНФЕРЕНЦИИ:

- *Принципы, лежащие в основе семантического представления знаний, и их унификация.*
  Типология знаний и особенности семантического представления различного вида знаний и метазнаний.
  Связи между знаниями и отношения, заданные на множестве знаний.
  Семантическая структура глобальной базы знаний, интегрирующей различные накапливаемые знания
- *Языки программирования, ориентированные на параллельную обработку семантического представления баз знаний*
- *Модели решения задач, в основе которых лежит обработка знаний, осуществляемая непосредственно на уровне семантического представления обрабатываемых знаний. Семантические модели информационного поиска, интеграции знаний, анализа корректности и качества баз знаний, сборки информационного мусора, оптимизации баз знаний, дедуктивного и индуктивного вывода в базах знаний, правдоподобных рассуждений, распознавания образов, интеллектуального управления. Интеграция различных моделей решения задач*
- *Семантические модели восприятия информации о внешней среде и отображения этой информации в базу знаний*
- *Семантические модели мультимодальных пользовательских интерфейсов интеллектуальных систем, в основе которых лежит семантическое представление используемых ими знаний, и унификация этих моделей*
- *Семантические модели естественно-языковых пользовательских интерфейсов интеллектуальных систем. Структура семантического представления лингвистических баз знаний, описывающих естественные языки и обеспечивающих решение задач понимания естественно-языковых текстов и речевых сообщений, а также задач синтеза естественно-языковых текстов и речевых сообщений, семантически эквивалентных заданным фрагментам баз знаний*
- *Интегрированные комплексные логико-семантические модели интеллектуальных систем, основанные на семантическом представлении знаний, и их унификация*
- *Различные технические платформы и варианты реализации интерпретаторов унифицированных логико-семантических моделей интеллектуальных систем, основанных на семантическом представлении знаний*
- *Средства и методы, основанные на семантическом представлении знаний и ориентированные на проектирование различных типовых компонентов интеллектуальных систем (баз знаний, программ, решателей задач, интерфейсов)*
- *Средства и методы, основанные на семантическом представлении знаний и ориентированные на комплексное проектирование различных классов интеллектуальных систем (интеллектуальных справочных систем, интеллектуальных обучающих систем, интеллектуальных систем управления, интеллектуальных*

робототехнических систем, интеллектуальных систем поддержки проектирования и др.)

- *Прикладные интеллектуальные системы, основанные на семантическом представлении используемых ими знаний*

ЦЕЛЬ И ФОРМАТ ПРОВЕДЕНИЯ КОНФЕРЕНЦИИ

Целью конференции является обсуждение проблем создания **открытой комплексной семантической технологии компонентного проектирования семантически совместимых гибридных интеллектуальных систем**. Этим определяется и формат её проведения, предполагающий широкое обсуждение различных вопросов создания указанной технологии и выставочные презентации докладов.

Выставочная презентация докладов даёт возможность каждому докладчику продемонстрировать результаты своей разработки на выставке. Формат проведения конференции предполагает точное время начала каждого доклада и точное время его выставочной презентации.

Важнейшей задачей конференции является привлечение к её работе не только учёных и аспирантов, но и студенческой молодежи, интересующейся проблемами искусственного интеллекта, а также коммерческих организаций, готовых сотрудничать с научными коллективами, работающими над интеллектуальными системами и созданием современных технологий и их проектированием.

УСЛОВИЯ УЧАСТИЯ В КОНФЕРЕНЦИИ

В конференции имеют право участвовать все те, кто интересуется проблемами искусственного интеллекта, а также коммерческие организации, готовые сотрудничать с научными коллективами, работающими над созданием современных технологий проектирования интеллектуальных систем.

Для участия в конференции OSTIS-2021 необходимо до 15 декабря 2020 года зарегистрироваться в системе СМТ, найти страницу конференции и на ней:

- подать **заявку** на конференцию OSTIS-2021. Каждое поле заявки обязательно для заполнения, в том числе указание того автора, кто будет представлять доклад. Заполняя регистрационную форму, Вы подтверждаете согласие на обработку Оргкомитетом конференции персональных данных, публикацию статей и информации об авторах в печатном и электронном виде. В заявке должна содержаться информация по каждому автору. К заявке доклада должны быть прикреплены **цветные фотографии** всех авторов статьи (это необходимо для публикации Программы конференции);
- загрузить **статью** для публикации в Сборнике материалов конференции OSTIS-2021. Статья на конференцию должна быть оформлена в соответствии с правилами оформления публикуемых материалов и занимать не менее 4 полностью заполненных страниц;
- загрузить **сканированный вариант письма о согласии** на публикацию и размещения передаваемых материалов в сети Интернет.

Если доклад представляется на конкурс докладов молодых учёных или на конкурс программных продуктов молодых учёных, это должно быть явно указано в заявке доклада.

Отбор статей для публикации в Сборнике и участия в работе конференции осуществляется рецензентами и редакционной коллегией сборника.

Заявки и статьи, оформленные без соблюдения предъявляемых требований, не рассматриваются.

До 10 февраля 2021 года, авторам статей, включённых в Программу конференции, направляются приглашения для участия в конференции.

Участие в конференции не предполагает организационного взноса.

ПОРЯДОК ПРЕДСТАВЛЕНИЯ НАУЧНЫХ СТАТЕЙ

Статьи (только по перечисленным выше направлениям) представляются в готовом для публикации виде (*http://proc.ostis.net -> Авторам*). Текст статьи должен быть логически законченным и содержать новые научные и практические результаты. От одного автора допускается не более двух статей.

После получения статьи, она отправляется на рецензирование и в срок до 25 января на сайте СМТ вы сможете ознакомиться с результатами рецензирования

Оргкомитет оставляет за собой право отказать в приеме статьи в случае, если статья не будет соответствовать требованиям оформления и тематике конференции, а также, если будет отсутствовать заявка доклада, соответствующая этой статье.

КОНКУРС ДОКЛАДОВ МОЛОДЫХ УЧЁНЫХ

Соавторами доклада, представляемого на конкурс докладов молодых учёных, могут быть учёные со степенями и званиями, но непосредственно представлять доклад должны авторы в возрасте до 35 лет, не имеющие степеней и званий.

Для того, чтобы принять участие в конкурсе научных докладов молодых учёных, необходимо:

1) заполнить заявку на участие в конференции, в которой чётко указать своё желание принять участие в данном конкурсе;
2) написать статью на конференцию и загрузить на сайте CMT;
3) заполнить, подписать, отсканировать и отправить по почте письмо о согласии;
4) лично представить доклад на конференции.

КОНКУРС ПРОЕКТОВ МОЛОДЫХ УЧЁНЫХ

Принимать участие в конкурсе проектов молодых учёных могут проекты прикладных интеллектуальных систем и систем, ориентированных на поддержку проектирования интеллектуальных систем, при этом представлять проект на конкурсе должен молодой учёный в возрасте до 35 лет, не имеющий учёной степени.

Для того, чтобы принять участие в конкурсе программных продуктов молодых учёных, необходимо:

1) заполнить заявку на участие в конференции), в которой чётко указать своё желание принять участие в данном конкурсе;
2) написать статью на конференцию и загрузить на сайте CMT;
3) лично представить доклад на конференции;
4) провести выставочную презентацию, разработанного программного продукта.

КОНКУРС СТУДЕНЧЕСКИХ ПРОЕКТОВ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

В конкурсе студенческих проектов могут принимать участие проекты, разработчиками которых являются студенты и магистранты высших учебных заведений, консультантами и руководителями проекта могут быть лица, имеющие научную степень и звание. Для того, чтобы принять участие в данном конкурсе, необходимо:

1) ознакомиться с положением о конкурсе студенческих проектов (http://www.conf.ostis.net);
2) заполнить заявку на участие в конкурсе студенческих проектов (http://www.conf.ostis.net);
3) подготовить описание проекта (http://www.conf.ostis.net).
4) выслать заявку на участие в конкурсе и описание проекта по электронному адресу конкурса студенческих проектов: ostis.stud@gmail.com.

ПУБЛИКАЦИЯ МАТЕРИАЛОВ КОНФЕРЕНЦИИ

Оргкомитет конференции предполагает публикацию статей, отобранных Программным комитетом по результатам их рецензирования, в Сборнике материалов конференции и на официальном сайте конференции http://conf.ostis.net и официальном сайте сборника http://proc.ostis.net**.**

По результатам рецензирования автор отправляет оргкомитету письмо о согласии, которое предусматривает дальнейшую возможность размещения статей, вошедших в сборник конференции, в открытом электронном доступе на иных ресурсах по усмотрению редакции сборника.

КЛЮЧЕВЫЕ ДАТЫ КОНФЕРЕНЦИИ

1 октября 2020г.	начало подачи материалов для участия в конференции
15 декабря 2020г. ***25 января 2021г.***	срок получения материалов для участия в конференции Оргкомитетом срок предоставления рецензий на статьи
30 января 2021г.	срок принятия решения о публикации присланных материалов и рассылки приглашений для участия в конференции и сообщение о включении статьи в Сборник материалов конференции OSTIS
20 февраля 2020г.	размещение на сайте конференции http://conf.ostis.net проекта программы конференции
22 февраля 2021г.	размещение на сайте конференции http://proc.ostis.net Сборника материалов и Программы конференции OSTIS-2021
24 февраля 2021г.	регистрация участников и открытие конференции OSTIS-2021
24-27 февраля 2021г.	работа конференции OSTIS-2021
01 марта 2021г.	публикация фоторепортажа и отчёта о проведённой конференции на сайте конференции: http://conf.ostis.net
15 марта 2021г.	загрузка материалов сборника конференции в РИНЦ

ФОРМИРОВАНИЕ ПРОГРАММЫ КОНФЕРЕНЦИИ

Программа конференции формируется Программным комитетом по результатам рецензирования, представленных статей, а также на основании подтверждения автора(-ов) статьи о прибытии на конференцию.

КОНТАКТНЫЕ ДАННЫЕ ОРГАНИЗАТОРОВ КОНФЕРЕНЦИИ OSTIS

Вся необходимая информация по предстоящей и предыдущих конференциях OSTIS находится на сайте конференции http://conf.ostis.net, а также на сайте материалов конференции http://proc.ostis.net.

По вопросам участия в конференции и решения спорных вопросов обращайтесь: ostisconf@gmail.com.

Методическая и консультативная помощь участникам конференции осуществляется только через электронную почту конференции.

Конференция проходит в Республике Беларусь, г. Минск.

Оргкомитет конференции находится на кафедре интеллектуальных информационных технологий Учреждения образования «Белорусский государственный университет информатики и радиоэлектроники (БГУИР) – г. Минск, ул. Платонова, 39, 5-ый учебный корпус БГУИР.