УДК 004.89

ГРАФ ЗНАНИЙ И МАШИННОЕ ОБУЧЕНИЕ КАК БАЗИС МЕТОДОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОБУЧЕНИИ

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

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

Н.А. Волорова
Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент

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

И. И. Пилецкий
Кандидат физико-математических наук, доцент БГУИР. В сфере IT более 49 лет. Участие в разработке нескольких десятков крупных проектов: главный конструктор проекта, главный архитектор программно-информационного обеспечения, руководитель проекта, начальник отдела, заведующий лабораторией (НИИ ЭВМ, Академия наук Беларуси, IBA, БГУИР). Автор десятков исследований, имеет более 106 публикаций.

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

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

Аннотация. Графовые технологии – это основа для создания интеллектуальных приложений. Граф знаний – одна из основных областей ИИ, который позволяет понимать предписывающую аналитику и приложения ИИ. Совместное применение графовых технологий, методов и алгоритмов машинного обучения позволяет получать скрытые зависимости и выполнять предиктивный анализ информации, получать ответы в режиме реального времени, реализовывать алгоритмы искусственного интеллекта. В статье приводятся методы построения графа знаний и применение машинного обучения при подготовке магистрантов по тематике «Обработка больших объемов информации», а также для получения экспертных данных при проведении исследовательских работ в университете.

Ключевые слова: интернет-источники, Big Data, Machine Learning, машинное обучение, Neo4j, Natural Language Processing, обработка естественного языка, графовые базы данных, графовые алгоритмы.

Введение.

В настоящее время основным источником данных и как следствие информации является интернет, интернет-источники. Данные могут быть получены как из социальных сетей, так и тематических сайтов (газет, журналов, библиотек, компаний и т. д.), содержащих различные публикации.

Пользователи интернет-ресурсов и социальных сетей могут самостоятельно выбирать интересные им направления и читать публикации интересных им людей, которым они симпатизируют. Связи пользователей интернет ресурсов, как правило, бывают достаточно сложными и представляют собой многоуровневую циклическую сеть.

Графы – один из самых мощных и гибких способов представления данных. Они обладают большой выразительной силой, то есть графы могут использоваться для моделирования большого числа систем в различных областях, включая социальные сети, физические системы [1, 2], графы знаний [3] и многие другие области исследований. Граф – это универсальная и выразительная структура, позволяющая моделировать всевозможные сценарии, от постройки космической ракеты до строительства системы дорог, от поставок продуктов питания до историй болезни населения, и многое другое.

Сама графованая база данных обладает рядом преимуществ и достоинств по сравнению другими БД. Как и реляционные БД, графовая БД обладает свойствами OLTP (On-Line Transaction Processing) и OLAP (On-Line Analytical Processing).

Графовые технологии обеспечивают организации транзакционных графовых хранилищ, интеллектуальный анализ данных и аналитическую обработку данных в реальном времени, поддерживают транзакции ACID (atomic, consistent, isolated и durable), что не обеспечивает ни одна NoSQL БД. Графовые технологии являются основой для построения интеллектуальных приложений, для применения алгоритмов искусственного интеллекта.

Основным назначением графовой базы данных является применение графовых алгоритмов для обработки полученных данных, выстраивание логических взаимосвязей и подготовка и выдача информации для пользователя.

Совместное применение графовых технологий, методов и алгоритмов машинного обучения 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 машинного обучения (например, применение нейронных сетей) положен графовый эмбеддинг. Данная технология позволяет выполнять всесторонний, глубокий и интеллектуальный анализ информации.

Графовые DB VS NoSQL Баз данных VS RDBMS.

2.1. Реляционная модель позволяет собирать данные, которые мы хотим хранить и анализировать, и разделяет ее на кортежи (строки) из которых строятся сущности (таблицы) которые в свою очередь связываются отношениями в зависимости от модели предметной области. Как правило строки хранят атомарные простые данные. Поэтому все операции можно интерпретировать как операции над кортежами и возвращение кортежей.

Реляционные базы данных позволяют манипулировать любой комбинацией строк из любой таблицы в рамках одной транзакции. Такие транзакции называются ACID: атомарные (atomic – все операции в транзакции либо успешны, либо для всех них выполняется откат), согласованные (consistent – по окончании транзакции база данных является структурно согласованной), изолированные (isolated – транзакции не мешают друг другу) и долговечные (durable – результаты применения транзакции не должны теряться, даже при сбоях). Эти свойства означают, что сразу по завершении транзакции ее данные согласуются и записываются на диск.

Рост связности для RDBMS приводит к увеличению соединений, которые снижают производительность и затрудняют внесение в базу данных обновлений. Поэтому при моделирования социальной сети, запросы будут становиться все более сложными и медленными, что может привести к полной деградации БД [4, 5].

2.2. NoSQL БД типа «ключ-значение», документные базы данных и семейства столбцов представляют собой агрегатно-ориентированные базы данных. Агрегат – это коллекция связанных объектов, которая интерпретируется как единое целое. Агрегаты облегчают работу баз данных на кластерах, поскольку агрегат представляет собой естественную единицу репликации и фрагментации. Кроме того, агрегаты упрощают разработку прикладных программ, которые часто

манипулируют данными с помощью агрегированных структур [6].

Для NoSQL вместо свойств ACID ввели свойства BASE: хранилище доступно большую часть времени (Basic availability), хранилище не обязано соблюдать очередность записей, и разные реплики не должны постоянно согласовываться (Soft-state), хранилище достигает согласованности с некоторой задержкой по времени (Eventual consistency). Принципы BASE значительно слабее гарантий ACID, и между ними нет прямого соответствия [6].

Значения в BASE-хранилище доступны (потому что это является основой масштабирования), но это не предлагает гарантированной согласованности реплик при записи.

2.3. Графовые базы данных позволяют хранить сущности и отношения между ними. Сущности моделируются узлами (nodes), которые имеют свойства. Узел интерпретируется как экземпляр объекта в приложении. Отношения моделируются ребрами (relationships or edges), которые могут иметь свойства. Ребра имеют направление; узлы организованы в соответствии с отношениями. БД отдельно хранит в специальном жестком формате структуру узлов и отношений, а данные (свойства) определяются как пара ключ-значение. Такое решение в базе Neo4J позволяет извлекать данные без обхода графа, с определенного узла (смещение) и по ключу извлекать данные (свойства).

В базе Neo4J которая полностью поддерживает транзакции ACID, данные всегда являются согласованными. Если база Neo4J работает на кластере, запись на ведущий узел синхронизируется с ведомыми узлами, которые всегда доступны для чтения. Операции записи на ведомые узлы всегда доступны и немедленно синхронизируются с ведущим узлом; остальные ведомые узлы не синхронизируются немедленно – они ждут, пока данные не будут распределены с ведущего узла. Графованая база данных, как и реляционные БД обладает свойствами OLTP и OLAP.

2.4. Пример.

Популярный пример заказа и доставки продуктов. Нужно учитывать наличие различной продукции на складе и ее доставку заказчику, в данном случае важен контент не только продуктов, но и заказчика, а также способы поставки продукции и доставки заказчику. Это не простая логистическая задача, которая традиционно решается с помощью RDBMS. Обычно для решения такой задачи достаточно 5-10 сущностей. Модель шаблона такой задачи приведен на рисунке 1.

На рисунке 2 приводится БД и модель для решения данной задачи с использованием графа знаний и графовой аналитики. Более полно миграцию RDBMS можно посмотреть на сайте Neo4j [7]. Сама графовая БД легко модифицируется в процессе эксплуатации. На данном рисунке приведена модель БД и слева ее сущности (узлы), отношения (связи) и атрибуты (свойства). Сами значения свойств хранятся отдельно. Доступ к данным обеспечивается с помощью не сложного декларативного SQL-подобного языка запросов Cypher [8].

Например, на языке запросов Cypher, создание отношения между двумя сущностями аналогичном как RDBMS приведено ниже:

MATCH (p: Product), (s: Supplier) WHERE p.supplierID = s.supplierID.

CREATE (s) - [: SUPPLIES] -> (p).

Выдача поставщиков продукции:

MATCH (c: Category {categoryName:»Produce»})<-- (: Product) <-- (s: Supplier).

RETURN DISTINCT s.companyName as ProduceSuppliers.

Результат: ProduceSuppliers.

«G'day», «Tokyo Traders», «Plutzer Lebensmittelgroßmärkte AG», «Mayumi's», «Grandma Kelly's Homestead».

Здесь важно отметить что, это довольно простая задача с точки зрения схемы RDBMS: 5-10 сущностей, связанных ключами. Но если для решения задачи необходимо использовать 100, 200 или более сущностей [4, 5], то сразу возникают следующие проблемы:

– Как отобразить объектную модель кода приложения на реляционную модель БД? Для этого необходимо применять специальные технологии программирования.

– Как обеспечить модификацию кода приложения и БД? При любых изменениях нужно перестраивать все окружение и как в этом случае обеспечить сопровождение приложения.

– Что делать с БД, когда необходимо добавить новые сущности или удалить старые? Это огромная проблема, решение которой лежит в создании новой БД и миграции данных из старой БД. Более подробно решения этой проблемы приведено в [5].

В силу специфики БД Neo4j в графовой БД этих проблем, как правило нет.

Более сложные проблемы возникают при моделировании социальных отношений с помощью RDBMS или NoSQL БД. Некоторые запросы к БД приводят к полной деградации выполнения запроса.

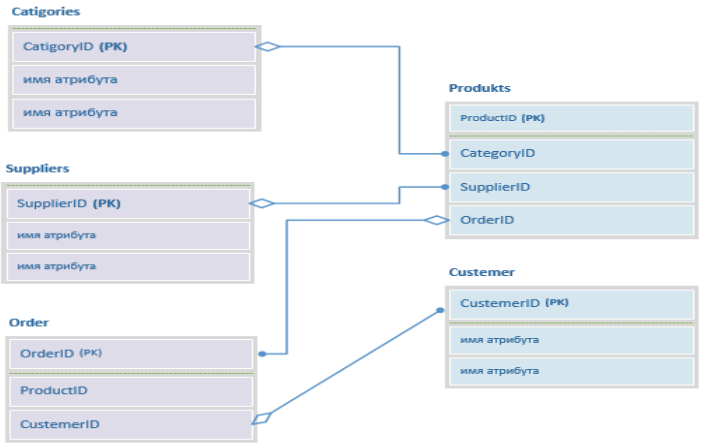


Рисунок 1. Реляционная модель шаблона задачи заказа продуктов

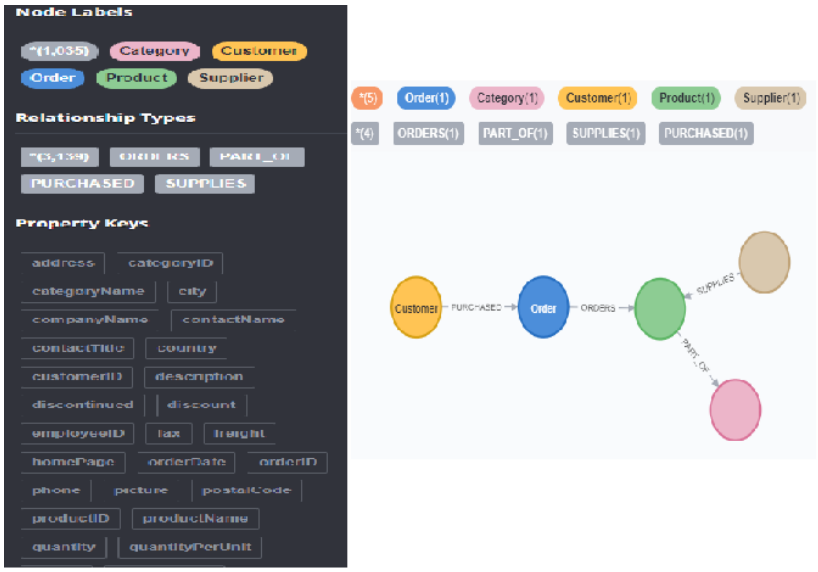


Рисунок 2. Задача заказа продуктов, БД и ее модель Neo4j

Граф знаний и машинное обучения.

3.1. Графы знаний являются основой Graph Data Scientist (GDS) и применяются для оптимизации различных рабочих процессов, для получения ответов на не простые интеллектуальные запросы. В общем случаи, графы знаний представляют собой взаимосвязанные наборы некоторых данных, которые описывают реальные сущности (люди, факты, вещи, детали и т. д.) их отношения друг с другом (схема самолёта, ракеты, схема дорог региона, модель физические сети или социальные сети, взаимосвязи сайтов и т. д.) в простом и понятном виде, в виде графовой модели. Граф знаний позволяет собирать и объединять данные в информацию, используя отношения данных для получения новых знаний [3].

При построении графовых БД сущности (узлы) и отношения как правило дополняются

различными характеристиками, свойствами, атрибутами, что составляет контент модели предметной области. Данный контент используется в системах искусственного интеллекта и графовой аналитике что бы быстро и точно дать ответ.

3.2. Графовые алгоритмы, как правило, классифицированы следующим образом: Pathfinnding, Centrality и Community Detection. Примеры классификации, визуализации и преобразования эмбеддинга можно найти на сайтах [9, 10]. Графовый эмбеддинг это методология представление свойств сущностей (узлов) и свойств отношений в графе как вектор свойств [9, 10] некого пространства размерностью, намного меньшей, чем их количество в графе. Выбор свойств для векторизации зависит от поставленной задачи. Полученные вектора свойств обычно используются различными алгоритмами в ML, что позволяет более глубоко анализировать графовую модель, применять предиктивную аналитику.

Так в приведенном примере 2.4. используя команды библиотеки APOC Neo4j для импорта данных или для экспорта (на пример, Load CSV или Export to CSV) можно в различном формате загружать или выгружать данные свойств узлов или отношений [7].

Пример выгрузки данных свойств для узла Product на рисунке 3, а для узла Supplier на рисунке. 4.

productID	productName	supplierID	categoryID	quantityPerUnit	unitPrice	unitsInStock
1	Chai	1	1	10 boxes x 20 bags	18.00	39
2	Chang	1	1	24 - 12 oz bottles	19.00	17
3	Aniseed Syrup	1	2	12 - 550 ml bottles	10.00	13
4	Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22.00	53
5	Chef Anton's Gumbo Mix	2	2	36 boxes	21.35	0
6	Grandma's Boysenberry Spread	3	2	12 - 8 oz jars	25.00	120
7	Uncle Bob's Organic Dried Pears	3	7	12 - 1 lb pkgs.	30.00	15
8	Northwoods Cranberry Sauce	3	2	12 - 12 oz jars	40.00	6
9	Mishi Kobe Niku	4	6	18 - 500 g pkgs.	97.00	29
10	Ikura	4	8	12 - 200 ml jars	31.00	31
11	Queso Cabrales	5	4	1 kg pkg.	21.00	22

Рисунок 3. Данные свойств узла Product

supplierID	companyName	contactName	contactTitle	address	city
1	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St.	London
2	New Orleans Cajun Delights	Shelley Burke	Order Administrator	P.O. Box 78934	New Orleans
3	Grandma Kelly's Homestead	Regina Murphy	Sales Representative	707 Oxford Rd.	Ann Arbor
4	Tokyo Traders	Yoshi Nagase	Marketing Manager	9-8 Sekimai Musashino-shi	Tokyo
5	Cooperativa de Quesos 'Las Cabras'	Antonio del Valle Saavedra	Export Administrator	Calle del Rosal 4	Oviedo
6	Mayumi's	Mayumi Ohno	Marketing Representative	92 Setsuko Chuo-ku	Osaka
7	Pavlova	Ltd.	Ian Devling	Marketing Manager	74 Rose St. Moonie Ponds
8	Specialty Biscuits	Ltd.	Peter Wilson	Sales Representative	29 King's Way
9	PB KnГ¤ckebrГ¶d AB	Lars Peterson	Sales Agent	Kaloadagatan 13	GГ¶teborg
10	Refrescos Americanas LTDA	Carlos Diaz	Marketing Manager	Av. das Americanas 12.890	Sao Paulo
11	Heli SГјГ¼waren GmbH & Co. KG	Petra Winkler	Sales Manager	TiergartenstraГ¸e 5	Berlin

Рисунок 4. Данные свойств узла Supplier

Далее нужно определить какие данные необходимо анализировать с помощью ML.

В общем случаи для построения преобразования эмбеддинга [9] вершин нужно:

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

– определить функцию подобия узлов, меру близости в графе (например, скалярное 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 их связности на графе. Например, это может быть: произведение Адамара или среднее арифметическое. На рисунке 5 приведен пример многоуровневой схемы для моделирования связей между белками в разных органах (а они ведут себя по-разному в зависимости от местоположения) [10].

Различные классы аналитических алгоритмов анализа графа знаний можно найти в

свободном доступе на сайте Neo4j. Эти алгоритмы позволяют найти кратчайшие пути, с учетом различных весовых критериев (например, расстояния или скорости); анализировать различные сети, выявлять наиболее важные узлы в сети, выявлять тенденции развития сети; специальный класс алгоритмов позволяет анализировать различные социальные сети, выявлять лидеров этих сетей, выявлять спамеров и потенциальных участников мошенничества.

Для определения взаимодействия графов и нейросетей можно использовать методы ML, которые находятся в свободном доступе: DeepWalk (word2vec), Node2Vec, 2D CNN, Graph Convolutional Networks.

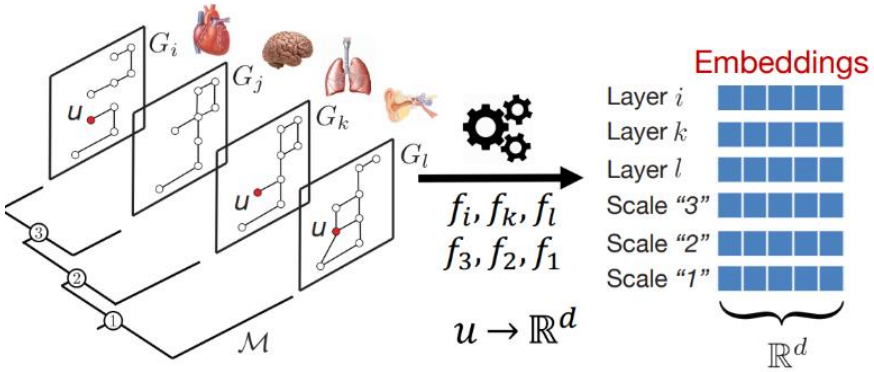


Рисунок 5. Многоуровневое преобразования графовой модели в векторное представление

Совместное использование аналитических алгоритмов графовых моделей и ML, позволяют получать скрытые зависимости и выполнять предиктивный анализ информации, получать ответы в режиме реального времени, реализовывать алгоритмы искусственного интеллекта (ИИ), отслеживать решения ИИ.

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.

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

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

Как уже упоминалось ранее, огромное количество графовых алгоритмов классифицированы на алгоритмы: Pathfinnding, Centrality и Community Detection (см. примеры [10-12]).

Pathfinnding. Класс алгоритмов поиска кратчайших путей, с учетом различных весовых критериев (например, расстояния или скорости) и методов поиска путей, например, найти самый быстрый маршрут для поездки, минимизировать трафик телефонных звонков (см. рис. 6 [10],). Ниже приведена схема применение методов анализа графовой модели и алгоритмов ML. На основании отобранных свойств, строятся векторы и затем оптимизируются отношения между ими.

Centrality. Данный класс алгоритмов (центральности) заключается в понимании того, какие узлы наиболее важные в сети. Эти алгоритмы позволяют определить, как быстро можно распространять информацию в различных группах и между группами сущностей, предсказать появления новых тенденций в этих группах, выявлять уязвимости и возможные цели атаки в сетях связи и транспорта (рисунок 7 [10],). Слева узлы имеют одинаковый цвет, если они принадлежат к одному сообществу, спарва узлы играют одинаковые роли в своих локальных окрестностях (рисунок 7.).

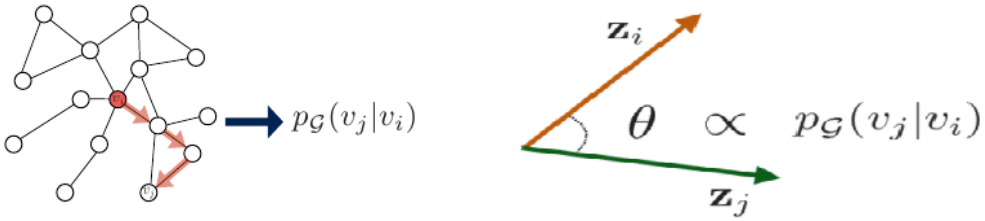


Рисунок 6. Пример графа поиска путей

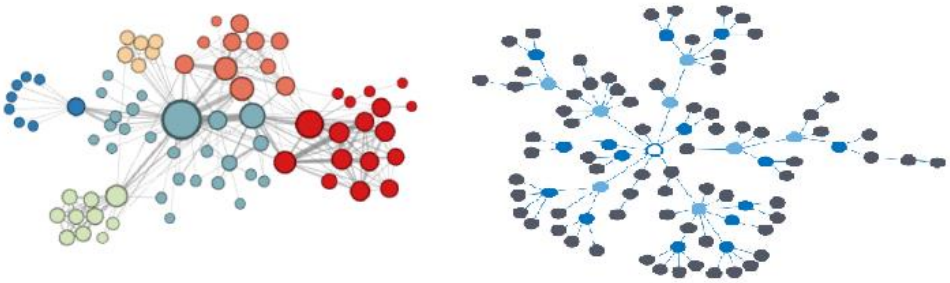


Рисунок 7. Примеры графов групп сущностей и их связей

Community Detection (обнаружение сообщества). Класс алгоритмов, позволяющий изучать различные социальные сети, выявлять лидеров этих сетей, определять количественные характеристик различных групп. Оценивать иерархии, предсказывать тенденции поведения к видоизменению в этих группах, выявлять спамеров и потенциальных участников мошенничества (рисунок. 8 [9, 13],). На данном рисунке приводится схема совместного использования графа знаний и алгоритмов анализа их свойств.

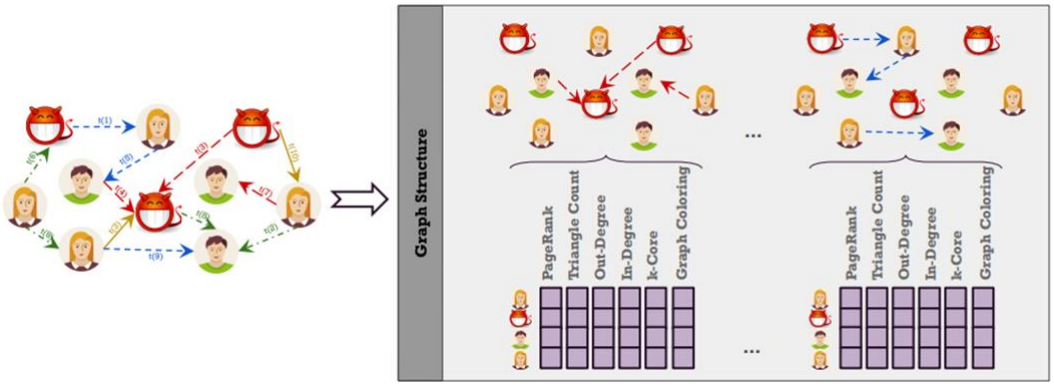


Рисунок 8. Пример графов социальных сетей, выявления социальных групп, спамеров

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

Графовая аналитика позволяет выявить закономерности в данных, например, в социальных данных. обнаружить сообщества или группу лиц, предсказать их поведение. Графическая визуализация помогает понять невидимые процессы ML алгоритмов, которые позволяют компьютерам учиться на примерах без явного программирования. Искусственные нейронные сети

и процесс глубокого обучения также используют графовые модели (рисунок 9 [10],).

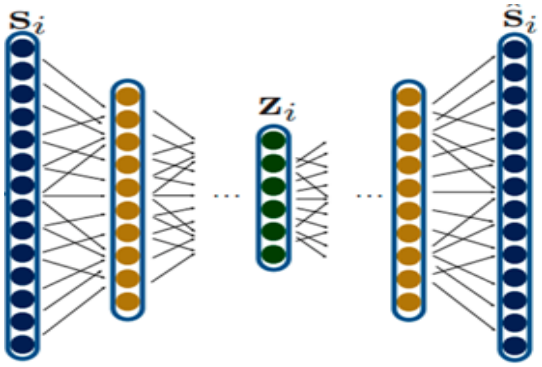


Рисунок 9. Пример графовой модели схемы искусственной нейронный сети

Примеры использования графа знаний.

В качестве примера использования графа знаний здесь рассмотрена «Система комплексного анализа данных интернет-источников» (ИС Анализа Данных), разработанная на кафедре информатики БГУИР и используемая при обучении магистрантов по курсу «Архитектурные решения для обработки больших объемов информации» и «Модели и методы обработки и анализа больших объемов информации». Более полная информация о данном направлении работ приведена в [14 – 16], в данной работе используется часть некоторых материалов из этих работ с учебной целью.

«ИС Анализа Данных» – предназначена для поддержки принятия обоснованных решений, на основе мониторинга и анализа данных из открытых Интернет источников, в том числе и научных публикаций. Выявления важных публикаций, ведущих специалистов и поиске экспертов определенных предметных областей. Создание многоцелевого, модифицируемого кластера в Университете для подготовки специалистов Data Scientist.

Система позволяет находить экспертов (авторитетов) в предметной области и выдавать оценку их рейтинга влиятельности.

Основу технологии разработки системы составляют методы и алгоритмы построения и обслуживания графовой модели различных социальных сетей авторов (блогеров) и их публикаций (в том числе и в СМИ), ссылок на их публикации и определение рейтинга конкретного автора (блогера) публикаций, определение тематик публикаций и классификация их по областям знаний. А для глубокого интеллектуального анализа применение методов аналитических алгоритмов анализа графовых моделей и использование ML для анализа тематик публикаций.

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

При необходимости набор модулей и компонент может быть расширен, а некоторые модули заменены новыми. Общая технология построения многофункционального комплекса по обработке данных и работы компонент, а именно чтения данных интернет-источников, фильтрации данных и векторизации, библиотеки ML модулей, хранилища и БД «граф знаний», приведена в более ранних публикациях [15, 16]. В «ИС Анализа Данных» применена новая архитектура построения многофункциональных комплексов как набор постоянно работающих компонент в виде отдельных серверов, изменена предметная область и система дополнена многофункциональным компонентом БД «граф знаний» и аналитическим компонентом подготовки и выдачи результата.

Компонент библиотека аналитических модулей, содержит набор модулей, которые осуществляют обработку данных, полученных из интернет источников с целью поиска

упоминаний о брендах, определения их тональности и формирования аналитических данных для передачи клиентскому модулю, а также содержит управляющие и служебные модули. В реализованной системе компонент библиотека аналитических модулей состоит из ML модулей: SVM и LDA, PLSA [17, 18].

Компонент графовая база данных и граф знаний [16] состоят из графовой БД, моделирующей предметную область, и программных модулей, позволяющих пополнять графовую БД данными из хранилища и извлекать из нее знания о запрашиваемых объектах, графовая модель является развитием графовой модели приведенной в [19]. «Граф знаний», динамически связывает в графической форме: названия, авторов, публикаций, тематики публикаций, ссылки на публикации и авторов, готовит и выдает различные отчеты. На рисунке 10 приведена структура графовой БД в хранилище и графовая модель предметной области.

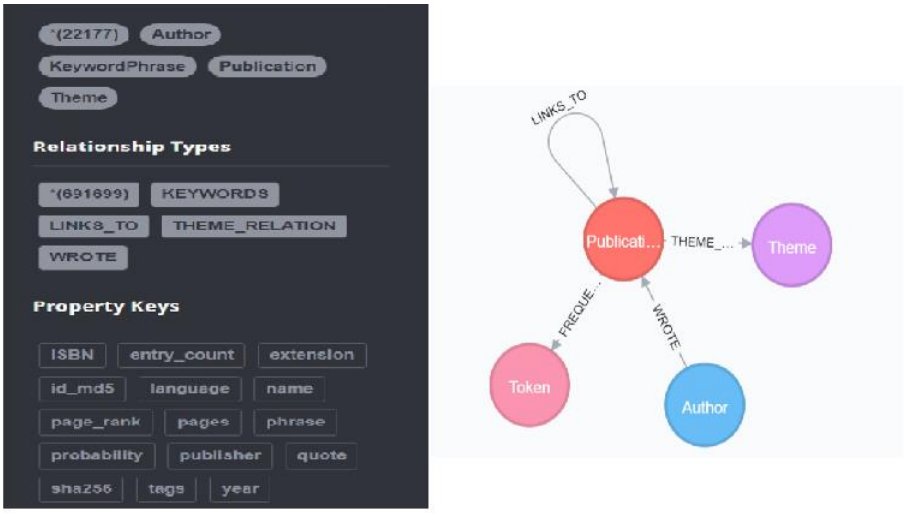


Рисунок 10. Модель БД в виде множества полей и графовая модель

Все основные данные содержатся в хранилище; структура одной из записей для документа приведена ниже:

Структура записи (Hash – primary key of publication, Title – title of publication, Author – author (authors) of publication, Year – publication date, Pages – number of pages, Publisher – publisher of publication, Language – primary language of publication, Topic – topic or topics of publication, Extension – extension of publication file, Tags – array of publication tags, Locator – name of the file).

Компонент графовая база данных добавляет данные из хранилища и регулярно модифицирует информацию в БД и графе знаний. Базой для получения знаний данного компонента является разработанная графовая модель предметной области. На рис. 10 приведена модель предметной области.

Сущности и связи (отношения) модели:

Автор – тот, кто опубликовал статью. Содержит данные о своем имени и о статьях, которые написал (отношение WROTE).

Публикация – публикация, написанная некоторым автором. Содержит в себе имя, год публикации (необходимо как минимум для генерации page-links для page_ranking алгоритмов), ID (sha256), ISBN, публикатор, количество страниц, язык, расширение файла, теги.

С публикацией соотносится следующая информация:

–ссылки на темы, к которым с определенной вероятностью она относится (вероятностная модель, отношение THEME_RELATION);

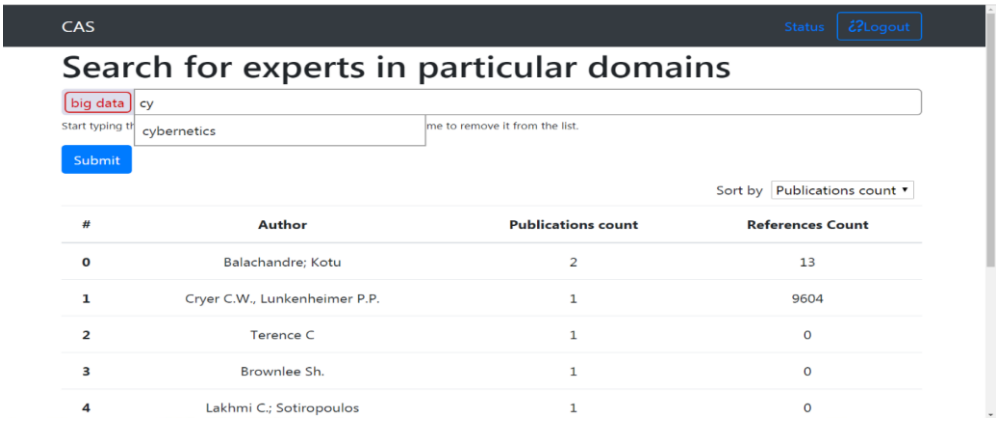
–ссылки на используемые в публикации источники (LINKS_TO, необходимо для page_ranking);

–ссылки на ключевые фразы, которые входят в ее текст (FREQUENCY, содержит количество вхождений, интерпретация мешка слов).

Тема – область знаний, к которой может относиться публикация. Определяется с помощью тематического анализа текста публикации, позволяет просмотреть все публикации, относящиеся к этой теме через THEME_RELATION.

Токен – сущность, которая представляет себя уникальным именем. Имеется возможность просмотреть все публикации, которые имеют в своем тексте вхождение ключевой фразы, также можно просмотреть количество вхождений (все это хранится в связи FREQUENCY).

Ниже приведены некоторые примеры выдачи аналитической информации (рисунки 11, 12, 13).



#	Author	Publications count	References Count
0	Balachandre; Kotu	2	13
1	Cryer C.W., Lunkenheimer P.P.	1	9604
2	Terence C	1	0
3	Brownlee Sh.	1	0
4	Lakhmi C.; Sotiropoulos	1	0

Рисунок 11. Поиск экспертов в выбранных областях знаний

Существуют различные алгоритмы определения наиболее важных статей, публикаций, блогеров; в «ИС Анализа Данных» используется алгоритм page_rank [20].

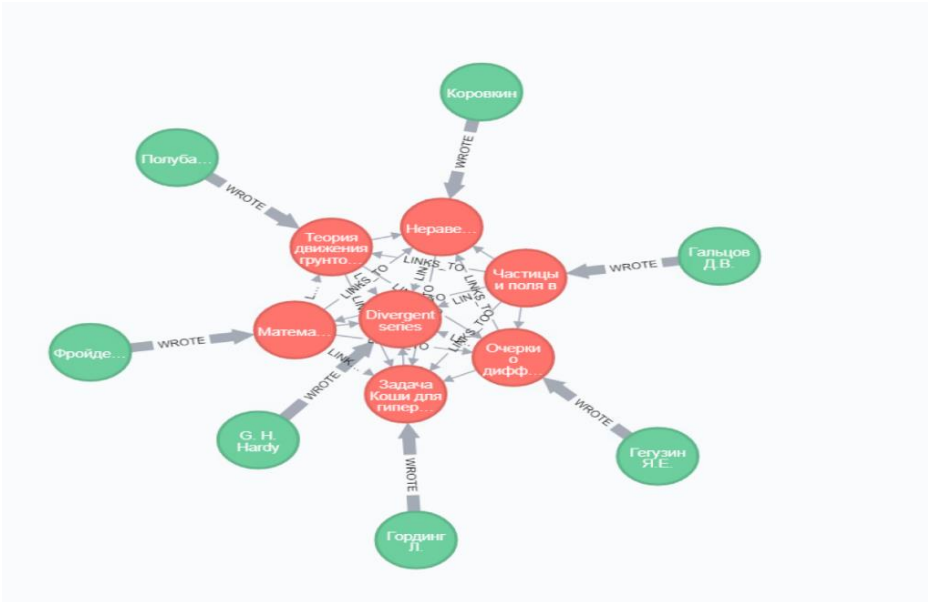


Рисунок 12. Выдача результатов по специфическому запросу с использованием page_rank и probability distribution

Комбинируя подобные запросы с вероятностью тем в анализируемом документе и ключевыми словами можно находить наиболее популярные статьи и авторов статей по указанному запросу на заданную тему. Например, запрос ниже (рисунок 12) иллюстрирует топ-7 отношений Author->Publication, где вероятность темы больше, чем 0.4, и Publication page_rank которых являются максимальными в данной области. На рис. 13 приведен анализ публикаций в области знаний за некоторый период времени.

Заключение.

«ИС Анализа Данных» – это инновационный научно-образовательный проект БГУИР, результаты которого используются при обучении магистрантов по тематике «Обработка больших объемов информации».

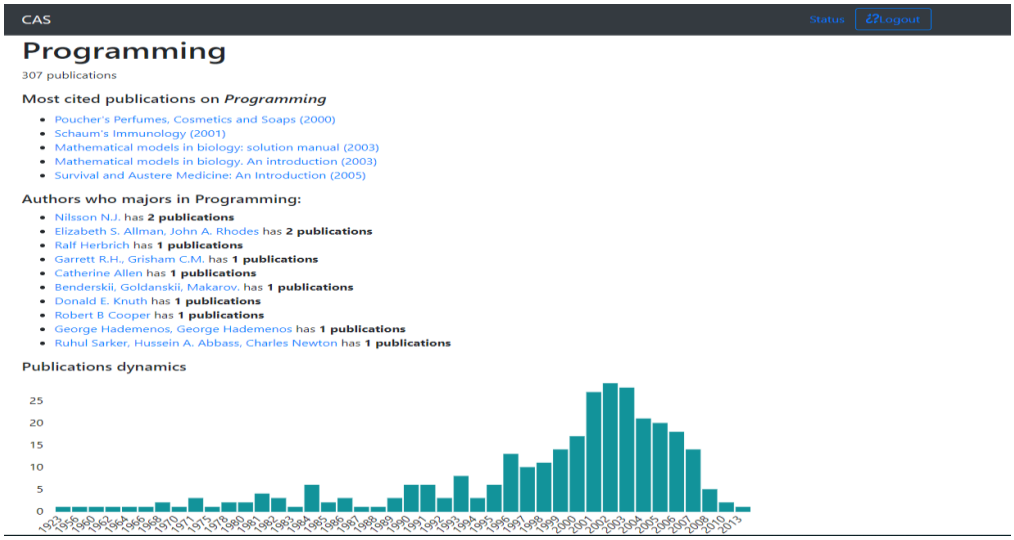


Рисунок 13. Страница области знаний «biotechnology»

Список литературы

[1] W. L. Hamilton, Z. Ying, and J. Leskovec, «Inductive representation learning on large graphs,»NIPS 2017, pp. 1024–1034, 2017.

[2] T. N. Kipf and M. Welling, «Semi-supervised classification with graph convolutional networks,» networks», ICLR 2017, 2017.

[3] T. Hamaguchi, H. Oiwa, M. Shimbo, and Y. Matsumoto, «Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach,» in IJCAI 2017, 2017, pp ICLR 2017, 2017.

[4] Пилецкий И.И., «Эволюционная технология разработки баз данных» //Доклады БГУИР, №3 (41), Минск, БГУИР, 2009 г. - С.107 -112.

[5] Пилецкий И.И., «Проектирование, разработка и сопровождение баз данных с использованием CASE средств» //Учебно-методическое пособие по курсу «Методы и технологии программирования», БГУИР, 2009, 129 стр.

[6] NoSQL // [Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/NoSQL/ Дата доступа: 22.02.2021.

[7] What Is Neo4J? // [Электронный ресурс] – Режим доступа: https://neo4j.com/product/neo4j-graph-database/ Дата доступа: 22.02.2021.

[8] Cypher // Электронный ресурс] – Режим доступа: http://goo.gl/W7Jh6x и http://goo.gl/ftv8Gx .

[9] Где и как врубиться в эмбеддинги графов. [Электронный ресурс] / Режим доступа: https://habr.com/ru/company/ods/blog/418727 / Дата доступа: 24.02.2021.

[10] William L. Hamilton, Rex Ying, J. Leskovec Representation Learning on Graphs: Methods and Applications // Режим доступа: https://www.semanticscholar.org/paper/Representation-Learning-on-Graphs%3A-Methods-and-Hamilton-Ying/ecf6c42d84351f34e1625a6a2e4cc6526da45c74/ Дата доступа: 25.02.2021.

[11] Graph Algorithms in Neo4jv // Режим доступа: https://neo4j.com/whitepapers/graph-algorithms-neo4j-ebook/?ref=pdf-white-paper-ai/ Дата доступа: 24.02.2021.

[12] The Neo4j Graph Algorithms User Guide v3.5//Copyright © 2019 Neo4j, Inc. neo4j.com/graph-algorithms-book/.

[13] Shobeir Fakhraei, Collective Spammer Detection in Evolving Multi-Relational Social // University of Southern California, Режим доступа: https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/ Дата доступа: 25.02.2021/ и https://cs.famaf.unc.edu.ar/~mdoming/seminarios/present_sem-5_dic_2018_Zigaran.pdf Дата доступа: 25.02.2021/.

[14] Пилецкий, И. И. Аналитический комплекс анализа данных из открытых интернет источников / И. И. Пилецкий, В. А. Прытков, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of

the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 – Minsk, BSUIR, 2018. – P. 193-199.

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

[16] Батура М.П., Пилецкий И.И., Прытков В.А., Волорова Н.А. Интеллектуальная система комплексного анализа данных интернет-источников // BIG DATA and Advanced Analytics = BIG DATAи анализ высокого уровня: сб. материалов VI Междунар. науч.-практ. конф (Республика Беларусь, Минск, 20-21 мая 2020 года).: в 3 ч. Ч. 2 / редкол. : В.А. Богуш [и др.]. – Минск : Бестпринт, 2020. – 428 с. р/220-241ISBN 978-985-90533-7-5. (ч. 1).

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

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

[19] Шпаков Н.Н., Черныш Н.Н., Пилецкий И.И.Граф знаний как средство анализа в системе комплексного анализа данных интернет - источников //«GLOBAL SCIENCE AND INNOVATIONS 2019: CENTRAL ASIA» атты VI Халықар. ғыл.-тәж. конф. материалдары (X ТОМ)/ Қҧраст.: Е. Ешім, Е. Абиев т.б.– Нур-Султан, May9-13, 2019 – 353 б. ISBN 978-601-341-186-6. С.120-123.

[20] The Pagepank algorithm neo4j [Электронный ресурс]. – Режим доступа: https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/ Дата доступа: 22.02.2021.

SYSTEM FOR COMPLEX ANALYSIS OF DATA FROM INTERNET SOURCES

M. P. BATURA
Head of the Research Laboratory 8.1 "New Learning Technologies" BSUIR, Doctor of Technical sciences, professor, academician of the International Academy Higher Education Sciences

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

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

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

Abstract. Graph technology is the foundation for building intelligent applications that enable more accurate predictions and faster decision making. The knowledge graph is one of the main areas of AI that enables us to understand prescriptive analytics and AI applications. The combined use of graph technologies, machine learning methods and algorithms allows you to obtain hidden dependencies and perform predictive analysis of information, receive answers in real time, and implement artificial intelligence algorithms. The article provides methods for constructing a graph of knowledge and the use of machine learning in the preparation of undergraduates on the topic "Processing large amounts of information", as well as for obtaining expert data during research at the university.

Keywords: Internet sources, Big data, analysis, Machine Learning, Natural Language Processing, Neo4j, graph databases, graph algorithms