УДК 004.6-024.11:004.738.5

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ ИНТЕРНЕТ- ИСТОЧНИКОВ

М.П. Батура
Заведующий лабораторией НИЛ 8.1 «Новые обучающие технологии» БГУИР, Доктор технических наук, профессор, академик «Международной академии наук высшей школы»

И.И. Пилецкий
Доцент кафедры информатики БГУИР, кандидатфизико-математических наук, доцент, старший научный сотрудник

В.А. Прытков
Проректор по учебной работе БГУИР, кандидат технических наук, доцент

Н.А. Волорова
Заведующая кафедройинформатики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: bmpbel@bsuir.by, ianmenski@gmail.com, prytkov@bsuir.by, volorova@bsuir.by

М.П. Батура

Заведующий лабораторией НИЛ 8.1 «Новые обучающие технологии» БГУИР, Доктор технических наук, профессор, академик «Международной академии наук высшей школы», заслуженный работник образования Республики Беларусь. Область научных исследований: Системный анализ, управление и обработка информации в технических и организационных системах. Опубликовано более 150 научных работ, в том числе 4 монографии, учебник, выдержавший три издания, 4 учебных пособия, имеет авторские свидетельства на изобретения.

И.И. Пилецкий

Кандидат физико-математических наук, доцент БГУИР. В сфере IT более 47 лет. Участие в разработке нескольких десятков крупных проектов: главный конструктор проекта, главный архитектор программно-информационного обеспечения, руководитель проекта, начальник отдела, заведующий лабораторией (НИИ информатики, Академия наук Беларуси, МБА, БГУИР). Автор десятков исследований, имеет более 95 публикаций.

В.А. Прытков

Проректор по учебной работе БГУИР, кандидат технических наук, доцент. Автор более 40 научных работ. Область научных интересов: обработка изображений и текстурный анализ, синтаксические методы обработки информации, анализ слабоструктурированных данных.

Н.А. Волорова

Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент. В сфере IT более 40 лет. Имеет более 140 публикаций, сферв научных интересов – модели сложных систем

Аннотация. В статье приводится описание инструментов мониторинга открытых интернет-источников с целью выявления экспертов в некоторой предметной области, определения тематик публикаций, оценки

популярности публикаций. Описываются принятые решения при разработке различных вариантов построении аналитического комплекса, полученные результаты его работы и применения методов искусственного интеллекта для проведения глубокого анализа данных.

Ключевые слова: интернет-источники, Big Data, мониторинг, анализ, Machine Learning, машинное обучение, Neo4j, Cassandra, Natural Language Processing, обработка естественного языка,

1. Введение

В настоящее все уже привыкли получать данные и информацию из интернет источников. В результате анализа данных интернет-источников полученная информация является базовой для принятия решений различными организациями.

Как правило, это неструктурированные текстовые данные, различные мультимедийные данные. Данные могут быть получены как из социальных сетей, так и тематических сайтов (газет, журналов, библиотек, компаний и т. д.), содержащих различные публикации (некоторые наиболее популярные платформы приведены на рисунке 1). Пользователи интернет-ресурсов и социальных сетей могут самостоятельно выбирать интересные им направления и читать публикации интересных им людей, которым они симпатизируют. В свою очередь создатели контента могут быть со своим контентом и своими подписчиками. Такие связи, как правило, бывают достаточно сложными и представляют собой многоуровневую циклическую сеть.

Несмотря на наличие различных систем для анализа данных из интернет-источников, данная тематика не только не утрачивает актуальности, а, напротив, становится все более востребованной. Поисковые системы помогают найти разнообразную информацию практически о любых явлениях в мире, деятельности организаций и людей, но основной проблемой является не то, как найти данные, *а, проблемой является то, как разобраться в этом многообразии источников данных и самих данных как превратить данные в информацию, а информацию – в знания для принятия разумных решений*.

Интерес к интеллектуальным системам анализа данных постоянно усиливается, поскольку создание новой техники и технологий становится все более наукоемким, причем эта тенденция характерна и для продукции массового рынка. Данный проект «Интеллектуальная система комплексного анализа данных интернет-источников» (ИС Анализа Данных) предполагает развитие системы «Система комплексного анализа данных интернет-источников» [2, 3, 4] расширении ее области применения и наделения компонент новой системы *функциональностью интеллектуальных систем, позволяющей выполнять глубокий интеллектуальный анализ исходных данных с применением методов искусственного интеллекта.*

Соответственно, компании, внедряющие новые технологии, а также придающие своей продукции интеллектуальные функции, обладают определенным конкурентным преимуществом. На глобальном уровне это проявляется в определении своей ниши в мировой экономике, ее расширении.

2. Назначение и цели

Назначение:

1. «ИС Анализа Данных» - предназначена для поддержки принятия обоснованных решений, на основе *мониторинга и анализа данных* из открытых Интернет источников, в том числе и научных публикаций.

2. Выявления важных публикаций, ведущих специалистов и *поиске экспертов* определенных *предметных областей*.

3. Создание многоцелевого, модифицируемого кластера в Университете для подготовки специалистов Data Scientist.

Рисунок *1*. – Разнообразие источников данных (источник: [1])

Цели и задачи

1. Сбор и анализ данных некоторой (некоторых) предметной области с целью:

– определения тематики и тональности публикаций, поиска экспертов (авторитетов) в предметной области, тематик их публикаций, а также просмотра их профилей;

– поиска новых областей исследований, тематик публикаций, определение областей к которым интерес уже пропал;

– просмотр различных индивидуальных параметров некоторой предметной области (публикации, снижение/увеличение интереса);

2. Разработке тематической масштабируемой системы, которая при минимальных модификациях может быть применена и в других областях, как корпоративных, так и правительственных, например, для *анализа СМИ или мониторинга* настроений в обществе и др.

Большой интерес представляют Задачи нахождения экспертов предметной области, которые имеют большую аудиторию подписчиков и на которую они могут влиять, их публикации и их связи.

Примеры задач: выявление лидера социального мнения, *группы лиц, связанных в соцсетях по некоторой тематике*, задачи *рекламы (маркетинга) определенного товара или группы товаров, поиска экспертов и документов* в некоторых научно-технических областях.

Социальные отношения могут быть явными или неявными, а социальные *сети помогают идентифицировать прямые и косвенные отношения между людьми, группами людей, и характер их взаимодействий.*

Перспективность проекта «ИС Анализа Данных» в том, что результаты могут быть использованы как для прикладных целей, например, определения наиболее перспективных и актуальных научных направлений с учетом конкретной специфики (для конкретной организации или страны, с учетом имеющегося задела и т. д.), определения экспертов в заданных предметных областях (для приглашения специалистов или формирования команды исследователей), так и для научно-образовательной цели – подготовки специалистов по анализу больших объемов неструктурированных данных (Data Scientist) на основе разработки и оптимизации аналитических ML-алгоритмов, применяемых в системе.

Модульность «ИС Анализа Данных» позволит при минимальной адаптации использовать данный программный комплекс для анализа и мониторинга различных явлений по заказу конкретных пользователей, например, анализа публикаций в «Твитере», изучения популярности «бренда», поиска блогеров в социальных сетях, изучения конкурентов, изучения рынка сбыта, поведения групп людей и т.д.

«ИС Анализа Данных» позволит не только помочь с проведением анализа и принятия решения, но и позволит сэкономить весьма дефицитный в условиях конкуренции ресурс –

время. В современной экономике важно не только дать качественный продукт с новыми свойствами, но и сделать это в числе первых.

Ценность работы заключается не только в создании «Интеллектуальной системы комплексного анализа данных интернет-источников», но и в создании в Университете многоцелевого, модифицируемого кластера для анализа данных интернет-источников и глубокой математической подготовки специалистов по анализу больших объемов неструктурированных данных, специалистов Data Scientist, внедрении результатов исследований в научно-образовательный процесс при подготовке магистрантов в области обработки больших объемов информации.

3. Технология построения «Интеллектуальной системы комплексного анализа данных интернет источников»

Основу технологии разработки системы должны составлять методы и алгоритмы построения и обслуживания графовой модели различных социальных сетей авторов (блогеров) и их публикаций (в том числе и в СМИ), ссылок на их публикации и определение рейтинга конкретного автора (блогера) публикаций, определение тематик публикаций и классификация их по областям знаний. И для глубокого интеллектуального анализа применение методов искусственного интеллекта на основе преобразований семантических данных из графовой модели в различные матричные представления для их анализа в ML.

Анализируя данные из социальных сетей (СМИ) можно выявить как прямые, так и скрытые отношения между людьми, группами людей, а также характер их взаимодействий. На основе определенных связей между субъектами группы можно сделать выводы об индивидуальных предпочтениях субъектов и прогнозировать выбор объекта предпочтения.

«ИС Анализа Данных» должна состоять из следующих компонент: сбора данных, фильтрации данных и составления «мешка слов» из N-грамм (векторизации), библиотеки аналитических модулей, хранилища данных, графовой базы данных и графа знаний, аналитического компонента, обеспечивающего, взаимодействие с пользователем и подготовки выдачи результата, клиентского модуля и универсальной интеграционной шины (управляющего компонента).

При необходимости набор модулей и компонент может быть расширен, а некоторые модули заменены новыми. Общая технология построения многофункционального комплекса по обработке данных и работы компонент, а именно чтения данных интернет-источников, фильтрации данных и векторизации, библиотеки ML модулей, хранилища и БД «граф знаний», приведена в более ранних публикациях [2, 3, 4]. В «ИС Анализа Данных» должна быть применена новая архитектура построения многофункциональных комплексов как набор постоянно работающих компонент в виде отдельных серверов, изменена предметная область и система дополнена многофункциональным компонентом БД «граф знаний» и аналитическим компонентом подготовки и выдачи результата.

Все взаимосвязи и взаимодействия компонент должны быть организованы на основе специально разрабатываемого управляющего компонента (универсальная шина) «ИС Анализа Данных». Данный компонент должен обладать функциональностью интеграции данных и приложений, реализовывать функции брокера, синхронного и асинхронного выполнения приложений. Управляющий компонент должен реализовать средства логирования, сбора статистики и мониторинга работы компонент системы.

Компонент скачивания публикаций и компонент извлечения текста из материалов должны использовать технологию Docker, инструменты мониторинга и предупреждения ошибок, развертывание и хранилище логов. Применять новую технологию извлечение медиа-данных из документов и обработки новых форматов документов (PostScript, заархивированные документы).

Компонент графовая база данных и граф знаний должен быть дополнен графовой моделью для поиска постов в социальных сетях на определённую тематику, и другими

графовыми моделями, которые могут быть максимально полезны и интересны пользователю, методами интеллектуального анализа графовой модели на основе применения ML.

Компонент хранилища данных должен быть дополнен новыми БД, соответствующими новым областям применения системы.

Библиотеки аналитических модулей новыми ML-модулями интеллектуального анализа данных.

3.1. **Ретроспективный анализ семейства СКА**

3.1.1. **«Аналитический комплекс анализа данных из открытых интернет источников»**

«Аналитический комплекс анализа данных из открытых интернет источников» (СКА) [2], позволяет получить информацию популярности некоторого бренда (компании, объекта, товара, события, явления и возможно некоторого человека) на основе анализа существующих данных из электронных газет, журналов, статей, социальных сетей с помощью различных аналитических методов и оценить отношение в Интернет-СМИ к этому явлению. Значимость проекта состоит в создании наращиваемого многомодульного комплекса, позволяющего анализировать большие данные из интернет источников различными аналитическими методами и выборе для применения наиболее подходящих для данного явления.

Аналитический комплекс из состоит компонент (см. рисунок 2): сбора данных, фильтрации данных и составление «мешка слов» из N-грамм (векторизации), хранилища данных, библиотеки аналитических модулей, подготовки выдачи результата, клиентского модуля.

При необходимости набор модулей компонент может быть расширен, а некоторые модули заменены новыми.

Сами компоненты состоят из функциональных модулей, на рисунке 2 приведена логическая схема взаимодействия компонент и модулей.

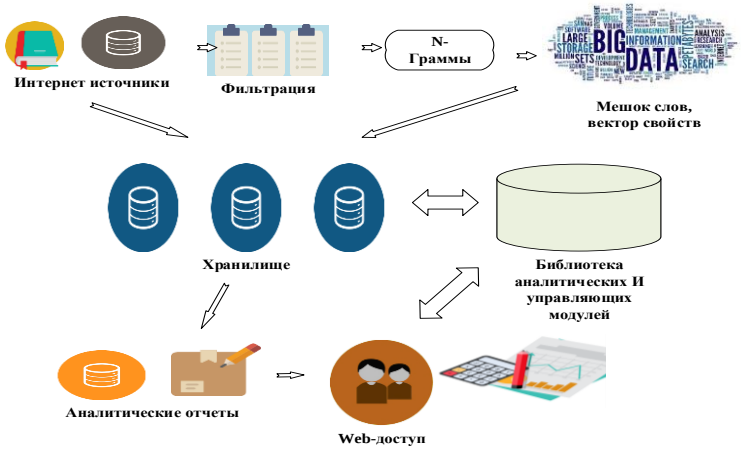


Рисунок *2*. – Логическая схема взаимодействия компонент

Аналитический комплекс выполняет свои функции в несколько этапов:

– Компонент сбора данных – выполняет целевое сканирование выбранных социальных сетей, новостных порталов, сайтов и помещает полученные документы в специальном виде в базу данных документов (в хранилище);

– Компонент обработки данных выполняет фильтрацию исходных данных (токенизацию, лемматизацию, стемминг, удаление стоп слов, перевод к нижнему регистру), построения «мешка слов» и векторизацию исследуемых наборов документов («мешок слов» в специальном виде хранится в хранилище);

– Компонент библиотека аналитических модулей, содержит набор модулей, которые осуществляют обработку данных, полученных из интернет источников с целью поиска упоминаний о брендах, определения их тональности и формирования аналитических данных для передачи клиентскому модулю, а также содержит управляющие и служебные модули. В реализованной системе компонент библиотека аналитических состоит из ML модулей: SVM и LDA, PLSA [4, 5]. Данные векторные и вероятностные модули апробированы. Применение комбинированного подхода к оценке некоторого явления позволяет наиболее достоверно его оценить, выполнить ретроспективный анализ прошедшего события, идентифицировать новое события и принять правильные решения;

– Компонент подготовки выдачи результата – подготавливает информацию в виде отчетов для пользователей аналитического комплекса;

– Компонент хранилища данных – содержит данные из социальных сетей, предварительно обработанные и размеченные данные, необходимые для построения классификатора, «мешок слов», информацию о брендах, а также служебную информацию, необходимую для работы других модулей системы.

Основной недостаток в данном решении состоит в том, что он разработан по общепринятой архитектуре для аналогичных комплексов анализа данных:

– нет общего архитектурного решения для всей системы в целом, каждый компонент представляет собой отдельную функциональность;

– все компоненты работают последовательно;

– анализ данных выполняется в отложенном режиме.

К достоинству данного архитектурного решения следует отнести удобство при модернизации и функциональному расширению компонент комплекса, т.к. компоненты состоят из отдельных модулей, которые реализуют его функциональность.

3.1.2. **Система комплексного анализа данных интернет-источников**

«Система комплексного анализа данных интернет-источников» (СКАД ИИ) [3], позволяет анализировать большие объемы данных из интернет-источников в области научных исследований и предназначенна для сбора информации о научных публикациях, построения графа знаний, что дает возможность определять экспертов предметной области, тематики их работ, их взаимосвязи, а также определять передовые научные направления.

Система позволяет находить экспертов (авторитетов) в предметной области и выдавать оценку их рейтинга влиятельности. Например, лучше прочитать три книги признанных экспертов в определённой области, чем десять книг дилетантов.

Перспективность проекта СКАД ИИ в том, что результаты могут быть использованы как для прикладных целей, например, определения наиболее перспективных и актуальных научных направлений с учетом конкретной специфики (для конкретной организации или страны, с учетом имеющегося задела и т. д.), определения экспертов в заданных предметных областях (для приглашения специалистов или формирования команды исследователей), так и для научно-образовательной цели – подготовки специалистов по анализу больших объемов неструктурированных данных (Data Scientist) на основе разработки и оптимизации аналитических ML-алгоритмов, применяемых в системе.

Модульность СКАД ИИ позволяет при небольших изменениях использовать данный программный комплекс для анализа и мониторинга различных явлений по заказу конкретных пользователей, например, как для СКА [2] изучения популярности «бренда», поиска блогеров в социальных сетях, изучения конкурентов, изучения рынка сбыта, поведения групп людей и т.д.

СКАД ИИ позволит не только помочь с проведением анализа и принятия решения, но и позволит сэкономить весьма дефицитный в условиях конкуренции ресурс – время. В современной экономике важно не только дать качественный продукт с новыми свойствами, но и сделать это в числе первых.

Ценность работы заключается не только в создании «Системы комплексного анализа данных интернет-источников», но и в создании в Университете многоцелевого, модифицируемого кластера для анализа данных интернет-источников и глубокой математической подготовки специалистов по анализу больших объемов неструктурированных данных, специалистов Data Scientist, внедрении результатов исследований в научно-образовательный процесс при подготовке магистрантов в области обработки больших объемов информации.

Facebook, Google, другие популярные сервисы построены на использовании графовых моделей данных. Facebook, например, использует не только информацию о людях, их именах, профессиях и т. д., но также сведения о взаимосвязях между людьми, которые представляют ещё большую ценность. Социальные отношения могут быть явными или неявными, а социальные сети помогают идентифицировать как прямые, так и косвенные отношения между людьми, группами людей, и характер их взаимодействий. Gartner утверждает, что способность использовать эти графы обеспечивает «устойчивое преимущество в конкурентной среде».

В отличии от известных поисковых систем (например, Facebook, Google) СКАД ИИ позволяет найти наиболее перспективные и актуальные научные направления и определить экспертов в заданных предметных областях.

Основу технологии разработки системы составляют методы и алгоритмы построения и обслуживания графовой модели социальной сети авторов и их публикаций, ссылок на их публикации и определение рейтинга конкретного автора публикаций, определение тематик публикаций и классификация их по областям знаний.

Анализируя данные из социальных сетей можно выявить как прямые, так и скрытые отношения между людьми, группами людей, а также характер их взаимодействий. На основе определенных связей между субъектами группы можно сделать выводы об индивидуальных предпочтениях субъектов и прогнозировать выбор объекта предпочтения.

Графовая модель социальной сети может быть построена на применении классической графовой модели, которая включает узлы и взаимосвязи, а также их свойства и метки.

Основным назначением графовой базы данных является применение графовых алгоритмов для обработки полученных данных, выстраивание логических взаимосвязей и подготовка и выдача информации для пользователя. Также ключевой особенностью таких БД является формирование очень гибких запросов, наподобие следующих: ***Какие авторы наиболее часто сотрудничали с автором данной популярной статьи****, является ли данный автор писателем только в области биологии,* ***или же он пишет еще и на темы математического анализа****, существуют ли математические публикации, которые по какой-то причине перекликаются с темой философии и т. д.*

СКАД ИИ состоит из следующих компонент (см. рисунок 3): сбора данных, фильтрации данных и составление «мешка слов» из N-грамм (векторизации), хранилища данных, библиотеки аналитических модулей, подготовки выдачи результата, клиентского модуля.

При необходимости набор модулей и компонент может быть расширен, а некоторые модули заменены новыми. Общая технология построения многофункционального комплекса по обработке данных из интернет источников и работы компонент, а именно чтения данных интернет-источников, фильтрации данных и векторизации, библиотеки и хранилища приведена в более ранних публикациях [2].

В *СКАД ИИ апробирована технология построения многофункциональных комплексов как набор постоянно работающих компонент в виде отдельных серверов, изменена предметная область и система дополнена компонентом БД «граф знаний» и компонентом подготовки и выдачи результата.*

Сами компоненты СКА состоят из набора функциональных модулей, на рисунке 3 приведена первоначальная логическая схема взаимодействия компонент и модулей. Данный

проект опирается на решения NLP & ML (см. рисунок 3, компоненты (2) и (3)), хранилище данных (компонент (4)), которое взаимодействует с компонентом граф – знаний (компонент (5)) и компонент взаимодействия с пользователем (компонент (6)). Для чтения данных их интернет источников используется компонент (1).

Достоинство данного решения состоит в том, что все сырые, обработанные данные и служебная информация собираются и хранятся в хранилище Hbase, которое используется для насыщения и построения графа – знаний. Хранилище может использоваться не только для накопления данных по текущей тематике, но для других областей применения, т.е хранилище может использоваться как озеро данных. Это значит, другие не графовые ML алгоритмы могут пополнять компонент (3) и использовать данные из хранилища.

Взаимодействие хранилища с графовой базой данных позволяет перестраивать и перезагружать граф - знаний, иметь множественное представление графа – знаний для различных областей применения.

Сама графованая база данных обладает рядом преимуществ и достоинств перед другими БД, она обладает свойствами OLTP & OLAP, поддерживает транзакции ACID (atomic, consistent, isolated и durable), чего не обеспечивает ни одна NoSQL БД.

Графовые технологии являются основой для построения интеллектуальных приложений, для применения алгоритмов искусственного интеллекта.

Но, как и для СКА (см. п. 3.1.1) данное решения уязвимо при возникновении сбоя при функционировании некоторых компонент. Обработка данных так же выполняется последовательно. И оказалось, что сама БД NoSQL – Hbase не обладает хорошими эксплуатационными характеристиками.

Рисунок *3*. – Первоначальная логическая схема взаимодействия компонент СКАД ИИ

3.1.3. **Модернизация система комплексного анализа данных интернет-источников**

В процессе реализации проекта СКАД ИИ были приняты важные дополнительный архитектурные решения (см. рисунок 4, новая архитектура СКАД ИИ):

– все компоненты функционируют как постоянно работающие самостоятельные сервера;

– для создания компонента хранилища БД Hbase была заменена на лидера хранилища типа " семейство столбцов" Cassandra;

– для обеспечения взаимодействия компонент как по данным, так и по управлению, разработан управляющий компонент, который выполняет роль интеграционной шины (аналог USB - enterprise service bus, для работы в среде Big Data);

– остановка работы одного из компонент не приводит к остановке работы всего комплекса;

– обеспечен многопользовательский режим доступа;

– существенно модернизированы компоненты чтения данных из интернет источников и компонент анализа и выдачи отчетов.

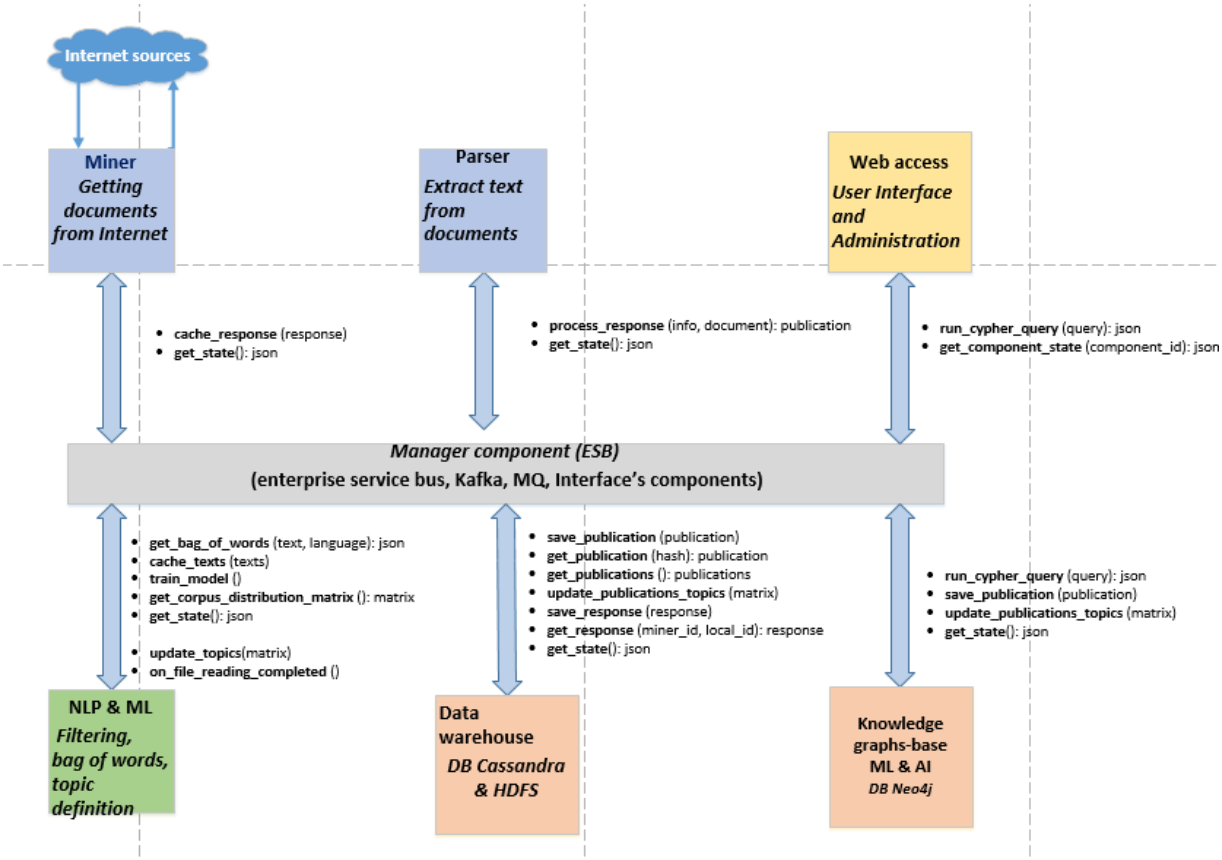
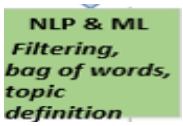


Рисунок *4*. – Новая архитектура СКАД ИИ

Компонент сбора данных – выполняет целевое сканирование выбранных социальных сетей, новостных порталов, сайтов и помещает полученные документы в исходном виде и специально обработанном виде в хранилище (в качестве хранилища используются HDFS и база данных документов Cassandra);

Компонент обработки данных выполняет фильтрацию исходных данных построение «мешка слов», извлечение из необработанных файлов публикаций текста, авторов, тематики, названия, литературу и др. Данные в специальном виде помещается в хранилище;

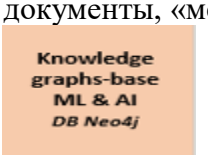


Компонент библиотека аналитических модулей содержит набор модулей, которые осуществляют обработку данных, полученных из интернет-источников, с целью определения тематик публикаций, и формирования аналитических данных для передачи клиентскому модулю, кроме того, содержит управляющие и служебные модули. Для тематического анализа текста применяется модифицированный алгоритм PLSA [5]. Тематический анализ с использованием EM-алгоритма позволяет выявить N самых важных тем во всём тексте, первоначально на основе

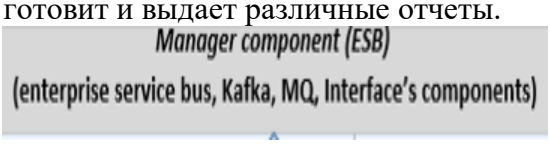
мешка слов, а в дальнейшем при пополнении корпуса документов, возможна корректировка классификатора документов. Здесь важно отметить, что для каждой темы определяется вероятность ее появления документе, что учитывается в графе знаний. Разработаны и апробированы модули построения и анализа графовых моделей, позволяющих получить знания для принятия решений [6, 7]. Область применения СКАД ИИ может быть расширена за счет пополнения компонента другими модулями, а применение комбинированного подхода к оценке некоторого явления позволит наиболее достоверно идентифицировать новые события и принять правильные решения;

Графический интерфейс обеспечивает взаимодействие с пользователем системы. Веб-интерфейс отправляет запросы к графу знаний через управляющий компонент, граф – знаний возвращает результат, который в виде различных отчетов выдается пользователю. Существуют различные алгоритмы определения наиболее важных статей, публикаций, блогеров [4, 7], в СКАД ИИ используется алгоритм page_rank [8].

Компонент хранилища данных – содержит данные из интернет-источников, предварительно обработанные и размеченные данные, необходимые для построения классификатора, «мешок слов», а также служебную информацию, необходимую для работы других модулей системы. В хранилище хранятся сырые данные с сайта, текст, фильтрованный текст, исходные документы, «мешок слов», тематика документов и служебная информация;



Компонент графовая база данных и граф знаний [9] состоят из графовой БД, моделирующей предметную область, и программных модулей, позволяющих пополнять графовую БД данными из хранилища и извлекать из нее знания о запрашиваемых объектах, графовая модель является развитием графовой модели приведенной в [6]. «Граф знаний», динамически связывает в графической форме: названия, авторов, публикаций, тематики публикаций, ссылки на публикации и авторов, готовит и выдает различные отчеты.



Управляющий компонент построен на базе Kafka, выполняет роль интеграционной шины, обеспечивает взаимодействие всех компонент, как по данным, так и по управлению. Является брокером СКАД ИИ и обеспечивает гарантированную доставку сообщений (MQ).

4. Графовые алгоритмы как база интеллектуального анализа информации

В настоящее время резко возрос интерес к применению графовых алгоритмов и аналитики для различных областей человеческой деятельности. Гибридная транзакционная и аналитическая обработка может потенциально переопределить способ выполнения некоторых бизнес-процессов, поскольку расширенная аналитика в реальном времени (например, планирование, прогнозирование и анализ «что, если») становится неотъемлемой частью самого процесса, а не отдельной выполняемой операцией после факта (Gartner research).

Здесь приведены некоторые примеры областей применения графовых алгоритмов, гибридной транзакционной и аналитической обработки данных:

– транзакции, маршрутизация, логистика, IoT, социальные сети, кибер-безопасность и аналитика;

– обнаружение мошенничества в реальном времени, графовая аналитика позволяет находить шаблоны построения связей и обнаружить мошенничество (например, в банковой сфере);

– выдавать рекомендации в реальном времени (например, в сфере услуг, соотносить данные о товаре, клиенте, предпочтениях, запасах, поставщике, логистике);

– мониторинг и управление различными физическими сложными сетями в реальном времени;

– идентификация и управление доступом клиентов, сотрудников к различным ресурсам в локальных и глобальных сетях в соответствии с их ролями;

– обеспечение работы различных приложений и функций социальные сетей, предсказание поведения социальных групп.

Графовые технологии – это основа для создания интеллектуальных приложений, позволяющая делать более точные прогнозы и быстрее принимать решения. Графы лежат в основе широкого спектра вариантов использования искусственного интеллекта (ИИ).

Так, граф знаний – одна из основных областей ИИ, который позволяет понимать предписывающую аналитику и приложения ИИ (например, обработка и понимание естественного языка (NLP, NLU), PageRank).

В общем случае, огромное количество графовых алгоритмов классифицированы на алгоритмы: Pathfinnding, Centrality и Community Detection (см. примеры [10]).

Pathfinnding. Класс алгоритмов поиска кратчайших путей, с учетом различных весовых критериев (например, расстояния или скорости), например, найти самый быстрый маршрут для поездки, минимизировать трафик телефонных звонков.

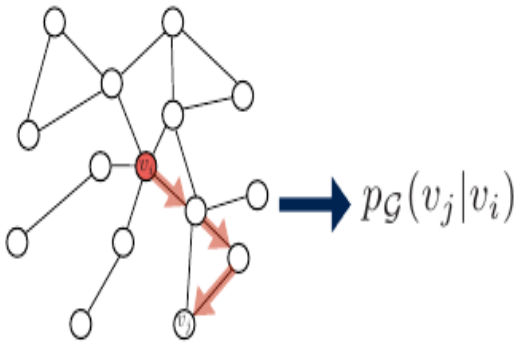


Рисунок 5. – Пример графа поиска путей

Centrality. Данный класс алгоритмов (центральности) заключается в понимании того, какие узлы наиболее важные в сети. Эти алгоритмы позволяют определить, как быстро можно распространять информацию в различных группах и между группами сущностей, предсказать появления новых тенденций в этих группах, выявлять уязвимости и возможные цели атаки в сетях связи и транспорта.

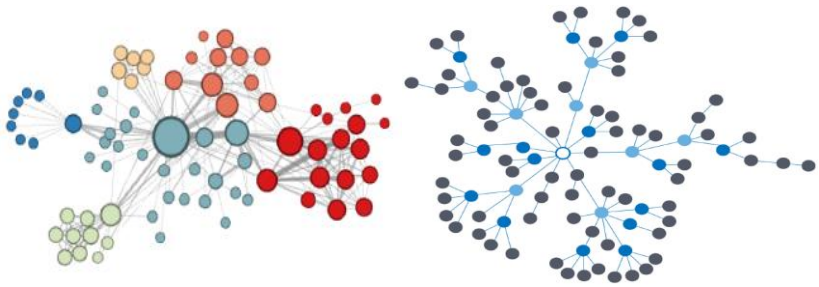


Рисунок 6. – Примеры графов групп сущностей и их связей

Community Detection (обнаружение сообщества). Класс алгоритмов, позволяющий изучать различные социальные сети, выявлять лидеров этих сетей, определять количественные характеристик различных групп. Оценивать иерархии, предсказывать тенденции поведения к видоизменению в этих группах, выявлять спамеров и потенциальных участников мошенничества.

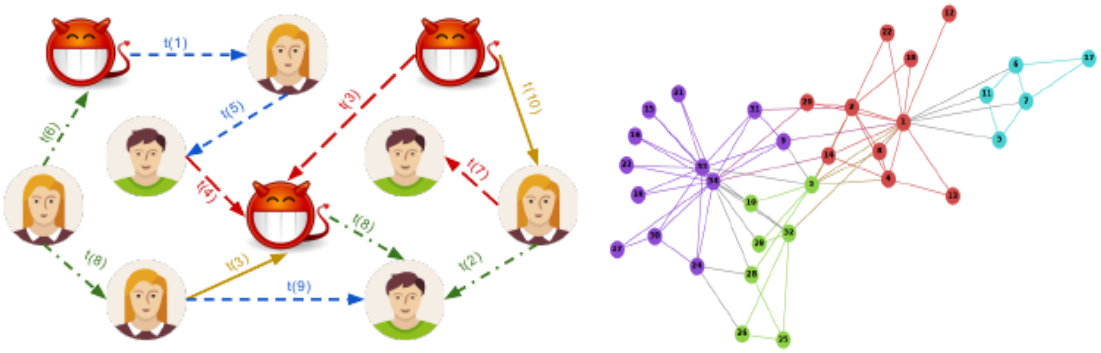


Рисунок 7. – Пример графов социальных сетей, выявления социальных групп, спамеров

Совместное использование информации графовых моделей и ML, позволяют получать скрытые зависимости и выполнять предиктивный анализ информации, получать ответы в режиме реального времени, реализовывать алгоритмы искусственного интеллекта, отслеживать решения ИИ. В настоящее время алгоритмы ИИ широко распространены для решения конкретной задачи, машина обучается выполнять какую-то задачу, например, автономное вождение автомобилей, управление различными дронами, автоматический поиск фото друзей на фотографиях и т.д. Для решения большинства из этих задач применяется графовая аналитика.

Графовая аналитика позволяет выявить закономерности в данных, например, в социальных данных. обнаружить сообщества или группу лиц, предсказать их поведение. Графическая визуализация помогает понять невидимые процессы ML алгоритмов, которые позволяют компьютерам учиться на примерах без явного программирования. Процесс глубокого обучения использует глубокие искусственные нейронные сети и ML в качестве моделей.

Так, на рисунке 8 и 9 приведены представления преобразования графовых данных для моделирования различных процессов в ML [11]. Для моделирования и предсказания из графовых моделей могут выполняться вырезки данных, а из отношений выбираться атрибуты, по которым нужно выполнять анализ данных.

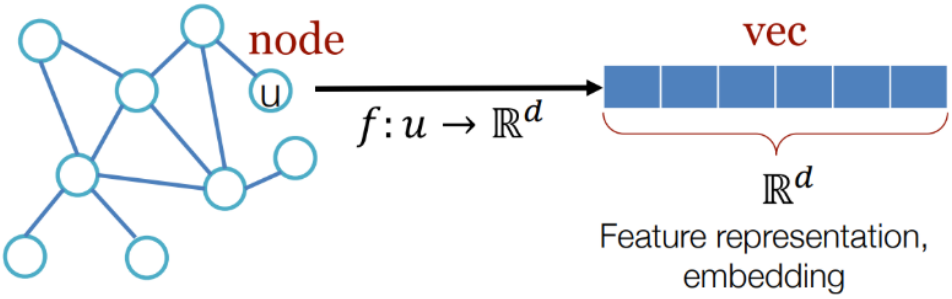


Рисунок *8.* – Схема преобразования атрибутов графовой модели в векторное представление

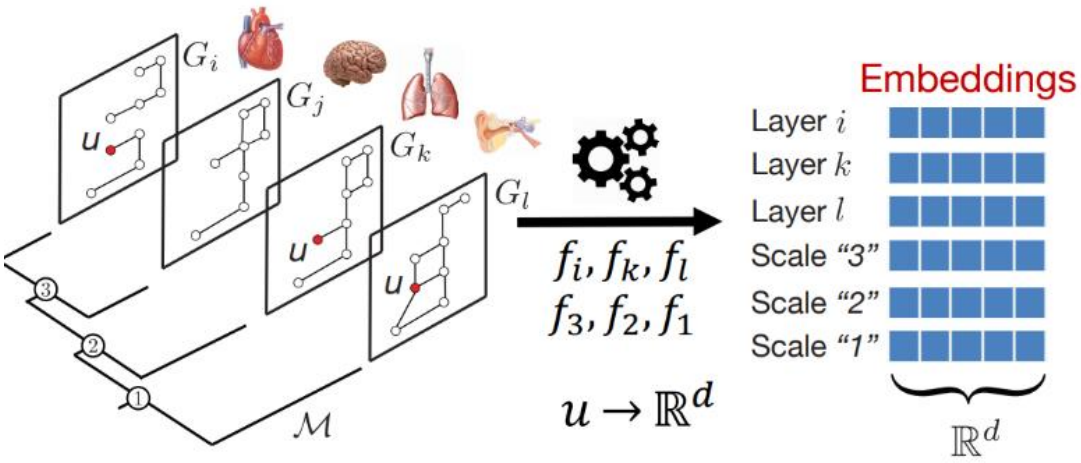


Рисунок *9*. – Многоуровневая схема преобразования атрибутов графовой модели в векторное представление

Искусственные нейронные сети и процесс глубокого обучения также используют графовые модели (см. рисунок 10).

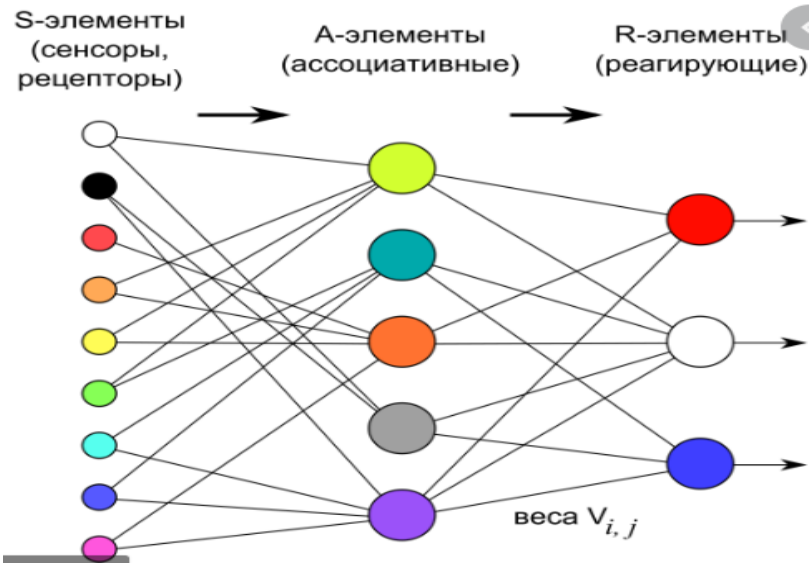


Рисунок *10*. – Пример графовой модели схемы искусственной нейронный сети

5. Примеры применения ИС Анализа Данных

В настоящее время (написание статьи) СКАД ИИ находится в опытной эксплуатации на ЦОД БГУИР на этой базе создается «ИС Анализа Данных».

В данном разделе приведены результаты работы СКАД ИИ с учетом важности публикаций и их тематик. Графовая модель позволяет получать знания о публикациях в различных аспектах, например, связанных с тематикой «биология». В таких запросах важно указывать порог вероятности тематики в статьях больше некоторой величины. Ниже приведены примеры получения информации из БД в виде графа знаний.

Данные получены с сайтов, который используется для публикаций научных работ: http://libgen.io, http://gen.lib.rus.ec/, https://arxiv.org

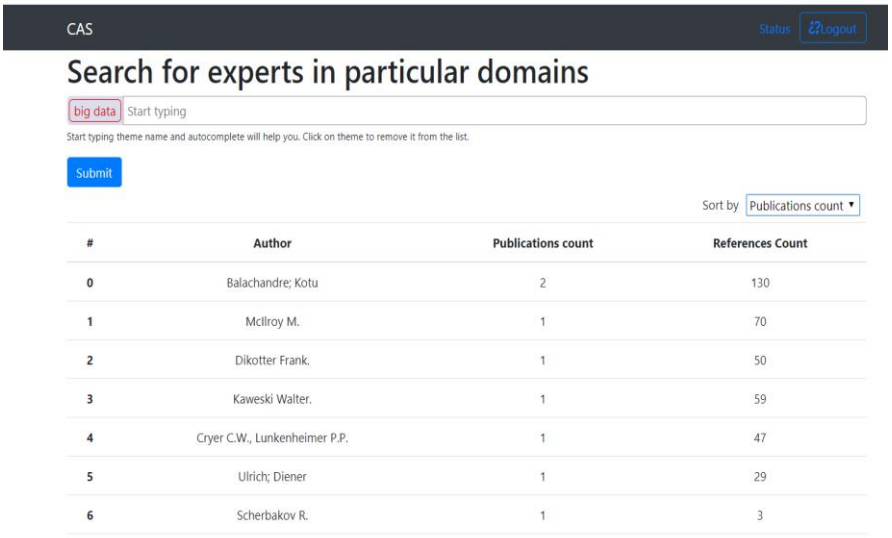


Рисунок *11*. – Эксперты в областях знаний (область – big data)

Показывается список экспертов в одной или нескольких областях знаний.

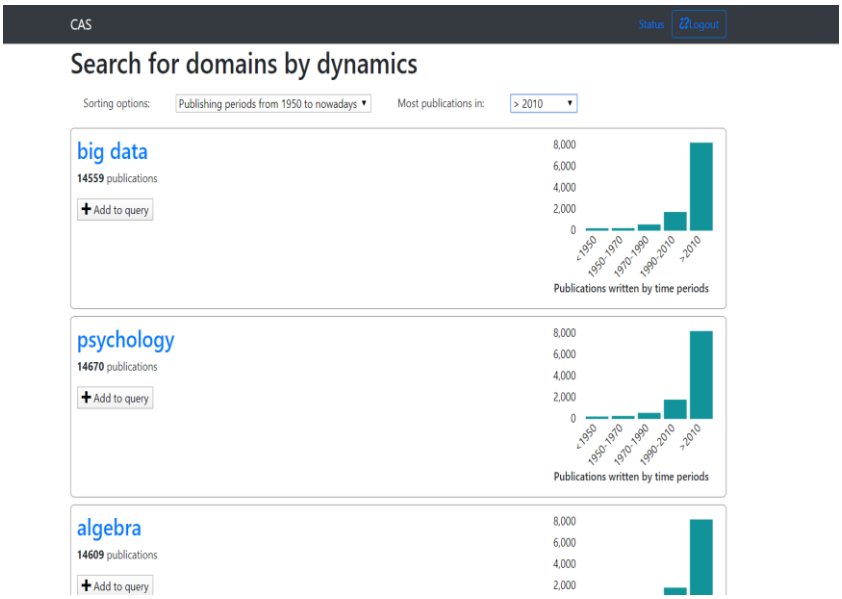


Рисунок *12*. – Гистограмма «областей знаний» с 1950 года

Приведены области знаний с количеством публикаций в каждой из них и столбчатой диаграммой распределения публикаций по периодам времени.

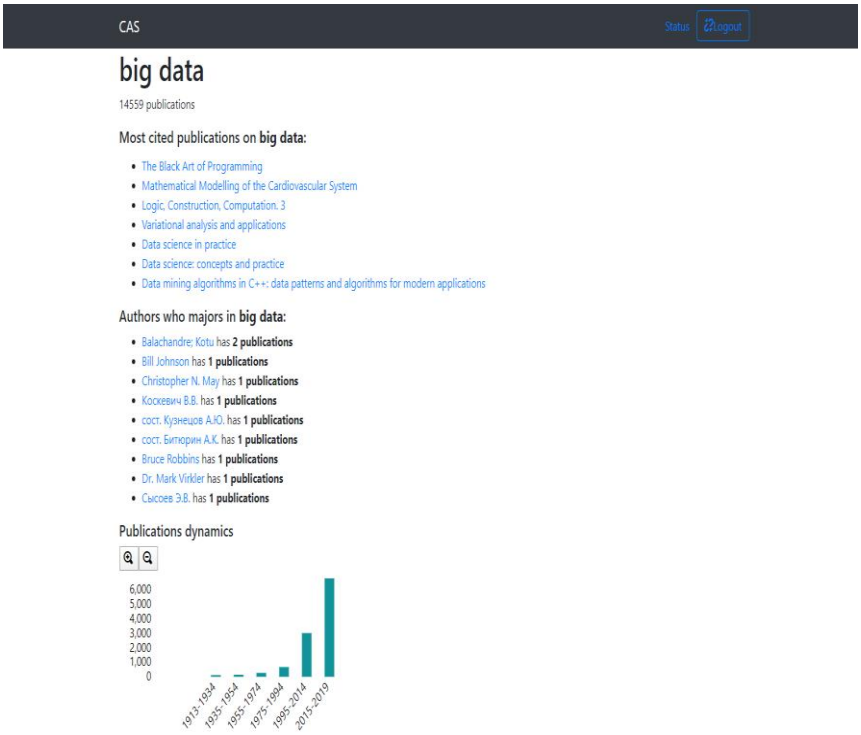


Рисунок *13*. – Подробная информация о области знаний

Страница со списком наиболее цитируемых работ, списком авторов с наибольшим числом публикаций в данной области, масштабируемой диаграммой распределения публикаций по периодам времени.

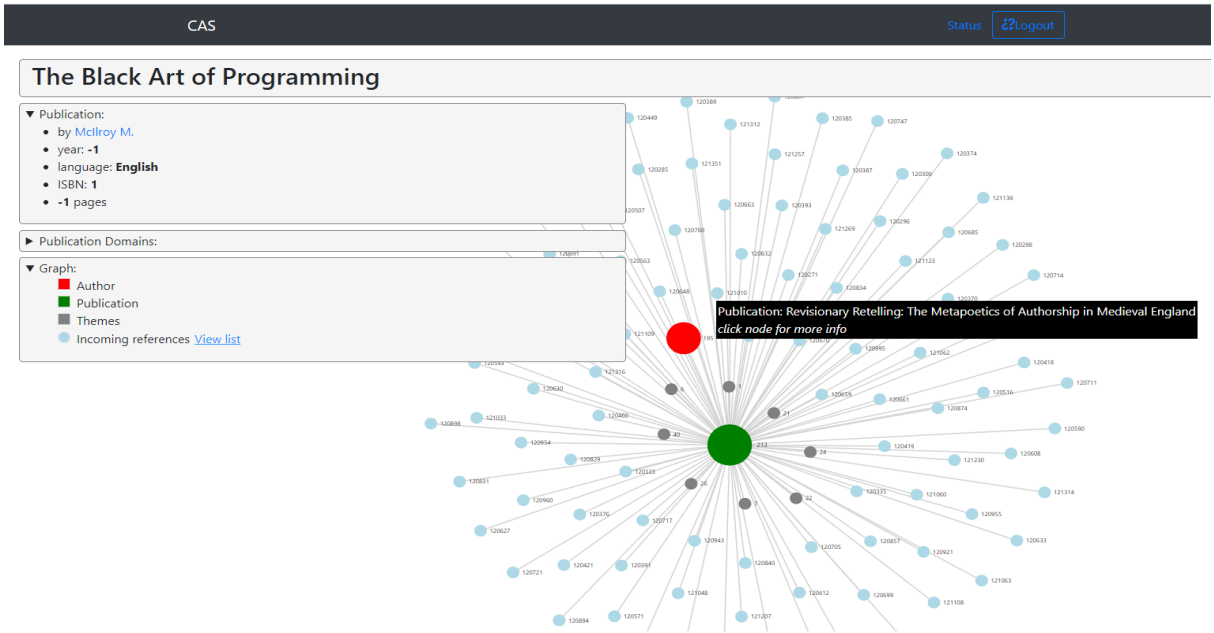


Рисунок *14*. – Подробная информация о конкретной публикации

Страница со списком тем, к которым относится публикация, со списком входящих (не более 50) и исходящих (не более 50) ссылок. Дополнительно – список ФИО и ID авторов. Из этой информации строится визуальный граф.

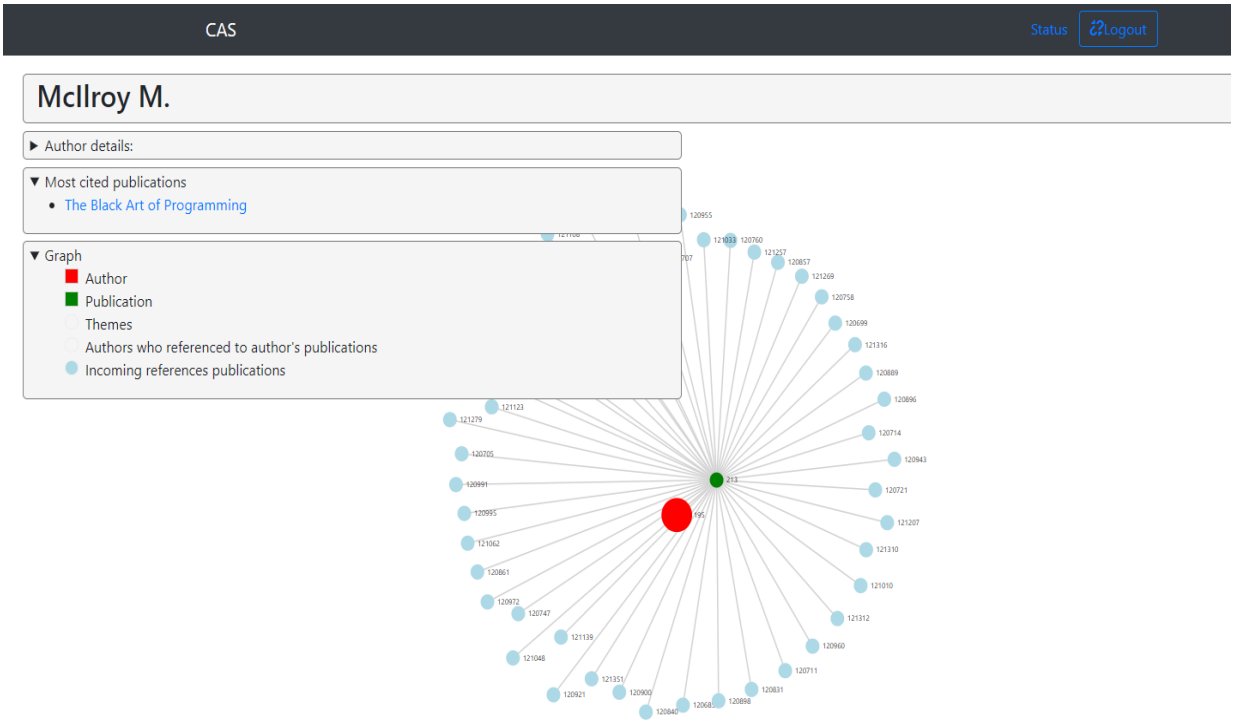


Рисунок *15.* – Подробная информация о авторе

Страница с информацией об авторе, со списком его наиболее цитируемых публикаций (не более 10) и исходящих (не более 50) ссылок для каждой из них. Для каждой публикации-«ссылки» дополнительно отправляется ее автор. Примечание. При большом количестве ссылок запросы 14 и 15 являются наиболее затратными по времени.

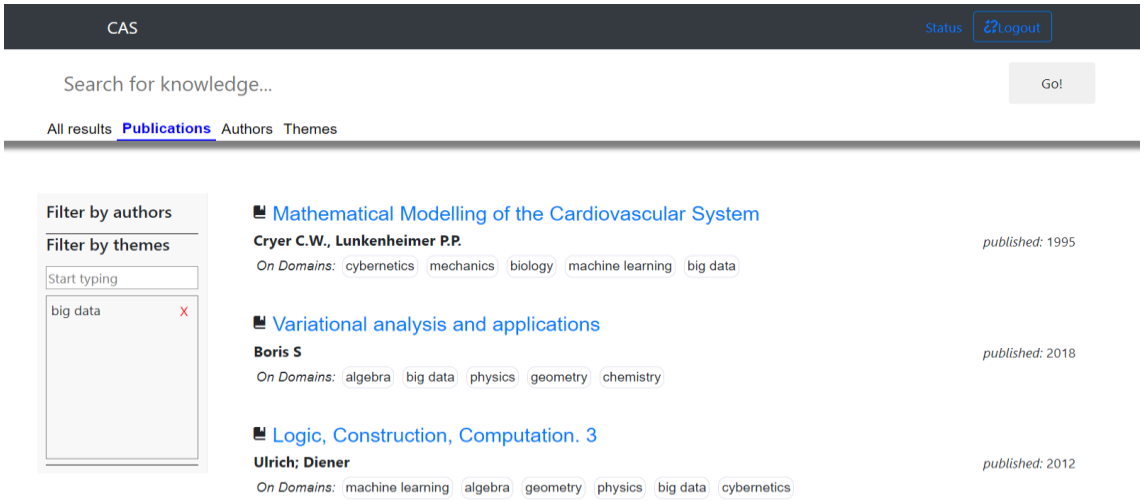


Рисунок *16.* –«Универсальный поиск» – поиск публикаций, связанных с темой «big data»

Страница со списком из 10 ответов-ссылок на подробности, удовлетворяющих запросу. К каждому из ответов, в зависимости от его типа (публикация/автор/тема) подгружаются дополнительные характеристики (количество публикаций, топ-5 тем, к которым относится публикация и т.д.). После получения первых 10 результатов возможен запрос на следующие 10, т.е. постраничная загрузка.

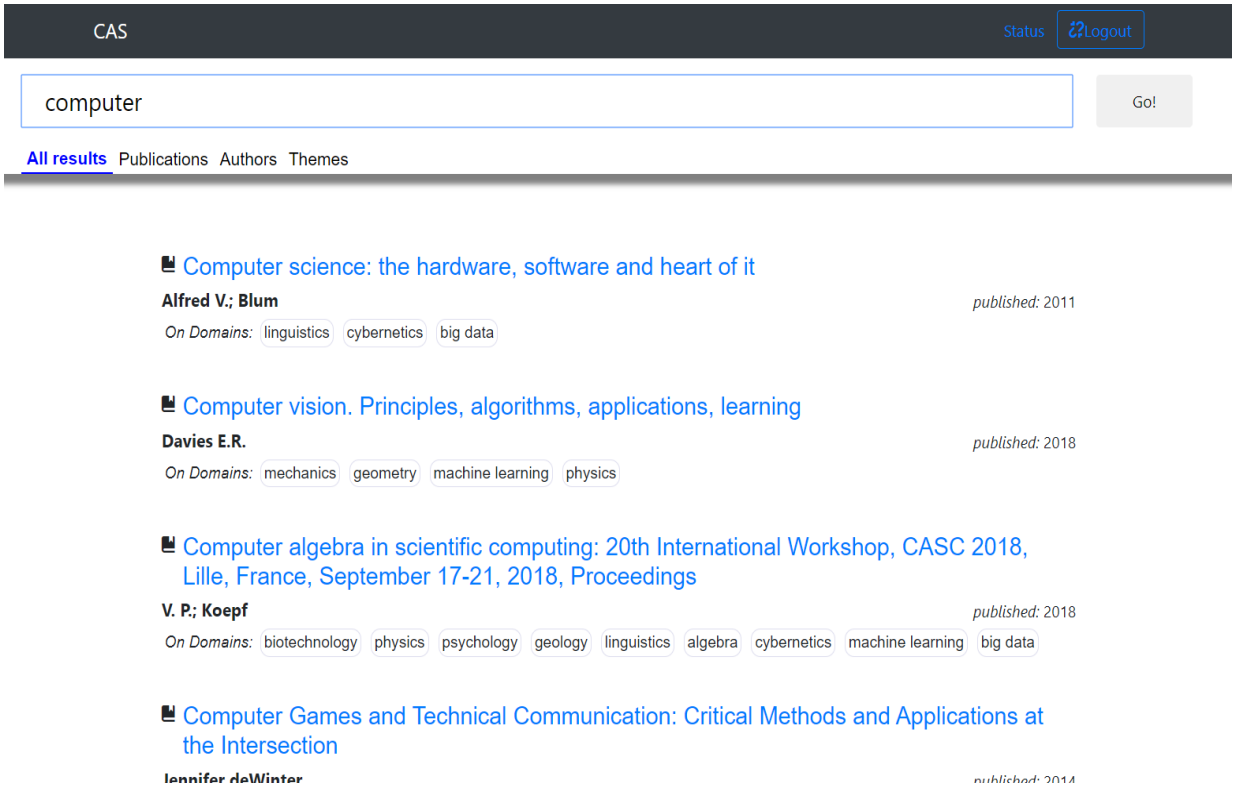


Рисунок 17. – Поиск по запросу «Computer»

Пример, получения более сложной интеллектуальной информации о данных из предметной области, на который скорее всего невозможно получить ответ из не графовой БД, приведен ниже.

Вопрос: *Какие авторы наиболее часто сотрудничали с автором данной популярной статьи, является ли данный автор писателем только в области биологии или же он пишет еще и на темы математического анализа, существуют ли математические публикации, которые по какой-то причине перекликаются с темой философии***.**

Запрос для известной публикации (ее ID=10) возвращает 10 авторов, которые имеют наибольшее число статей, ссылающихся на данную статью, выстроенных по убыванию числа ссылающихся статей. Плюс, возвращается сама публикация с ID=10 и ее автор (см. рисунок 18).

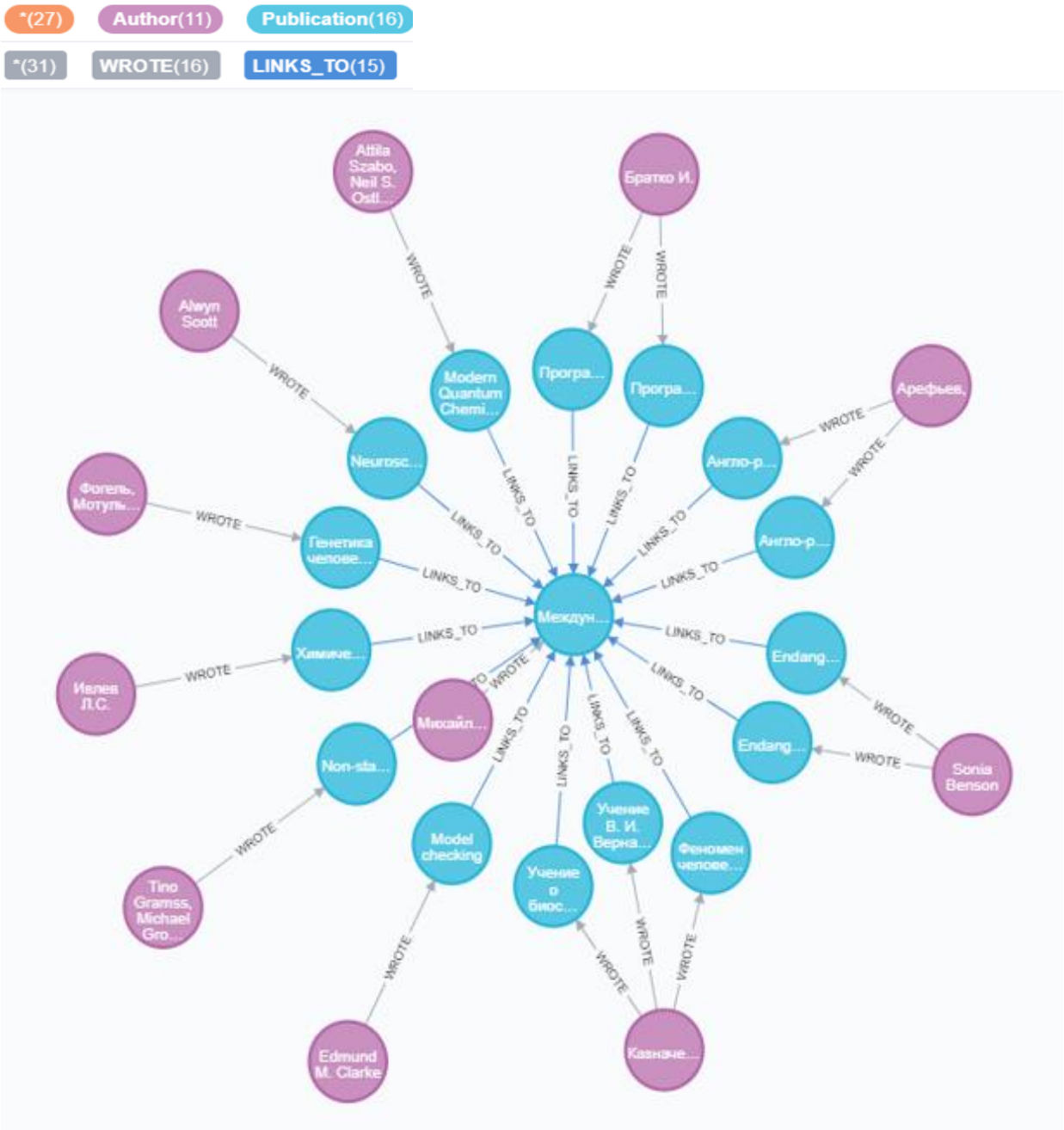


Рисунок *18*. – Какие авторы наиболее часто сотрудничали с автором данной популярной статьи

Запрос для известного автора (ID=8220) возвращает его публикации, которые не имеют связей с темой под названием “biology”. Из всех связей, обозначающих отношение тематики (THEME_RELATION) берутся лишь те, вероятность правильности отношения статьи к конкретной тематике у которых более 0.7 (см. рисунок 19).

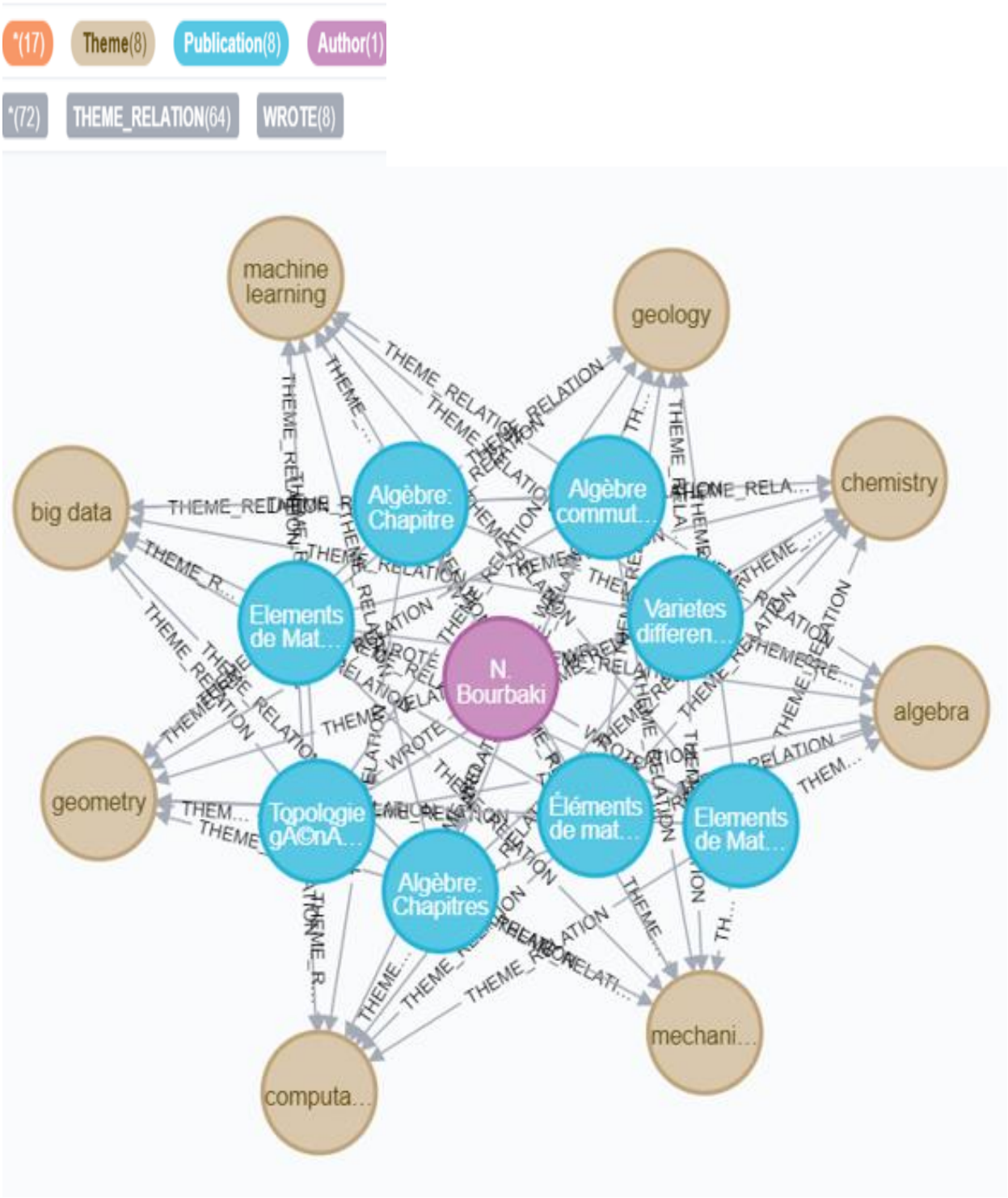


Рисунок *19*. – Является ли данный автор писателем только в области биологии или же он пишет еще и на темы математического анализа

Запрос выбирает публикации, относящиеся к теме “algebra” с вероятностью, большей 0.7, имеющие также отношение к другим тематикам (с вероятностью большей 0.7). И, для построения графа, запрос возвращает тему “algebra” и, для каждой публикации, список других тем, к которым она относится (см. рисунок 20).

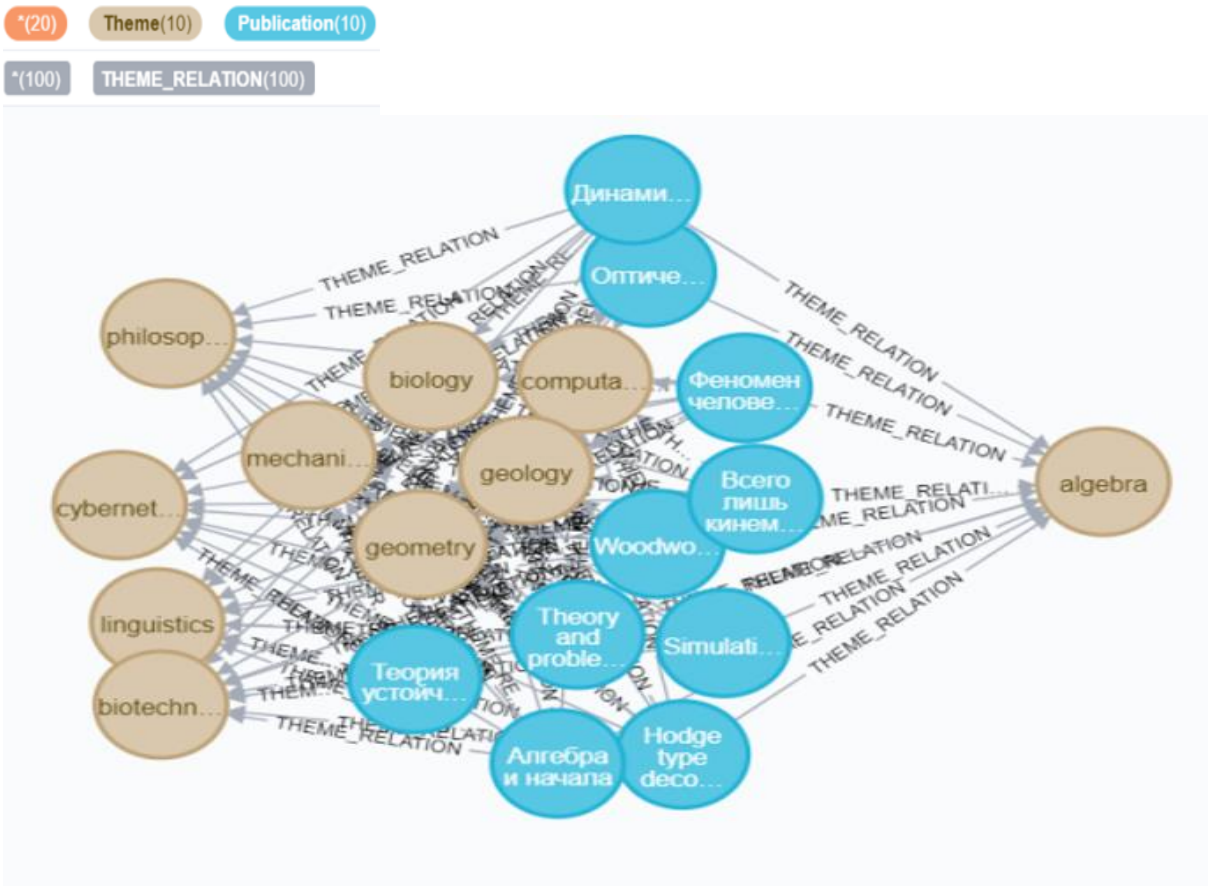


Рисунок *20.* – Существуют ли математические публикации, которые по какой-то причине перекликаются с темой философии

6. Заключение

«Интеллектуальная система комплексного анализа данных интернет-источников» – это инновационный научно-образовательный проект БГУИР. Результаты выполнения проекта используются при обучении магистрантов по тематике «Обработка больших объемов информации», подготовке специалистов «Data Scientist», а также для получения экспертных данных при проведении исследовательских работ в университете. В настоящее время над проектом работает коллектив студентов, которые уже получили навыки работы в команде над большим проектом и знания по извлечению данных из интернет-источников, их обработке и анализу с помощью NLP и ML алгоритмов.

Основными результатами СКАД ИИ являются:

– интеллектуальное взаимодействие графовой базы данных с ML, расширение системы алгоритмами ИИ;

– разработка ESB СКАД ИИ, которая обеспечивает интеграцию данных и приложений (серверов). Созданная архитектура позволяет модернизировать и функционально наращивать СКАД ИИ в процессе эксплуатации. Данная архитектура построения многофункциональных комплексов позволяет работать СКАД ИИ как набор постоянно работающих компонент в виде отдельных серверов;

– компоненты и комплексы могут быть добавлены или заменены новыми;

– сами компоненты можно нарастить функционально, дополнить компоненты «хранилище», граф-знаний, «библиотека модулей» векторными алгоритмами и алгоритмами нейровычислений.

Компонент графовая база данных и граф знаний может быть дополнен графовой моделью для поиска постов в социальных сетях на определённую тематику, и другими графовыми моделями, которые могут быть максимально полезны и интересны пользователю, методами интеллектуального анализа графовой модели на основе применения ML.

Компонент «хранилище» данных может быть дополнен новыми БД, соответствующими новым областям применения системы.

Библиотеки аналитических модулей новыми ML-модулями интеллектуального анализа данных

В момент подготовки данной статьи в реализации проекта СКАД ИИ и его модернизации принимали участие студенты БГУИР кафедры информатики: Гутковский В.Н., Сернацкий В. И., Судникович К. И., Черныш Н. Н., Шпаков Н. И.

Список литературы

[1.]Data Never Sleeps 7.0 [Электронный ресурс] / Режим доступа: https://www.domo.com/learn/data-never-sleeps-7 Дата доступа: 14.02.2020.

[2.]Пилецкий, И. И. Аналитический комплекс анализа данных из открытых интернет источников / И. И. Пилецкий, В. А. Прытков, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 – Minsk, BSUIR, 2018. – P. 193 – 199.

[3.]Батура М.П., Пилецкий И.И., Прытков В.А., Волорова Н.А., Козуб В.Н. Система комплексного анализа данных интернет источников // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов V Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13–14 марта 2019 года). В 2 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. – Минск : БГУИР, 2019. – 379 с. ISBN 978-985-543-484-0 (ч. 2)..p/172-187

[4.]Романов А.А., Пилецкий И. И. Классификация тональности текстовых документов с помощью метода опорных векторов. Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2017 -06 мая 2017.

[5.]Чугаинов К. В., Пилецкий И. И. Методы тематической кластеризации новостных статей. Научно-практические исследования №2 (ISSN 2541-9528) – Омск: Дельта, – 2017 с. 295 – 298.

[6.]Прытков В.А. Нормализация словоформ при анализе репозитория университета с использованием графовой базы данных / В.А. Прытков, И.И. Пилецкий, Н.А. Волорова // Сб. мат. V межд. Науч.-практ. конф. “Big Data and Advanced Analytics (Big Data и анализ высокого уровня)”. - Минск, 2019. - с.209-220

[7.]В. Н. Козуб, И. И. Пилецкий. Использование алгоритмов обработки естественного языка и графовых баз данных для построения рекомендательной системы / // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 / editorial board: M. Batura [etc.]. – Minsk, BSUIR, 2018. – P. 274 – 277.

[8.]Pagepank algo neo4j. [Электронный ресурс] – Режим доступа: https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/ Дата доступа: 22.01.2019

[9.]Шпаков Н.Н., Черныш Н.Н., Пилецкий И.И. ГРАФ ЗНАНИЙ КАК СРЕДСТОВО АНАЛИЗА В СИСТЕМЕ КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ ИНТЕРНЕТ ИСТОЧНИКОВ//*«GLOBAL SCIENCE AND INNOVATIONS 2019: CENTRAL ASIA»* атты VI Халықар. ғыл.-тәж. конф. материалдары (X ТОМ)/ Қҧраст.: Е. Ешім, Е. Абиев т.б.– Нур-Султан, May 9-13, 2019 – 353 б. ISBN 978-601-341-186-6. С.120-123

[10.] Representation Learning on Graphs: Methods and Applications. [Электронный ресурс] / Режим доступа: https://www.semanticscholar.org/paper/Representation-Learning-on-Graphs%3A-Methods-and-Hamilton-ing/ecf6c42d84351f34e1625a6a2e4cc6526da45c74 Дата доступа: 24.02.2020

[11.] Где и как врубиться в эмбеддинги графов. [Электронный ресурс] / Режим доступа: https://habr.com/ru/company/ods/blog/418727/ Дата доступа: 24.02.2020

SYSTEM FOR COMPLEX ANALYSIS OF DATA FROM INTERNET SOURCES

M. P. BATURA
Head of the Research Laboratory 8.1 "New Learning Technologies" BSUIR, Doctor of Technical sciences,

I.I. PILETSKI
PhD
Associate Professor of Informatics Department of the BSUIR

V.A. PRYTKOV
PhD
Vice-rector for education BSUIR, Associate Professor

N.A. VOLARAVA
PhD
Head of the Informatics Department of the BSUIR, Associate Professor

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: bmpbel@bsuir.by, *ianmenski@gmail.com*, *prytkov@bsuir.by, volorova@bsuir.by*

Abstract. The article describes the monitoring tools for open Internet sources in order to identify experts in a certain subject area, determine the topics of publications, and assess the popularity of publications. Describes the decisions taken in the development of different variants of the construction of analytical complex. The obtained results of the complex and the application of artificial intelligence methods for conducting in-depth data analysis are presented.

Keywords: Internet sources, Big Data, monitoring, analysis, Machine Learning, Natural Language Processing, Neo4j, Hbase, Cassandra.