http://dx.doi.org/10.35596/1729-7648-2023-21-4-101-109

Оригинальная статья
Original paper

УДК 519.684.6; 004.021

МЕТОДОЛОГИЯ ПОСТРОЕНИЯ ПРОТОТИПА СИСТЕМЫ КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ ТЕМАТИЧЕСКИХ САЙТОВ

И. И. ПИЛЕЦКИЙ, М. П. БАТУРА, Н. А. ВОЛОРОВА

Белорусский государственный университет информатики и радиоэлектроники
(г. Минск, Республика Беларусь)

Поступила в редакцию 10.04.2023

Аннотация. Одно из современных направлений получения информации для принятия обоснованных решений – анализ данных из открытых интернет-источников и СМИ, содержащих множество публикаций. Критически важно не только получение достоверной информации, но и время для ее анализа. Разработана и апробирована комплексная методология быстрого построения прототипа системы комплексного анализа тематических сайтов. Создана технология взаимосвязанных методов, методологий и инструментов по построению графовой базы данных, графа знаний, анализа данных с использованием методов и моделей машинного обучения с предоставлением аналитических результатов пользователям. Разработанные технологии могут применяться для анализа данных известных мировых сайтов с целью построения прототипа системы комплексного анализа информации интернет-источников.

Ключевые слова: система комплексного анализа, RDF-схема, RDF-словари и онтологии, машинное обучение, графовые базы данных, графовые алгоритмы, PageRank.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Благодарность. Авторы выражают благодарность студентам Белорусского государственного университета информатики и радиоэлектроники П. А. Зорко и О. А. Кулевич за апробацию технологических решений.

Для цитирования. Пилецкий, И. И. Методология построения прототипа системы комплексного анализа данных тематических сайтов / И. И. Пилецкий, М. П. Батура, Н. А. Волорова // Доклады БГУИР. 2023. Т. 21, № 4. С. 101–109. http://dx.doi.org/10.35596/1729-7648-2023-21-4-101-109.

METHODOLOGY FOR BUILDING A PROTOTYPE SYSTEM FOR COMPLEX DATA ANALYSIS OF THEMATIC SITES

IVAN I. PILETSKI, MIKHAIL P. BATURA, NATALIA A. VOLARAVA

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Submitted 10.04.2023

Abstract. One of the modern directions of obtaining information for making informed decisions is the analysis of data from open Internet sources, the analysis of media containing hundreds of thousands of publications. It is critically important not only to obtain reliable information, but also the time needed to obtain and analyze it. The purpose of the research in this work is the development and testing of a complex methodology for quickly building a prototype of a system for complex analysis of thematic sites. A technology of interconnected methods, methodologies, and tools for building a graph database, a knowledge graph, data analysis using methods and models of machine learning with the provision of analytical results to users has been created. The main task of this work is to use these technologies to analyze data from well-known world sites in order to build a prototype of a systems for complex analysis of data from Internet sources.

Keywords: complex analysis system, RDF schema, RDF dictionaries and ontologies, machine learning, graph databases, graph algorithms, PageRank.

Conflict of interests. The authors declare no conflict of interests.

Gratitude. The authors are grateful to the students of the Belarusian State University of Informatics and Radioelectronics P. A. Zorko and O. A. Kulevich for approbation of technological solutions.

For citation. Piletski I. I., Batura M. P., Volarava N. A. (2023) Methodology for Building a Prototype System for Complex Data Analysis of Thematic Sites. *Doklady BGUIR*. 21 (4), 101–109. http://dx.doi.org/10.35596/1729-7648-2023-21-4-101-109 (in Russian).

Введение

Общепринятое представление данных в форме HTML не позволяет отразить семантику сайта и значения данных. Однако во многих ситуациях важно иметь многомерные данные, такие как статистика, различные документы и публикации, чтобы их можно было связать с соответствующими наборами данных и концепциями для получения знаний о конкретной предметной области. Это аналогично, как получать знания, используя информацию и ее интерпретацию из конкретной базы данных (БД).

Одно из сложных современных направлений – предоставление знаний с помощью специальных глобальных словарей предметных областей, классификаторов, мета-описаний, специальных языков и методологий их применения. Данная методология используется для создания и описания содержимого известных сайтов: Wikipedia, DBpedia, TerMef and French Bioloinc Portal, TerMef, BIOLOINC, WikiData, Scientific Research Publishing, IEEE Xplore, SpringerOpen, научной электронной библиотеки «КиберЛенинка», сайта с научными публикациями Semantic Scholar, крупных организаций и др. Многие сайты применяют специальную абстрактную модель описания ресурса Resource Description Framework (RDF – среда описания ресурса, имеет вид тройки «субъект – предикат (или свойство) – объект (значение свойства)»[1]. Такой подход позволяет выполнять описание знаний в тематических предметных словарях и обмениваться этими знаниями с другими сайтами. RDF-ресурс может быть представлен любой сущностью – информационной или неинформационной. Например, это может быть описание некоторого сайта или реального объекта.

К наиболее распространенным RDF-словарям относятся: FOAF (Friend-of-a-Friend) – словарь для описания людей, их деятельности, отношения с другими людьми и объектами; SKOS (Simple Knowledge Organization System) – словарь для представления таксономий и слабоструктурированных знаний; DC (Dublin Core) – определяет общие атрибуты метаданных; BIBO (Bibliographic Ontology) – используется как онтология цитирования, классификации документов; SIOC (Semantically-Interlinked Online Communities) – семантическая технология соединенных онлайн-сообществ; DOAP (Description of a Project); Music Ontology. RDF-словари и онтологии OWL (Web Ontology Language – язык представления веб-онтологий) применяют абстрактную модель RDF- и RDFS-описания (RDFS – RDF Schema) ресурса. Онтология – это конкретное формальное представление того, что означают термины в той области, в которой они используются. RDF-информация может быть представлена в нескольких форматах – Turtle, N-Triples, JSON-LD, RDF/XML, TriG и N-Quads, TriG*[2]. Данные с таких веб-сайтов можно использовать для быстрого построения прототипа графовой БД для дальнейшего более глубокого анализа данных сайта, что и применено в статье.

Для веб-сайтов, которые хранят данные в JSON-LD, в коде страницы можно найти скрипт JSON-LD. Такой сайт можно сериализовать в графовую базу данных на основе информации, хранящейся в формате JSON. Кроме RDF и JSON можно сериализовать и другие типы данных. Так, данные на общедоступном популярном сайте Wikipedia хранятся в формате Turtle и тоже могут быть сериализованы в графовую БД. Для сайта Wikidata информация о данных хранится на самом сайте.

Именами сущностей RDF являются имена классов, предикатов, индивидуумов, представленные URI (Uniform Resource Identifier) – универсальным идентификатором ресурса, который идентифицирует логический или физический ресурсы. Обычно они выражаются использованием компактной записи, и префикс определяет URI пространства имен. RDF представляет собой абст-

[1] The RDF Data Cube Vocabulary (w3.org).

[2] https://www.w3.org/2013/dwbp/wiki/RDF_AND_JSON-LD_UseCases.

рактную модель, обеспечивающую способ разбиения знаний на дискретные части, и позволяет обмениваться информацией. Для описания групп взаимосвязанных ресурсов и отношений между ресурсами применяется RDFS[3], для отношений между людьми можно использовать schema.org. Расширение RDFS фиксирует понятие встроенной тройки, на которую ссылаются с помощью строк <<….. >>, а семантика ребра следует за ними в терминах онтологии схемы (xsd:date). Например, <<:boy :hasGirlfriend :girl>> :startDate «2014-04-14»^^xsd:date. Данная нотация прямо связана с понятием «граф знаний», что позволяет генерировать графовые базы данных, знаний и строить по запросам графы знаний. Для доступа к данным таких сайтов (в формате Turtle) можно использовать специально разработанный язык SPARQL Protocol and RDF Query Language[4]. Но гораздо важнее построить графовую БД с целью комплексного и глубокого анализа данных.

Графовая база данных и IT-среда для анализа данных

Существуют различные методы формализации знаний. В статье рассматриваются сайты, которые используют описание ресурса с помощью семантических троек RDF. Данные методология и технология описания сайтов позволяют выполнить генерацию (построение) графовой БД из троек RDF. Такая тематическая графовая БД содержит базу знаний сайта в виде графа знаний, что позволяет применять различные аналитические алгоритмы машинного обучения (ML) для более глубокого анализа данных сайта [1–3].

IT-среда и технология ее применения позволяют построить графовую БД тематического сайта. В качестве основных компонент IT-среды для построения графовой БД веб-сайта используются графовая система управления базами данных Neo4j Desktop[5] и ее расширения. Это специальные плагины Neosemantics (n10s), APOC (Awesome Procedures on Cypher) и GDS (Graph Data Science Library), а также вспомогательные функции, написанные на языках Python и Cypher.

Узлы в графе RDF могут быть либо ресурсами, представленными уникальным идентификатором ресурса URI (например, общеизвестные URL-адреса), литералами (например, такими же, как в XML), либо вспомогательными пустыми узлами. Типы ребер называются предикатами. При построении графовой БД выполняется преобразование троек RDF («S-субъект – P-предикат – O-объект») в графовую БД с узлами, отношениями и свойствами узлов и отношений. Рассмотрим примеры демонстрации принятых технологических решений.

Пример 1. Построение графовой базы данных из данных сайта[6] (данные хранятся в RDF)

При выполнении специального запроса получим граф, изображенный на рис. 1. Граф содержит четыре узла трех различных типов (артист, альбом, песня) и три связи двух видов (композитор, вхождение песни в альбом (в свойствах связи номер дорожки песни в альбоме)). У каждого из узлов и отношений свои уникальные id и uri.

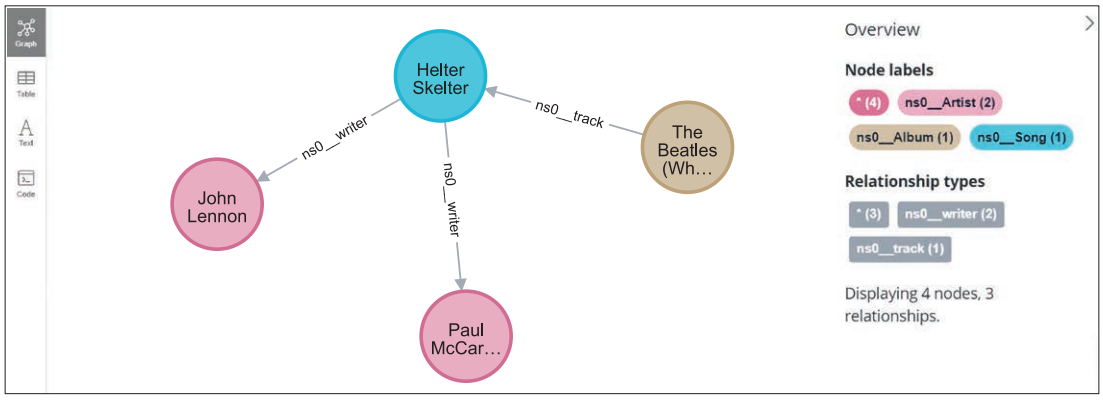

Рис. 1. Представление сайта, полученное по ссылке[6]
Fig. 1. Representation of the site obtained from the link[6]

Предметная область системы комплексного анализа искусственного интеллекта (СКА ИИ) включает в себя следующие сущности и их свойства:

[3] https://www.w3.org/2000/01/rdf-schema, schema.org (https://schema.org/docs/gs.html).

[4] https://www.w3.org/TR/rdf-sparql-protocol/.

[5] https://neo4j.com/product/neo4j-graph-database/.

[6] https://github.com/jbarrasa/datasets/blob/master/rdfstar/beatles-hs.ttl.

• User (sch_Users): name – имя пользователя системы СКА ИИ и его характеристики;

• Article (sch_Article): title – название статьи, year – год выпуска статьи, venue – награды, n_citation – цитирование статьи, abstract – описание статьи, theme – тематика статьи, references – ссылки, pagerank – рейтинг статьи по pagerank;

• Author (sch_Person): name – имя и фамилия автора, articles – статьи, которые принадлежат автору, popularity – количество статей этого автора, pagerank – рейтинг автора по pagerank.

Структура графовой БД, которая отражает предметную область разрабатываемой системы, приведена на рис. 2.

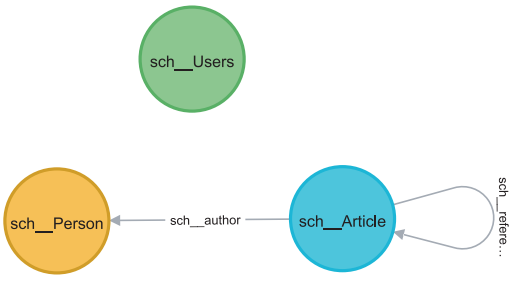

Рис. 2. Структура графовой базы данных для анализа публикаций
Fig. 2. Structure of graph database for publications analysis

Начальная структура БД имеет следующие свойства: sch_author – связывает статью и автора, который написал ее, sch_references – связывает первую и последующую статью, на которую ссылается первая статья. Для применения технологических решений используются данные из предметной области широко известных сайтов: IEEE Xplore, научной электронной библиотеки «КиберЛенинка», Semantic Scholar и SpringerOpen.

Пример 2. Построение графа свойств на основе данных, встроенных в веб-страницы (JSON-LD)

Рассмотрим сайт со статьями на научные темы Semantic Scholar и технологические решения для заполнения данными графовой базы данных СКА ИИ, структура которой приведена на рис. 2. Как правило, данные, нужные для сериализации и построения графовой БД, находятся в коде страницы сайта в script-элементе с типом application/ld+json. Для извлечения информации с веб-страницы используется процедура библиотеки APOC – apoc.load.html, а для визуализации статей сайта по тематике Big Data и анализа – плагин Neosemantics и специальный запрос.

На рис. 3 приведен скриншот общего представления обновленной графовой базы данных СКА ИИ, полученной в результате загрузки данных из указанного сайта. Типы узлов и связей получены с этого сайта с помощью технологии быстрого построения тематической графовой БД. После загрузки данных образовались узлы и связи со свойствами, JSON-LD использует schema.org[3]:

• узлы: Organization – организация; Article – статья или публикация; ScholarlyArticle – научная статья или публикация; Person – человек, в случае со статьями – автор или издатель статьи; ListItem – является элементом какого-то списка; BreadcrumbList – это ListItem, состоящий из цепочки связанных веб-страниц; ImageObject – изображение чего-то (графический файл); _GraphConfig – содержит в себе все настройки и конфигурацию плагина Neosemantics; Resource – все остальные узлы, которые не попали под какую-либо из категорий;

• связи: sameAs – идентичные элементы; image – URL-адрес или полностью описанный ImageObject; author – автор публикации или статьи; publisher – издатель; mainEntityOfPage – указывает на страницу с основным описываемым объектом; item – данные какой-то сущности; itemListElement – какая-то сущность списка; url – URL-адрес элемента; contentUrl – URL-адрес на медиаобъект.

Созданная IT-среда Neo4j Desktop и технологические решения позволяют анализировать, какие веб-страницы и элементы связаны друг с другом и какой связью, а также объединять несколько похожих графовых баз данных (к примеру, различные сайты со статьями) в одну графовую БД. Так, в графовую базу данных СКА ИИ к данным Semantic Scholar были добавлены данные с сайта SpringerOpen. Полученную тематическую графовую БД этого сайта (или сайтов) будем использовать для дальнейшего анализа данных с применением графа знаний и алгоритмов ML. Узлы, ребра графовой БД и граф знаний с помощью технологии embedding (вложений) преобразовываются в векторное представление некоторого пространства для применения алгоритмов ML.

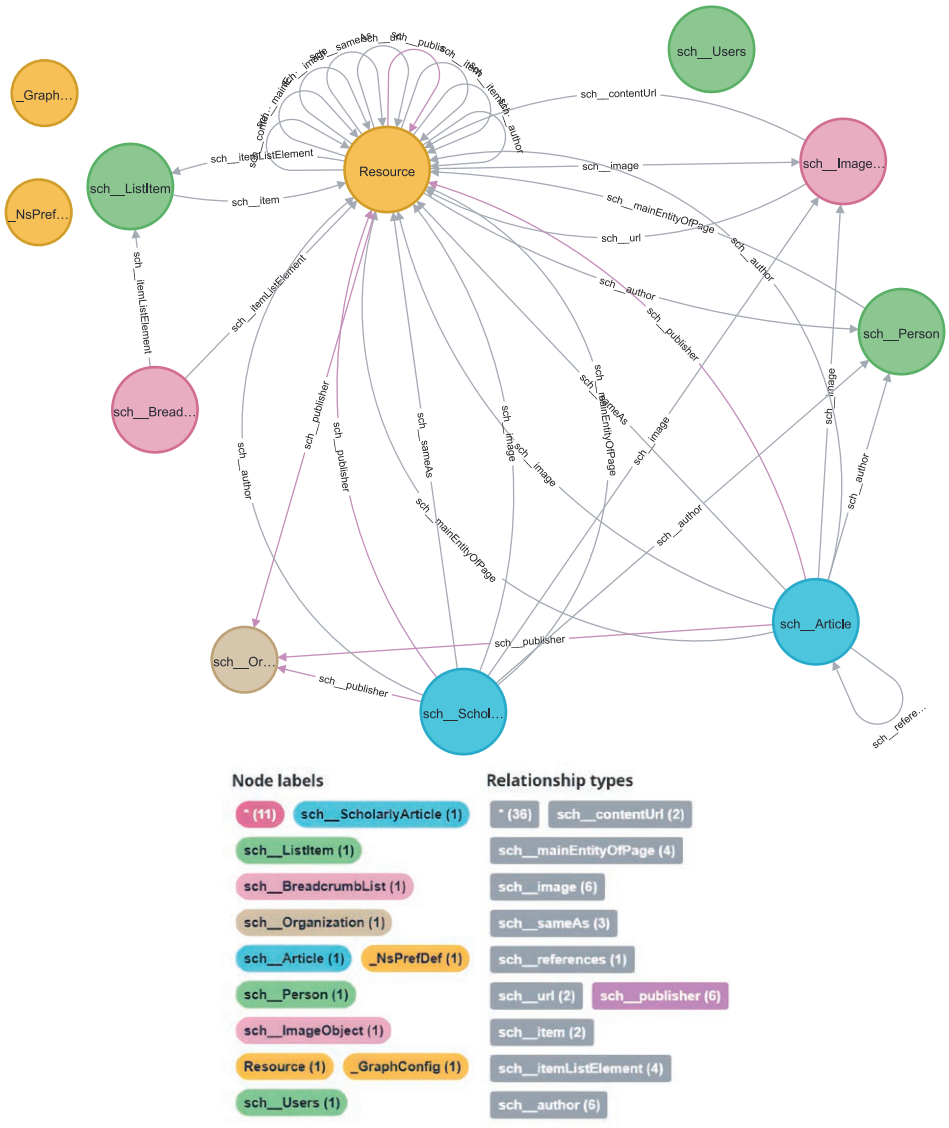

Рис. 3. Скриншот общего представления обновленной графовой базы данных
Fig. 3. Screenshot of general representation of the updated graph database

Граф знаний

Граф знаний (KG – Knowledge Graph) – ориентированный граф, узлы которого – это сущности и литеральные значения (литералы), а ребра – отношения между этими сущностями [4]. KG – естественная модель данных во многих реальных ситуациях. KG фиксирует все полезные отношения и с помощью вложения (embedding) графа объединяет огромное количество знаний (области) в векторное представление более низкого измерения. Вложения графов – это проекции узлов и ребер в непрерывное низкоразмерное пространство.

Известные примеры баз и графов знаний – Google Knowledge Graph, DBPedia (огромный набор данных из 228 млн вещей и онтологии – люди, места, фильмы, книги, организации, виды, болезни и т. д.), Geonames (содержит 12 млн географических объектов), Wordnet (лексическая база данных английского языка, содержащая определения и синонимы), FactForge (открытые данные и новостные статьи о людях, об организациях и о местах), Wikidata (междометный граф знаний, содержит описание множества фактов с богатым контекстом и ссылками).

Примеры графов знаний, где используются графовая база знаний СКА ИИ, и специальные запросы к БД приведены ниже.

Пример 3. Найти все статьи, которые написал Хонг Мэй

Полученый граф знаний, представленный на рис. 4, позволяет определить тематику публикаций ученого с мировым именем.

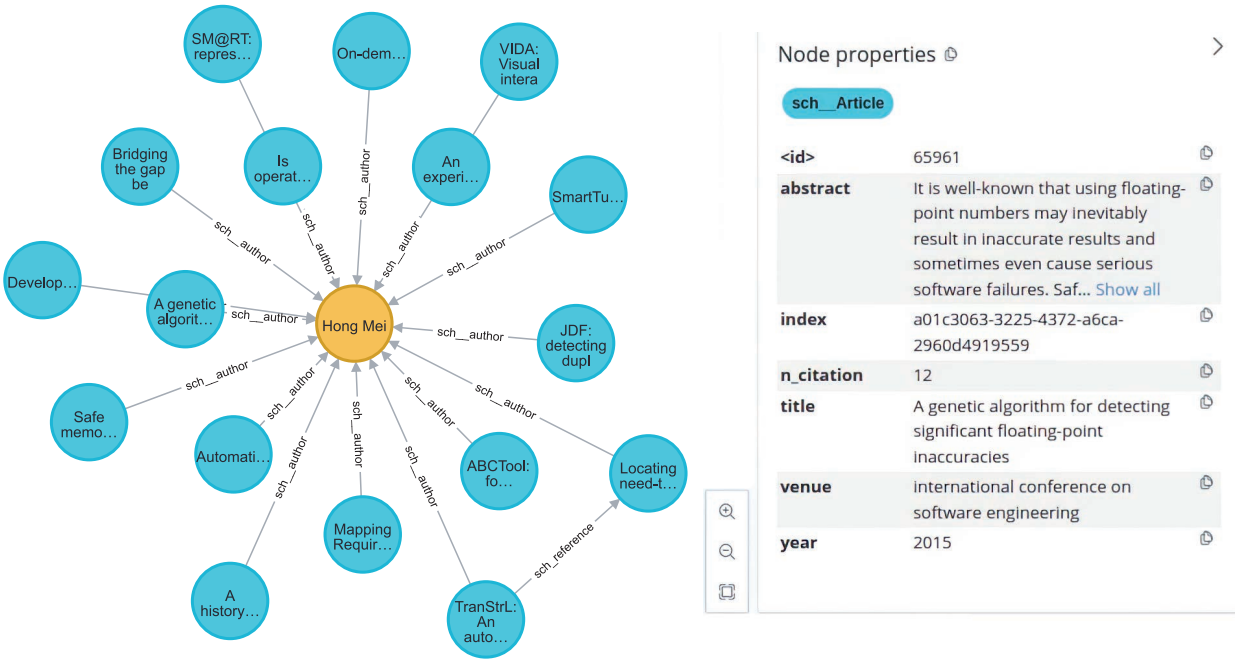

Рис. 4. Статьи, написанные автором Хонг Мэйем, и тематика публикаций
Fig. 4. Articles written by the author Hong Mei and topics of publications

Запрос к базе данных следующий:
MATCH (ar:sch__Article)-[:sch__author]->(ath:sch__Person)
WHERE ath.name = ‘Hong Mei’
RETURN ar, ath.
Для статьи ABCTool на рис. 4 справа приведены уточненные характеристики.

Пример 4. Поиск авторов по названию статьи
Запрос к базе данных следующий:
MATCH (ar:sch__Article)-[:sch__author]->(ath:sch__Person)
WHERE ar.sch__name = ‘Big Data Storage’
RETURN ar, ath.

Результат: статья ‘Big Data Storage’ написана авторами Jorg Daubert, Herman Ravkin, Mario Linchaka, M. Strohbach. Скриншот результата поиска авторов по названию статьи изображен на рис. 5.

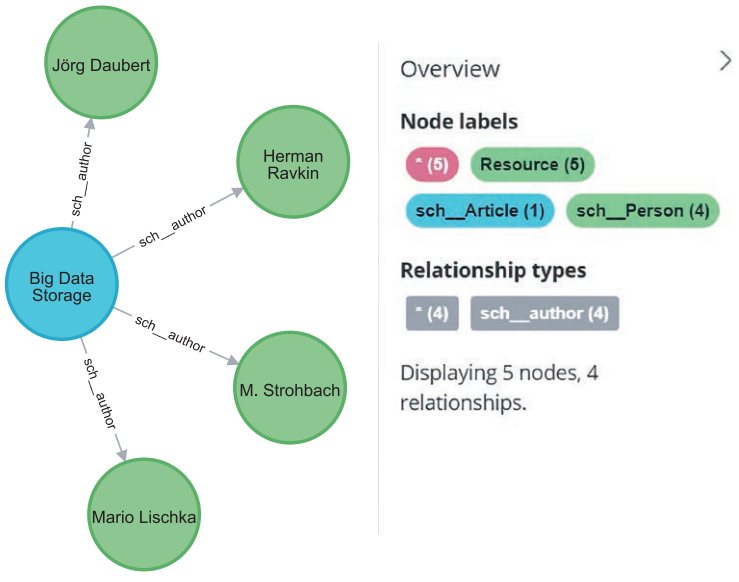

Рис. 5. Скриншот результата поиска авторов по названию статьи
Fig. 5. Screenshot of authors search result by article title

Пример 5. Издание Journal of Big Data сайта SpringerOpen и статьи, опубликованные в нем

Скриншот результата поиска издания Journal of Big Data и статей, опубликованных в нем, приведен на рис. 6. Выделенная статья отражает тематику публикации.

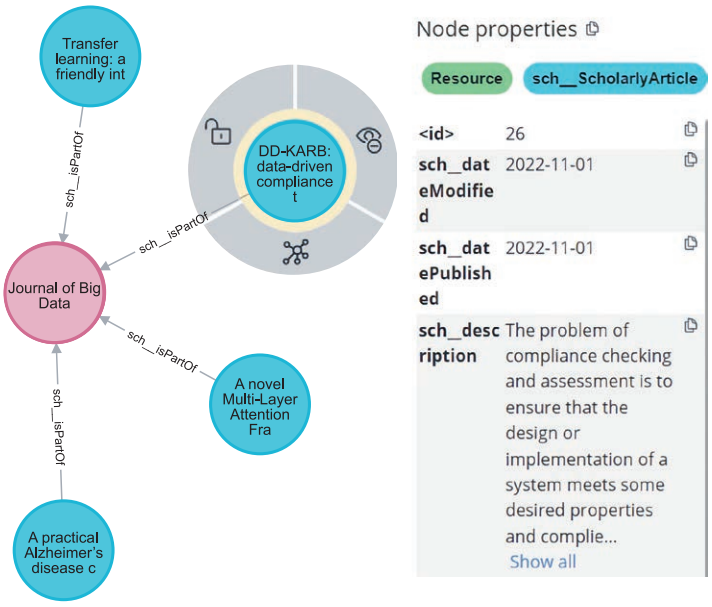

Рис. 6. Скриншот результата поиска издания Journal of Big Data и статей, опубликованных в нем
Fig. 6. Screenshot of the search result of the Journal of Big Data and articles published in it

Анализ данных

В качестве предметной области тематических сайтов в СКА ИИ можно использовать сайты научных публикаций, такие как Semantic Scholar, SpringerOpen, IEEE Xplore, «КиберЛенинка», и графовую базу данных, приведенную на рис. 2. Анализ наиболее важных статей и авторов выполнен с помощью алгоритма PageRank из библиотеки Graph Data Science. Вычисленное значение нормализируется и записывается в свойство pagerank вершины ArticlesGraph (рис. 2), т. е. свойство pagerank определено для всех публикаций в рассматриваемой БД. Фрагмент гистограммы популярности статей приведен на рис. 7: по оси X – название статей, по оси Y – их популярность.

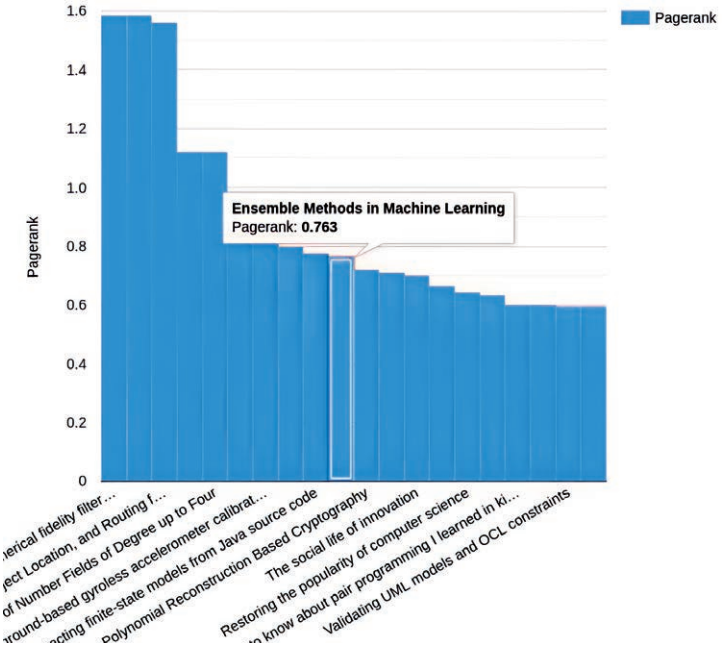

Рис. 7. Фрагмент гистограммы популярности статей
Fig. 7. Fragment of Histogram of articles popularity

Пример 6. Совместное применение графовых технологий и алгоритмов машинного обучения – на наборе данных европейских дорог (Neo4j Desktop[5])

Для демонстрации технологических решений совместного применения вложений и ML используется набор данных европейских дорог[7], который содержит 894 города и 1250 дорог (дополнительная информация по технологическим решениям приведена в статье «Граф знаний и машинное обучение как IT-среда интеллектуального анализа данных интернет-источников»[8]). Структура загружаемых данных в графовую БД выглядит следующим образом: *road_number,origin_country_code,origin_reference_place,destination_country_code,destination_reference_place,distance,watercrossing.*

Для городов (узлы – Place) строится векторное представление, что позволяет применять алгоритмы ML для более глубокого анализа данных в графовых БД. На рис. 8 приведен результат выполнения ML-алгоритма кластеризации.

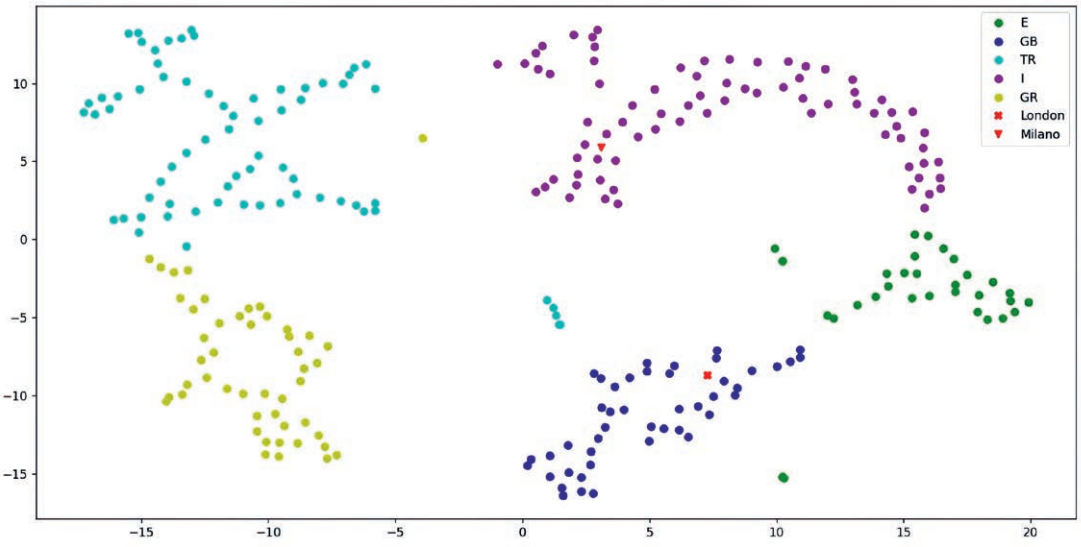

Рис. 8. Результат работы алгоритма кластеризации
Fig. 8. The result of the clustering algorithm operation

Заключение

1. Разработана и апробирова комплексная методология последовательного применения взаимосвязанных методов и инструментов по построению графовой базы данных, графа знаний, анализа данных с использованием векторного преобразования графовых данных, методов и моделей машинного обучения и предоставления аналитических результатов пользователям. Создана и апробирована IT-среда для быстрого построения тематической графовой базы данных из данных сайтов, продемонстрировано применение графа знаний и показано, что это естественная модель данных во многих реальных ситуациях.

2. Применена технология преобразования (embedding) графов (графовых данных) в непрерывное низкоразмерное векторное представление, что позволяет анализировать содержимое графовых баз данных с помощью алгоритмов машинного обучения.

3. Предлагаемая комплексная методология применяется при создании системы комплексного анализа искусственного интеллекта в Белорусском государственном университете информатики и радиоэлектроники для анализа публикаций известных мировых сайтов.

Список литературы / References

1. Diestel R. (2017) *Graph Theory*. Berlin, Springer-Verlag Publ.
2. Needham M., Hodler Amy E. (2019) *Graph Algorithms*. Sebastopol, O'Reilly Media.

[7] https://raw.githubusercontent.com/neo4j-examples/graph-embeddings/main/data/roads.csv.
[8] https://libeldoc.bsuir.by/handle/123456789/46990.

3. Hamilton W. L., Rex Ying, Leskovec J. (2017) *Representation Learning on Graphs: Methods and Applications*. Stanford, Stanford University. (9), 1–25.
4. Portisch Jan, Heist Nicolas, Paulheim Heiko (2022) Knowledge Graph Embedding for Data Mining VS. Knowledge Graph Embedding for Link Prediction – Two Sides of the Same Coin? *Semantic Web*. (1), 1–24. DOI 10.3233/SW-212892.

Вклад авторов

Авторы внесли равный вклад в написание статьи.

Authors' contribution

The authors contributed equally to the writing of the article.

Сведения об авторах

Пилецкий И. И., к. ф.-м. н., доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники

Батура М. П., д. т. н., профессор, заведующий научно-исследовательской лабораторией 8.1 «Новые обучающие технологии» Белорусского государственного университета информатики и радиоэлектроники

Волорова Н. А., к. т. н., доцент, заведующая кафедрой информатики Белорусского государственного университета информатики и радиоэлектроники

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 17 293-23-01
E-mail: piletski@bsuir.by
Пилецкий Иван Иванович

Information about the authors

Piletski I. I., Cand. of Sci., Associate Professor at the Department of Informatics of the Belarusian State University of Informatics and Radioelectronics

Batura M. P., Dr. of Sci. (Tech.), Professor, Head of the Research Laboratory 8.1 "New Learning Technologies" of the Belarusian State University of Informatics and Radioelectronics

Volarava N. A., Cand. of Sci., Associate Professor, Head of the Department of Informatics of the Belarusian State University of Informatics and Radioelectronics

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 17 293-23-01
E-mail: piletski@bsuir.by
Piletski Ivan Ivanavich