УДК 519.684.6;004.021
https://doi.org/10.37661/1816-0301-2024-21-1-105-120

Оригинальная статья
Original Paper

Система комплексного анализа данных тематических сайтов ИСКАД ИИ

И. И. Пилецкий✉, М. П. Батура, Н. А. Волорова, П. А. Зорко, А. О. Кулевич

Белорусский государственный университет информатики и радиоэлектроники,
ул. П. Бровки, 6, Минск, 220013, Беларусь
✉*E-mail: ianmenski@gmail.com*

Аннотация

Цели. В настоящее время основным источником получения информации является Интернет. Огромный объем информации, доступной в сети, делает актуальной задачу всестороннего анализа данных из открытых интернет-источников. Цель работы заключается в создании многоцелевого, модифицируемого кластера для глубокого анализа данных интернет-источников, основными задачами которого являются выявление наиболее важных публикаций в некоторой предметной области и их тематический анализ, определение лидера научного направления и тенденций развития направлений деятельности и взаимодействия групп людей.

Методы. Для решения поставленной задачи была разработана методология построения многоцелевого кластера с использованием технологий быстрого построения тематической графовой базы данных, графа знаний, методов и моделей машинного обучения для глубокого анализа данных.

Результаты. Разработана Система комплексного анализа данных тематических сайтов ИСКАД ИИ, апробированы методология быстрого построения тематической графовой базы данных и комплексная технология глубокого анализа данных интернет-источников и известных мировых сайтов.

Заключение. Создана среда информационных технологий для быстрого построения тематических графовых баз данных. Результаты применения технологии быстрого построения графовых баз данных показаны на примерах работы ИСКАД ИИ.

Ключевые слова: тематические сайты, большие данные, метод машинного обучения, анализ данных, графовая база данных, граф знаний, база данных Neo4j

Для цитирования. Система комплексного анализа данных тематических сайтов ИСКАД ИИ / И. И. Пилецкий [и др.] // Информатика. – 2024. – Т. 21, № 1. – С. 105–120.
https://doi.org/10.37661/1816-0301-2024-21-1-105-120

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 08.12.2023
Подписана в печать | Accepted 12.01.2024
Опубликована | Published 29.03.2024

System of complex data analysis of thematic sites ISCAD IS

Ivan I. Piletski✉, Michal P. Batura, Natalia A. Volorova, Polina A. Zorko, Alexei O. Kulevich

Belarusian State University
of Informatics and Radioelectronics,
st. P. Brovki, 6, Minsk, 220013, Belarus
✉*E-mail: ianmenski@gmail.com*

Abstract

Objectives. Currently, the main source of information is the Internet. The huge amount of information available on the Internet makes it urgent to comprehensively analyze data from open Internet sources. The goal of this work is to create a multi-purpose, modifiable cluster for in-depth analysis of data from Internet sources, the main objectives of which are to identify the most important publications in a certain subject area, thematic analysis of these publications, identifying the leader of a scientific direction and determining trends in the development of areas and interaction of groups of people.

Methods. To solve this problem, a methodology was developed for constructing a multi-purpose cluster using technologies for quickly constructing a thematic graph database, a knowledge graph, methods and models of machine learning for in-depth analysis of data.

Results. A system for comprehensive analysis of data from thematic sites ISKAD IS has been developed, a methodology for quickly constructing a thematic graph database and a comprehensive technology for in-depth analysis of data from Internet sources and analysis of data from the most important well-known world sites have been tested.

Conclusion. An IT environment has been created for the rapid construction of thematic graph databases. The results of using the technology for quickly constructing graph databases are shown using examples of the work of ISKAD IS.

Keywords: thematic sites, Big Data, machine learning method, analysis of data, graph database, knowledge graph, database Neo4j

For citation. Piletski I. I., Batura M. P., Volorova N. A., Zorko P. A., Kulevich A. O. *System of complex data analysis of thematic sites ISCAD IS*. Informatika [*Informatics*], 2024, vol. 21, no. 1, pp. 105–120 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-1-105-120

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

Введение. В настоящей работе используются материалы по разработке программных комплексов анализа данных из интернет-источников, полученные авторами ранее [1–3] при реализации Системы комплексного анализа данных тематических сайтов ИСКАД ИИ. Все работы выполнялись в Белорусском государственном университете информатики и радиоэлектроники на протяжении нескольких лет.

По данным Gartner group за 2023 г., большинство известных ИТ-компаний разрабатывали или имели аналитические средства анализа данных. Проблемами существующих средств являются их тяжеловесность, сложность модифицирования и адаптации к изменению тематики области применения. Последние данные трендов Gartner в области ИТ "Gartner Identifies Top 10 Data and Analytics Technology Trends for 2021" показывают возрастающую роль графовых технологий. Так, к 2025 г. графовые технологии будут использоваться в 80 % инноваций в области данных и аналитики по сравнению с 10 % в 2021 г., что даст возможность быстро принимать решения в организации[1].

Одним из сложных современных направлений является представление знаний с помощью специальных глобальных словарей предметных областей, метаописаний и специальных языков, а также методологий их применения. Многие важные мировые тематические сайты, такие как EBSCO, ScienceDirect, SpringerLink, ACM Digital Library, IEEE Xplore, CiteSeerX, Google

[1]Gartner [Electronic resource]. – Mode of access: https://www.gartner.com/en/newsroom/press-releases/2021-03-16-gartner-identifies-top-10-data-and-analytics-technologies-trends-for-2021/. – Date of access: 18.10.2023.

Scholar, Semantic Scholar, libgen: Library Genesis, Medium, КиберЛенинка, SpringerOpen, Wikipedia, Wikidata и др., используют специальную технику описания ресурса RDF (Resource Description Framework, среда описания ресурса)[2].

RDF представляет собой абстрактную модель, обеспечивающую способ разбиения знаний на дискретные части и позволяющую обмениваться информацией. RDF – это модель обмена данными, которая описывает, как данные сериализуются и как ими обмениваются. Модель RDF не описывает, как данные хранятся и организуются, она предназначена для обмена информацией (импорта и экспорта). Такой подход позволяет описывать знания в тематических предметных словарях и обмениваться этими знаниями с другими сайтами. Словари RDF и онтологии OWL (Web Ontology Language, язык представления веб-онтологий) применяют абстрактные модели RDF и RDFS описания ресурса. Онтология – это конкретное формальное представление того, что означают термины в той области, в которой они используются. Данные для импорта и экспорта на RDF-сайтах могут быть представлены в нескольких форматах: JSON-LD, Turtle, N-Triples, RDF/XML, TriG и N-Quads, TriG.

Разработанная методология описания сайтов позволяет применять специальную технологию построения (генерации) графовой БД из описания RDF-данных. Такая тематическая графовая БД содержит базу знаний сайта в виде графа знаний, что дает возможность применять различные аналитические алгоритмы ML для более глубокого анализа данных сайта [4–6].

Целью настоящей работы являются апробация и тестирование методологии [2] разработки многоцелевого, модифицируемого кластера (семейства программного обеспечения) для анализа данных интернет-источников (например, научных публикаций, социальных сетей, СМИ). Такой анализ позволяет выявлять наиболее важные публикации в некоторой предметной области (например, в космических исследованиях, здравоохранении, социальной сфере), определять тематику этих публикаций, выявлять лидера научного направления, предсказывать тенденции развития направлений и взаимодействия групп людей.

Разработанное программное обеспечение Системы комплексного анализа данных тематических сайтов ИСКАД ИИ позволит реализовать поставленные цели и выполнить анализ публикаций в предметной области. В качестве предметной области для Системы комплексного анализа данных тематических сайтов ИСКАД ИИ могут быть использованы важные мировые сайты, в которых применяется специальная техника описания ресурса RDF.

1. Методы построения тематических сайтов

1.1. Среда описания ресурса. Одним из способов формализации знаний является применение какого-либо доступного стандартного языка. Для формального описания знаний в тематических словарях наиболее широко используются схема RDF, язык веб-онтологий (Web Ontology Language, OWL) для онтологий и простая система организации знаний (Simple Knowledge Organization System, SKOS) для схем таксономической классификации. Каждый из словарей допускает разные уровни выразительности: от базового определения категорий и отношений до таксономий и более сложных конструкций, например сложных классов. Такой подход позволяет применять различное программное обеспечение для реализации методологии [7].

Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а ребра отображают отношения.

Для доступа к данным мировых тематических сайтов можно использовать специально разработанный язык SPARQL Protocol and RDF Query Language. ИТ-специалистами разработано множество различных редакторов, помогающих строить простые и сложные запросы на языке SPARQL [7]. Применение языка SPARQL позволяет получать результат в виде простого скалярного или несложного структурированного значения. Однако этот язык не дает возможность решить главную задачу, которая заключается том, что по множеству описаний RDF необходимо построить тематическую графовую БД с целью дальнейшего глубокого анализа данных сайта и длительного исследования данных в БД.

---

[2]Среда описания ресурса (RDF): понятия и абстрактный синтаксис [Электронный ресурс]. – Режим доступа: https://www.w3.org/2007/03/rdf_concepts_ru/Overview.html. – Дата доступа: 18.10.2023.

1.2. Правила преобразования троек RDF в графовую БД. Граф RDF – это набор триплетов или операторов (субъект, предикат, объект), где и субъект, и предикат являются ресурсами, а объект может быть либо другим ресурсом, либо литералом. Литералы не могут быть предметом других утверждений. Ресурсы однозначно идентифицируются URI. Существуют три основных правила сериализации троек RDF (субъект – предикат – объект) в графовую БД[3]. Данные преобразования являются частью методологии [2] построения ИСКАД ИИ и реализуются с помощью ИТ-среды в графовой БД Neo4j.

Правило 1. Узел в Neo4j, представляющий ресурс RDF, помечен :Resource и будет иметь свойство uri с URI-ресурса:

(S,P,O) => (:Resource {uri:S}).

Правило 2. Предикаты троек отображаются в свойствах узла в Neo4j, если объект тройки является литералом:

(S,P,O) && isLiteral(O) => (:Resource {uri:S, P:O}).

Правило 3. Предикаты троек отображаются на отношения в Neo4j, если объект тройки является ресурсом:

(S,P,O) && !isLiteral(O) => (:Resource {uri:S})-[:P]->(:Resource {uri:O}).

1.3. Графовые технологии и машинное обучение. Основа совместного применения графовых технологий и методов машинного обучения, используемых в ИСКАД ИИ, описана в работах [1, 3]. Графовая БД (узлы и отношения) содержит неструктурированные данные, так как они представлены в реальном мире, но для решения задач с помощью машинного обучения нужно преобразовать пространство, где находится граф, в другое пространство для машинного обучения – векторное, для которого применимы известные алгоритмы машинного обучения (например, node2vec или GraphSAGE). Данное преобразование выполняется с помощью сложной методологии выделения вектора свойств, называемого включением (embedding) [8]. Графовые включения – это представление узлов и отношений в графе как вектора свойств. В качестве значений вектора свойств могут быть выбраны некоторые атрибуты вершин и отношений. Такая методология совместного применения графовых технологий и методов машинного обучения позволяет создавать технологии глубокого анализа данных интернет-источников. Конкретные технологические решения и примеры применения их в ИСКАД ИИ приведены в разд. 2.

2. Результаты построения системы ИСКАД ИИ

2.1. Система комплексного анализа данных тематических сайтов ИСКАД ИИ. Существуют различные варианты архитектурных решений построения систем анализа данных интернет-источников. Так, в публикациях [1, 2] при разработке интеллектуальной системы комплексного анализа данных интернет-источников выполнен ретроспективный анализ трех вариантов архитектурных решений, определены структурные компоненты и их функции. Основное архитектурное решение заключается в том, что система должна состоять из следующих компонентов: сбора данных, фильтрации данных и составления «мешка слов» из N-грамм (векторизации), библиотеки аналитических модулей, хранилища данных, графовой БД и графа знаний, аналитического компонента, обеспечивающего взаимодействие с пользователем и подготовку выдачи результата, клиентского модуля и универсальной интеграционной шины (управляющего компонента). При необходимости набор модулей и компонентов может быть расширен, а некоторые модули заменены новыми. Были последовательно разработаны и реализованы три варианта данной системы. В третьем варианте приняты важные дополнительные архитектурные решения: все компоненты функционируют как постоянно работающие самостоятельные серверы; в качестве хранилища скаченных данных использовалась БД лидера хранилища типа «семейство столбцов» Cassandra; для анализа данных применялась графовая БД, моделирующая предметную область (данные поступали из хранилища). Для обеспечения взаимодействия компонентов использовался управляющий компонент, который выполнял роль

[3]Neosemantics (n10s): Neo4j RDF & Semantics toolkit [Electronic resource]. – Mode of access: https://neo4j.com/labs/neosemantics/. – Date of access: 18.10.2023.

интеграционной шины (разработан на базе интеграционной шины Kafka). При такой архитектуре остановка работы одного из компонентов не приводила к остановке работы всего комплекса и можно было легко выполнять его модернизацию. Однако данное фундаментальное архитектурное решение не позволяет быстро перестроить систему на новую тематику и построить многоцелевой, модифицируемый кластер семейства тематических графовых БД.

Анализ аналогов известных сайтов: КиберЛенинка, Semantic Scholar, SpringerOpen, Medium – позволил принять решение о разработке мультиплатформенного решения для анализа данных различных предметных областей, с помощью которого можно быстро получать информацию о предметной области с мировых крупных сайтов на базе быстрого построения графовой(ых) БД предметной области и выполнять более глубокий анализ данных предметной области.

2.2. Архитектура ИСКАД ИИ. Система комплексного анализа данных тематических сайтов ИСКАД ИИ позволяет анализировать данные с различных тематических сайтов и предназначена для сбора информации о научных публикациях, генерации графовой БД, построения графа знаний, преобразования свойств узлов и отношений графовой БД в векторное представление с целью применения алгоритмов ML для более глубокого анализа данных. Такой комплексный подход к анализу данных дает возможность определять передовые научные направления и экспертов в предметных областях, тематику их работ и взаимосвязи.

Основными компонентами ИСКАД ИИ являются: получение данных из интернет-источников; графовая БД и граф знаний; извлечение свойств из графовой БД и их анализ с помощью алгоритмов ML, а также интеграция (специальный веб-сайт ИСКАД ИИ). Компонент «извлечение свойств из графовой БД» обладает дополнительной функциональностью и может использовать технологию включений [8], что позволяет строить векторы свойств меньшей размерности для более глубокого анализа данных. Общая функциональная архитектура ИСКАД ИИ показана на рис. 1.

Компонент «получение данных из интернет-источников» выполняет различные операции над данными, включая очистку, структурирование, нормализацию и приведение их к единому формату. Это позволяет обеспечить качество и консистентность данных, а также подготовить их для дальнейшего анализа. Получение данных осуществляется в специальной ИТ-среде и обеспечивает быстрое построение тематических графовых БД, использующих RDF-описание данных в различных форматах (рис. 1, п. 1 и п. 2).

Компонент «графовая БД и граф знаний» реализуется с помощью ИТ-среды и графовой БД Neo4j (рис. 1, п. 3), которая является лидером среди графовых БД на протяжении последних 10 лет[4]. Она обладает исключительными свойствами горизонтального масштабирования, с ростом данных не деградирует, работает в десятки и сотни раз быстрее, чем реляционная БД, и обеспечивает требования ACID (atomicity, consistency, isolation, durability) и соответствие спецификациям JTA, JTS и XA. Графовая БД Neo4j обеспечивает работу с миллионами узлов и отношений, а также доступ к данным как на языке запросов Cypher (соответствует требованиям CRUD (create, read, update, delete)), так и на популярных языках программирования. Она позволяет получать графы знаний и графически визуализировать наборы данных и результаты запросов. В процессе доступа скачивания данных с сайтов компонентом «получение данных» графовая БД строится и модифицируется автоматически.

Компонент «извлечение свойств из графовой БД и их анализ с помощью алгоритмов ML» (рис. 1, п. 4–6) позволяет реализовывать запросы пользователей, выдавать тематические графы знаний и обеспечивает анализ данных в графовых БД, преобразовывая свойства в узлах и ребрах (отношениях) в векторное представление с целью применения для дальнейшего анализа данных алгоритмов ML [8, 9].

Компонентом «интеграция» является веб-сайт ИСКАД ИИ (рис. 1, п. 7), который дает возможность работать другим компонентам системы и предоставляет доступ к получению аналитических данных пользователям системы. Веб-сайт обеспечивает пользователей возможностью создавать, обновлять и взаимодействовать с тематическими графовыми БД. Он

[4]Neo4j Graph Database [Electronic resource]. – Mode of access: https://neo4j.com/product/neo4j-graph-database. – Date of access: 18.10.2023.

предоставляет удобный интерфейс для выполнения запросов к графу знаний, получения и выдачи репортов. Веб-сайт способствует интеграции с различными источниками данных в системе, такими как БД, файловые хранилища или внешние API, чтобы получать необходимую информацию для отчетов.

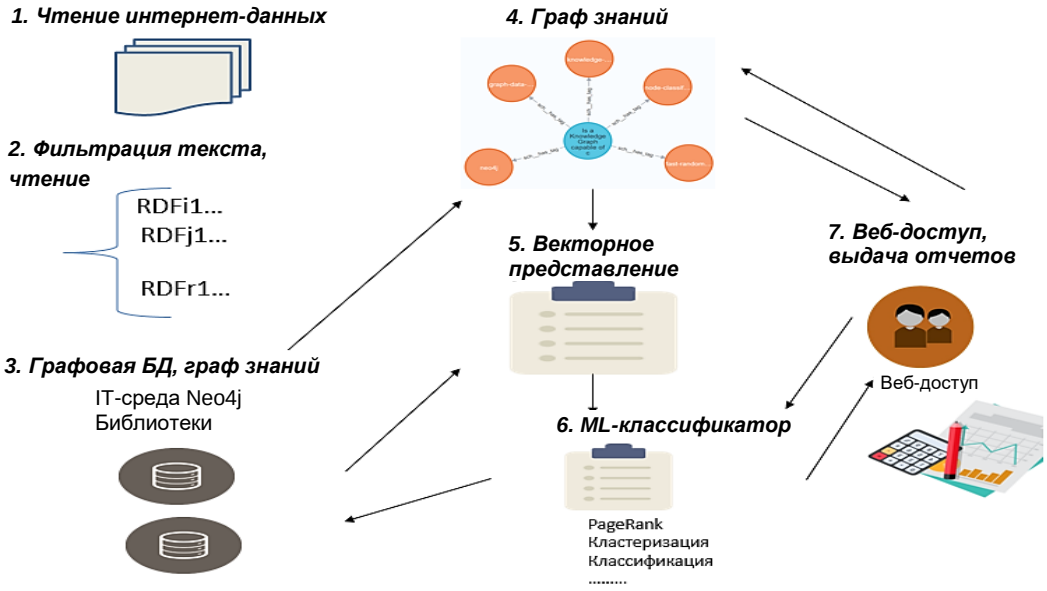


Рис. 1. Общая функциональная архитектура ИСКАД ИИ

Fig. 1. General functional architecture of ISKAD IS

Зарегистрированный пользователь может иметь доступ к просмотру публикаций различных предметных областей, поиску наиболее цитируемых авторов предметной области, просмотру параметров некоторой предметной области и различной информации об авторе публикации. Пользователь может видеть имя, количество публикаций и сами публикации автора, а также гистограммы по авторам и статьям. Гистограммы по авторам и статьям строятся с помощью алгоритма PageRank. При этом пользователь сам выбирает, какая гистограмма ему нужна и какое количество статей или авторов должно в нее входить.

Администратор веб-сайта может добавлять или удалять пользователей и предоставлять им расширенные права, управлять состоянием БД, менять структуру и содержимое БД и делать замену на сайте одной БД на другую.

Веб-сайт является центральным компонентом, который облегчает управление данными, взаимодействие пользователей и получение информации из системы. Клиент-серверная архитектура была использована для реализации компонента. Она является распространенным подходом к разработке веб-приложений. Такой подход обеспечивает интеграцию и эффективность в работе с пользователями и компонентами системы. Данное решение и представленная методология реализованы при разработке проекта БГУИР «ГПНИ по теме "Интеллектуальная система комплексного анализа данных интернет-источников (ИСКАД ИИ)"». В настоящей работе продемонстрирована реализация основных технических решений.

2.3. ИТ-платформа для ИСКАД ИИ. В качестве основных компонентов ИТ-среды для построения ИСКАД ИИ используются графовая СУБД Neo4j Desktop, специальные библиотеки (плагины), расширяющие возможности анализа данных в графовой БД (Neosemantics (n10s), библиотека APOC (Awesome Procedures on Cypher), библиотека Neo4j Graph Data Science (GDS)) и фреймворки для разработки веб-сайта ИСКАД ИИ.

2.4. Инфологическая модель графовой БД. В ИСКАД ИИ разработаны БД и ее инфологическая модель. При создании данной модели за основу бралась предметная область проекта, которая во многом совпадает с предметной областью, приведенной в работе [2]. Важно отметить, что шаблон графовой БД может быть определен исходя из назначения БД для анализа данных предметной области и он может отличаться от приведенного в настоящей статье. Сама графовая БД может модифицироваться в процессе своей работы. Главными сущностями и атрибутами предметной области ИСКАД ИИ являются: User (User) – пользователь ИСКАД ИИ, Article (sch__Article) – статья или публикация, Author (sch__Person) – автор статьи или публикации, Tag (sch__Tag) – ключевые слова, которые относятся к статье, List (sch__List) url – ссылка на список со статьями. Общая схема БД, сущности и их атрибуты представлены на рис. 2.

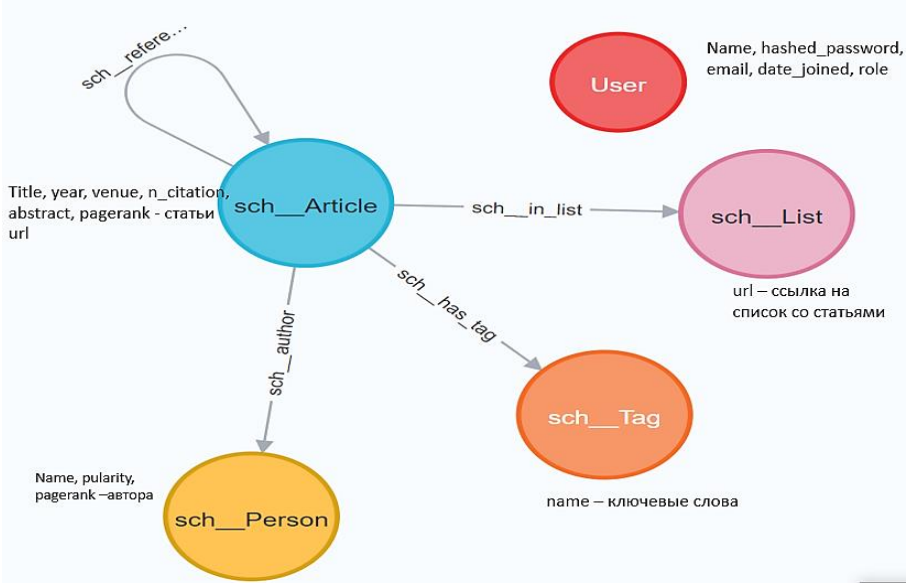


Рис. 2. Общая схема БД

Fig. 2. General Database Schema

2.5. Компонент «получение данных из интернет-источников». Для реализации методологии ИСКАД ИИ в графовую БД системы последовательно были загружены данные с сайтов SpringerOpen, Semantic Scholar, КиберЛенинка и Medium.

Извлечение информации с веб-страницы и RDF выполняется процедурами библиотек APOC n10s apoc.load.html и n10s.rdf.import. Для этого в командах следует указать URL-адрес страницы и CSS-подобный селектор, чтобы выбрать конкретный требуемый элемент. Визуализация и анализ RDF выполняются процедурами плагина Neosemantics. Код типичных команд загрузки данных с сайта SpringerOpen представлен ниже. Объем загружаемых и преобразуемых данных определяется кодом команд:

```
CALL apoc.load.html("https://journalofcloudcomputing.springeropen.com/articles/10.1186/2192-113X-1 18" , { jsonld: 'head script[type="application/ld+json"]'})
 YIELD value
UNWIND ["https://cybersecurity.springeropen.com/articles/10.1186/s42400-023-00144-1", "https://cybersecurity.springeropen.com/articles/10.1186/s42400-023-00141-4", "https://cybersecurity.springeropen.com/articles/10.1186/s42400-023-00140-5", "https://cybersecurity.springeropen.com/articles/10.1186/s42400-023-00138-z", "https://cybersecurity.springeropen.com/articles/10.1186/s42400-020-00050-w"] as page
CALL apoc.load.html(page, { jsonld: 'head script[type="application/ld+json"]'}) YIELD value
CALL n10s.rdf.import.inline(value.jsonld[0].data, "JSON-LD")  yield terminationStatus, triplesLoaded, triplesParsed, extraInfo
RETURN page, terminationStatus, triplesLoaded, triplesParsed, extraInfo
```

На рис. 3 представлена графовая БД ИСКАД ИИ, полученная с помощью технологии быстрого построения тематической графовой БД.

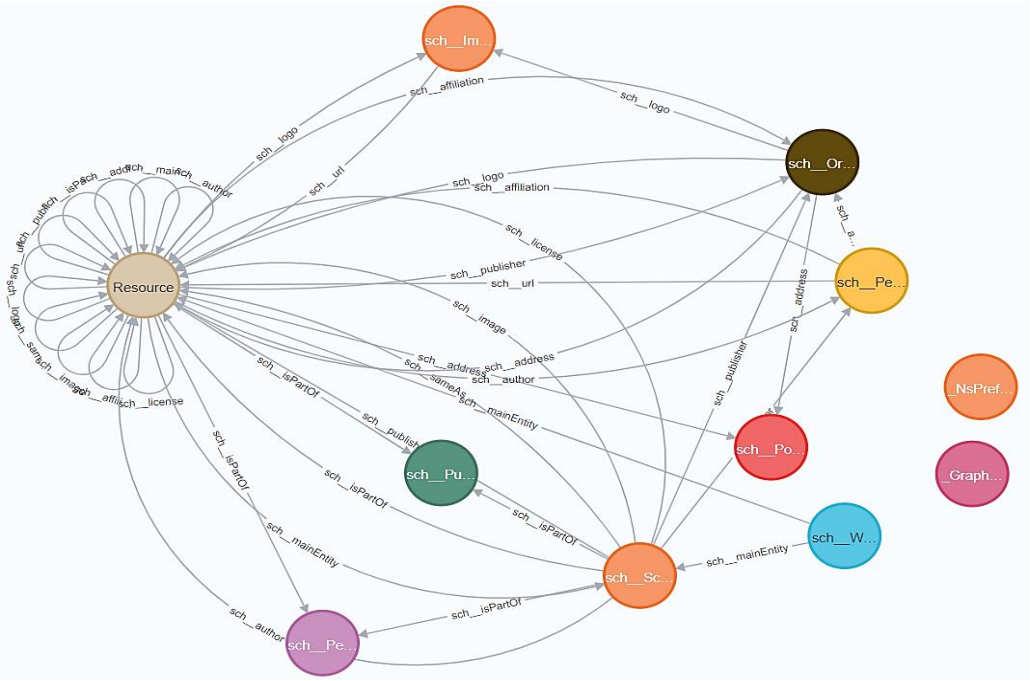


Рис. 3. Общее представление графовой БД сайта SpringerOpen

Fig. 3. General presentation of the graph database of the SpringerOpen site

С помощью описанной технологии в графовую БД ИСКАД ИИ были загружены данные с сайтов SpringerOpen, Semantic Scholar, КиберЛенинка и Medium. Важно отметить, что данные с различных сайтов были объединены автоматически в одной графовой БД (для тестирования загружено около 250 000 документов).

2.6. Компонент «графовая БД и граф знаний». Графы знаний – это особый тип графов с упором на контекстное понимание. Они представляют собой взаимосвязанные наборы фактов, которые описывают объекты, события или вещи реального мира и их взаимосвязи в формате, понятном человеку и машине. В графах знаний используется принцип организации, позволяющий пользователю (или компьютерной системе) рассуждать о лежащих в их основе данных. Принцип организации дает дополнительный уровень метаданных, который добавляет связанный контекст для поддержки рассуждений и получения знаний. Принцип организации делает сами данные более интеллектуальными, а не блокирует инструменты для понимания данных внутри кода приложения. В свою очередь, это одновременно упрощает системы и поощряет их широкое повторное использование [8]. Граф знаний (KG, Knowledge Graph) – ориентированный граф, узлы которого представляют собой сущности и литеральные значения (литералы), а ребра – отношения между этими сущностями [9].

Приведем примеры применения графов знаний. В них используется техника выполнения запросов drill-down, позволяющая уточнять полученные данные. Результаты выполнения запросов представлены на рис. 4–6.

Пример 1. Найдем все статьи одного из главных и самых успешных исследователей БД Neo4j и графов знаний Томаза Братанича (Tomaz Bratanic). Запрос будет иметь ограничение в 10 узлов из-за большого количества статей, написанных автором (рис. 4):

```
MATCH (a:sch__Article)-[:sch__author]->(au:sch__Person)
WHERE au.name = "Tomaz Bratanic"
RETURN a, au LIMIT 10
```

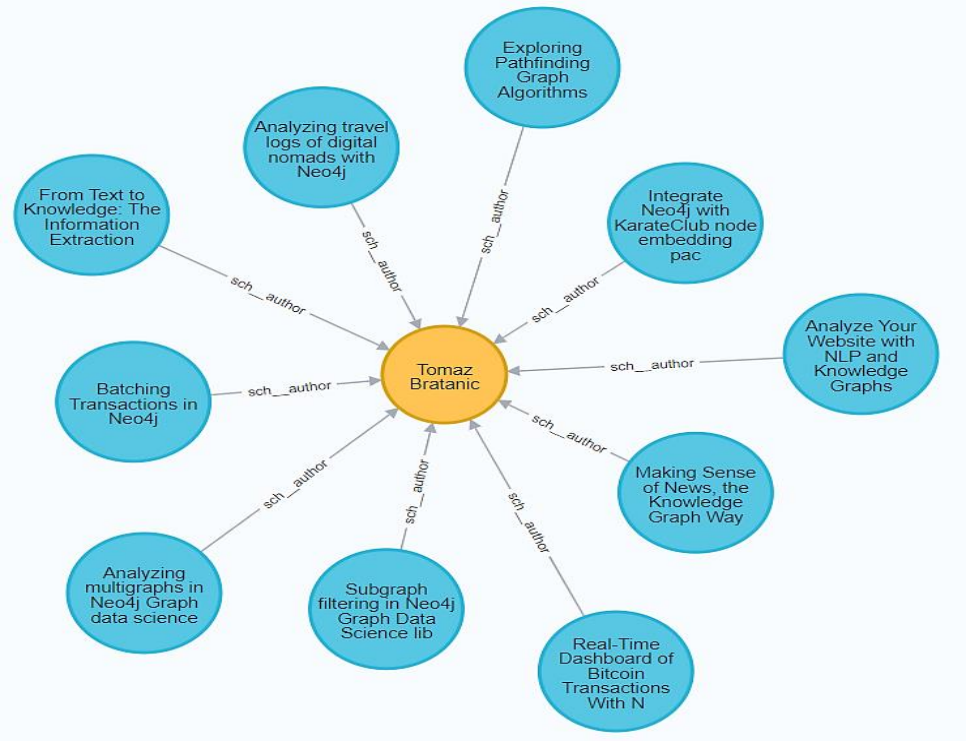


Рис. 4. Результат выполнения запроса из примера 1

Fig. 4. Example 1 of query result

На данном этапе можно отметить главную тематику статей автора: все, что связано с анализом данных, графами и графовыми алгоритмами, особенно с БД Neo4j.

Пример 2. Найдем авторов, которые пишут статьи на тему Neo4j (рис. 5):

```
MATCH (t:sch__Tag)<-[:sch__has_tag]-(a:sch__Article)-[:sch__author]->(au:sch__Person)
WHERE t.name = "neo4j"
RETURN t, au LIMIT 10
```

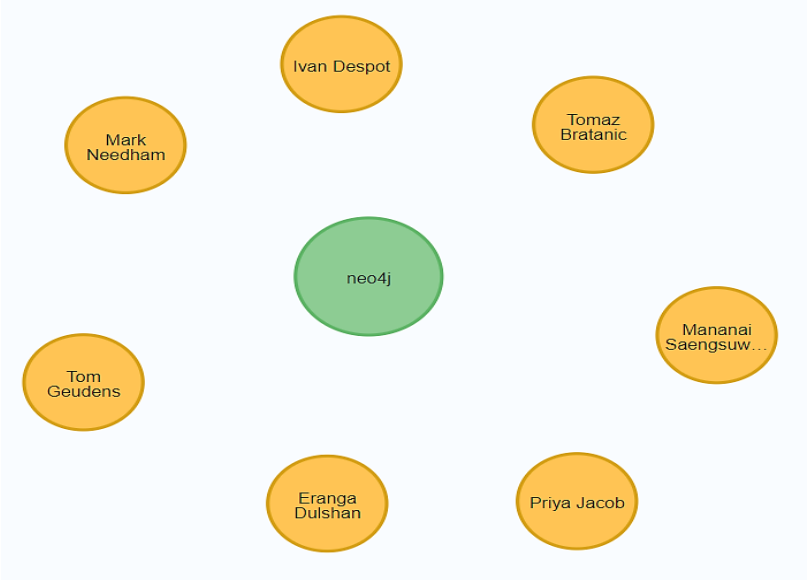


Рис. 5. Результат выполнения запроса из примера 2

Fig. 5. Example 2 of query result

Пример 3. Найдем статьи Томаза Братанича в соавторстве с еще одним исследователем использования графовых БД Марком Нидхемом (Mark Needham) (рис. 6):

```
MATCH (au1:sch__Person)<-[:sch__author]-(a:sch__Article)-[:sch__author]->(au2:sch__Person)
WHERE au1.name = "Tomaz Bratanic" AND au2.name = "Mark Needham"
RETURN au1, au2, a
```

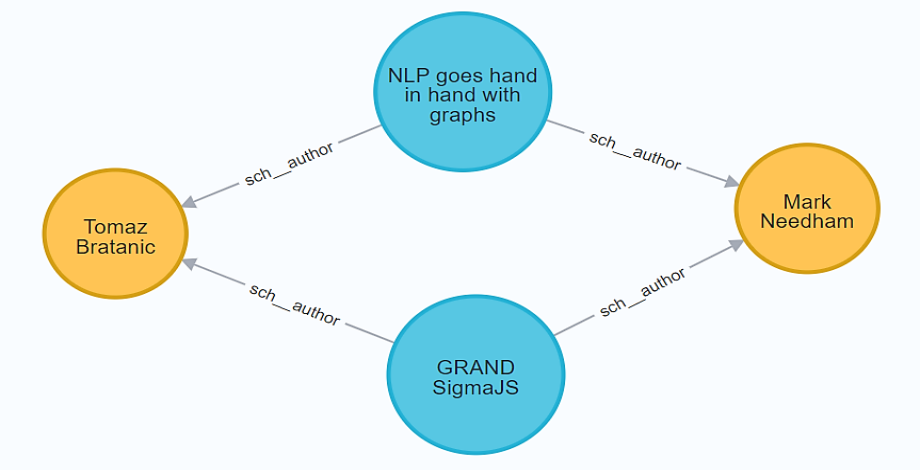


Рис. 6. Результат выполнения запроса из примера 3

Fig. 6. Example 3 of query result

В рассмотренной графовой БД нашлись две статьи, которые были написаны совместно этими авторами. Дополнительно можно уточнить ключевые слова, которые относятся к данным статьям.

2.7. Компонент «извлечение свойств из графовой БД и их анализ с помощью алгоритмов ML». Данный компонент состоит из двух модулей: модуля создания модели машинного обучения предметной области и модуля анализа свойств графовой БД с помощью алгоритмов ML, подготовки и выдачи репортов. Модель машинного обучения позволяет распознавать закономерности взаимодействия авторов публикаций, их рейтинг, рейтинг статей и предметных областей. Рейтинги определяются с помощью специального алгоритма PageRank. Однако в БД предметной области отсутствуют некоторые важные свойства узлов и отношений между ними (специфика исходных интернет-сайтов), которые не позволяют применять алгоритмы ML.

Одна из проблем заключается в отсутствии связей между соавторами. Для создания таких связей объединяются те авторы, которые совместно написали статьи. Это задача прогнозирования существования связи между двумя объектами, и она функционально решается в модуле создания модели машинного обучения. С помощью модели машинного обучения с предсказанием связей производится прогнозирование соавторства.

Другая проблема заключается в том, что не все загруженные данные отражают тематику опубликованных статей, около половины статей не содержат теги (Tag (sch__Tag) – ключевые слова статьи). Данные с сайтов SpringerOpen, Semantic Scholar, КиберЛенинка и Medium неоднородные и неотфильтрованные, что затрудняет применение общих решений алгоритмов ML.

Сама графовая БД (узлы и отношения) содержит набор разнотипных данных, к которым невозможно применить алгоритмы ML. Данные свойства в узлах и ребрах (отношениях) были преобразованы в векторное представление, что позволило использовать алгоритмы ML.

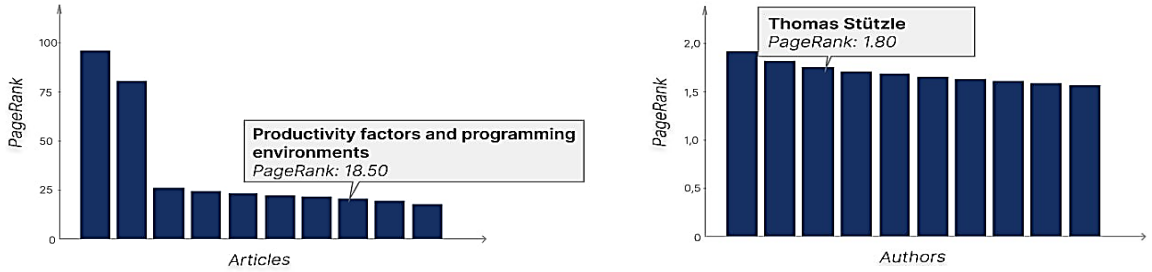
Все указанные проблемы были решены в модуле создания модели машинного обучения предметной области с помощью специально разработанных функций. Полученная обновленная модель графовой БД была протестирована и подготовлена для анализа данных. Модуль анализа

свойств графовой БД с помощью алгоритмов ML состоит из многих разработанных функций, которые применяются для анализа данных, содержащихся в подготовленной и модифицированной графовой БД ИСКАД ИИ.

Расчет PageRank каждой статьи, содержащейся в графовой БД, выполнен с помощью процедур, предоставляемых библиотекой Graph Data Science Library. Первый шаг – это выполнение процедуры построения графа in-memory из целевых сущностей статей, второй шаг – вычисление PageRank статей на основе собранных данных in-memory.

Так как набор данных не имеет между авторами прямых связей, необходимых для выполнения алгоритма PageRank, для расчет PageRank каждого автора публикаций, содержащихся в графовой БД, применяется алгоритм создания связей между авторами, аналогичный приведенному выше. Первый шаг – это выполнение процедуры создания графа in-memory из целевых сущностей авторов публикаций, для которых будет определяться PageRank, второй шаг – вычисление PageRank статей авторов на основании данных in-memory.

Диаграммы популярности статей и авторов (рис. 7) получены в модуле при настройке компонента для работы в комплексе ИСКАД ИИ. Основной поток получения отчетов в ИСКАД ИИ предусмотрен через компонент «интеграция» и веб-доступ ИСКАД ИИ (см. разд. 2.8).



a) *b)*

Рис. 7. Популярность статей (*a*); авторов (*b*)

Fig. 7. Articles (a) and authors (b) popularity

Анализ графа знаний выполнен с помощью алгоритмов вычисления графовых включений и алгоритмов машинного обучения.

Векторное представление или графовые включения (graph embedding) – технология, которая отображает узлы в графе в виде плотных векторов низкой размерности и позволяет аналогичным узлам в исходном графе (разные методы имеют разные определения сходства) быть похожими в пространстве выражений низкой размерности. Полученные векторы используются в различных алгоритмах ML для более глубокого анализа данных в графовой БД и графах знаний, например для решения таких задач, как классификация узлов, прогнозирование ссылок, визуализация или реконструкция исходного графа, и других алгоритмах.

В ИСКАД ИИ решена задача создания алгоритма, предсказывающего теги для статей, которые в графовой БД их еще не имеют. Анализ качества документов, полученных из интернет-источников, не позволяет выдавать полную и достоверную информацию по предметным областям, по публикациям и их авторам. Часть документов не содержит теги. Анализ данных графовой БД показал, что примерно половина узлов вообще не содержит теги.

Для прогнозирования тегов с использованием библиотеки GraphSAGE и языка программирования Python построена модель преобразования графа в векторное представление с помощью алгоритма генерации графовых включений. Входными данными для GraphSAGE является вектор свойств узлов. GraphSAGE поддерживает графы с несколькими типами узлов, где каждый тип узла имеет разные представляющие его функции. В настоящей статье применен алгоритм моночастичной проекции. Моночастичная проекция позволяет на основе графа с двумя типами узлов вывести из него граф с одним типом узлов. Библиотека Neo4j предоставляет пользователям алгоритм построения моночастичной проекции с помощью алгоритма

сходства узлов Node Similarity из библиотеки GDS. Алгоритм Node Similarity сравнивает наборы узлов на основе узлов, которые связаны между собой. Два узла считаются похожими, если у них много общих соседей.

Как и ранее для определения PageRank, для реализации алгоритма создается граф in-memory, который будет содержать свойство (атрибут) узла openaiEmbedding включения слов, созданного на основе данных свойства (атрибута) title статьи с помощью алгоритма text-embedding-ada-002. Построение графа in-memory выполняется с помощью специального кода. К полученному графу применяется алгоритм сходства узлов с установленным значением topK, равным 1000, для создания связи с как можно большим количеством статей в графе.

Алгоритм Node Similarity создает несколько компонентов связности. Для нахождения большего из них, который содержит практически все необходимые узлы, применяется алгоритм Weakly Connected Components.

Для реализации алгоритма классификации требуется выбрать теги, которые необходимо предсказать. Выберем теги, которые встречаются как минимум в 200 статьях, и добавим свойство target в выбранные узлы.

Анализ результатов работы алгоритмов позволяет сделать выводы, что полученные теги действительно отражают суть выбранных статей. Ниже приведен пример выдачи тегов:

1. Introduction to Data Mesh adoption in Adidas – motivation and takeaways --- [data],
2. Things to Do When You Feel Ruled by Time --- [productivity],
3. A Data Science project start to finish --- [coding, programming, python, python3, software-development],
4. Time series anomaly detection – in the era of deep learning --- [data-science, machine-learning],
5. How to Optimize Your Apache Spark Application with Partitions --- [spark],
6. Rule Execution with SHACL --- [knowledge-graph],
7. Language & Cognition: re-reading Jerry Fodor --- [data-science, machine-learning],
8. The Jobs Of The Future --- [leadership].

2.8. Компонент «интеграция и веб-доступ ИСКАД ИИ». Веб-сайт является центральным компонентом, который облегчает управление данными, взаимодействие пользователей и получение информации из системы. Все компоненты ИСКАД ИИ взаимодействуют через веб-сайт, который предоставляет единый интерфейс для пользователей и обеспечивает согласованность данных и операций между компонентами системы.

Веб-сайт обеспечивает взаимодействие с компонентами «получение данных из интернет-источников», «графовая БД и граф знаний», «извлечение свойств из графовой БД и их анализ с помощью алгоритмов ML». Для пользователей системы веб-сайт реализует функцию регистрации пользователей; предоставляет доступ к просмотру публикаций различных предметных областей, поиску наиболее цитируемых авторов предметной области, просмотру параметров некоторой предметной области, просмотру различной информации об авторе публикации. Пользователь может получать гистограммы по авторам и статьям с применением алгоритма PageRank и др.

В системе также есть администраторы, которые обладают функциями управления ею. Разработка веб-сайта выполнена по классической двухзвенной клиент-серверной архитектуре, в которой клиентский компьютер взаимодействует напрямую с сервером без участия промежуточных узлов или компонентов. Клиент-серверная архитектура является распространенным подходом к разработке веб-приложений. Она представляет собой модель, в которой приложение разделяется на две основные составляющие: клиентскую и серверную.

2.9. Примеры работы веб-сайта. При входе на сайт пользователь с ролью «гость» попадает на стартовую страницу, которая содержит надпись BSUIR Science Work. В начале страницы есть кнопки перехода на страницу регистрации и авторизации. После регистрации и авторизации пользователь получает доступ к режимам выдачи отчетов и управления работой системы с помощью кнопок DATA, STATISTICS и DATA MANIPULATION. При нажатии кнопки DATA пользователь переходит на страницу со всеми статьями и авторами, которые находятся в графовой БД ИСКАД ИИ.

На страницах выдачи информации есть функция фильтрации по авторам и статьям, пользователь также может искать статью по ее названию, введя нужный текст в поле search. На рис. 8

отображены такие данные по статьям как, как заголовок статьи, год издания, краткое описание статьи, PageRank, издание, цитирование, на рис. 9 – информация о конкретной выбранной статье. Для просмотра авторов пользователь должен нажать на вкладку Author. Изначально показывается 10 авторов. Для получения более подробной информации об авторе нужно выделить его, и в сплывающем окне появится необходимая информация, например как на рис. 10, где указано, сколько статей из тех, которые имеются на сайте (в рассматриваемом случае одна), написал именно этот автор, и ниже они приведены. Чтобы убрать всплывающее окно с данными о статье, можно нажать на кнопку Close или на любое место затемненной области вокруг окна. Нажав на кнопку STATISTICS, пользователь попадает на страницу для сбора статистики по имеющимся на сайте данным.

uid	title	year	url	abstract	pagerank	venue	n_citation
97961	The multinotch, part IV: Extra precision via Δ coefficients	2022		In [1], we presented a new digital filter architecture, the multinotch, which minimized the computational latency while preserving numerical accuracy even in the presence of severe quantization. While this method is far more accurate than discretizing polynomial filters, it can still be susceptible to problems caused by a sample rate which is significantly higher than the frequencies of the features that the filter is trying to implement. This paper presents a modification, called Δ coefficients, which preserve all the positive properties of the multinotch while dramatically increasing the numerical accuracy over a large frequency range.	1.585258472082883	advances in computing and communications	50
132663	The multinotch, part X: A low latency, high numerical fidelity filter for mechatronic control systems	2023		Control of lightly damped mechatronic systems is often accomplished in practice with a PID-like controller in series with a filter to limit the effects of high frequency resonances. The high frequency filtering is often limited	1.585258472082883	advances in computing and communications	8

Рис. 8. Фильтрованные данные по статьям

Fig. 8. Filtered data by article

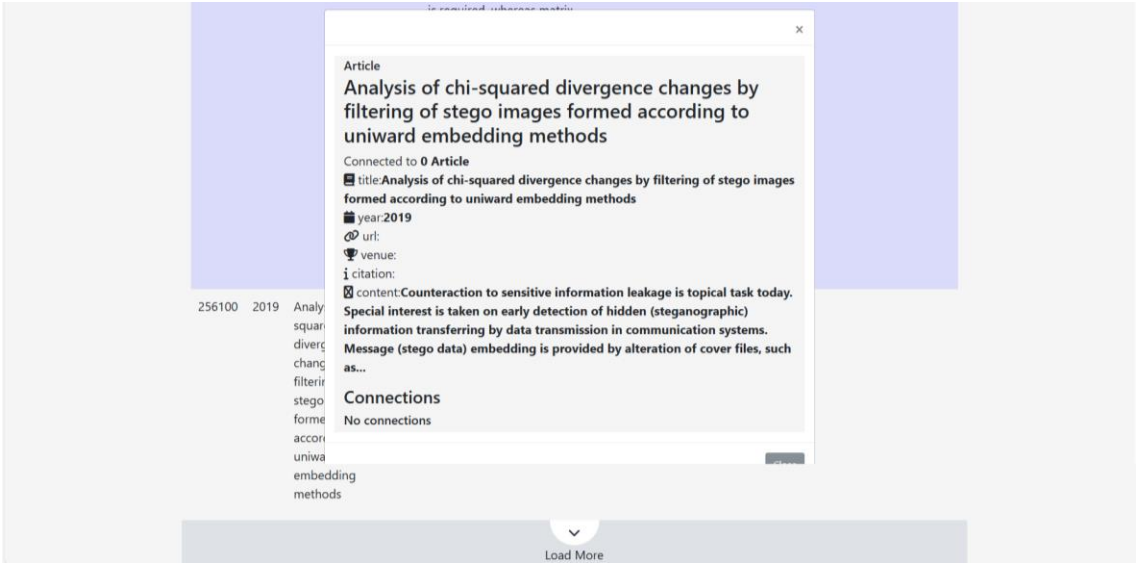


Рис. 9. Информация о выбранной статье

Fig. 9. Information about the selected article

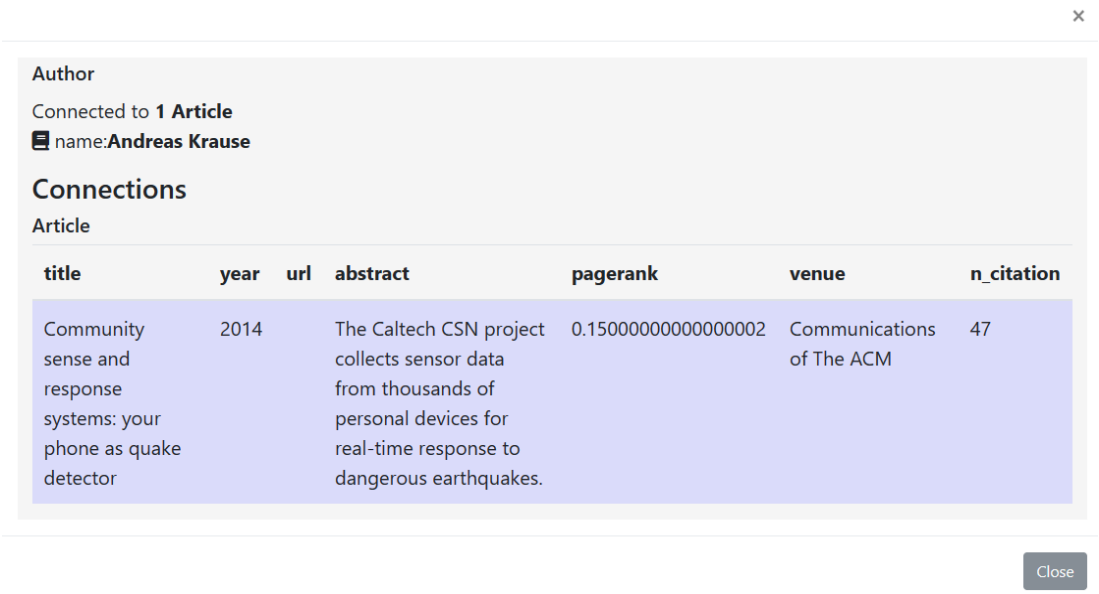


Рис. 10. Фильтрованные данные по автору

Fig. 10. Filtered data by author

Загружать данные пользователь может двумя способами. Первый способ – нажать на кнопку DATA MANIPULATION, затем на кнопку Choose file, выбрать нужный файл с расширением txt, где находятся ссылки на статьи, которые пользователь хочет добавить в графовую БД проекта, и нажать на кнопку SUBMIT. Второй способ – во второе поле вставить готовые ссылки на статьи, которые пользователь хочет добавить в графовую БД проекта, и нажать на кнопку SUBMIT.

Заключение. В статье разработана и апробирована комплексная технология последовательного применения взаимосвязанных методов, методологий и инструментов по построению графовой БД, графа знаний, анализа данных с использованием векторного преобразования графовых данных, методов и моделей машинного обучения и предоставления аналитических результатов пользователям. Создана и апробирована ИТ-среда для быстрого построения тематической графовой БД из данных сайтов и продемонстрировано применение графа знаний. Использована технология преобразования графов (графовых данных) в непрерывное низкоразмерное векторное представление, что позволяет анализировать содержимое графовых БД с помощью алгоритмов ML.

Представленная технология реализована в ИСКАД ИИ и применяется в БГУИР при анализе публикаций известных мировых сайтов, а также при проведении занятий с магистрантами. В дальнейшем при использовании ИСКАД ИИ необходимо предварительно проводить анализ загружаемых данных в графовую БД на их полноту. Не следует совмещать данные с различной структурой в одной графовой БД.

Вклад авторов. *И. И. Пилецкий* выполнил анализ предметной области, разработал методику и технологию быстрого протопирования тематических графовых БД, а также методологию углубленного анализа графовой БД. *М. П. Батура* руководил выполнением всего проекта, проанализировал полученные результаты на соответствие функциональным требованиям ИСКАД ИИ. *Н. А. Волорова* выполнила анализ интернет-источников предметной области, подготовила требования к разработке ИСКАД ИИ, организовала технологию реализации и тестирование системы. *П. А. Зорко* разработала компонент извлечения свойств из графовой БД и осуществила их анализ с помощью алгоритмов ML. *А. О. Кулевич* разработал ПО получения данных из интернет-источников, графовую БД и граф знаний.

Список использованных источников

1. Интеллектуальная система комплексного анализа данных интернет-источников / М. П. Батура [и др.] // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов VI Междунар. науч.-практ. конф., Минск, 20–21 мая 2020 г. : в 3 ч. Ч. 1 / редкол.: В. А. Богуш [и др.]. – Минск : Бестпринт, 2020. – С. 220–241.

2. Пилецкий, И. И. Графовые технологии в интеллектуальной системе комплексного анализа данных интернет-источников / И. И. Пилецкий, М. П. Батура, Л. Ю. Шилин // Доклады БГУИР. – 2020. – Т. 18, № 5. – С. 89–97.

3. Граф знаний и машинное обучение как ИТ-среда интеллектуального анализа данных интернет-источников / М. П. Батура [и др.] // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. науч. ст. VIII Междунар. науч.-практ. конф., Минск, 11–12 мая 2022 г. / Бел. гос. ун-т информатики и радиоэлектроники ; редкол.: В. А. Богуш [и др.]. – Минск, 2022. – С. 330–344.

4. Diestel, R. Graph Theory / R. Diestel. – Berlin : Springer-Verlag, 2017. – 448 p.

5. Needham, M. Graph Algorithms / M. Needham, A. E. Hodler. – Sebastopol : O'Reilly Media, 2019. – 265 p.

6. Hamilton, W. L. Representation learning on graphs: Methods and applications / W. L. Hamilton, R. Ying, J. Leskovec // IEEE Data Engineering Bulletin. – 2017. – Vol. 40, no. 3. – P. 52–74.

7. Ovcinnikova, J. Visual diagrammatic queries in ViziQuer: Overview and implementation / J. Ovcinnikova, A. Sostaks, K. Cerans // Baltic J. of Modern Computing. – 2023. – Vol. 11, no. 2. – P. 317–350.

8. Portisch, J. Knowledge graph embedding for data mining vs. knowledge graph embedding for link prediction – two sides of the same coin? / J. Portisch, N. Heist, H. Paulheim // Semantic Web. – 2022. – Vol. 13, no. 3. – P. 399–422. https://doi.org/10.3233/SW-212892

9. Barrasa, J. Knowledge Graphs / J. Barrasa, A. E. Hodler, J. Webber. – Sebastopol : O'Reilly Media, 2021. – 85 p.

References

1. Batura M. P., Piletski I. I., Prytkov V. A., Volorova N. A. *Intelligent system for comprehensive analysis of data from Internet sources.* BIG DATA i analiz vysokogo urovnja : sbornik materialov VI Mezhdunarodnoj nauchno-prakticheskoj konferencii, Minsk, 20–21 maja 2020 g. : v 3 chastjah. Chast' 1 [*BIG DATA and Advanced Analytics : Collection of Materials of the VI International Scientific and Practical Conference, Minsk, 20–21 May 2020 : in 3 Parts. Part 1*]. Ed. board: V. A. Bogush [et al.]. Minsk, Bestprint, 2020, pp. 220–241 (In Russ.).

2. Piletski I. I., Batura M. P., Shilin L. Yu. *Graph technologies in an intelligent system for complex analysis of data from Internet sources*. Doklady Belorusskogo gosudarstvennogo universiteta informatiki i radioèlektroniki [*Doklady BGUIR*], 2020, vol. 18, no. 5. pp. 89–97 (In Russ.).

3. Batura M. P., Piletsky I. I., Volorova N. A., Zorko P. A., Kulevich A. O. *Knowledge graph and machine learning as an IT environment for mining data from Internet sources.* BIG DATA i analiz vysokogo urovnja : sbornik nauchnyh statej VIII Mezhdunarodnoj nauchno-prakticheskoj konferencii, Minsk, 11–12 maja 2022 g. [*BIG DATA and Advanced Analytics : Collection of Scientific Articles of the VIII International Scientific and Practical Conference, Minsk, 11–12 May 2022*]. Ed. board: V. A. Bogush [et al.]. Minsk, 2022, pp. 330–344 (In Russ.).

4. Diestel R. *Graph Theory*. Berlin, Springer-Verlag, 2017, 448 p.

5. Needham M., Hodler A. E. *Graph Algorithms*. Sebastopol, O'Reilly Media, 2019, 265 p.

6. Hamilton W. L., Ying R., Leskovec J. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 2017, vol. 40, no. 3, pp. 52–74.

7. Ovcinnikova J., Sostaks A., Cerans K. Visual diagrammatic queries in ViziQuer: Overview and implementation. *Baltic Journal of Modern Computing*, 2023, vol. 11, no. 2, pp. 317–350.

8. Portisch J., Heist N., Paulheim H. Knowledge graph embedding for data mining vs. knowledge graph embedding for link prediction – two sides of the same coin? *Semantic Web*, 2022, vol. 13, no. 3, pp. 399–422. https://doi.org/10.3233/SW-212892

9. Barrasa J., Hodler A. E., Webber J. *Knowledge Graphs*. Sebastopol, O'Reilly Media, 2021, 85 p.

Информация об авторах

Пилецкий Иван Иванович, кандидат физико-математических наук, доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники.

Батура Михаил Павлович, доктор технических наук, профессор, заведующий лабораторией «Новые обучающие технологии» Белорусского государственного университета информатики и радиоэлектроники.

Волорова Наталья Алексеевна, кандидат технических наук, доцент, старший научный сотрудник лаборатории «Новые обучающие технологии» Белорусского государственного университета информатики и радиоэлектроники.

Зорко Полина Александровна, магистрант кафедры информатики Белорусского государственного университета информатики и радиоэлектроники.

Кулевич Алексей Олегович, магистрант кафедры информатики Белорусского государственного университета информатики и радиоэлектроники.

Information about the authors

Ivan I. Piletski, Ph. D. (Phys.-Math.), Assoc. Prof. of the Department of Informatics of Belarusian State University of Informatics and Radioelectronics.

Michal P. Batura, D. Sc. (Eng.), Prof., Head of the Laboratory "New Educational Technologies" of Belarusian State University of Informatics and Radioelectronics.

Natalia A. Volorova, Ph. D. (Eng.), Assoc. Prof., Senior Researcher of the Laboratory "New Educational Technologies" of Belarusian State University of Informatics and Radioelectronics.

Polina A. Zorko, Master's Student of the Department of Informatics of Belarusian State University of Informatics and Radioelectronics.

Alexei O. Kulevich, Master's Student of the Department of Informatics of Belarusian State University of Informatics and Radioelectronics.