МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙКОЙ ФЕДЕРАЦИИ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»

Современные технологии и средства построения графа знаний

Учебно-методическое пособие

Москва 2023

УДК 004:3:004.89:004.822
ББК 32.813
С56

Современные технологии и средства построения графа знаний: *Учебно-методическое пособие* / *А.А. Артамонов, Р.Р. Тукумбетова, К.В. Ионкина, М.С. Улизко.* – М.: НИЯУ МИФИ, 2023. – 44 с.

Представлены теоретические концепции, лежащие в основе графа знаний вопросы, приведены примеры наиболее известных графов знаний, рассмотрены области современного применения графов знаний, а также наиболее востребованные и перспективные технологии построения и ведения графов знаний.

Предназначено для студентов старших курсов всех факультетов НИЯУ МИФИ, а также может быть использовано преподавателями при проведении ими практических занятий.

Рецензент канд. техн. наук М.А. Григорьева

ISBN 978-5-7262-2925-6

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В настоящее время перед исследователями остро встают вопросы разработки новых методов хранения, систематизации, формализации и автоматической обработки накопленного объема знаний. При этом на первый план выходит необходимость установления и сохранения семантических связей между объектами. Одной из структур, позволяющих решать подобные задачи, являются графы знаний.

Термин «граф знаний» был введен Google в 2012 г. для базы знаний, которая начала использоваться для повышения эффективности работы поисковой системы компании. Графы знаний представляют знания в виде объектов (сущностей) и семантических связей между ними. Широкое распространение графы знаний получили в таких направлениях, как искусственный интеллект, глубокое обучение, аналитика больших данных и т.д. Основными характеристиками графа знаний являются:

1) расширяемость (возможность хранения регулярно обновляющихся данных различных форматов);
2) возможность анализа данных посредством составления запросов;
3) наличие семантических связей между объектами графа знаний;
4) интерпретация фактов и вывод новых знаний.

С помощью графа знаний за счет установления семантических связей между его объектами и понятиями становится возможным решение трудноформализуемых задач интеллектуальной обработки данных. В пособии изложены и существующие современные технологии и средства построения и ведения графа знаний.

РАЗДЕЛ 1. ТЕОРЕТИЧЕСКИЕ КОНЦЕПЦИИ И МОДЕЛИ, ЛЕЖАЩИЕ В ОСНОВЕ ГРАФА ЗНАНИЙ

Прежде чем ввести определение графа знаний (Knowledge Graph), следует определить разницу между данными, информацией и знаниями. Данные обычно представляют собой набор фактов. После обработки, фильтрации и преобразования данные обретают структуру, тем самым преобразовываясь в информацию. Понимание, которое может быть получено из этой информации, называется *знанием*.

1.1. Определение графа знаний. Методология построения

Ориентированный помеченный граф – четырехплет

$$G = (N, E, L, f),$$

где N – набор узлов; $E \subseteq N \times N$ – набор ребер; L – набор меток; $f: E \rightarrow L$ – функция из множества ребер E в множество меток.

Присвоение метки B ребру $E = (A, C)$ можно рассматривать как триплет (A, B, C) и визуализировать, как показано на рис. 1.1.

Рис. 1.1. Граф триплета (A, B, C)

Граф знаний – ориентированный помеченный граф, в котором значения, зависящие от предметной области, привязаны к узлам и ребрам. В качестве узлов могут выступать человек, компания, компьютер и т.д. Метки ребер фиксируют отношения между узлами, например дружеские отношения между двумя людьми, отношения с клиентами между компанией и человеком или сетевое соединение между двумя компьютерами и т.д. [1].

Графы знаний имеют разное применение в зависимости от данных. например направленный помеченный граф, в котором узлами являются люди, а ребра фиксируют родительскую связь, или направленный помеченный граф, в котором узлами являются классы объектов (например, книга, учебник и т.д.), а ребра фиксируют связь подкласса. Такой граф также известен как *таксономия*.

В некоторых моделях данных в триплете (*A*, *B*, *C*) *A*, *B*, *C* обозначают субъектом, предикатом и объектом триплета соответственно. Более того многие графы знаний в настоящее время представляют извлеченные факты в форме триплета субъект-предикат-объект (Subject-Predicate-Object, SPO), что соответствует стандарту, предписанному RDF (Resource Description Framework)[1] [2].

Граф знаний хранит знания в машиночитаемой форме и предоставляет средства для сбора, организации, обмена, поиска и использования информации. Информация может быть добавлена в граф знаний пользователем как в ручном режиме, так и при помощи автоматизированных и полуавтоматических методов. Независимо от метода добавления знаний ожидается, что записанная информация может быть понята и проверена людьми.

Множество математических расчетов по графу может быть сведено к навигации по нему. Например, в графе знаний, который отражает дружеские отношения, чтобы вычислить друзей друзей человека *A*, мы можем перемещаться по графу от узла *A* ко всем узлам *B*, связанным с ним отношением, помеченным как друг, а затем рекурсивно ко всем узлам *C*, связанными связью дружбы с узлами *B*.

Для того чтобы проиллюстрировать построение графа знаний, рассмотрим в качестве примера предложение: «Александр Демьяненко играл персонажа Шурика в фильме «Операция «Ы» и другие приключения Шурика», а также персонажа Димы Горина в фильме «Карьера Димы Горина».

Если представленные в табл. 1.1 факты визуализировать, то получится граф знаний (рис. 1.2).

[1] RDF – модель для представления данных. RDF представляет утверждения о ресурсах в виде, пригодном для машинной обработки.

Таблица 1.1

Триплеты SPO

Субъект	Предикат	Объект
Александр Демьяненко	игралВ	«Операция «Ы» и другие приключения Шурика»
Александр Демьяненко	играл	Шурик
Шурик	персонажВ	«Операция «Ы» и другие приключения Шурика»
«Операция «Ы» и другие приключения Шурика»	является	фильм
Александр Демьяненко	игралВ	«Карьера Димы Горина»
Александр Демьяненко	играл	Дима Горин
Дима Горин	персонажВ	«Карьера Димы Горина»
«Карьера Димы Горина»	является	фильм

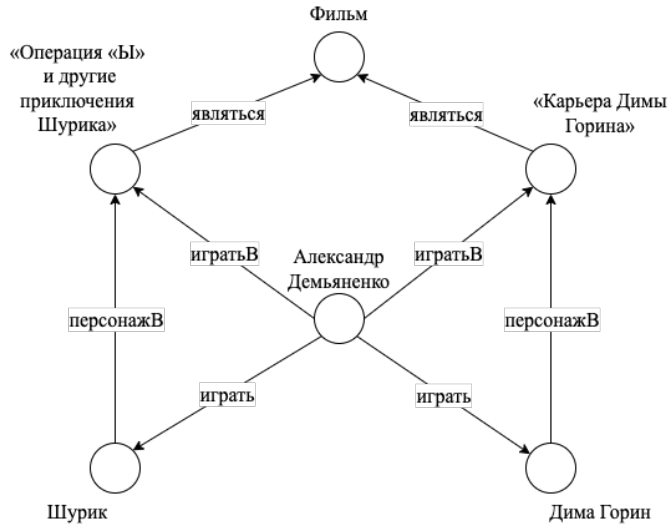


Рис. 1.2. Граф знаний

На рис. 1.2 можно заметить, что один и тот же узел в разных триплетах может выступать в роли субъекта и объекта. Например, в триплете (Шурик-персонажВ-«Операция «Ы» и другие приключения Шурика») Шурик выступает в роли субъекта, однако в триплете (Александр Демьяненко-играл-Шурик) Шурик является объектом.

Методология построения графа знаний с помощью автоматизированных методов включает в себя четыре шага.

1. Извлечение триплетов (фактов) SPO из текста. Этот шаг включает в себя извлечение сущностей (субъектов и объектов) и извлечение связей (предикатов). Для этого используют методы обработки естественного языка (NLP), например, такие как синтаксический анализ зависимостей (dependency parsing).

2. Распознавание и связывание сущностей. Это шаг, который сопоставляет А.С. Демьяненко, Александра Демьяненко, Александр Сергеевич Демьяненко и т.д. с одним объектом – Александром Демьяненко.

3. Устранение неоднозначности и сохранение триплетов SPO в базе данных графа. Здесь факт, представленный как SPO, преобразуется в вид субъекта, который связан с объектом через отношения, описанные предикатом.

4. Обработка графа для того, чтобы заполнить недостающие связи, произвести кластеризацию сущностей и т.д.

1.2. Google Knowledge Graph

Граф знаний Google – база знаний, используемая Google и его службами для улучшения результатов своей поисковой системы с помощью информации, собранной из различных источников. Информация предоставляется пользователям в информационном окне рядом с результатами поиска. Эти информационные окна были добавлены в поисковую систему Google в мае 2012 г., сначала в Соединенных Штатах Америки, а к концу года были введены на международном уровне. Google назвал эти информационные окна, которые появляются справа или сверху от результатов поиска, «панелями знаний» (рис. 1.3) [3].

Информация, содержащаяся в графе знаний Google, после запуска росла быстрыми темпами, утроившись за семь месяцев и охватив 570 млн объектов и 18 млрд фактов. К середине 2016 г. Google

сообщил, что хранит 70 млрд фактов и ответил «примерно на треть» из 100 млрд запросов, которые они обрабатывали в месяц. К маю 2020 г. это число выросло до 500 млрд фактов о 5 млрд организаций.

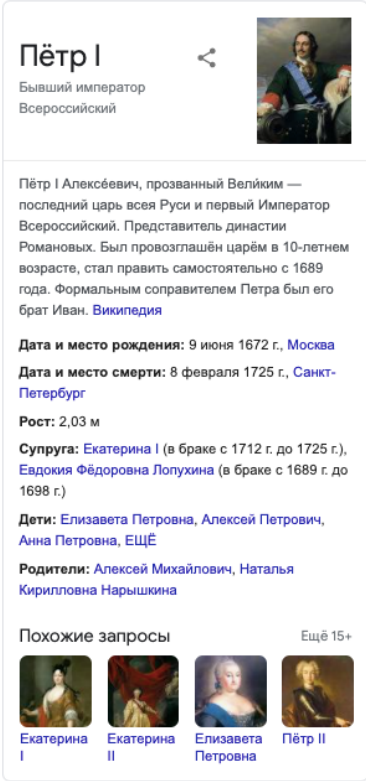


Рис. 1.3. Панель знаний Петра I (май 2021 г.)

Официальной документации о том, как реализован граф знаний Google, нет в открытом доступе. Согласно Google, информация графа знаний получена из многих источников, включая Всемирную книгу фактов ЦРУ [2], Викиданные [3] и Википедию. Граф знаний

[2] *Всемирная книга фактов ЦРУ* – справочник о странах мира, составляемый Центральным разведывательным управлением США.

[3] *Викиданные* – база знаний, созданная Фондом Викимедиа. Используется для обеспечения централизованного хранения данных, которые могут содержаться в статьях Википедии.

используется для ответов на прямые устные вопросы в Google Assistant и голосовых запросах Google Home. Граф знаний критиковали за предоставление ответов без ссылки на источник. Google утверждает, что панель знаний обновляется автоматически на основе информации, доступной в интернете и в других различных источниках.

1.3. DBpedia

DBpedia – проект, направленный на извлечение структурированной информации из данных, представленных в различных проектах Викимедиа[4]. Эта структурированная информация напоминает общедоступный граф знаний, доступный каждому во всемирной паутине (World Wide Web, WWW). DBpedia позволяет пользователям семантически запрашивать отношения и сущности в ресурсах Викимедии [4].

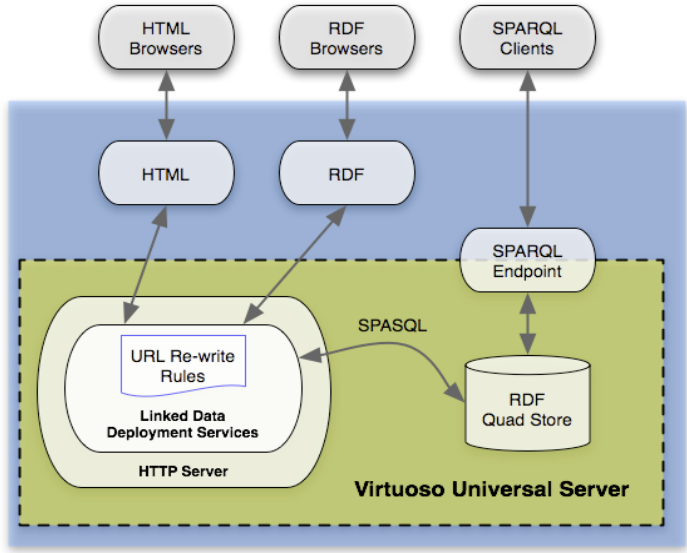


Рис. 1.4. Архитектура предоставления данных DBpedia

[4] *Фонд Викимедиа* – некоммерческая благотворительная организация, которая поддерживает инфраструктуру для работы ряда многоязычных краудсорсинговых вики-проектов, таких как Википедия, Викисловарь, Викигид и т.д.

DBpedia использует стандарт RDF для представления извлеченной информации, т.е. информация представлена в виде фактов SPO. В этой паутине фактов можно перемещаться с помощью стандартных веб-браузеров, автоматических поисковых роботов или создавать сложные запросы с помощью языков запросов, подобных SQL (например, SPARQL), рис. 1.4.

1.4. Social Graph (Facebook Entities Graph)

Социальный граф – модель социальной сети в виде графа, представляющая социальные отношения между субъектами. В качестве узлов графа выступают такие социальные объекты, как пользовательские профили с различными атрибутами (например: имя, день рождения, родной город), сообщества, медиаконтент и т.д., а в качестве ребер – социальные связи между объектами.

В качестве примера приведем фрагмент социального графа на платформе Facebook. На графе отображено в каких отношениях состоят разные социальные объекты (рис. 1.5).

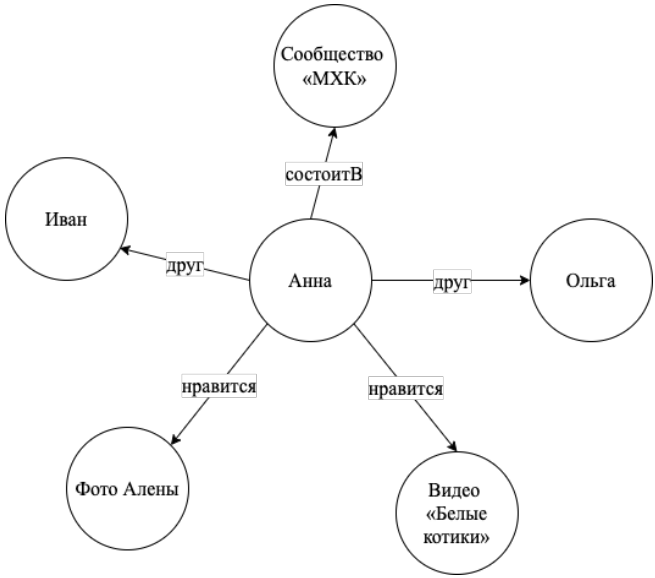


Рис. 1.5. Фрагмент социального графа на платформе Facebook

Пользователь Анна находится в дружеских отношениях с пользователями Иван и Ольга, при этом Иван и Ольга не являются друзьями, но у них есть общий друг Анна. Анне понравилась фотография Алены, а также видео «Белые котики». Кроме того, Анна состоит в сообществе «МХК».

Термин «социальный граф» был впервые использован на конференции Facebook F8 24 мая 2007 г. в рамках представления платформы Facebook для объяснения того, как платформа намерена использовать отношения/связи между людьми для улучшения взаимодействия между пользователями онлайн. Определение было расширено и теперь относится к социальному графу всех пользователей интернета.

С момента объяснения концепции социального графа Марк Цукерберг, один из основателей Facebook, часто говорил, что *цель Facebook* – предложить социальный граф веб-сайта Facebook другим веб-сайтам, чтобы отношения/связи пользователей могли использоваться на веб-сайтах, находящихся вне контроля Facebook [6].

Для того чтобы извлекать информацию из социального графа Facebook, компания создала *Facebook Graph API* – инструмент, который позволяет взаимодействовать с данными, представленными на платформе Facebook.

1.5. NELL

Never-Ending Language Learning System (NELL) – семантическая система машинного обучения, разработанная исследовательской группой в Университете Карнеги-Меллона. NELL поддерживается грантами DARPA[5], Google, NSF[6] и CNPq[7], часть системы работает на суперкомпьютерном кластере, предоставляемом Yahoo!.

[5] *DARPA* (The Defense Advanced Research Projects Agency) – Управление перспективных исследовательских проектов Министерства обороны США – Управление Министерства обороны США, отвечающее за разработку новых технологий для использования в интересах вооруженных сил.

[6] *NSF* (National Science Foundation) – Национальный научный фонд – независимое агентство при правительстве США, отвечающее за развитие науки и технологий.

[7] *CNPq* (The Brazilian National Council for Scientific and Technological Development) – Национальный совет по научному и технологическому развитию Бразилии – организация федерального правительства Бразилии при Министерстве науки и технологий, занимающаяся продвижением научных и технологических исследований и формированием кадрового потенциала для исследований в стране.

Входные данные для NELL включают в себя:

1) исходную онтологию, определяющую сотни категорий (например, человек, спортивная команда, фрукты, эмоции) и отношений (например, играетВКоманде (спортсмен, спортивная команда), играетНаИнструменте (музыкант, инструмент);

2) от 10 до 15 начальных примеров каждой категории и отношения.

Учитывая эти входные данные, а также набор из 500 миллионов веб-страниц и доступ к остальной части сети через API-интерфейсы поисковых систем, NELL работает 24 ч в сутки для выполнения двух текущих задач.

1. Извлечение новых экземпляров категорий и отношений. Другими словами, поиск новых экземпляров входных категорий (например, «Барак Обама» – человек и политик) и нахождение пары фраз, которые являются экземплярами входных отношений (например, пара «Александр Овечкин» и «Вашингтон Кэпиталз» является экземпляром отношения играетВКоманде). Эти новые экземпляры добавляются к растущей базе знаний структурированных убеждений.

2. Ежедневное повышение качества чтения данных. NELL использует различные методы для извлечения фактов из Интернета. Извлеченные факты перепроверяются, используя растущую базу знаний в качестве набора обучающих примеров. Подтвержденные факты с высоким уровнем убежденности вносятся в базу знаний убеждений [7].

NELL была запрограммирована ее разработчиками так, чтобы иметь возможность идентифицировать базовый набор фундаментальных семантических отношений между предопределенными категориями данных. Система NELL была запущена в начале 2010 г. исследовательской группой Карнеги–Меллона и с тех пор работает круглосуточно, просматривая сотни миллионов веб-страниц в поисках связей между информацией, которую она уже знает, и тем, что она находит в процессе поиска для того, чтобы установить новые связи, имитируя то, как люди изучают новую информацию. Например, встретив пару слов «Пайкс–Пик»[8], NELL заметит, что оба слова написаны с заглавной буквы, и сделает вывод из второго слова, что

[8] *Пайкс-Пик* (на англ. Pikes Peak) – гора в центральной части США, штат Колорадо.

это название горы[9], а затем построит связи со словами, которые окружают эти два слова, чтобы выявить другие связи.

Цель NELL и других таких систем семантического обучения, как система Watson, разработанная компанией IBM, заключается в разработке программных средств, которые будут способны отвечать на вопросы пользователей, заданные на естественном языке, без вмешательства человека.

В настоящее время NELL накопила базу знаний более чем 2,8 млн убеждений из 1 186 различных категорий данных. По состоянию на май 2021 г. последние собранные факты по проекту датированы февралем 2019 г. [8].

1.6. Semantic Web

Semantic Web (рус. – Семантическая паутина) представляет собой надстройку над Всемирной паутиной, стандартизирующую представление информации в таком виде, в котором она может быть обработана машинным способом. На данный момент она не полностью реализована, однако имеются тенденции (например, Google Knowledge Graph, Facebook Social Graph), которые свидетельствуют о том, что интернет в будущем будет устроен в виде Семантической паутины. Тогда машины будут способны не только читать данные в сети, но и понимать их.

Когда машины смогут понимать данные в сети, то интернет превратиться из платформы обмена информацией в платформу знаний. Компьютеры будут помогать пользователю находить контент, который лучше соответствует потребностям пользователя, и принимать более правильные решения, предлагая неочевидные идеи, а также лучше определять тенденции, аномалии, пробелы и т. д. в различных областях.

В настоящее время, данные в интернете представлены на естественном языке (например, на английском языке). Чтобы реализовать Семантическую паутину, необходимо стандартизировать способ представления информации. Стандартизация позволит машинам понимать информацию. RDF (Resource Description Framework) и OWL (Web Ontology Language) – несколько шагов, предпринятых

[9] *Peak* – перевод с англ. «гора»

для стандартизации. Язык RDF позволяет описывать структуру семантической сети в виде графа. Каждому узлу и каждому ребру графа можно назначить отдельный URI. Утверждения в RDF имеют вид так называемого триплета «субъект – предикат (связь) – объект». Утверждения, записанные на языке RDF, можно интерпретировать с помощью онтологий. Для создания онтологий используется язык OWL. Онтологии создаются для получения логических заключений из данных [9].

Представление информации в сети виде графа знаний, например создание Google Knowledge Graph, Facebook Knowledge Graph, Bing Knowledge Graph, является одним из шагов на пути реализации Семантической паутины.

РАЗДЕЛ 2. ОБЛАСТИ СОВРЕМЕННОГО ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ ПОСТРОЕНИЯ И ВЕДЕНИЯ ГРАФОВ ЗНАНИЙ, ИХ НАЗНАЧЕНИЕ

За последние 10 лет (2011–2020 гг.) в реферативных базах данных было опубликовано более 1640 публикаций по разработкам в области применения технологий построения и ведения графов [10].

На рис. 2.1 представлены десять тематических областей, к которым относится наибольшее количество выявленных публикаций: 46,3 % публикаций относится к компьютерным наукам; 14,5 % – к математике. В рамках этих двух направлений разрабатываются технологии создания и ведения графов знаний. В то же время другие направления, представленные на круговой диаграмме, характеризуют области применения графов знаний: они применяются в инженерных науках (например, в проекте Manufacturing Open Knowledge Graph, структурирующем общедоступную информацию о проектировании и производстве различной продукции), в системах поддержки принятия решений, в социальных науках (социальные графы) и т. д.

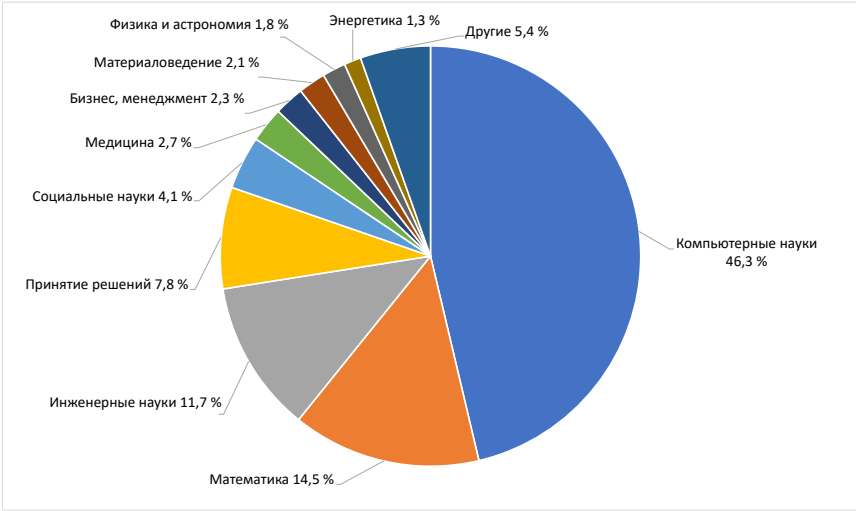


Рис. 2.1. Распределение публикаций по направлениям исследований

В табл. 2.1 приведены десять наиболее популярных тематик, связанных с графами знаний, за последние 10 лет. Большая часть публикаций относится к направлениям, связанным с процессом извлечения информации для построения графов знаний (распознавание именованных сущностей, извлечение неявной информации, обработка естественного языка и др.). Также распространены тематики, связанные с созданием систем взаимодействия с пользователем на основе графов знаний (вопросно-ответные системы, системы рекомендаций, расширение запросов и т.д.). При этом тематика, посвященная системам рекомендаций, характеризуется наибольшей цитируемостью публикаций и обладает более высокой актуальностью по сравнению с другими рассматриваемыми тематиками.

Таблица 2.1

Десять наиболее популярных тематик, связанных с построением и использованием графов знаний

№	Тематика	Доля от общего числа публикаций, %
1	Анализ тональности; распознавание именованных сущностей; извлечение неявной информации	17,0
2	SPARQL; обработка запросов; **Semantic Web**	1,6
3	Вопросно-ответные системы; естественно-языковой интерфейс; представление значений	1,5
4	Совместная фильтрация; системы рекомендаций; неявная обратная связь	1,5
5	Неоднозначность смысла слова; именованная сущность; **Wordnet**	1,0
6	**Генная онтология**; семантическое сходство; SPARQL	0,9
7	Распознавание именованных сущностей; обработка естественного языка; анонимизация	0,9
8	Псевдорелевантная обратная связь; расширение запросов; поиск документов	0,8
9	Краудсорсинг; Turk; назначение задач	0,7
10	Тематическая модель; распределение Дирихле; классификация текстов	0,6

Кроме того, анализ тематик позволил выявить примеры проектов, использующих графы знаний (в табл. 2.1 выделены полужирным шрифтом).

1. Semantic Web (описание проекта дано в разд. 1).

2. Wordnet – лексическая база данных английского языка, представляющая собой электронный словарь-тезаурус и набор семантических сетей. Wordnet разработан Принстонским университетом и выложен в открытый доступ в 1995 г. [11].

3. «Генная онтология» (англ. – Gene Ontology, GO) – проект, посвященный созданию унифицированной терминологии для аннотации генов всех биологических видов. Создание «Генной онтологии» началось в 1998 г. Проект курируется Консорциумом GO (GOC) и финансируется Национальным исследовательским институтом генома человека, подведомственным Национальным институтам здравоохранения США [12].

На рис. 2.2 приведено распределение количества публикаций по годам. Первые упоминания графов знаний появились в публикациях в середине 2000-х г., однако до 2013 г. таких публикаций были единицы. После того, как в 2012 г. Google объявил о создании собственного графа знаний, интерес к теме начал расти, а количество публикаций стремительно увеличиваться, особенно в последние 5 лет. В период с 2016 по 2020 г. ежегодное количество публикаций по данной теме выросло практически в 10 раз. Следует отметить, что в 2018 г. консалтинговое агентство Gartner включило графы знаний в график общественного интереса (Цикл Гартнера) в качестве новой ключевой технологии, которая позволяет более гибко отображать сложные области знаний [13].

Научно-исследовательская деятельность по данному направлению ведется в 71 стране мира (рис. 2.3). Явным лидером по количеству публикаций является Китай (817 публикаций), за которым следуют США (276), Германия (154), Великобритания (80), Австралия (54). Можно предположить, что лидирующие позиции данных стран объясняются высоким уровнем развития информационных технологий в них.

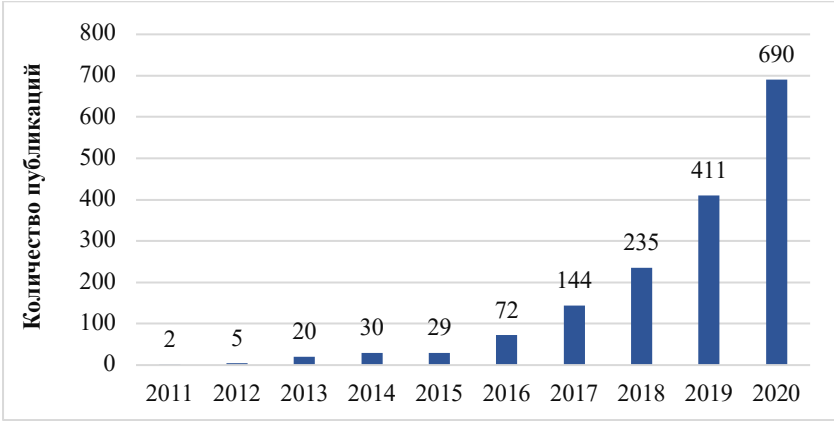


Рис. 2.2. Распределение количества публикаций
по годам за период 2011–2020 гг.

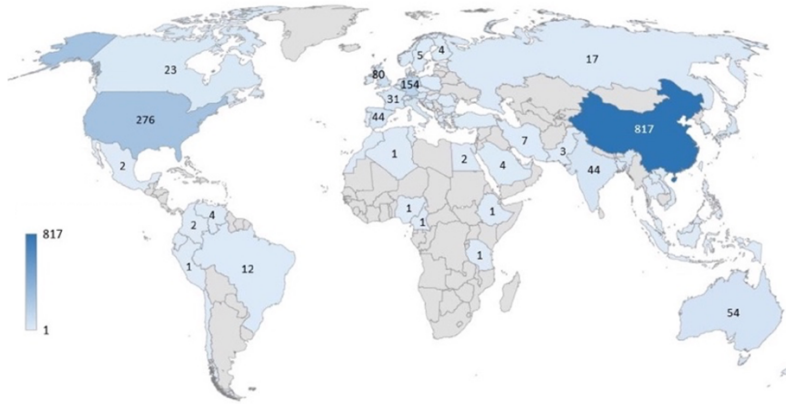


Рис. 2.3. Распределение количества
публикаций по странам

При этом Китай не только является лидером по количеству публикаций, но для него также характерен наиболее резкий рост публикационной активности. В США и Германии наблюдается не такой

стремительный, но тем не менее стабильный прирост ежегодного количества публикаций, связанных с построением и использованием графов знаний (рис. 2.4).

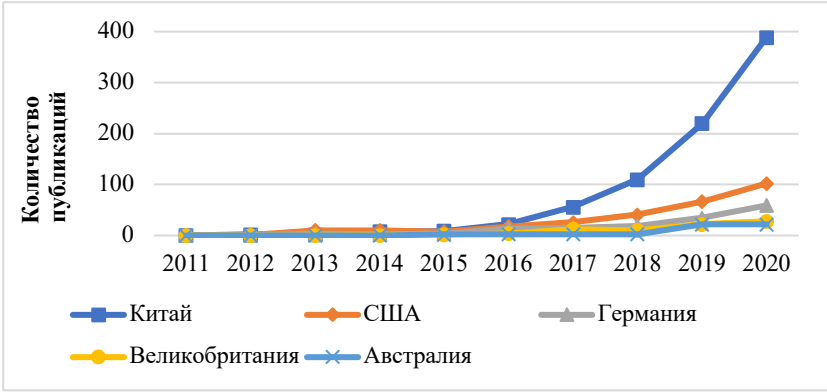


Рис. 2.4. Динамика публикационной активности в странах-лидерах по количеству публикаций (2011–2020 гг.)

В табл. 2.2 представлено 15 организаций с наибольшим количеством публикаций. Принимая во внимание долю публикаций от общего числа, которая приходится на каждую из представленных организаций, можно сделать вывод, что в исследованиях по данному направлению нет ярко выраженного лидера. Однако следует отметить, что первые девять позиций занимают китайские организации (научно-исследовательские институты и высшие учебные заведения). Среди лидеров по количеству публикаций только две организации корпоративного сектора – IBM (22 публикации) и Microsoft (20 публикаций). При этом публикации последней цитируются значительно активнее других (56,4 цитирования на одну публикацию), что свидетельствует об их актуальности и важности с точки зрения научного сообщества. Что касается российских организаций, то наибольшее количество статей по данной тематике было опубликовано Санкт-Петербургским национальным исследовательским университетом информационных технологий, механики и оптики (ИТМО, 6 публикаций).

Таблица 2.2

Пятнадцать организаций с наибольшим количеством публикаций

№	Название организации	Количество публикаций
1	Китайская академия наук (КНР)	63
2	Университет Китайской академии наук (КНР)	40
3	Университет Цинхуа (КНР)	35
4	Пекинский университет (КНР)	33
5	Министерство образования (КНР)	30
6	Харбинский технологический институт (КНР)	29
7	Пекинский университет почты и телекоммуникаций (КНР)	27
8	Тяньцзиньский университет (КНР)	27
9	Оборонный научно-технический университет Народно-освободительной армии Китая (КНР)	25
10	Университет Южной Калифорнии (США)	23
11	Институт автоматизации Китайской академии наук (КНР)	22
12	IBM (США)	22
13	Уханьский университет (КНР)	22
14	Боннский университет (Германия)	21
15	Microsoft (США)	20

В табл. 2.3 представлен топ-10 организаций, финансирующих исследования по анализируемой тематике. Из десяти лидирующих организаций семь имеют непосредственное отношение к китайскому правительству, что свидетельствует, с одной стороны, об интересе Китая к технологиям построения и ведения графов знаний, с другой стороны – о готовности государства поддерживать исследования в данной области. Наибольшее количество публикаций было сделано при поддержке Национального фонда естественных наук КНР.

Таблица 2.3

Топ-10 финансирующих организаций

№	Название организации	Количество публикаций
1	Национальный фонд естественных наук (КНР)	356
2	Национальная программа ключевых исследований и разработок (КНР)	84
3	Европейская комиссия (ЕС)	61
4	Национальный научный фонд (США)	58
5	Министерство науки и технологий (КНР)	55
6	Национальная программа фундаментальных исследований «973» (КНР)	53
7	Фонды поддержки фундаментальных исследований для центральных университетов (КНР)	50
8	Министерство образования (КНР)	29
9	Агентство перспективных оборонных исследовательских проектов DARPA (США)	17
10	Министерство финансов (КНР)	17

В результате анализа информации о конференциях, в рамках которых был сделан ряд публикаций по графам знаний, было обнаружено, что подавляющее большинство этих конференций являются общетематическими и посвящены информационным технологиям в целом. Однако была выявлена одна специализированная конференция – «Международная конференция по графам знаний Института инженеров электротехники и электроники» (IEEE ICKG), которая посвящена непосредственно графам знаний и их применению. Данная конференция проводится в Китае с 2010 г. На протяжении 8 лет ее принимал Хэфэйский технологический университет, затем, начиная с 2018 г., организаторами становились и другие крупные университеты и научно-исследовательские организации Китая. В рамках конференции учитываются все аспекты, связанные с графами знаний, включая алгоритмы, программное обеспечение, платформы и приложения для построения графа знаний. IEEE ICKG привлекает исследователей и разработчиков из таких областей, как инженерия знаний, аналитика больших данных, статистика, машинное обучение, распознавание образов, интеллектуальный анализ данных, визуализация знаний, высокопроизводительные вычисления и др.

В 2021 г. двенадцатая по счету конференция прошла в Гонконгском университете науки и технологий. В качестве основных тем были выбраны:

1) машинное обучение и графы знаний;
2) формирование рассуждений посредством графов знаний;
3) аналитика и практическое применение графов знаний;
4) графы знаний и обработка естественного языка;
5) графы знаний для «понятного» искусственного интеллекта;
6) мультимодальные графы знаний;
7) социальные сети и изучение представлений данных;
8) графы знаний и культурное наследие;
9) графы знаний для систем геопространственной информации;
10) графы знаний предметных областей;
11) графы знаний для образования.

Рассматривая патентную активность за 10 лет, необходимо отметить, что в начале рассматриваемого периода (2011–2013 гг.) патенты получали в основном американские компании Google и IBM; в последующие четыре года (2013–2016 гг.) к ним присоединилась компания Microsoft Technology Licensing и в меньшей степени – китайская организация Beijing Baidu Netcom Science Technology. За период 2016–2019 гг. ситуация в целом оставалась без изменений, однако за последние два года лидерские позиции стали занимать китайские организации при меньшем участии американских. Это может свидетельствовать о том, что, несмотря на ведущие позиции корпораций из США по количеству патентов, наблюдается тенденция к все более активному участию китайских компаний и меньшему вовлечению бывших американских лидеров в разработки в области графов знаний.

РАЗДЕЛ 3. ОБЗОР НАИБОЛЕЕ ВОСТРЕБОВАННЫХ И/ИЛИ НАИБОЛЕЕ ПЕРСПЕКТИВНЫХ ТЕХНОЛОГИЙ И ПРОГРАММНЫХ СРЕДСТВ

По сравнению с другими информационными системами, ориентированными на знания, отличительные особенности графов знаний заключаются в их практическом сочетании структур представления, процессов управления информацией и алгоритмов поиска [14]. Поэтому в настоящее время многие компании и научные группы начинают создавать свои собственные графы знаний для организации информации, данных и знаний в целом. Исходя из этого, появляются и программные средства для их создания, хранения и визуализации.

3.1. Описание графовых СУБД как одной из основ графов знаний

Граф знаний строится на основе базы знаний, которая, в свою очередь, основывается на собранной информации из текста на веб-страницах, базах данных, аудио- и видеоконтенте. Базовый конвейер процесса построения графа знаний представлен на рис. 3.1 [15].

Рис. 3.1. Базовый конвейер процесса построения графа знаний

Также процесс построения базы знаний может быть рассмотрен с точки зрения построения графа: процесс сбора, процесс хранения, процесс отображения (рис. 3.2) [16]. Ключевым аспектом, который рассматривается в этом разделе, является слой хранилища данных, в качестве которого чаще всего выступает некоторая база данных. Если учитывать, что граф знаний представляет из себя набор

взаимосвязанных сущностей (с помощью триплетов), то для хранения структуры такого вида используются графовые базы данных или базы данных с семантической моделью данных.

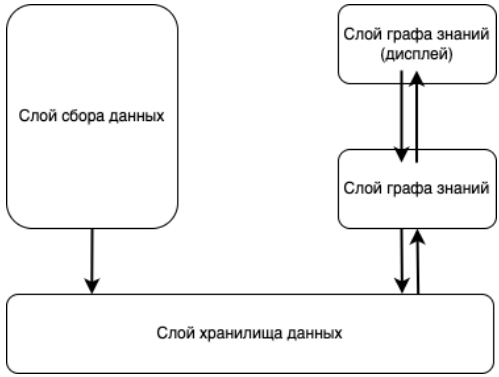


Рис. 3.2. Концепция построения графа знаний

Поскольку оба типа баз данных относятся к NoSQL (нереляционным базам данных), то рассмотрим процесс развития NoSQL баз данных. Этот процесс может быть развит по фазам, представлен в табл. 3.1 [17].

Таблица 3.1

Развитие NoSQL баз данных

Фаза	Описание
1	1972–2002 гг. Объектно-ориентированная модель, OLAP-анализ, реляционная СУБД. 2002–2007 гг. Полу-структурированные модели баз данных. Объектно-ориентированные базы данных. Продвинутые реляционные СУБД
2	2007–2009 гг. Хранение данных по типу «ключ-значение» (Redis и Riak). Oracle RDF
3	2009–2012 гг. Колоночное хранение (Cassandra, Yahoo Hadoop и т.д.). Документно-ориентированное хранение (Marklogic, MongoDB и т.д.)
4	2015 г.–н.в. Графовые базы данных (Neo4j и др.)

Следует отметить, что на текущий момент не существует классической СУБД (системы управления базы данных) с семантической моделью данных [18], поэтому далее приведено описание NoSQL СУБД (*Neo4j, Amazon Neptune, ArangoDB, OrinentDB, Cassandra*), которые могут быть использованы для реализации графа знаний. Также рассчитывается оценка графовых СУБД (*neo4j, Amazon Neptune, ArangoDB, OrinentDB*) между собой на основе многокритериального анализа для возможности выбора наиболее подходящей в зависимости от решаемой задачи.

Neo4j – графовая система управления базами данных, написанная на языке программирования Java. Является одной из самых популярных графовых СУБД, хранит данные в виде узлов и ребер между ними, для обращения с базами данных используется REST API и язык запросов Cypher.

Amazon Neptune – быстрый, надежный, полностью управляемый сервис графовой базы данных, который упрощает создание и запуск приложений, работающих с наборами тесно связанных данных. В основе Amazon Neptune лежит высокопроизводительное ядро графовой базы данных, оптимизированное для хранения большого количества (порядка миллиардов) взаимосвязей и выполнения запросов к графу с задержками на уровне миллисекунд. Amazon Neptune поддерживает популярные модели графов Property Graph и RDF W3C, а также их соответствующие языки запросов Apache TinkerPop Gremlin и SPARQL, что позволяет создавать запросы для эффективной навигации по наборам сложносвязанных данных. Neptune лежит в основе таких примеров использования графов, как сервисы рекомендаций, системы выявления мошенничества, графы знаний, разработка лекарственных препаратов и сетевая безопасность.

ArangoDB относится к мультимодельным СУБД, т.е. может быть отнесена как к графовым, так и документно-ориентированным, и к типу «ключ-значение». СУБД предназначена для построения высокопроизводительных приложений с использование SQL-подобного языка запросов (AQL) или расширений на JavaScript. Возможна реализация сервиса с REST/Web API.

OrientDB – система управления NoSQL базами данных с открытым исходным кодом, написанная на языке программирования Java. Это мультимодельная база данных, поддерживающая графы, документы, ключи/значения и объектные модели, но отношения в ней

регулируются так же, как и в графовых базах данных с прямыми связями между записями. Язык запросов построен на SQL.

Apache Cassandra – нереляционная отказоустойчивая распределенная СУБД, рассчитанная на создание высокомасштабируемых и надежных хранилищ огромных массивов данных, представленных в виде хэша. Проект был разработан на языке Java. Эта СУБД относится к гибридным NoSQL-решениям, поскольку она сочетает модель хранения данных на базе семейства столбцов с концепцией ключ-значение.

3.2. Рассмотрение программных средств для построения графов знаний

Для построения графов необходимы как инструменты для семантического анализа текстовой информации, так и для визуализации представленных данных.

3.2.1. LOD2 Stack

LOD2 Stack (The Linked Data Technology stack – стек технологий связанных данных) является проектом, который включает программные средства и компоненты, которые охватывают весь жизненный цикл связанных данных: хранение, интеграция, связывание, публикации, анализа и визуализации данных [19]. Основные компоненты LOD2 имеют открытый исходный код. Стек разработан таким образом, чтобы быть универсальным; для всех функциональных возможностей определяются четкие интерфейсы, которые позволяют подключать альтернативные сторонние реализации.

Архитектура LOD2 Stack основана на трех основных принципах: интеграция и развертывание программного обеспечения с использованием системы упаковки Debian, использование стандартизированных словарей для доступа к базе знаний и интеграции между различными инструментами, интеграция пользовательских интерфейсов на основе веб-приложений с поддержкой REST API [20].

LOD2 Stack включает методы для работы с данными на различных этапах жизненного цикла связанных данных с помощью следующих особенностей и концепций:

1) хранение данных (используются технологии хранения с динамической оптимизацией запросов, адаптивное кэширование соединений, оптимизированная обработка и масштабируемость кластера данных);

2) настройка (облегчает создание богатых семантических баз знаний, используя семантическую технологию);

3) взаимосвязь данных (используются связывающие подходы, дающие высокую точность и отзыв, которые настраиваются автоматически или с обратной связью с конечным пользователем);

4) классификация (необработанные данные связываются и интегрируются с онтологиями верхнего уровня, сохраняя целостность данных);

5) оценка качества данных (используются методы оценки качества на основе таких характеристик, как происхождение, контекст, охват или структура);

6) эволюция данных (данные могут быть динамичны, поэтому используются методы выявления проблем в базах знаний и автоматического предложения стратегий их устранения);

7) исследование данных – методы поиска, просмотра, исследования и визуализации различных видов связанных данных (т.е. пространственных, временных, статистических).

Таким образом, *LOD2 Stack* – проект, который определяет и способствует развитию стеку технологий для реализации связывания данных, что влечет за собой улучшение методов создания графов знаний. Данный проект не предлагает готовых решений, но определяет и развивает основополагающие концепции.

3.2.2. PoolParty Semantic Suite

PoolParty Semantic Suite представляет собой модульный масштабируемый программный пакет, который состоит из девяти основных независимых модулей: управление таксономией и тезаурусом; выделение текста и сущности; управление онтологией; концептуальная маркировка; интеграция данных; управление связанными данными; семантический поиск; рекомендательная система; аналитика и визуализация [21]. Одни из основных блоков программного пакета – семантический модуль (менеджер тезаурусов), программные модули для извлечения и анализа текстовой информации.

Семантический пакет PoolParty (PPTM – PoolParty Thesaurus Manager) – инструмент для создания и поддержки многоязычных систем организации знаний, предназначенный для использования простым пользователем, не имеющим опыта работы в семантической сети или специальных технических навыков, предоставляет функциональные возможности для публикации словарей в виде связанных открытых данных [22]. PoolParty Extractor предлагает API (программный интерфейс), предоставляющий алгоритмы интеллектуального анализа текста, основанные на семантических моделях знаний. Программный интерфейс имеет методы для автоматического анализа документов, который извлекает значимые фразы, категории именованных сущностей или другие метаданные. Извлекаемая информация или схемы метаданных могут быть сопоставлены с тезаурусом SKOS (стандартизированная модель организации знаний для семантической паутины), который используется в качестве единой семантической модели [23].

На официальном сайте платформы есть демонстрация одного из методов PoolParty Extractor, возможность тестирования выполнения одной из задач интеллектуального анализа текста. Например, на текст «PoolParty Semantic Suite is the most complete and advanced semantic middleware platform on the global market. It uses innovative means to help organizations build and manage enterprise knowledge graphs as a basis for their AI strategy» интеллектуальный анализ выдал результат нахождения основных терминов и описания тематики текста с оценкой, представлена на рис. 3.3.

Компонент преобразования данных PoolParty был создан и высоко развит в рамках проекта LOD2 Stack, который был ключевым для того, чтобы сделать жизненный цикл связанных данных полностью работоспособным и готовым продуктом. PoolParty Semantic Suite улучшил свои программные интерфейсы.

Таким образом, PoolParty Semantic Suite можно настроить в зависимости от задач организации и использовать только те модули, которые необходимы для их решения. Платформа является масштабируемой и гибкой, использует основные стандарты, которые позволяют работать с различными форматами данных.

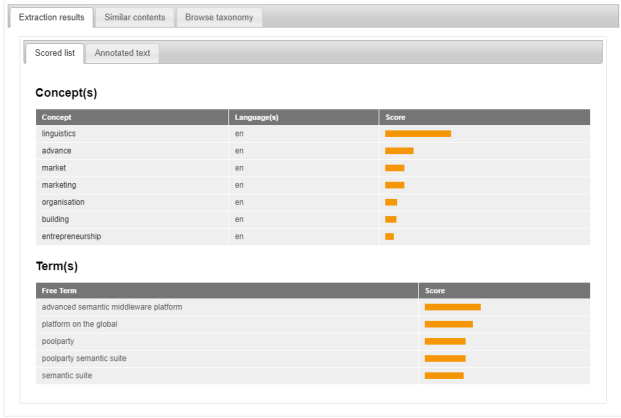


Рис. 3.3. Результат анализа текста методами инструмента PoolParty Extractor

3.2.3. VKGBuilder

Представление графа знаний можно разделить на общий граф знаний (GKG – general knowledge graph) и вертикальный граф знаний (VKG – vertical knowledge graph). Источник данных GKG более широк, чем VKG, и в нем больше сущностей и отношений. VKG основан на данных конкретной отрасли или темы. VKGBuilder используется для создания вертикальных графов знаний [24]. Инструмент VKGBuilder состоит из трех модулей: модуля интеграции знаний, модуля хранилища знаний и модуля доступа к знаниям.

Интеграция знаний – основной модуль для построения VKG с тремя основными компонентами: импортирование и извлечение данных, настройка схемы данных, обнаружение конфликтов и несоответствие схем. Модуль интеграции знаний работает со структурированными, слабоструктурированными и неструктурированными данными. Структурированные данные из внутренней реляционной базы данных импортируются и преобразуются в стандартизированный формат. Слабоструктурированные данные дополнительно проходят этап автоматизированной обработки для извлечения основных свойств и значений определенных объектов. Для неструктурированных данных дополнительно используются методы

интеллектуального анализа текста для обнаружения недостающих отношений между сущностями или заполнения значений свойств данной сущности. После извлечения или импорта семантических данных из различных источников выполняется интеграция данных для построения интегрированного графа знаний, в ходе которого исправляются конфликты с выбранной схемой данных [25].

Хранилище знаний представляет собой виртуальную графовую базу данных, которая объединяет нереляционные базы данных, хранилища в памяти и инвертированные индексы (структура данных, где для каждого слова перечислены все документы, в которых оно встретилось). Такая архитектура хранения используется для поддержки быстрого доступа к данным вертикального графа знаний. Модуль доступа предоставляет различные графические и программные интерфейсы для конечных пользователей.

Таким образом, *VKGBuilder* – инструмент для реализации вертикального графа данных, который способен описать отраслевые или тематические знания и использовать конкретные сущности рассматриваемой области.

3.2.4. GNOSS

GNOSS является платформой управления графами знаний и включает: GNOSS Knowledge Graph Builder и GNOSS Sherlock. Данная платформа – коммерческое решение и используется для взаимодействия с пользователями в определенном ресурсе, улучшая поиск информации и предоставляя интересующий контент [26].

GNOSS Knowledge Graph Builder находит взаимосвязь между сущностями и оптимизирует результаты поиска, а также предоставляет конкретный и релевантный контекст, обогащенную информацию по каждому вопросу и предлагает рекомендации в соответствии с интересами пользователей.

GNOSS Sherlock предоставляет компиляцию API, которая позволяет обрабатывать, понимать, интерпретировать и анализировать структурированные и неструктурированные тексты, также включает в себя чат-бота и инструмент бизнес-анализа. GNOSS Sherlock выполняет следующие задачи: обработку естественного языка, автоматическое создание графиков знаний и систем запросов, анализ и

рекомендации, основанные на распознавании сущностей и связывании, а также обнаружение знаний.

Граф знаний, построенный с помощью GNOSS, объединяет изолированную и разнородную информацию из различных систем вместе с информацией, полученной пользователями платформы. Также позволяет людям и машинам задавать вопросы, отображать значимые результаты и облегчает рассуждения о них для дальнейшего уточнения поиска. График знаний предоставляет конкретный и релевантный контекст, обогащенную информацию по каждому вопросу и предлагает рекомендации, которые соответствуют интересам пользователей.

На официальном сайте компании есть возможность попробовать методы компонента GNOSS Sherlock[10]. Например, в демонстрации был использован следующий текст:

> National Research Nuclear University MEPhI (Moscow Engineering Physics Institute) (Russian: Национальный исследовательский ядерный университет "МИФИ" / НИЯУ МИФИ or МИФИ) is one of the most recognized technical universities in Russia. MEPhI was founded in 1942 as the Moscow Mechanical Institute of Munitions (Московский механический институт боеприпасов, ММИБ), but it was soon renamed the Moscow Mechanical Institute. Its original mission was to train skilled personnel for the Soviet military and atomic programs. It was renamed the Moscow Engineering Physics Institute (Московский инженерно-физический институт) in 1953, which was its name until 2009.
>
> By the Order of the Government of Russia, issued by the Russian Government on April 8, 2009 (#480-r) on behalf of Russian President's Decree of October 7, 2008 (#1448) "On the pilot project launching on creating National Research Universities" the university MEPhI was granted this new status. The university was reorganized. The aim of the university existence is now preparing the specialists by giving them higher professional, post-graduation professional, secondary professional and additional professional education, as well as educational and scientific activities

При обработке данного текста инструментов автоматизированными способами получен результат семантического анализа, а также построен граф знаний (рис. 3.4 и 3.5).

Два компонента платформы могут использоваться и отдельно, если для решения задач организации достаточны только методы интеллектуального анализа. Таким образом, *платформа GNOSS –*

[10] https://sherlock.gnoss.ai/en/Demo.

коммерческое решение, направленное на улучшение качества информационного ресурса на основе запросов пользователей и их интересов.

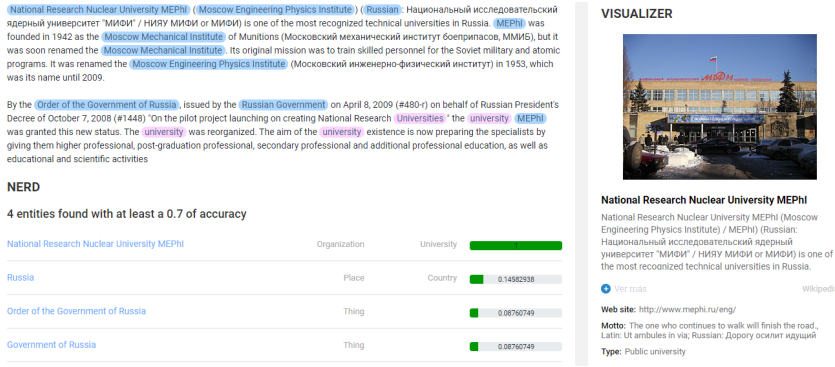


Рис. 3.4. Семантический анализ текста методами инструмента GNOSS Sherlock

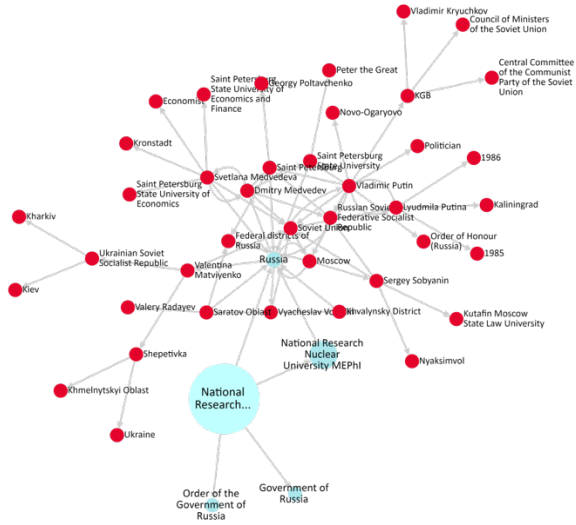


Рис. 3.5. Граф знаний на основе текста методами инструмента GNOSS Sherlock

3.3. Пример построения графа знаний

Как сказано в разд. 1 и 3.1, граф знаний может быть построен на основании имеющихся данных в глобальной сети Интернет и некоторой предварительно построенной схеме базы знаний.

Построение в данном случае может быть реализовано с помощью следующей схемы [5].

1. Извлечение данных с информационных источников и их сохранение в хранилище данных.
2. Извлечение сущностей и связей.
3. Построение онтологии графа знаний.
4. Хранение графа знаний.

Извлечение данных может быть осуществлено с помощью различных инструментов (например, методами web-scraping на Python, разработкой своего скраппера данных и т.д.). Извлечение сущностей и взаимосвязей может происходит следующими способами [27, 28]:

- вручную;
- алгоритмом TF-IDF (выделение сущностей);
- алгоритмом k-means (выделение связей);
- естественной обработкой текста (NLP);
- машинным обучением;
- другими.

Получение базы знаний может происходить двумя способами: выбором существующей базы знаний (например, приведены в разд. 3.1) или разработкой своей собственной. Во втором случае может понадобиться инструмент построения онтологий – классический программным обеспечением является Protégé.

В качестве примера реализована задача построения графа знаний политиков США на основе информации, представленной из одной из социальных сетей. Сбор информации пользователей социальной сети осуществлялся с помощью метода web-scraping и методами извлечения и анализа текста PDF-документов.

На странице пользователя представлены следующие блоки с данными: основная информация, опыт работы, образование, достижения, дополнительная информация. В качестве объектов выбраны

следующие данные, которые могут быть представлены в анкете пользователя:

- имя пользователя (Person);
- учебные заведения (Education);
- организации (Organisation);
- контактная информация (Contact);
- должность (Position);
- отрасль, в которой работает пользователь (Industry);
- интересы (Interest);
- навыки (Skill);
- другое (Other).

Таким образом, с каждой анкеты пользователя социальной сети извлекалась информация данных объектов, сохранялась в формате JSON для последующей обработки и загрузки в графовое хранилище.

В качестве графового хранилища выбрана база данных Neo4j. Для загрузки, кроме узлов, необходимо определить тип ребер. Ребра определяются на основе особенностей собранных данных, в рассматриваемой случае выбраны следующие ребра для связи объектов:

- учится(-лся) в (EDUCATED_AT);
- работает(-л) в (WORKED_IN);
- работает(-л) на должности (WORKED_AS);
- обладает навыком (SKILLED_WITH);
- имеет следующую контактную информацию (CONTACT_BY);
- имеет опыт в отрасли (INDUSTRY_KNOWLEDGE_OF).

После определения узлов и ребер разработана программа на языке программирования Python для создания и загрузки собранных данных в графовое хранилище. Таким образом, по результату работы программы создается граф (рис. 3.6).

Граф знаний может использоваться как инструмент поиска объектов, так и инструмент для анализа информации из собранных данных, в данном случае политиков США. Анализ информации графа позволяет, например, выявлять некоторые шаблоны карьерного продвижения целевых пользователей социальной сети. Например, если на графе видно, что одна персона занимает значимые должности в нескольких государственных органах, поэтому можно прогнозировать, что в будущем она займет ключевую позицию в другом госоргане. Таким примером может выступать Марк Эспер, который

занимал должности вице-президента Торговой палаты США, а затем занял пост министра обороны США, что отображено на графе (рис. 3.7).

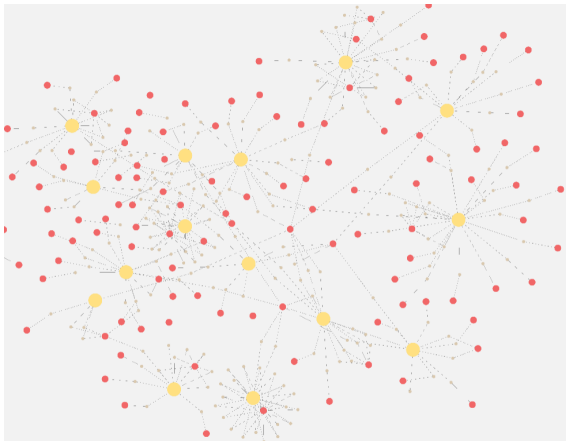


Рис. 3.6. Созданный граф в базе данных Neo4j
на основе информации пользователей социальной сети

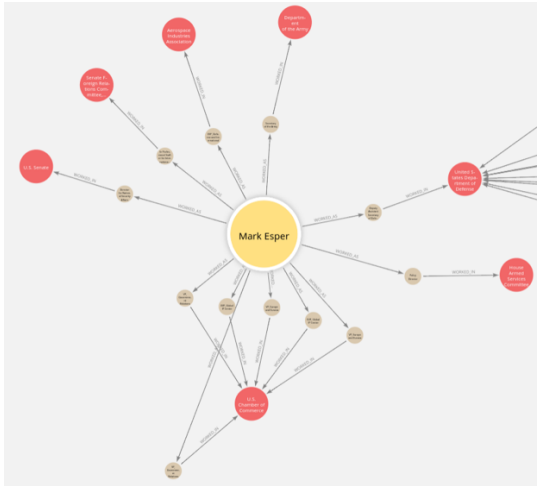


Рис. 3.7. Пример шаблона карьерного продвижения
по ключевым должностям в разных государственных органах

Графовая база данных Neo4j имеет интерфейс для составления запросов. Запросы могут использоваться при дополнительном анализе собранных данных и позволят выявить скрытые связи между группой объектов. Запросы в Neo4j пишутся на языке запросов Cypher. Цель Cypher – предоставить человеко-читаемый язык запросов к графовым базам данных, похожий на язык запросов реляционных баз данных SQL.

Запрос для получения сведений о карьере людей, которые когда-либо учились в образовательном учреждении под названием «Ecole nationale d'Administration» будет иметь вид:

MATCH (o:Organisation)-[:WORKED_IN]-(p:Position)-[:WORKED_AS]-(n:Person)-[:EDUCATION_AT]-(e:Education) WHERE e.name = "Ecole nationale d'Administration" RETURN o, p, n, e

Результат выполнения данного запроса представлен на рис. 3.8.

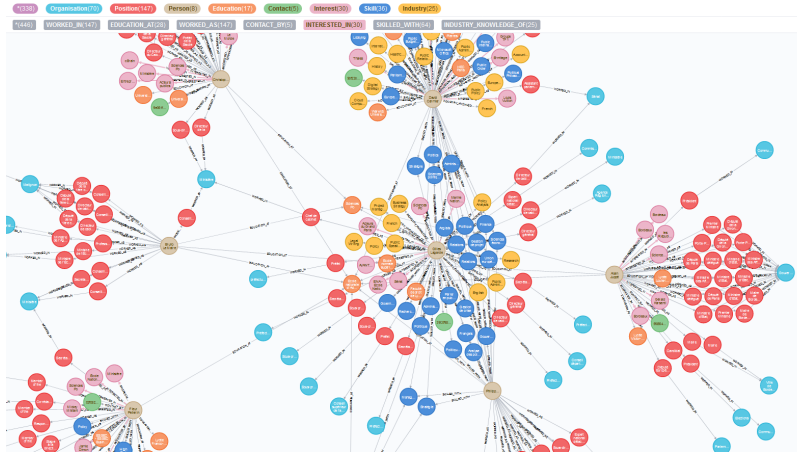


Рис. 3.8. Сведения о карьере людей, которые закончили университет «Ecole nationale d'Administration»

Таким образом, описан процесс построения графа знаний на примере графовой базы данных Neoj4 в комплекте с представленными

инструментами и программными обеспечениями, что предоставляет возможность анализа собранных данных с помощью визуального представления в виде графов, а также используя язык запросов Cypher.

ЗАКЛЮЧЕНИЕ

В пособие даны определения и методология проектирования графа знаний. Рассмотрены области применения технологии графа знаний на таких примерах, как Google Knowledge Graph, Facebook Social Graph, DBpedia и т.д. Приведено описание проекта Semantic Web, в основе которого лежит технология графа знаний. Распространение применения графов знаний для структурирования информации в сети, в том числе такими технологическими гигантами, как Google, Facebook и др., является свидетельством того, что в будущем информация в интернете будет представлена в виде «семантической паутины».

Анализ динамики публикационной и патентной активности показал динамичный рост количества ежегодно публикуемых статей и регистрируемых патентов, особенно выраженный в последние несколько лет, что свидетельствует об увеличивающемся интересе к данному направлению и его высокой значимости.

Технологии графов знаний преимущественно относятся к области компьютерных наук, в рамках которой разрабатываются методы создания и ведения таких графов. Выявлены сферы, в которых используются информационно-аналитические системы на основе графов знаний, а именно: инженерия и проектирование, здравоохранение и науки о жизни, менеджмент и принятие решений, социальная сфера и образование. Необходимо отметить, что графы знаний предположительно рассматриваются некоторыми странами, в частности США и КНР, в качестве технологии двойного назначения, так как соответствующие разработки финансируются и ведутся военными ведомствами и организациями.

Анализ программных продуктов выявил, что один их возможных способов создания графа знаний может быть применение графовой СУБД, которая при необходимости может использовать для построения базы знаний средства создания онтологий. Также можно сделать вывод, что создание графов знаний в программной части продолжится как для графовых СУБД, так и для СУБД/сервисов с семантической моделью данных.

Рассмотрены основные программные обеспечения, которые позволяют проводить интеллектуальный анализ текста на естественном языке, выделяя основные понятия и отношения между ними. С

развитием применения графа знаний в различных областях создаются проекты по стандартизации и определению стека технологий, например LOD2 Stack. С помощью данного проекта разрабатываются и дополняются готовые решения. Организацией Semantic Web реализован модульный масштабируемый программный пакет PoolParty Semantic Suite, который имеет методы и готовые решения для создания графа знаний. А такие платформы, как GNOSS, позволяют использовать данный подход при работе с клиентами в бизнесе, улучшая поиск информации и таргетирование информации.

Технология хранения, анализа и поиска информации с помощью графа знаний только начинает развиваться. Однако, несмотря на относительно недавнее появление данной технологии, уже имеются платные и с открытым исходным кодом программные обеспечения и инструменты. Это позволяет применять графы знаний в различных научных областях, что способствует ее развитию и дальнейшему широкому применению. В рамках пособия описана реализации графа знаний политиков США на примере использования графовой базы данных Neo4j. На примере показано, что граф способствует анализу собранных данных с помощью визуального представления и языка запросов, позволяет выявлять скрытые связи и прогнозировать события в рамках решаемой задачи.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ И ИНТЕРНЕТ-ИСТОЧНИКОВ

1. Group R.W. Resource Description Framework (RDF). 25 февраля 2014. Available: https://www.w3.org/RDF/. [Дата обращения: май 2021.]
2. Chaudhri V. K., Chittar N., Genesere M. The Stanford AI Lab Blog 10 мая 2021. Available: http://ai.stanford.edu/blog/introduction-to-knowledge-graphs/. [Дата обращения: май 2021.]
3. Ehrlinger L. Towards a Definition of Knowledge Graphs SEMANTiCS 2016, 2016.
4. Lehmann J., Isele R., Jakob M. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia Semantic Web. 2012. T. 1. P. 1–5. 2012.
5. DBpedia About DBpedia 2021. Available: https://www.dbpedia.org/about/. [Дата обращения: май 2021.]
6. Marlow C., Ugander, J. The Anatomy of the Facebook Social Graph Research Gate. 2011.
7. University C. M. Read the Web. Project Overview 2019. Available: http://rtw.ml.cmu.edu/rtw/overview. [Дата обращения: май 2021.]
8. University C.M. NELL 22 февраля 2019. Available: https://twitter.com/cmunell. [Дата обращения: май 2021.]
9. Group S.W.W. SEMANTIC WEB 2015. Available: https://www.w3.org/standards/semanticweb/#:~:text=The%20term%20%E2%80%9CSemantic%20Web%E2%80%9D%20refers,SPARQL%2C%20OWL%2C%20and%20SKOS. [Дата обращения: май 2021.]
10. Поиск документов 2021. Available: https://www.scopus.com/. [Дата обращения: 11.01.2021.]
11. Тезаурус WordNet Available: http://db4.sbras.ru/elbib/data/show_page.phtml?20+1531. [Дата обращения: 20.05.2021.]
12. The Gene Ontology Resource 2021. Available: http://geneontology.org/. [Дата обращения: 20.05.2021.]
13. Knowledge Graphs are on the rise 2019. Available: https://2019.semantics.cc/knowledge-graphs-are-rise. [Дата обращения: 20.05.2021.]
14. Jose Manuel Gomez-Perez, Jeff Z. Pan Guido Vetere. Honghan Wu. Enterprise Knowledge Graph: An Introduction. Exploiting Linked Data and Knowledge Graphs in Large Organisations. 2017. P. 1–14.

15. Nayantara Jeyaraj (Taro). Conceptualizing the Knowledge Graph Construction Pipeline». Available: https://towardsdatascience.com/ conceptualizing-the-knowledge-graph-construction-pipeline-33edb25ab831 [Дата обращения: 25.05.2021.]

16. Yu Yaozu, Zhang Jiangen. Constructing government procurement knowledge graph based on crawler data / Journal of Physics: Conference Series. 1693(1):012032.

17. Bhattacharyya A., Chakravarty D. Graph Database: A Survey. 2020 International Conference on Computer, Electrical and Communication Engineering (ICCECE 2020). 2020. P.1–8.

18. Graph Database Management (GDBMS) Platform». Available: http://www.gabormelli.com/RKB/Graph_Database_Management_(GDBMS)_Platform. [Дата обращения: 26.04.2021.]

19. LOD2 StackAvailable: https://lod2.eu/stack/. [Дата обращения: 24.05.2021.]

20. Sören Auer. Introduction to LOD2. Linked Open Data – Creating Knowledge Out of Interlinked Data. 2014 P. 1–17.

21. Grinko O., Kupriyanovsky V., Pokusaev O., Volokitin Y., Ponkin I., Namiot D., Redkina A. The ontologization of European Union data as a transition from a data economy to a knowledge economy // International Journal of Open Information Technologies. 2018. 6(11). P. 65–84.

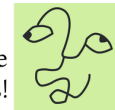
22. Nagy H., Pellegrini T., Schandl T., Blumauer A., Mader C. Realizing thesaurus based uses cases with the PoolParty Suite. Bid-textos universitaris de biblioteconomia i documentacio. 2011. 27.

23. Martínez-González M.M., Alvite-Díez M.L. A semantic web methodological framework to evaluate the support of integrity in thesaurus tools // Journal of Information Science. 2020. 46(3). P. 378–391.

24. Ruan T., Wang H., Hu F. VKGBuilder-A Tool of Building and Exploring Vertical Knowledge Graphs. // In International Semantic Web Conference (Posters & Demos). 2014, October. P. 445–448.

25. Leijie F., Yv B., Zhenyuan Z. Constructing a vertical knowledge graph for non-traditional machining industry // 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) 2018, March. P. 1–5.

26. Knowledge Graph – GNOSS. Available: https://www.gnoss.com /en/knowledge-graph. [Дата обращения: 27.04.2021.]



27. Bai J., Che L. Construction and Application of Database Micro-course Knowledge Graph Based on Neo4j. 2021. ACM International Conference Proceeding Series, PartF168982, 3450798.

28. Lu R., Cai Z., Zhao S. A Survey of Knowledge Reasoning based on KG. 2019. IOP Conference Series: Materials Science and Engineering 569(5), 052058.

Современные технологии и средства
построения графа знаний

Учебно-методическое пособие

Редактор М.В. Макарова

Подписано в печать 16.01.2023. Формат 60x84 1/16
Уч.-изд. л. 2,75. Печ. л. 2,75. Тираж 100 экз.
Изд. № 031-1. Заказ № 6.

Национальный исследовательский ядерный
университет «МИФИ»
Типография НИЯУ МИФИ.
115409, Москва, Каширское ш., 31.