

Метод формирования графа знаний математических уравнений

Е. К. Липачёв^[0000-0001-7789-2332], Б. Р. Мурадымов^[0009-0004-1187-8158]

*Казанский федеральный университет,
Институт информационных технологий и интеллектуальных систем,
Университет Иннополис, г. Иннополис*

Аннотация. Предложен метод построения графа знаний математических уравнений. В качестве источников знаний при проектировании и формировании графа знаний использовались научно-образовательный сайт EqWorld «Мир математических уравнений» и онтология профессиональной математики OntoMath^{PRO}. Разработана система программных инструментов, обеспечивающих полный цикл формирования графа знаний. Произведено сохранение графа знаний в хранилище данных с возможностью выполнения SPARQL-запросов.

Ключевые слова: граф знаний, математическое уравнение, математическая онтология, представление математического знания.

Method of Forming a Knowledge Graph of Mathematical Equations

Е. К. Lipachev^[0000-0001-7789-2332], B. R. Muradymov^[0009-0004-1187-8158]

*Institute of Information Technologies and Intelligent Systems,
Kazan Federal University,
Innopolis University, Innopolis, Russia*

Abstract. We propose an approach to constructing a knowledge graph of mathematical equations. As sources of knowledge when designing and forming a knowledge graph, we use the scientific and educational site EqWorld “The World of Mathematical Equations” and the professional mathematics ontology OntoMath^{PRO}. We have developed a system of software tools that provide a full cycle of forming a knowledge graph. We have organized the storage of the knowledge graph in a data warehouse with the ability to execute SPARQL queries.

Keywords: knowledge graph, mathematical equation, mathematical ontology, representation of mathematical knowledge.

1. Введение

Графы знаний являются основой многих информационных систем, использующих доступ к структурированным знаниям.

Термин «граф знаний» был введен в официальном блоге Google для обозначения использования семантических знаний (определенных как «вещи, а не строки») в веб-поиске [1]. Этот термин также используется для обозначения баз знаний сети, таких как DBpedia. В широком смысле, любое представление некоторых знаний в виде графа можно рассматривать как граф знаний. Однако не существует общепринятого определения того, что такое граф знаний. Например, в [2] предложено вместо формального определения графа знаний ограничиться минимальным набором характеристик графов знаний, с помощью которых можно отличать графы знаний от других наборов знаний.

Можно определить граф знаний как направленный граф, состоящий из узлов и ребер, где один узел указывает на сущность (реальный объект или абстрактное понятие), а ребро между двумя узлами передает семантическое отношение между двумя сущностями. Структура описания ресурсов (Resource Description Framework, RDF) и графы маркированных свойств (Labeled Property Graphs, LPG) – два основных способа представления и управления графами знаний (см., например, [3, 4]).

Обзоры исследований по графам знаний приведены в [3–5]. В статье [5] перечислены семь важных направлений исследований по моделированию и построению графов знаний, а также методам объединения и обогащения знаний.

Отметим работы, связанные с построением и применением графов знаний в математике. В работе [6] рассматривается технология формирования графа знаний современных приложений в математике на примере онтологии семантической библиотеки LibMeta. Целью исследования [7] является создание семантической модели предметной области междисциплинарного научного журнала и использование графа знаний, созданного на основе онтологии библиотеки LibMeta. В работе [8] описывается процесс построения графа знаний для коллекции математических статей на русском языке из журнала «Известия ВУЗов. Математика». Для построения графа знаний используется разработанная онтология представления графа знаний.

В настоящей работе представлен подход построения графа знаний математических уравнений, основанный на извлечении знаний из научно-образовательного сайта EqWorld и онтологии профессиональной математики OntoMath^{PRO}.

2. Постановка задачи

Определение [3, 4]. *Граф знаний определяется в виде кортежа $G=(E, R, T, D)$, где E – множество вершин, представляющих сущности, R – множество отношений, связывающих сущности, T – множество RDF-триплетов $(s, p, o) \in$*

$E \times R \times E$, s – субъект, p – предикат, o – объект, D – множество описаний сущностей и отношений.

Целью настоящей работы является разработка метода формирования графа знаний математических уравнений. В качестве первоначальной основы для извлечения данных, включаемых в граф знаний, используется информация об уравнениях, размещенная на научно-образовательном сайте EqWorld, а также описания концептов и отношений онтологии профессиональной математики OntoMath^{PRO}.

Первоочередными задачами являются определение типов узлов и отношений графа знаний, создание программных инструментов извлечения и импорта знаний из внешних источников, разработка RML-правил отображения структурированных данных в RDF, генерация RDF-триплетов и построение графа знаний на их основе, организация хранилища графа знаний с обеспечением возможности выполнения запросов к сформированному графу.

3. Формирование данных для графа знаний

3.1. Извлечение знаний об уравнениях с сайта EqWorld

Международный научно-образовательный сайт EqWorld «Мир математических уравнений» (<https://eqworld.ipmnet.ru/indexr.htm>) содержит порядка 2000 веб-страниц и предоставляет информацию о примерно 800 уравнениях, распределенных по типам в 38 разделах и 60 подразделах. Каждое уравнение на веб-странице снабжено гиперссылкой на PDF-документ с кратким описанием уравнения.

Для извлечения знаний с веб-страниц сайта разработан алгоритм автоматического поиска и обработки информации об уравнениях, преобразования полученных данных в TeX-представление и формирования объекта уравнения.

Далее приведены основные шаги алгоритма.

Шаг 1. Генерация набора URL-адресов веб-страниц EqWorld.

Шаг 2. Инициализация результирующего списка.

Шаг 3. Цикл по набору URL-адресов веб-страниц.

Шаг 3.2. Получение HTML-страницы по текущему URL-адресу.

Шаг 3.3. Извлечение базового типа уравнений, представленного на странице.

Шаг 3.4. Цикл обработки каждой группы уравнений, представленной на странице с текущим URL-адресом.

Шаг 3.4.1. Инициализация списка типов обрабатываемой группы уравнений.

Шаг 3.4.2. Извлечение подтипа группы уравнений, если таковой имеется.

Добавление базового типа и полученного подтипа в список типов текущей группы уравнений.

Шаг 3.4.3. Цикл обработки каждого уравнения в группе.

Шаг 3.4.3.1. Извлечение названия уравнения.

Шаг 3.4.3.2. Извлечение ссылки на связанный с уравнением PDF-документ.

Шаг 3.4.3.3. Извлечение символьной записи уравнения.

Шаг 3.4.3.4. Формирование L^AT_EX-представления уравнения на основе извлеченной символьной записи.

Шаг 3.4.3.5. Извлечение HTML-представления уравнения.

Шаг 3.4.3.6. Инициализация объекта уравнения.

Шаг 3.4.3.7. Добавление полученного объекта уравнения в результирующий список.

Завершение цикла 3.4.3.

Завершение цикла 3.4.

Завершение цикла 3.

Шаг 4. Возврат результирующего списка, содержащего объекты уравнений.

3.2. Извлечение знаний из онтологии OntoMath^{PRO}

OntoMath^{PRO} относится к классу предметных онтологий и целью её разработки является классификация и систематизация основных понятий профессионального математического знания [9–11].

Для извлечения информации о классах уравнений и их отношениях из онтологии OntoMath^{PRO}, объединения полученных знаний и объектов знаний, извлеченных с сайта EqWorld, с последующей их интеграцией в граф знаний, разработан алгоритм, основные шаги которого приведены далее.

Предварительно произведена векторизация концептов онтологии с помощью TF-IDF для последующего связывания извлеченных уравнений с концептами на основе косинусной меры сходства.

Шаг 1. Цикл обработки каждого уравнения входного списка.

Шаг 1.1. Извлечение названия текущего уравнения. При отсутствии названия происходит переход к шагу 1.4.

Шаг 1.2. Формирование запроса к онтологии на поиск концепта с соответствующим названием.

Шаг 1.3. Если концепт найден, происходит присваивание текущему уравнению в качестве типа URI найденного концепта, а также дополнение уравнения информацией, найденной в онтологии.

Шаг 1.4. Цикл обработки списка подтипов уравнения. Цикл производится с конца списка, что позволяет осуществлять поиск от частного типа уравнения к общему.

Шаг 1.4.1. Получение очередного типа уравнения.

Шаг 1.4.2. Формирование запроса к онтологии на поиск концепта с соответствующим названием типа.

Шаг 1.4.3. Если концепт найден, то выполняются те же действия, что и на шаге 1.3, кроме перехода к шагу 1.4.

Шаг 1.4.3. Если подходящий концепт не был найден, то происходит объединение название уравнения со всем типами, к которым оно принадлежит, формируя документ уравнения.

Шаг 1.4.4. К документу уравнения применяется текстовая предобработка. Приведение к нижнему регистру, удаление стоп-слов, токенизация, стемминг, удаление дублирующихся токенов.

Шаг 1.4.5. Векторизация документа уравнения с помощью TF-IDF.

Шаг 1.4.6. Вычисление косинусной меры сходства между векторным представлением текущего уравнения и векторными представлениями всех концептов онтологии.

Шаг 1.4.7. Текущему уравнению присваивается концепт онтологии с наибольшей мерой косинусного сходства между векторным представлениями.

Завершение цикла 1.4.

Завершение цикла 1.

Шаг 2. Возврат результирующего списка с уравнениями, размеченный в терминах концептов онтологии.

3.3. Генерация RDF-триплетов

После извлечения из источников знаний информации о математических уравнениях производится их согласование, и как результат формируется XML-файл, содержащий структурированный набор метаданных о каждом уравнении. Например, каждое уравнение имеет название с указанием языка (xml-тэги `LabelText` и `LangTag`), иерархия типов уравнения (тело xml-тэга `Types`), строковое, HTML и LaTeX представления (тэги `StringForm`, `HtmlForm`, `LatexForm`), URI родительского и присвоенного уравнению концепта онтологии (`SubClassOf`, `LinkedOmpConcept`) и так далее. Фрагмент файла приведён на Рис. 1.

```

▼<ArrayOfEquation xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  ▼<Equation>
    <URI>0209421c-ce31-45c4-907b-0915dc964247</URI>
    ▼<Labels>
      ▼<Label>
        <LabelText>Линейное уравнение</LabelText>
        <LangTag>ru</LangTag>
      </Label>
      ▼<Label>
        <LabelText>Linear equation</LabelText>
        <LangTag>en</LangTag>
      </Label>
      ▼<Label>
        <LabelText>Линейное уравнение</LabelText>
        <LangTag>ru</LangTag>
      </Label>
    </Labels>
    ▼<Types>
      <string>Алгебраические уравнения</string>
    </Types>
    <SubClassOf>http://ontomathpro.org/omp2#E1904</SubClassOf>
    <LinkedOmpConcept>http://ontomathpro.org/omp2#E1907</LinkedOmpConcept>
    <StringForm>ax+b=0.</StringForm>
    <Reference>https://eqworld.ipmnet.ru/en/solutions/ae/ae0101.pdf</Reference>
    ▼<Comments>
      <string>Линейное уравнение – это уравнение, у которого полная степень составляющих его
      многочленов равна 1.</string>
    </Comments>
    <LatexForm>\[ax+b=0\]</LatexForm>
    <HtmlForm><i>ax</i> + <i>b</i> = 0.</HtmlForm>
  </Equation>
  ▼<Equation>
    <URI>0209421c-ce31-45c4-907b-0915dc964247</URI>

```

Рис 1. Фрагмент XML-файла с извлеченной информацией об уравнениях

В процессе формирования RDF-триплетов используется язык RML (RDF Mapping Language) [12]. Генерация RDF-триплетов осуществляется с помощью вызова RML-процессора, в который передаются: XML-файл с описанием уравнений, созданный на предыдущих этапах (разделы 3.1 и 3.2), а также система RML-правил отображения в виде карт троек, определяющих субъект, предикат и объект. Фрагмент разработанных RML-правил приведен на Рис. 2.

```

<#EquationMapping> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "equations.xml" ;
    rml:iterator "/ArrayOfEquation/Equation";
    rml:referenceFormulation ql:XPath;
  ];

  rr:subjectMap [
    rr:template "http://www.eqgraph.ru/{URI}";
  ];

  rr:predicateObjectMap [
    rr:predicate rdf:type;
    rr:objectMap [
      rml:reference "LinkedOmpConcept";
      rr:termType rr:IRI;
    ]
  ];

```

Рис 2. Фрагмент карты троек, в которой определены логический источник, карта субъекта и карта предиката-объекта

Результатом работы RML-процессора является файл в формате turtle, содержащий сгенерированные триплеты. Фрагмент файла приведен на Рис.3.

```

<http://www.eqgraph.ru/4adf18fa-9cb3-4d32-a9bd-572cc6aefd86> a <http://ontomathpro.org/omp2#E1906>;
eq:htmlForm "<i>ax</i><sup>3</sup> + <i>bx</i><sup>2</sup> + <i>cx</i> + <i>d</i> = 0.";
eq:latexForm "\\[ax^{3}+bx^{2}+cx+d=0\\]";
eq:reference "https://eqworld.ipmnet.ru/en/solutions/ae/ae0103.pdf";
eq:stringForm "ax^(3)+bx^(2)+cx+d=0.";
rdfs:comment "http://ru.wikipedia.org/wiki/%D0%9A%D1%83%D0%B1%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%";
rdfs:label "Cubic equation"@ru, "Кубическое уравнение"@ru;
rdfs:subClassOf <http://ontomathpro.org/omp2#E1904> .

```

Рис. 3. Фрагмент файла, содержащего сгенерированные RDF-триплеты

Приведем пошаговое описание алгоритма генерации RDF-триплетов.

- Шаг 1. Формирование системы RML-правил отображения структурированных данных об уравнениях.
- Шаг 2. Создание конфигурации RML-процессора с указанием пути к XML-файлу с описанием уравнений и пути к файлу, содержащему RML-правила.
- Шаг 3. Вызов RML-процессора, настроенного на конфигурацию, сформированную на предыдущем шаге.
- Шаг 4. Формирование и отправка запроса к хранилищу на запись сгенерированных RDF-триплетов в виде графа знаний.

4. Программная реализация и тестирование

Программное решение реализовано с помощью языка C# и платформы .NET в виде 3 основных модулей:

- **ParserService** – модуль извлечения информации об уравнениях с веб-сайта EqWorld с использованием библиотеки HTMLAgilityPack

(<https://html-agility-pack.net/>). Результатом работы модуля является список объектов класса `Equation`, которые содержат информацию о уравнениях в 10 свойствах класса;

- **LinkingService** – модуль обогащения знаний из онтологии `OntoMathPRO` с использованием функционала библиотеки машинного обучения `Accord.Net` (<https://accord-framework.net/>). В модуле реализован метод, формирующий SPARQL-запросы к онтологии на извлечение информации об уравнениях; **RdfGenerationService** – модуль генерации RDF-триплетов? предназначен для вызова RML-процессора, конфигурация которого настроена на обработку XML-файла уравнений в соответствии с разработанные RML-правилами отображения. Результатом работы модуля является множество сгенерированных RDF-триплетов.

Помимо основных модулей, реализованы вспомогательные сервисы, такие как: **SparqlService** – формирование, отправка и обработка SPARQL-запросов, **VectorizationService** – формирование векторного представления концептов онтологии, **TextPreprocessor** – набор методов обработки текстов, **EquationFormsBuilder** – формирование строкового, HTML и L^AT_EX представлений математических уравнений.

В настоящей версии созданный граф знаний математических уравнений содержит информацию о 449 уравнениях, полученных из 25 разделов портала `EqWorld` и обогащенный знаниями из онтологии `OntoMathPRO`.

Заключение

Представлен граф знаний математических уравнений, спроектированный и реализованный на основе разработанного метода извлечения знаний из научно-образовательного сайта `EqWorld` «Мир математических уравнений» и онтологии профессиональной математики `OntoMathPRO`. Дальнейшее развитие графа знаний предполагает пополнение графа информацией о новых классах уравнений и связей между ними, а также с другими объектами знаний. Планируется расширение математической онтологии, обеспечивающее согласование с информацией об уравнениях, представленной на сайте `EqWorld`. Например, в онтологии `OntoMathPRO` дифференциальное уравнение Риккати представлено одним классом, а на сайте `EqWorld` размещены страницы с описанием 16 подтипов указанного уравнения.

В представленной реализации отождествление уравнений с концептами онтологии производится на основе названия и типа уравнения. При этом структура самого уравнения не анализируется и при связывании не используется. Таким образом, возникает задача разработки алгоритма сопоставления уравнений и концептов онтологии на основе структуры математического выражения.

Ещё одной задачей является создание системы наименований для математических уравнений в графе знаний, например, из 797 уравнений сайта EqWorld для 628 уравнений отсутствуют названия и они представлены как подтипы общего класса уравнений.

Литература

1. *Singhal A.* Introducing the Knowledge Graph: things, not strings // Google Official Blog. 2012. URL: <https://blog.google/products/search/introducing-knowledge-graph-things-not/>, дата обращения 07.07.2025
2. *Paulheim H.* Knowledge graph refinement: A survey of approaches and evaluation methods // Semantic Web. 2017. V. 8. P. 489–508. <https://doi.org/10.3233/SW-160218>
3. *Ji S., Pan S., Cambria E., Marttinen P., Yu P.S.* A Survey on Knowledge Graphs: Representation, Acquisition, and Applications // IEEE Transactions on Neural Networks and Learning Systems. 2022. V. 33. No. 2. P. 494–514. <https://doi.org/10.1109/TNNLS.2021.3070843>
4. *Hogan A. et al.* Knowledge Graphs. Springer Cham, 2022. 237 p. <https://doi.org/10.1007/978-3-031-01918-0>
5. *Peng, C., Xia, F., Naseriparsa, M. et al.* Knowledge Graphs: Opportunities and Challenges // Artif Intell Rev. 2023. V. 56. P. 13071–13102. <https://doi.org/10.1007/s10462-023-10465-9>
6. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* Ontological Approach to a Knowledge Graph Construction in a Semantic Library // Lobachevskii J. Math. 2023. V. 44. No. 6. P. 2229–2239. <https://doi.org/10.1134/S1995080223060471>
7. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* From Texts to Knowledge Graph in the Semantic Library LibMeta // Lobachevskii J. Math. 2024. V. 45. No. 5. P. 2211–2219. <https://doi.org/10.1134/S1995080224602625>
8. *Гизатуллин Б.Т., Невзорова О.А.* К построению графа знаний коллекции математических статей // Научный сервис в сети Интернет: труды XXV Всероссийской научной конференции (18-21 сентября 2023 г., онлайн). — М.: ИПМ им. М.В.Келдыша, 2023. — С. 94-103. <https://doi.org/10.20948/abrau-2023-28>. <https://keldysh.ru/abrau/2023/theses/28.pdf>
9. *Kirillovich A.V., Nevzorova O.A., Lipachev E.K.* OntoMathPRO 2.0 Ontology: Updates of Formal Model // Lobachevskii J. of Math. 2022. V. 43. No. 12. P. 3504–3514. <https://doi.org/10.1134/S1995080222150136>.
10. *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>
11. *Елизаров А.М., Кириллович А.В., Липачёв Е.К., Невзорова О.А.* Новые компоненты онтологии OntoMathPRO представления математического знания // Научный сервис в сети Интернет. 2023. № 25. С. 141–151. <https://doi.org/10.20948/abrau-2023-32>

12. RDF Mapping Language (RML). Unofficial Draft 20 June 2024. URL: <https://rml.io/specs/rml/>

References

1. *Singhal A.* Introducing the Knowledge Graph: things, not strings // Google Official Blog. 2012. URL: <https://blog.google/products/search/introducing-knowledge-graph-things-not/>, last access 07.07.2025
2. *Paulheim H.* Knowledge graph refinement: A survey of approaches and evaluation methods // Semantic Web. 2017. V. 8. P. 489–508. <https://doi.org/10.3233/SW-160218>
3. *Ji S., Pan S., Cambria E., Marttinen P., Yu P.S.* A Survey on Knowledge Graphs: Representation, Acquisition, and Applications // IEEE Transactions on Neural Networks and Learning Systems. 2022. V. 33. No. 2. P. 494–514. <https://doi.org/10.1109/TNNLS.2021.3070843>
4. *Hogan A. et al.* Knowledge Graphs. Springer Cham, 2022. 237 p. <https://doi.org/10.1007/978-3-031-01918-0>
5. *Peng, C., Xia, F., Naseriparsa, M. et al.* Knowledge Graphs: Opportunities and Challenges // Artif Intell Rev. 2023. V. 56. P. 13071–13102. <https://doi.org/10.1007/s10462-023-10465-9>
6. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* Ontological Approach to a Knowledge Graph Construction in a Semantic Library // Lobachevskii J. Math. 2023. V. 44. No. 6. P. 2229–2239. <https://doi.org/10.1134/S1995080223060471>
7. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* From Texts to Knowledge Graph in the Semantic Library LibMeta // Lobachevskii J. Math. 2024. V. 45. No. 5. P. 2211–2219. <https://doi.org/10.1134/S1995080224602625>
8. *Gizatullin B.T., Nevzorova O.A.* Towards Building the Knowledge Graph for a Collection of Mathematical Articles // Nauchnyi servis v seti Internet: trudy XXV Vserossiiskoi nauchnoi konferentsii (18–21 sentiabria 2023 g.). — M.: IPM im. M.V.Keldysha, 2023. — P. 94-103. <https://doi.org/10.20948/abrau-2023-28>.
9. *Kirillovich A.V., Nevzorova O.A., Lipachev E.K.* OntoMathPRO 2.0 Ontology: Updates of Formal Model // Lobachevskii J. of Math. 2022. V. 43. No. 12. P. 3504–3514. <https://doi.org/10.1134/S1995080222150136>.
10. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* OntoMathPRO: Ontology of Mathematical Knowledge // Dokl. Math. 2022. Vol. 106 (3). P. 429-435. <https://doi.org/10.1134/S1064562422700016>
11. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* New components of the OntoMathPRO ontology for representing math knowledge // Nauchnyi servis v seti Internet: trudy XXV Vserossiiskoi nauchnoi konferentsii (18–21 sentiabria 2023 g.). — M.: IPM im. M.V.Keldysha, 2023. — P. 144-151. <https://doi.org/10.20948/abrau-2023-32>.
12. RDF Mapping Language (RML). Unofficial Draft 20 June 2024. URL: <https://rml.io/specs/rml/>