http://dx.doi.org/10.35596/1729-7648-2020-18-5-89-97

Оригинальная статья
Original paper

УДК 519.684.6;004.021

ГРАФОВЫЕ ТЕХНОЛОГИИ В ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЕ КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ ИНТЕРНЕТ-ИСТОЧНИКОВ

ПИЛЕЦКИЙ И.И., БАТУРА М.П., ШИЛИН Л.Ю.

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

Поступила в редакцию 9 июля 2020

Аннотация. Целью работы, изложенной в статье, является рассмотрение и демонстрация применения графовых технологий для глубокого анализа данных. В статье рассматривается интеллектуальная система комплексного анализа данных интернет-источников и возможные направления ее дальнейшего развития. Данная система представляет собой многоцелевой кластер с использованием технологий построения графа знаний, методов и моделей машинного обучения для глубокого анализа данных интернет-источников (например, научных публикаций, социальных сетей, СМИ). Целью анализа является выявление наиболее важных публикаций в некоторой области (например, в робототехнике, космических исследованиях, здравоохранении, в социальной сфере), тематический анализ этих публикаций, выявление лидера научного направления, предсказание тенденций развития направлений и взаимодействия групп людей. При разработке данной системы были применены вероятностные алгоритмы машинного обучения и методы построения и обслуживания графовой модели социальной сети авторов и их публикаций, определение рейтинга конкретного автора публикаций, определение тематик публикаций и классификация их по областям знаний. Основой для создания интеллектуальных приложений являются графовые технологии, которые позволяют делать более точные прогнозы. Совместное применение методов и алгоритмов машинного обучения с графовыми технологиями позволяет получать скрытые зависимости и выполнять предиктивный анализ информации, получать ответы в режиме реального времени, реализовывать алгоритмы искусственного интеллекта. В основу методов совместной работы с графовыми технологиями и машиного обучения (например, применение нейронных сетей) положен графовый эмбеддинг. Данная технология позволяет выполнять всесторонний, глубокий и интеллектуальный анализ информации. Приведены аналитические отчеты, полученные с помощью графовых технологий в интеллектуальной системе комплексного анализа данных интернет-источников.

Ключевые слова: интернет-источники, мониторинг, машинное обучение, обработка естественного языка, графовые базы данных, графовые алгоритмы, pagerank.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Пилецкий И.И., Батура М.П., Шилин Л.Ю. Графовые технологии в интеллектуальной системе комплексного анализа данных интернет-источников. Доклады БГУИР. 2020; 18(5): 89-97.

GRAPHIC TECHNOLOGIES IN AN INTELLIGENT SYSTEM OF COMPLEX ANALYSIS OF DATA FROM INTERNET SOURCES

IVAN I. PILETSKI, MIKHAIL P. BATURA, LEANID Y. SHYLIN

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

Submitted 9 July 2020

Abstract. The purpose of the work outlined in the article is to review and demonstrate the use of graph technologies for deep data analysis. The first part of the article discusses the Intelligent System for the Comprehensive Analysis of Internet Sources Data and its possible directions for its further development. This system is a multi-purpose cluster using technologies for constructing a knowledge graph, methods and models of machine learning for in-depth analysis of data from Internet sources (for example, scientific publications, social networks, media). The purpose of the analysis is to identify the most important publications in a certain area (for example, in robotics, space research, healthcare, in the social sphere), thematic analysis of these publications, to identify the leader of a scientific direction and to predict trends in the development of directions and interaction of groups of people. When developing this system, we utilized probabilistic machine learning algorithms and methods for constructing and maintaining a graph model of the social network of authors and their publications, determining the rating of a particular author, determining the topics of publications and classifying them by areas of knowledge. The basis for the creation of intelligent applications is graph technology, which allows you to make predictions that are more accurate. The combined application of methods and algorithms of machine learning with graph technologies allows you to get hidden dependencies and perform predictive analysis of information, get answers in real time, and implement artificial intelligence algorithms. Methods of collaboration with graph technologies and a learning machine (for example, using neural networks) are based on graph embedding. This technology allows you to perform a comprehensive, deep and intelligent analysis of information. At the end of the article, there are analytical reports obtained using graph technologies in the Intelligent System for Complex Analysis of Internet Sources Data.

Keywords: Internet sources, monitoring, data analysis, Machine Learning, Natural Language Processing, graph databases, graph algorithms, pagerank, graph technologies.

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

For citation. Piletski I.I., Batura M.P., Shylin L.Y. Graph technologies in an intelligent system of complex analysis of data from Internet sources. Doklady BGUIR. 2020; 18(5): 89-97.

Что такое ИСКАД ИИ?

Интеллектуальная система анализа данных интернет-источников (ИСКАД ИИ) предназначена для поддержки принятия обоснованных решений на основе *мониторинга и анализа данных* из открытых интернет-источников, обеспечения достоверности цифровой информации. Данная система представляет собой многоцелевой кластер с использованием технологий построения графа знаний, методов и моделей машинного обучения (Machine Learning – ML) для глубокого анализа данных интернет-источников (например, научных публикаций, социальных сетей, СМИ). Целью анализа является выявление наиболее важных публикаций в некоторой области (например, в робототехнике, космических исследованиях, здравоохранении, в социальной сфере), тематический анализ этих публикаций, выявление лидера научного направления, предсказание тенденций развития направлений и взаимодействия групп людей. В данной статье частично используются результаты, ранее полученные авторами и опубликованные в материалах международной конференции[1].

[1] Батура М.П., Пилецкий И.И., Прытков В.А., Волорова Н.А. Интеллектуальная система комплексного анализа данных интернет-источников: сб. материалов VI Междунар. науч.-практ. конф. Минск: БГУИР; 2020;1:220-241.

Анализируя данные из социальных сетей, можно выявить как прямые, так и скрытые отношения между людьми, группами людей, а также характер их взаимодействий. Примеры задач: выявление лидера профессионального или 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.

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.

Основным назначением графовой базы данных (БД) является применение графовых алгоритмов для обработки полученных данных, выстраивание логических взаимосвязей и подготовка и выдача информации для пользователя. Сама графованая БД обладает рядом преимуществ по сравнению с другими БД, она обладает свойствами OLTP и 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.

В ИСКАД ИИ применена новая архитектура построения многофункциональных комплексов как набор постоянно работающих компонент в виде отдельных серверов. Все взаимосвязи и взаимодействия компонент организованы на основе специально разработанного управляющего компонента (универсальней шины) ИСКАД ИИ. Данный компонент обладает функциональностью интеграции данных и приложений, реализовывает функции брокера, синхронного и асинхронного выполнения приложений. Управляющий компонент реализовывает средства логирования, сбора статистики и мониторинга работы компонент системы. Компонент скачивания публикаций и компонент извлечения текста из материалов использует технологию Docker, инструменты мониторинга и предупреждения ошибок, применяет технологию извлечения медиа-данных из документов и обработки новых форматов документов (PostScript, заархивированные документы). Компонент графовая БД и граф знаний позволяет анализировать и выдавать информацию в аналитическом виде для принятия обоснованных решений. Компонент хранилища данных разработан на основе применения NoSQL СУБД, который может содержать различные тематические БД, соответствующие новым областям применения системы. Взаимодействие хранилища с графовой БД позволяет перестраивать и перезагружать граф знаний, иметь множественное представление графа знаний для различных областей применения. Библиотека аналитических модулей может пополняться новыми ML-модулями интеллектуального анализа данных.

ИСКАД ИИ позволяет не только помочь с проведением анализа и принятием решения, но и дает возможность 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.

Графовые технологии как база интеллектуального анализа информации

В настоящее время резко возрос интерес к применению графовых алгоритмов и аналитики для различных областей человеческой деятельности. Графовые технологии – это основа для создания интеллектуальных приложений, позволяющих делать более точные прогнозы и быстрее принимать решения. Графы лежат в основе широкого спектра вариантов использования искусственного интеллекта (ИИ). 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).

Граф знаний – одна из основных областей ИИ, который позволяет понимать предписывающую аналитику и приложения ИИ (например, обработка и понимание естественного языка (Natural Language Processing – NLP, Natural-language understanding – NLU), определение PageRank).

В общем случае огромное количество графовых алгоритмов классифицированы на алгоритмы: Pathfinnding, Centrality и Community Detection [1, 2]. Примеры классификации, визуализации и преобразования эмбеддинга можно найти на сайтах[2].

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.

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 появление новых тенденций в этих группах, выявлять уязвимости и возможные цели атаки в сетях связи и транспорта.

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

Есть ряд задач, для решения которых может быть применена ML и различные алгоритмы без учителя, например, вероятностные алгоритмы или нейронная сеть, которые взаимодействуют с графами.

Совместное использование информации графовых моделей и ML позволяют получать скрытые зависимости и выполнять предиктивный анализ информации, получать ответы в режиме реального времени, реализовывать алгоритмы ИИ, отслеживать решения ИИ. В настоящее время алгоритмы ИИ широко распространены для решения конкретной задачи, 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. Процесс глубокого обучения использует глубокие искусственные нейронные сети и ML в качестве моделей. В основу методов совместной работы с графами и ML технологиями (например, применение нейронных сетей) положен графовый эмбеддинг.

Графовый эмбеддинг это представление узлов и отношений в графе как вектор свойств. Графовые эмбеддинги – это способ представления графов для задач машинного обучения с помощью функции преобразования. В качестве значений вектора свойств могут быть выбраны некоторые атрибуты вершин и отношений. В зависимости от поставленной задачи эти свойства или атрибуты узлов и ребер могут быть разными. Данные ML технологии могут быть комбинированы с другими методами для улучшения аналитических результатов и нахождения скрытых зависимостей.

При работе с такими технологиями используются функции энкодер и декодер, функции кодирования и восстановления списка ребер и вершин по полученному представлению графа. Функция декодер позволяет визуализировать полученный результат в графическом виде, а также делать предсказание. Например, для взаимодействия графов и нейросетей можно использовать методы, которые находятся в свободном доступе: DeepWalk (word2vec), Node2Vec, 2D CNN, Graph Convolutional Networks.

Для построения преобразования эмбеддинга вершин необходимо:

– задать функцию соответствия преобразования узла u в вектор R^d;

– определить функцию подобия узлов, меру близости в графе (например, скалярному произведению двух узлов);

– оптимизировать параметры функции подобия.

Для построения преобразования эмбеддинга для ребер нужно задать функцию, которая для любой пары вершин u и v построит векторное представление R^d вне независимости от их связности на графе. Например, это может быть произведение Адамара или среднее

[2]Материалы по построению эмбеддинга доступны на сайтах: https://arxiv.org/abs/1709.05584, https://arxiv.org/abs/1708.02218, https://arxiv.org/abs/1609.02907.

арифметическое. На рис. 1 приведен пример многоуровневой схемы для моделирования связей между белками в разных органах (а они ведут себя по-разному в зависимости от местоположения) [3].

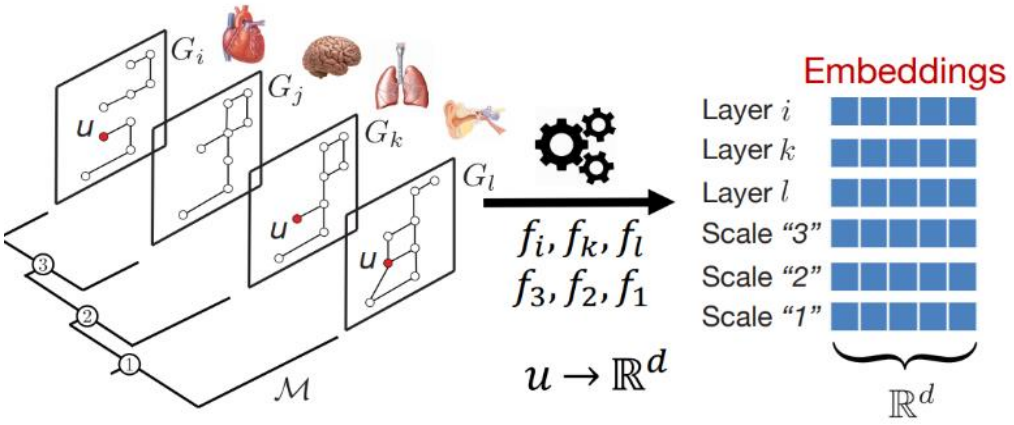


Рис. 1. Многоуровневое преобразование графовой модели в векторное представление
Fig. 1. Multilevel transformation of a graph model into a vector representation

Но при построении преобразования эмбеддинга для нейронных сетей существует проблема – это стандартная размерность входа. При работе с графами количество вершин может быть произвольным, поэтому необходимо преобразовывать матрицу связности до заданной размерности.

Применение графовых технологий в ИСКАД ИИ

В данном разделе приведена небольшая часть результатов работы ИСКАД ИИ с учетом важности публикаций и их тематик.

Графовая база данных ИСКАД ИИ содержит следующие сущности (см. рис. 2.):

– Author – тот, кто опубликовал статью;

– Publication – публикация, подготовленная автором. Содержит данные о публикации и ссылки на темы. Имеет ссылки на статьи, на которые ссылаются (LINKS_TO), необходимые для вычисления PageRank[3] и определения эксперта в некоторой области знаний;

– Theme – тема, к которой может относиться публикация (темы публикаций определяются вероятностными алгоритмами ML);

– Token – сущность, которая представляет себя уникальным именем. Имеется возможность просмотреть все публикации, которые имеют в своем тексте вхождение ключевой фразы (хранится в связи FREQUENCY).

Объекты описанных сущностей связываются между собой следующими отношениями:

– WROTE – связывает автора и его статью;

– THEME_RELATION – отношение, обозначающее связь публикации с определенной тематикой;

– LINKS_TO – отношение между двумя объектами публикаций, обозначающее, что в тексте одной публикации есть ссылка на текст другой публикации;

– FREQUENCY – отношение, связывающее публикацию и токен. В дополнительном свойстве «entry_count» сохраняется число вхождений данного токена в текст публикации.

Схема структуры базы данных компонента графа знаний представлена на рис. 2.

[3]Алгоритм вычисления PageRank доступен на официальном сайте: neo4j https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/.

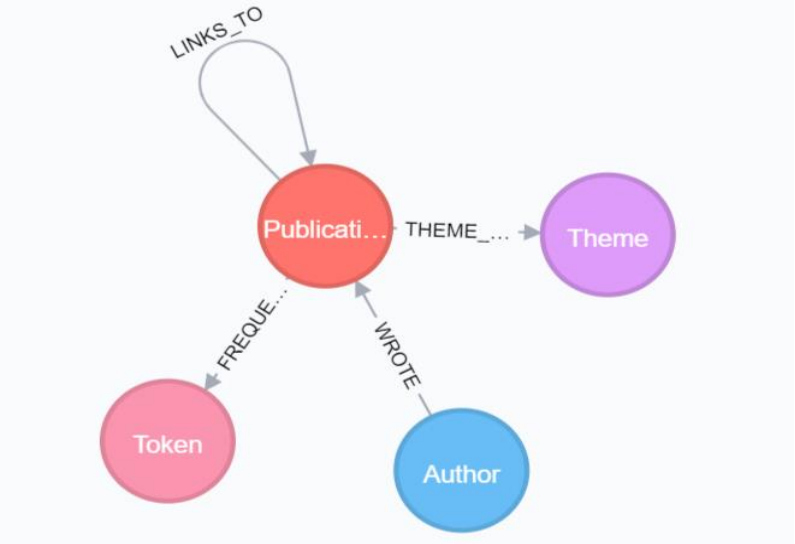
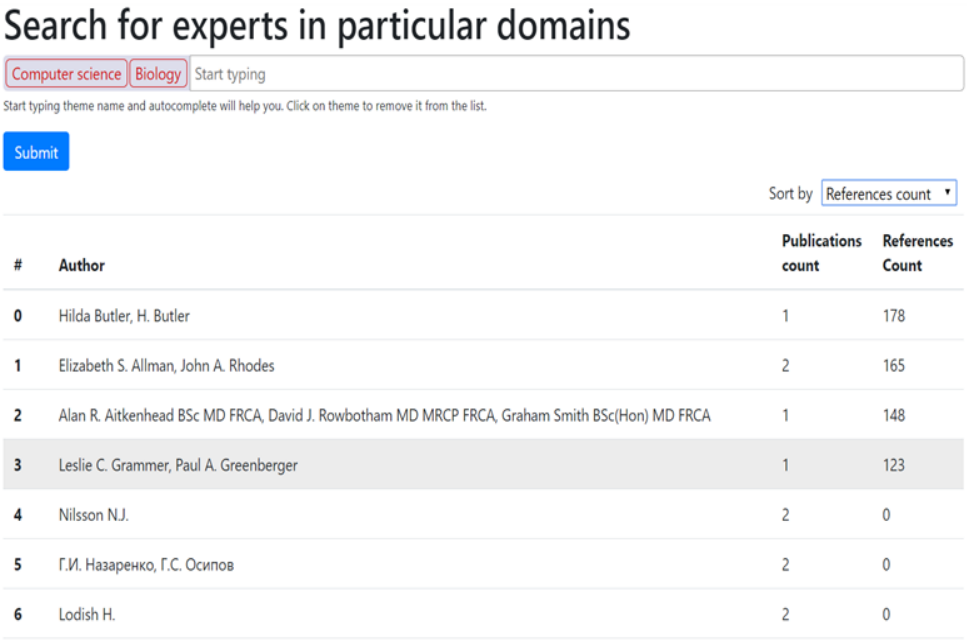


Рис. 2. Структура БД графа знаний
Fig. 2. Knowledge Graph Database Structure

Графовая модель позволяет получать знания о публикациях в различных аспектах, например, связанных с тематикой Computer science, Biology или Big Data. В таких запросах важно указывать порог вероятности тематики в статьях больше некоторой величины. Ниже приведены примеры получения аналитической информации из графовой базы данных. Необходимо отметить важное свойство всех отчетов – это drill down, т. е. информацию в отчете можно уточнять по каждой позиции простым нажатием курсора, по каждому узлу графа знаний. Данные получены с сайта, который используется для публикаций научных работ: https://arxiv.org, http://libgen.io, http://gen.lib.rus.ec/.

Поиск авторов, которых можно было бы назвать экспертами в каких-то областях знаний, основан на использовании рассчитанного для каждой публикации коэффициента PageRank, количества публикаций авторов и количества ссылок на публикации отдельного автора (рис. 3), динамика публикаций в областях знаний приведена на рис. 4.



#	Author	Publications count	References Count
0	Hilda Butler, H. Butler	1	178
1	Elizabeth S. Allman, John A. Rhodes	2	165
2	Alan R. Aitkenhead BSc MD FRCA, David J. Rowbotham MD MRCP FRCA, Graham Smith BSc(Hon) MD FRCA	1	148
3	Leslie C. Grammer, Paul A. Greenberger	1	123
4	Nilsson N.J.	2	0
5	Г.И. Назаренко, Г.С. Осипов	2	0
6	Lodish H.	2	0

Рис. 3. Выбор «экспертов» предметной области
Fig. 3. The choice of subject area experts

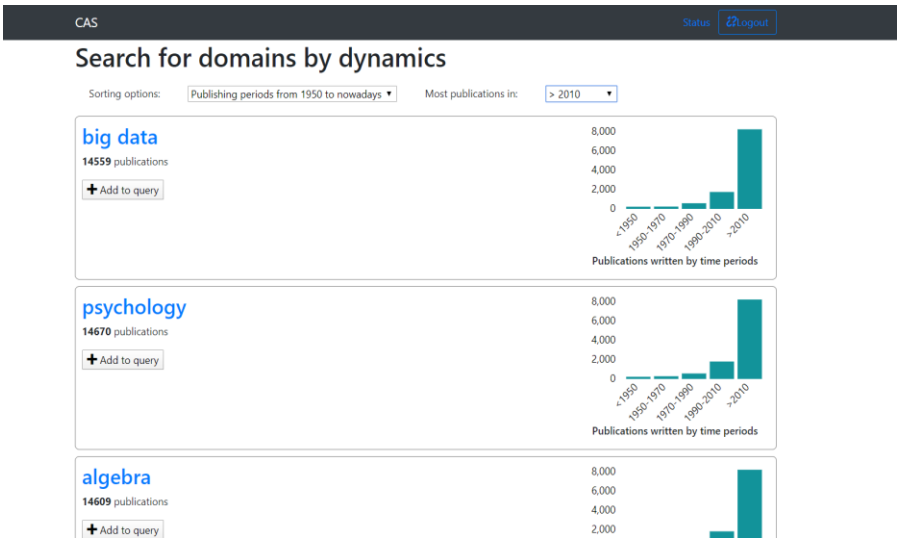


Рис. 4. Области знаний и число публикаций
Fig. 4. Knowledge Areas and Number of Publications

Информация об области знаний название области знаний, десять наиболее цитируемых публикаций, относящихся к ней, десять авторов, имеющих наибольшее число публикаций в данной области знаний, приведена на рис. 5.

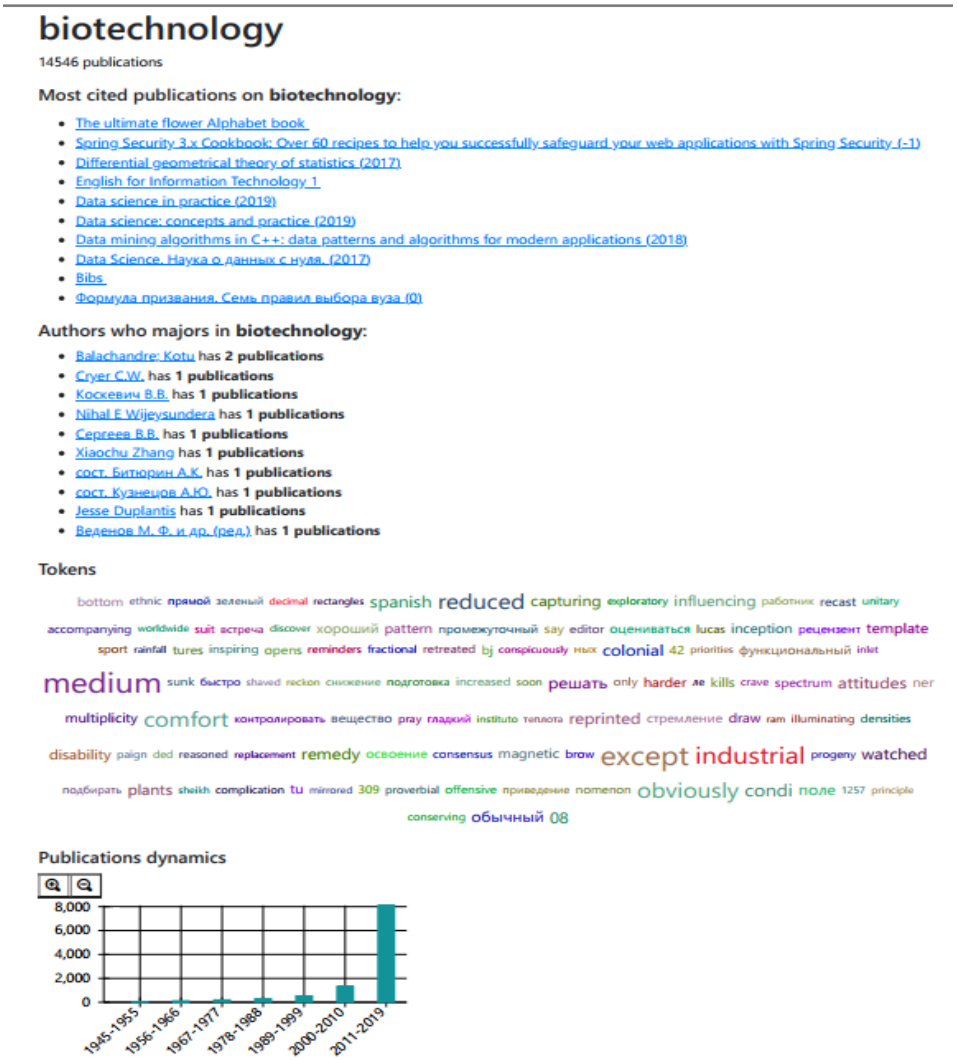


Рис. 5. Подробная информация о области знаний
Fig. 5. Detailed information about the field of knowledge

На рис. 6, 7 совмещены текстовая информация и граф, являющийся фрагментом графа знаний. На рис. 8 приведена подробная информация о публикациях отдельного автора. Важно еще раз отметить, что вся информация может быть уточнена с помощью техники drill down.

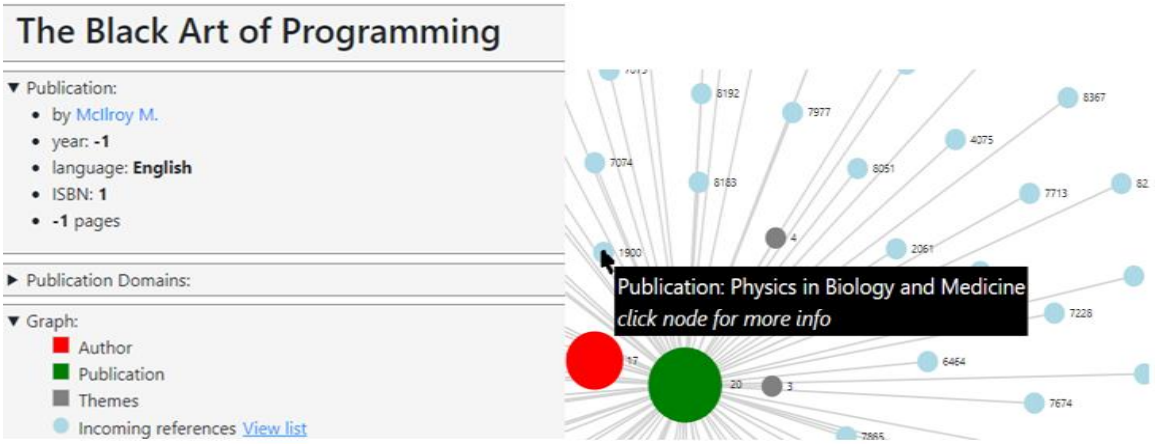


Рис. 6. Информация о конкретной публикации
Fig. 6. Information about a particular publication

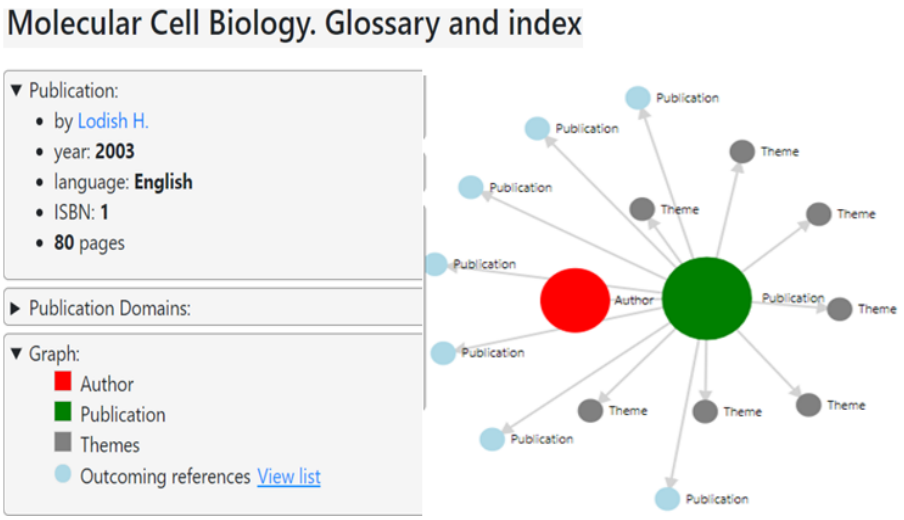


Рис. 7. Страница отдельной публикации
Fig. 7. Separate Publication Page

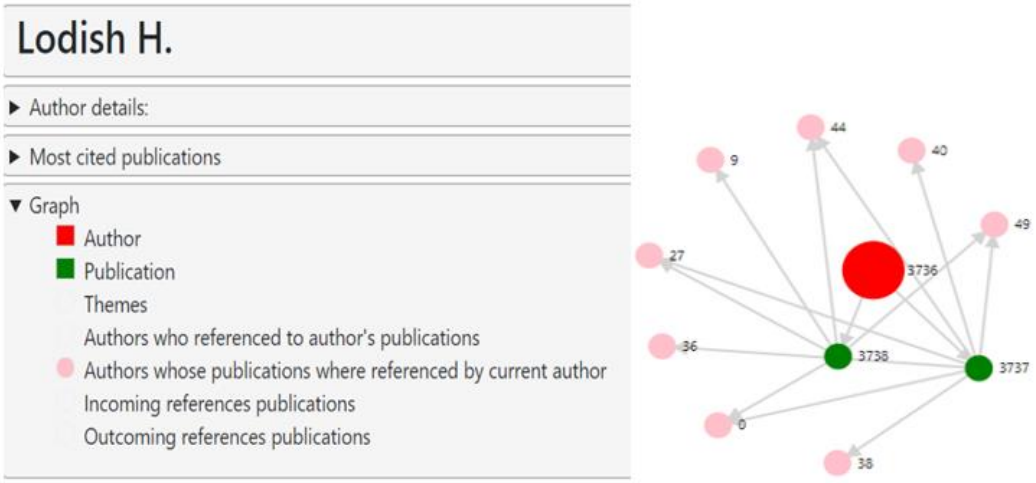


Рис. 8. Страница информации отдельного автора
Fig. 8. Page Information of Individual Author

Заключение

В статье рассмотрены графовые технологии в ИСКАД ИИ, целью создания которой является выявление экспертов в некоторой предметной области, определение тематик публикаций, оценка их популярности. В современном мире крайне важным фактором является время, поэтому быстрое получение достоверного источника информации для принятия правильного решения является весьма актуальным. Проанализированы графовые технологии для создания интеллектуальных приложений на основе совместного применения графовых алгоритмов и ML с целью применения их в ИСКАД ИИ, что позволяет делать аналитические прогнозы и принимать более точные решения. Приведены реализованные технологические решения в проекте ИСКАД ИИ и показаны полученные аналитические результаты. Приведенные результаты работы 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. Объем статьи не позволяет продемонстрировать более богатые возможности системы. Авторы статьи выражают благодарность студенту кафедры информатики БГУИР Н.Н. Черныш за помощь в подготовке рисунков.

References

1. Diestel R. *Graph Theory*. Berlin: Springer-Verlag; 2017.
2. Needham M., Hodler Amy E. *Graph Algorithms*. Sebastopol: O'Reilly Media; 2019.
3. Hamilton W.L., Rex Ying, Leskovec J. *Representation Learning on Graphs: Methods and Applications*. Stanford: Stanford University; 2017; 9:1-25.

Сведения о вкладе авторов

Все авторы в равной степени внесли вклад в написание статьи.

Author contribution

All authors equally contributed to the writing of the article.

Сведения об авторах

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

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

Шилин Л.Ю. д.т.н., профессор, декан факультета информационных технологий и управления Белорусского государственного университета информатики и радиоэлектроники.

Information about the authors

Piletski I.I., PhD, Associate Professor of the Department of Informatics Department of Belarusian State University of Informatics and Radioelectronics.

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

Shylin L.Y., D.Sci., Professor, Dean of the Faculty of Information Technologies and Control of Belarusian State University of Informatics and Radioelectronics.

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6,
Белорусский государственный университет
информатики и радиоэлектроники
тел. +375-29-632-32-35;
e-mail: bmpbel@bsuir.by
Батура Михаил Павлович

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovka, str., 6,
Belarusian State University
of Informatics and Radioelectronics
tel. +375-29-632-32-35;
e-mail: bmpbel@bsuir.by
Batura Mikhail