К построению графа знаний коллекции математических статей

Б.Т. Гизатуллин[1], О.А. Невзорова[1]

[1] *Казанский Федеральный Университет*

Аннотация. В работе описывается процесс построения графа знаний для коллекции математических статей на русском языке из журнала «Известия ВУЗов. Математика». Коллекция содержит около 1100 документов формата LaTeX. Для построения графа знаний используется разработанная онтология представления графа знаний. Из коллекции выделяются базовые сущности статей: коды УДК, авторы, названия статей, использованные формулы, даты публикации статей, аффилиации авторов, ссылки на другие работы. Каждая выделенная сущность статей записывается в граф знаний через специальное отношение. Также на коллекции проводится тематическое моделирование с использованием метода латентного размещения Дирихле, для которого были подобраны оптимальные гиперпараметры. Выделенные тематики документов записываются через связи в граф знаний. Математические термины статей извлекаются с использованием онтологии OntoMath^{PRO} и также включаются в граф знаний. Для построения графа знаний разработаны программные инструменты, позволяющие создавать граф знаний для любой научной коллекции, удовлетворяющей изученным шаблонам представления статей.

Ключевые слова: Построение графа знаний, Связанные данные, Тематическое моделирование, Математическая статья

Towards Building the Knowledge Graph for a Collection of Mathematical Articles

B.T. Gizatullin[1], O.A. Nevzorova[1]

[1] *Kazan Federal University*

Abstract. This paper describes the process of creating a knowledge graph for a collection of mathematical articles in the Russian language, gathered from the "Izvestiya VUZov. Matematika" journal. The collection consists of approximately 1100 documents in LaTex format. The work involves constructing an ontology for the collection of mathematical articles, which will serve as the basis for the created knowledge graph. Various article objects are

extracted from the collection, including universal decimal classification codes, authors, titles, used formulas, articles publication dates, authors affiliations and references to other works. Each object is recorded through a specific relationship in the knowledge graph. Thematic modeling is also performed on the collection using the latent Dirichlet allocation method, for which optimal hyperparameters are selected. The document themes are recorded in the knowledge graph through relationships. An interesting approach is used for extracting mathematical terms. In this work, mathematical entities are identified in the documents using the OntoMath^{PRO} ontology. During the knowledge graph construction process, tools were developed that allow the creation of a knowledge graph on any collection that meets the patterns of the original collection. The resulting knowledge graph can serve as a foundation for various research purposes and the development of intelligent systems, that can be used by researchers, journals, as well as students.

Keywords: Knowledge graph construction, Linked Data, Topic modeling, Mathematical Paper

Введение

Современное научное сообщество активно публикует большое количество математических статей, представляющих собой ценные знания исследователей со всего мира. Однако с увеличением объема научной литературы всё сложнее становится эффективно анализировать, управлять и использовать научную информацию. Представление математических знаний в формализме графа знаний позволяет улучшить представление и обработку больших математических коллекций. В статье подробно описан процесс построения графа знаний для коллекции математических статей научного журнала на русском языке. Граф знаний позволит систематизировать и структурировать различные данные, извлекаемые из математических статей. Граф знаний позволит устанавливать связи между статьями на основе их содержания, что поможет выявить, в том числе, и скрытые связи и провести более глубокий анализ предметной области.

В качестве коллекции, на которой проводилось построение графа, использовался набор из 1114 научных статей в формате LaTeX из журнала «Известия высших учебных заведений. Математика» за 1997-2005 гг.

Близкие работы

Граф знаний — это граф данных, предназначенный для накопления и передачи знаний о реальном мире. Узлы такого графа представляют интересующие объекты, а ребра – отношения между этими объектами.

Два типа графов знаний, применимых на практике, выделены в [1]: открытые графы знаний и графы знаний предприятия. Открытые графы знаний публикуются в Интернете, что делает их содержание доступным для общественности. Наиболее яркими примерами открытых графов являются DBpedia [2], Freebase [3], Wikidata [4], YAGO [5] и др.

Корпоративные графы знаний обычно являются внутренними для компаний и применяются в коммерческих целях [6]. Примеры отраслей, использующих корпоративные графы знаний, включают поисковые системы (например, Bing, Google), коммерцию (например, Airbnb, Amazon, eBay, Uber), социальные сети (например, Facebook, LinkedIn) и т.д.

Открытые графы знаний также создаются для конкретных областей. Наибольший интерес для данной работы представляет построение графа знаний на основе публикаций, относящихся к академической литературе в различных областях (например, OpenCitations [7], SciGraph [8], Microsoft Academic Knowledge Graph [9], открытый граф знаний для платформы публикаций [10]).

В настоящей работе представлен новый открытый граф знаний, разработанный на основе коллекции русскоязычных математических статей научного журнала. Этот граф знаний использует при построении онтологию OntoMath^{PRO} [11, 12]. Наша основная цель — разработать модель графа математических знаний, которую можно в дальнейшем использовать для представления различных коллекций математических статей, а также для разработки инструментов аналитики графовых данных.

Выделение типов объектов и связей

Перед построением графа знаний необходимо определить, какие типы объектов и связей будут в нем присутствовать. Для описания такой структуры построена специальная онтология представления математических статей MathCollectionOntology (mco). Эта онтология является минимальной для построения необходимого графа знаний. Ниже представлены некоторые классы онтологии MathCollectionOntology, их описания и соотносимые классы внешних онтологий. Для выравнивания использованы онтологии Dublin Core (http://purl.org/dc/terms/), онтология BIBO (http://purl.org/ontology/bibo/) и онтология DBPedia (https://www.dbpedia.org/resources/ontology/) .

Таблица 1. Классы онтологии представления MathCollectionOntology

Описание класса	Класс онтологии представления	Выравнивание на внешние онтологии
Коллекция математических статей	mco:collection	owl:equivalentClass bibo:Collection
Произвольный текст	mco:text	owl:equivalentClass bibo:Article
Автор статьи	mco:creator	rdfs:subClassOf dcterms:Agent
Математическая статья (подкласс класса text)	mco:article	rdfs:subClassOf bibo:AcademicArticle;
Аффилиация автора	mco:affiliation	owl:equivalentClass dbo:Organisation

Указанные классы определяют типы объектов при построении графа знаний. Ниже описаны свойства, определенные в онтологии представления, которые используются для задания отношений между объектами графа знаний, где возможно, в конце пунктов, добавлено выравнивание свойств на внешние онтологии:

- mco:has_omp_term – связывает объект типа «article» с литералом, представляющим русскоязычный лейбл класса из онтологии OntoMath^{PRO} (термин из онтологии), под лейблом понимается название концепта (представленного классом онтологии). Свойство указывает, что данный термин встречается в математической статье.
- mco:use_formula – связывает литерал-формулу с объектом типа «article».
- mco:has_lda_topic – связывает объект типа «article» с темой (литерал-слово), полученной после применения тематического моделирования методом латентного размещения Дирихле на коллекции. (rdfs:subPropertyOf dcterms:subject).
- mco:has_publication_date – связывает объект типа «article» с литералом, обозначающим дату публикации объекта. (rdfs:subPropertyOf dcterms:date).
- mco:has_title – связывает объект типа «article» с литералом-заголовком. (owl:equivalentProperty dcterms:title).
- mco:contains – связывает объект типа «collection» объектом типа «article». (owl:equivalentProperty dcterms:hasPart).
- mco:creator_name – связывает объект типа «creator» с литералом, задающим имя автора.
- mco:has_udc_code – связывает объект типа «article» с литералом, задающим код универсальной десятичной классификации (УДК) статьи. (rdfs:subPropertyOf dcterms:subject).
- mco:created_by – связывает объект типа «text» с объектом типа «creator». (owl:equivalentProperty dbo:creator).
- mco:has_bibliograhic_citation – указывает ссылку статьи на некоторый текст, связывает объект типа «article» с объектом типа «text». (owl:equivalentProperty dcterms:References, bibo:cites).
- mco:has_affilation – указывает аффилиацию автора, связывает объект типа «creator» с объектом типа «affiliation». (owl:equivalentProperty dbo:affiliation).
- mco:affilation_name – связывает объект типа «affiliation» с литералом, задающим название аффилиации.

Таким образом, была создана онтология, состоящая из 6 классов (включая базовый класс) и 12 отношений.

Извлечение объектов и связей из коллекции

Далее описывается процесс выделения различных сущностей из статей математической коллекции. Схема процесса создания графа знаний для коллекции представлена на рис. 1.

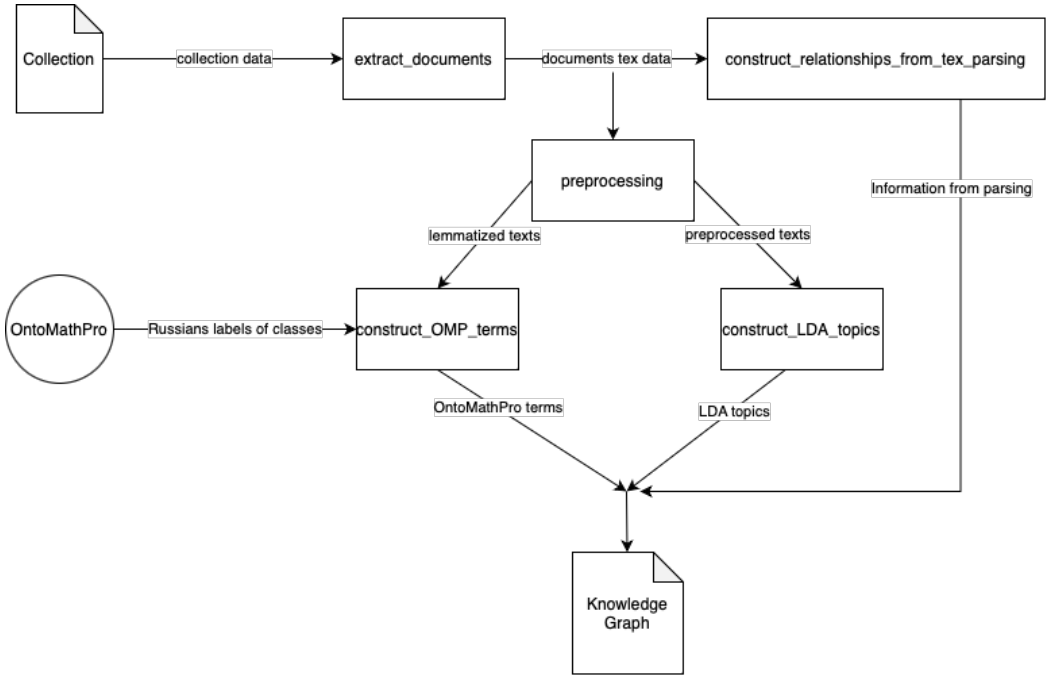

Рис. 1. Схема создания графа знаний для коллекции математических статей

На вход подается коллекция, состоящая из файлов математических статей в формате LaTeX, эти статьи открываются в нужной кодировке (подбирается верная кириллическая кодировка из списка возможных). Затем из LaTeX кода извлекаются необходимые сущности: коды УДК, даты публикации статьи, авторы, ссылки, заголовки, используемые формулы, аффилиации авторов. Проводится стандартная предобработка текстов статей (удаление стоп-слов, токенизация, лемматизация). Далее из лемматизированных текстов извлекаются математические термины с помощью онтологии OntoMath^{PRO}, а также слова, которые характеризуют тематику текстов.

Ниже представлено более подробное описание процесса получения математических терминов и тем статей коллекции.

Для выделения объектов для связи «mco:has_omp_term» из статей выделяются математические понятия, которые присутствуют в онтологии OntoMath^{PRO}. Для этого из онтологии отбираются лейблы – названия концептов (классов) на русском языке. Сохраняются оригинальные версии названий концептов и их версии после лемматизации. Далее, для каждого лемматизированного текста проверяется, присутствует ли в нем тот или

иной лемматизированный лейбл. Если присутствует, то оригинальная версия найденного лейбла добавляется в список терминов для документа.

Для выделения объектов отношений «mco:has_lda_topic», было проведено тематическое моделирование текстов коллекции с использованием латентного размещения Дирихле. Была обучена модель, для которой были подобраны оптимальные гиперпараметры, при подборе гиперпараметров поиском по сетке максимизировалась метрика CV Coherence. Для исходной коллекции при 11 темах метрика была наибольшей. Также определялись гиперпараметры, связанные с формированием словаря: ограничение на максимальную долю документов, в которых может встречаться терм, чтобы он был сохранен в словаре (метрика была наибольшей при значении 0.3), а также определенное число документов, в которых терм должен быть обязательно представлен (метрика была наибольшей при значении 10).

Для каждой темы были отобраны 10 лучших слов, которые ее характеризуют. Для каждого документа проверяется, присутствует ли то или иное слово, которое характеризует наиболее вероятную тему этого документа, в тексте самого документа, и, если присутствует, то это слово добавляется в список слов-тематик этого документа. Выделенные слова-тематики записывались в граф знаний через свойство «mco:has_lda_topic».

Таблица 2 показывает распределение количества статей по выявленным темам коллекции, а также представляет пять наиболее характерных слов для каждой темы.

Таблица 2. Распределение тематик в коллекции

Номер темы	Число статей, отнесенных к этой теме	Характерные для этой темы слова (top 5, после лемматизации)
1	148	выпуклый, функционал, обобщённый, найтись, вариационный.
2	36	вершина, граф, многогранник, шаг, ребро.
3	293	кривая, собственный, краевой, поверхность, коши.
4	89	устойчивость, алгоритм, процесс, устойчивый, время.
5	159	подпространство, банахов, приближение, мера, приблизить.
6	78	схема, разностный, сетка, итерационный, краевой.
7	65	алгебра, базис, многообразие, тождество, решётка.
8	105	поле, связность, многообразие, тензор, кривизна.

9	33	управление, семейство, оптимальный, функционал, процедура.
10	58	группа, модуль, подгруппа, кольцо, полугруппа.
11	50	Многообразие, структура, группа, расслоение, слой.

Некоторые статистические параметры построенного графа

Всего в графе знаний построено 203481 связи. Количество связей внутри графа по некоторым типам связей показано в Таблице 3.

Таблица 3. Распределение количества связей по некоторым типам связей

Тип связи	Число связей
has_omp_term	81465
has_lda_topic	5298
has_udc_code	1229
use_formula	56935
has_bibliograhic_citation	10545

Среди всех статей коллекции, наибольшее количество связей «mco:has_omp_term», то есть наибольшее количество используемых понятий из русскоязычных лейблов классов онтологии OntoMath^{PRO}, получилось равным 184, а наименьшее – 18. Количество различных цитируемых текстов – 7974, количество уникальных авторов среди текстов, на которые ссылаются статьи сборника – 5924. Также выявлен текст с наибольшим количеством ссылок среди статей коллекции. Этот текст упоминается 36 раз. Количество уникальных авторов в коллекции – 947.

Заключение

В результате проведенного исследования была построена онтология представления графа знаний математической коллекции, в которой определены основные классы и отношения.

Граф знаний построен на основе коллекции математических документов, из статей которой в формате LaTeX извлекались все требуемые объекты.

На коллекции математических статей проведено тематическое моделирование с использованием латентного размещения Дирихле, для которого подобраны оптимальные гиперпараметры. Слова, характеризующие тематики документов, записывались через связи в граф знаний.

С помощью онтологии OntoMath^{PRO} выделены математические термины статей. Были посчитаны различные статистики построенного графа знаний.

В процессе работы созданы программные инструменты, позволяющие построить граф знаний на любой коллекции математических статей, которые подходят под шаблоны исследуемой коллекции.

Построенный граф знаний может служить основой для разнообразных научных и прикладных исследований, таких как рекомендательные системы, системы поиска рецензентов и близких работ, системы назначения кодов УДК математическим статьям.

Дальнейшие исследования включают изучение и анализ построенного графа знаний, расширение выделяемых из статей сущностей, расширение количества коллекций, а также добавление коллекций на английском языке, и создание связей на основе содержания между исследованиями на разных языках.

Благодарности. Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 21-11-00105.

Литература

1. Hogan, A., Gutierrez, C., Cochez, M., et al.: Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge, 237 p. Springer Cham (2022).
2. Lehmann, J., Isele, R., Jakob, M., et al. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal, 6(2), 167–195 (2015).
3. Bollacker, K., Cook, R., Tufts, P.: Freebase: a shared database of structured general human knowledge. In: Proceedings of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1962–1963 (2007). AAAI Press.
4. Vrandečić, D., and Krötzsch, M.: Wikidata: A free collaborative knowledge base. Communications of the ACM, 57(10), pp. 78–85 (2014).
5. Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G., and Weikum, G.: YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In: Srinivasan, S., Ramamritham, K., Kumar, A., et al. (eds.) Proc. of the 20th International Conference on World Wide Web, pp. 229–232, ACM Press, India, Hyderabad (2011).
6. Noy, N. F., Gao, Y., Jain, A., et al.: Industry-scale knowledge graphs: Lessons and challenges. Communications of the ACM 62(8), 36–43 (2019).
7. Peroni, S., Shotton, D. M., and Vitali, F.: One Year of the OpenCitations Corpus: Releasing RDF-Based Scholarly Citation Data into the Public Domain. In: d'Amato, C., Fernández, M., Tamma, V., et al. (eds.), The Semantic Web – ISWC – 16th International Semantic Web Conference,

Proceedings, Part II (Lecture Notes in Computer Science), vol. 10588, pp. 184–192. Springer, Cham (2017).
8. Iana, A., Jung, S., Naeser, P., et al.: Building a conference recommender system based on SciGraph and WikiCFP. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., et al. (eds.) Semantic Systems. The Power of Al and Knowledge Graphs. Lecture Notes in Computer Science, vol. 11702, pp. 117–123. Springer, Cham (2019).
9. Färber, M.: The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In: The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science, vol. 11779, pp. 113-129. Springer, Cham (2019).
10. Nevzorova, O., Zhistlov N. et al.: Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics. In: The Semantic Web – ISWC 2013. ISWC 2013. Lecture Notes in Computer Science, vol. 8218, pp. 379–394. Springer, Berlin, Heidelberg (2013).
11. Nevzorova, O., Zhiltsov, N., Kirillovich, A., Lipachev, E.: OntoMath^{PRO} Ontology: A Linked Data Hub for Mathematics. In: Klinov, P., Mouromtsev, D. (eds.) Knowledge Engineering and the Semantic Web. KESW 2014. Communications in Computer and Information Science, vol. 468, pp. 105-119. Springer, Cham (2014).
12. Елизаров А.М., Кириллович А.В., Липачёв Е.К., Невзорова О.А. Онтология математического знания OntoMathPRO // Доклады Российской академии наук. Математика, информатика, процессы управления. 2022. Т. 507. № 1. С. 29–35. https://doi.org/10.31857/S2686954322700011.

References

1. Hogan, A., Gutierrez, C., Cochez, M., et al.: Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge, 237 p. Springer Cham (2022).
2. Lehmann, J., Isele, R., Jakob, M., et al. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal, 6(2), 167–195 (2015).
3. Bollacker, K., Cook, R., Tufts, P.: Freebase: a shared database of structured general human knowledge. In: Proceedings of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1962–1963 (2007). AAAI Press.
4. Vrandečić, D., and Krötzsch, M.: Wikidata: A free collaborative knowledge base. Communications of the ACM, 57(10), pp. 78–85 (2014).
5. Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G., and Weikum, G.: YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In: Srinivasan, S., Ramamritham,

K., Kumar, A., et al. (eds.) Proc. of the 20th International Conference on World Wide Web, pp. 229–232, ACM Press, India, Hyderabad (2011).
6. Noy, N. F., Gao, Y., Jain, A., et al.: Industry-scale knowledge graphs: Lessons and challenges. Communications of the ACM, 62(8), 36–43 (2019).
7. Peroni, S., Shotton, D. M., and Vitali, F.: One Year of the OpenCitations Corpus: Releasing RDF-Based Scholarly Citation Data into the Public Domain. In: d'Amato, C., Fernández, M., Tamma, V., et al. (eds.), The Semantic Web – ISWC – 16th International Semantic Web Conference, Proceedings, Part II (Lecture Notes in Computer Science), vol. 10588, pp. 184–192. Springer, Cham (2017).
8. Iana, A., Jung, S., Naeser, P., et al.: Building a conference recommender system based on SciGraph and WikiCFP. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., et al. (eds.) Semantic Systems. The Power of Al and Knowledge Graphs. Lecture Notes in Computer Science, vol. 11702, pp. 117–123. Springer, Cham (2019).
9. Färber, M.: The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In: The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science, vol. 11779, pp. 113-129. Springer, Cham (2019).
10. Nevzorova, O., Zhistlov N. et al.: Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics. In: The Semantic Web – ISWC 2013. ISWC 2013. Lecture Notes in Computer Science, vol. 8218, pp. 379–394. Springer, Berlin, Heidelberg (2013).
11. Nevzorova, O., Zhiltsov, N., Kirillovich, A., Lipachev, E.: OntoMath^{PRO} Ontology: A Linked Data Hub for Mathematics. In: Klinov, P., Mouromtsev, D. (eds.) Knowledge Engineering and the Semantic Web. KESW 2014. Communications in Computer and Information Science, vol. 468, pp. 105-119. Springer, Cham (2014).
12. Elizarov A., Kirillovich, A., Lipachev, E., Nevzorova, O. OntoMathPro – Ontology of Mathematical Knowledge // Doklady Mathematics. 2022. Vol. 507(1). Pp. 29–35. https://doi.org/ 10.31857/S2686954322700011.