Модель семантического поиска в коллекциях математических документов на основе онтологий♣

© Е.В. Биряльцев, А.М. Елизаров, Н.Г. Жильцов, В.В. Иванов,
О.А. Невзорова, В.Д. Соловьев
НИИММ им. Н.Г. Чеботарева Казанского (Приволжского) федерального университета

Аннотация

Предложена модель семантического поиска в электронных коллекциях математических документов. Рассмотрены вопросы представления математических документов на основе онтологий, классификации и формальной интерпретации поисковых запросов с учетом семантики исходных текстов.

1 Введение

Поиск по математическим документам [18] – актуальная и быстроразвивающаяся область исследований. Современные математические поисковые системы условно можно разделить на две группы [16].

К первой относятся системы поиска научных публикаций [3, 5], а также поисковые интерфейсы крупнейших научных коллекций [14, 19, 20]. Они предлагают сервис полнотекстового поиска по ключевым словам и индексируют значительные объемы актуальных научных статей в области математики, представленных в форматах PDF или LaTeX.

Отличительная особенность поисковых систем второй группы состоит в том, что они используют семантику математической нотации и реализуют поиск по формулам или выражениям [1, 6, 10]. Стоит отметить, что поисковые системы обеих групп недостаточно принимают во внимание важную особенность большинства математических документов – их структурированность. В данной работе описывается модель поиска, которая учитывает структуры математического документа и объектов математического знания.

Известные в России семантические поисковые системы, как правило, не работают с естественнонаучными публикациями или используют иной подход. Например, метапоисковик Nigma [23] предлагает сервис для решения математических уравнений и поиск концептуальных объектов в виде таблиц. Поисковик EXACTUS [25] ориентирован на обработку запросов на естественном языке и не специализируется на научных коллекциях.

Основная цель подхода, представленного в настоящей статье, – использовать вышеуказанные особенности математических текстов для расширения возможностей стандартного полнотекстового поиска. Раздел 2 раскрывает специфику математических текстов в контексте задачи поиска. Разделы 3 и 4 описывают формат семантической разметки и методы ее получения, которые ориентированы на разные аспекты представления исходных математических документов. В разделе 5 приводятся классификация и формальная интерпретация поисковых запросов, учитывающих семантику математических текстов.

2 Структура математического документа

Большинство математических документов, особенно научных публикаций, имеет четкую логическую структуру: выделяются главы, определения, формулировки теорем, доказательства, следствия, заключения и т. д. Зачастую структурные элементы выделяются явно, например, с помощью стилей формата PDF или тэгов языка LaTeX. В последнее десятилетие активно развивались подходы к представлению логической структуры математических документов в целях различных приложений. В частности, выделяется работа [7], в которой авторы представили онтологию DRa (Document Rhetorical aspect ontology), специфицирующую, помимо структурных элементов, еще и отношения логического следования между ними, в частности, отношения *использует* (теорема использует определение), *обосновывает* (доказательство обосновывает теорему), *ссылается* (пример ссылается на теорему) и т. д. Другой известный подход – формат OMDoc и онтология OMDoc [8, 12]. OMDoc имеет три уровня – формул, утверждений и теорий. Логической структуре документа отвечает уровень утверждений, который полно описывает семантические отношения между структурными элементами. Например, формализуются утверждения вида «доказательство *доказывает* теорему», «пример *относится* к определению», «символ *имеет* определение».

В контексте задачи поиска эксплицитная форма представления структурных элементов позволит выполнять семантические поисковые запросы, которые достаточно сложны для рассмотренных выше поисковых систем. Необходимо отметить и то, что содержимое математических документов – описа-

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

ния объектов математического знания – имеет особую внутреннюю структуру. Учет семантических связей между терминами, упоминающимися в математических текстах, также позволяет рассчитывать на повышение точности поиска.

Подход с привлечением терминологических ресурсов достаточно распространен. Например, в работе [22] описана процедура автоматизируемого получения тезауруса на основе коллекций естественно-научных текстов уровня школьных и университетских учебных программ.

Теорема 1. *Пусть G — конечная группа, в которой для любой подгруппы A выполняется неравенство $(*)$. Тогда $G = K\lambda H$, где H — силовская 2-подгруппа группы G, подгруппа K абелева, $|H/C_H(K)| \leq 2$ и либо группа H абелева, либо $H/Z(H)$ — четверная или диэдральная группа.*

Следствие. Если G — неабелева нильпотентная группа, в которой условие $(*)$ выполняется для любой неабелевой подгруппы A, то $G = K \times H$, где H — неабелева силовская p-подгруппа, строение которой определено в теореме 3, а группа K абелева.

Доказательство. Из леммы и теоремы 1 следует, что в группе G только одна силовская подгруппа может быть неабелевой. □

Рис. 1. Фрагмент математического документа

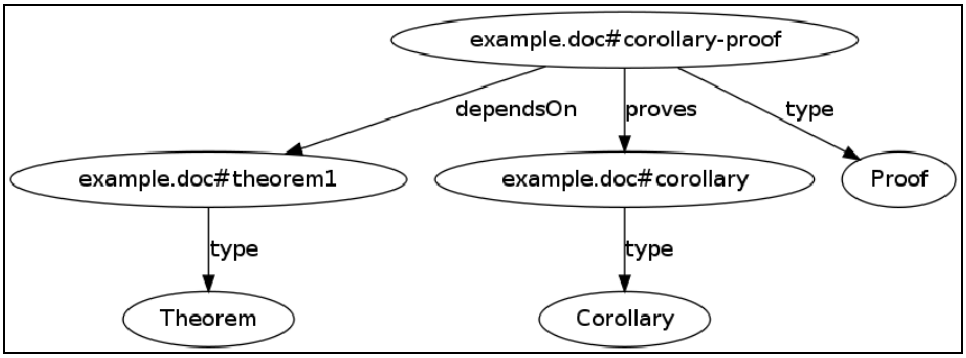

Рис. 2. RDF граф структурных элементов математического документа

3 Семантическая разметка математического документа

Предлагаемая семантическая разметка основывается на известных ресурсах и технологиях Семантического веба и специфицирует форму представления математических документов, учитывающую структуры математического документа и объектов математического знания. В качестве онтологии структуры математического документа выбрана OMDoc. Данная онтология, выраженная на языке OWL-DL, концептуально описывает типовые структурные элементы (теоремы, леммы, доказательства, формулы, определения) и отношения между ними.

Возьмем в качестве примера следующий фрагмент математической научной публикации [21] (рис. 1).

Разметка структурных элементов и их отношений для данного примера на языке RDF/N3 выглядит следующим образом:

<example.doc#theorem1> <rdf:type> <omdoc:Theorem>.
<example.doc#corollary> <rdf:type> <omdoc:Corollary> .
<example.doc#corollary-proof> <rdf:type> <omdoc:Proof> .
<example.doc#corollary-proof> <omdoc:proves> <example.doc#corollary> .
<example.doc#corollary-proof> <omdoc:dependsOn> <example.doc#theorem1> .

Компоненты триплетов с префиксом *example.doc* – сгенерированные URI структурных элементов математического документа example.doc, компоненты с префиксом *omdoc* – концепты и отношения онтологии OMDoc, наконец, *rdf:type* – отношение, определенное в языке RDFS. В виде RDF-графа структура из примера показана на рис. 2.

Для представления объектов математического знания – терминов и научных результатов математических теорий – предлагается использовать подход на основе контролируемых словарей, которые специфицируют математические термины и отношения между ними. Примеры таких ресурсов – DBPedia [4] или математический тезаурус Кембриджского университета [13]. Связывание различных ресурсов производится с помощью дополнительных отношений, выраженных с привлечением онтологии SKOS [2] – онтологии представления контролируемых словарей.

Определим дополнительное отношение *hasMention* («упоминает») таким образом, что его доменом является концепт *omdoc:MathematicalKnowledgeItem* и дипазоном – кон-

цепт *skos:Concept*. С помощью этого отношения, например, можно выразить следующий факт для рассматриваемого фрагмента:

<example.doc#corollary> <hasMention> <dbpedia:Nilpotent_group> .

Таким образом, конкретный структурный элемент «следствие» содержит упоминание термина «нильпотентная группа», определенного в DBPedia. Эта информация дополняет рассмотренный ранее RDF граф структурных элементов исходного документа (рис. 3).

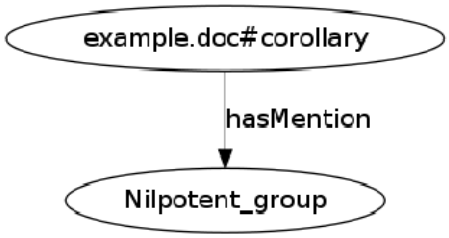

Рис. 3. Терминологическая разметка

4 Семантическое аннотирование математических документов

Получение семантической разметки для исходных математических документов – центральная задача, решаемая при моделировании семантического поиска. Разрабатываемые авторами методы семантического аннотирования принадлежат двум смежным направлениям – обработка размеченных и неразмеченных исходных текстов.

4.1 Обработка размеченных документов

Получение структурной разметки. LaTeX – один из самых популярных форматов представления математических публикаций. Стандартные средства LaTeX и специального пакета команд AMS-LaTeX позволяют размечать такие структурные элементы, как главы, формулировки теорем, доказательства, формулы и т. д. Пакет sTeX [9] расширяет эти возможности и предлагает средства для разметки не только структурных элементов, но и отношений между ними, например, для указания связи между доказательством и теоремой. Данный пакет примечателен тем, что разработаны средства [11] для генерации RDF-представления документа, аннотированного средствами sTeX, с помощью онтологии OMDoc. Таким образом, для математического документа, аннотированного с помощью команд из данного пакета, можно автоматически получить эксплицитное представление его структуры. Для произвольного LaTeX-документа ставится задача получения мэппинга LaTeX-тэгов на онтологию структуры математического документа. Формат мэппинга фиксирует утилита LaTeXML [15]. В качестве подходов для решения этой задачи исследовались методы на основе близости имен тэгов, а также заголовков структурных элементов. Трудности при определении типа элемента состоят в следующем:

- объявление тэгов может быть вынесено в отдельный стилевой файл; в этом случае будут недоступны для обработки полезные конструкции вида *\newtheorem{defns}{\hskip\parindent Определение}[section];*
- авторы математических публикаций часто сокращают имена тэгов, например, встречаются следующие варианты аннотирования теорем: theorem, thm, thms, thmnonum и т. д.

В качестве канонических имен для определения типов структурных элементов рассматривались текстовые наименования концептов онтологии OMDoc. Эксперименты, проведенные с разными алгоритмами близости строк, показали, что для анализа имен тэгов оптимален строковый алгоритм N-gram. Данный алгоритм подсчитывает меру близости строк на основе количества общих подпоследовательностей длины N (обычно N=3) и возвращает число от 0 до 1. Эта мера показывает результаты мэппинга на уровне 85 % точности и 96 % полноты при значении меры, равном 0,26.

Анализ семантических отношений между структурными элементами – более сложная задача, требующая отдельного исследования. Одним из ресурсов для ее решения является популярный механизм меток/ссылок в LaTeX. Лингвистический анализ контекстов таких ссылок позволит определять тип отношений.

Получение терминологической разметки. Терминологическое аннотирование выполнено на основе лингвистических технологий онтолингвистической системы «OntoIntegrator» [24]. Обработка содержимого структурных элементов включает сегментацию текста на предложения, распознавание объектов текста (выделение формул, числовых последовательностей, слов, знаков препинания, аббревиатур и др.), распознавание именных групп, в том числе содержащих термины прикладной онтологии, распознавание сложных синтаксических конструкций (групп сочинительного сокращения) и другие процедуры (например, выделение и классификация омонимов). В экспериментах, проведенных на корпусе статей по теории групп из коллекции журнала «Известия высших учебных заведений. Математика», использовалась прикладная онтология (список терминов по теории групп из Wikipedia [17]), содержащая 79 терминов и терминологических словосочетаний. Терминологическое покрытие строилось в границах именных групп, выделяемых по различным синтаксическим моделям и анализом составляющих именных групп на принадлежность терминологическому списку. Например, в рассматриваемом фрагменте статьи выделены однословные и многословные именные группы на основании соответствующих синтаксических моделей (*конечная группа; силовская 2-подгруппа группы; четверная группа; диэдральная группа; неабелева нильпотентная группа и др.)*. Необходимо отметить, что

математические тексты содержат большое число слов с формульно-префиксными частями (*p-подгруппа, 2-подгруппа*), а также формульно-постфиксными частями (*группа G, подгруппа K)*. В качестве префикса могут быть использованы произвольные формулы и выражения. Такие объекты не содержатся в словаре системы и обрабатываются специальными методами, которые отсекают левый формульный префикс и работают по синтаксической модели правой части слова. Обработка слов с формульно-постфиксными частями производится по синтаксической модели именной группы с аббревиатурой (*группа G, подгруппа K*).

4.2 Обработка неразмеченных документов

Обработка исходно-неразмеченных текстовых документов связана с автоматическим семантическим аннотированием, которое предлагается выполнить на основе метода лексико-синтаксических шаблонов (ЛСШ) для выделения начальных и финальных текстовых позиций структурных элементов. Для решения данной задачи необходимо:

- выделить множества ЛСШ структурных элементов на основе корпусных исследований коллекций математических текстов;
- описать лексический состав и синтаксические модели ЛСШ;
- разработать методы распознавания ЛСШ в математических текстах, учитывающие потенциальную многозначность, которая в ряде случаев непосредственно коррелирует с распознаванием типа ЛСШ (начальный или финальный классы).

5 Классификация поисковых запросов

Проблема поиска по структуре математических документов формулируется в терминах фактологического поиска в некоторой базе знаний. В рамках предлагаемого подхода база знаний представляет собой RDF-хранилище, которое содержит факты, извлеченные из исходных документов, в смысле представления, описанного в разделе 3. В данном разделе рассматриваются классификация и формальная интерпретация поисковых запросов, обрабатываемых в рамках предлагаемой модели поиска. Язык SPARQL выбран для описания как де-факто стандарт языка запросов к RDF-графам. Классификация конкретизирует каждый тип поисковых запросов по характеру объектов, использованных в запросе, а также с точки зрения применения определенных терминологических словарей.

Первый тип запросов – поиск структурных элементов математического документа с использованием отношений между ними. Это базовый тип запросов, при формулировании которых используется только онтология структуры математического документа в качестве терминологии. Пример поискового запроса данного типа – «найти доказательства теорем». На языке SPARQL данный запрос формулируется следующим образом:

```
SELECT ?p WHERE {?p a omdoc:Proof.
?p omdoc:proves omdoc:Theorem}
```

Второй тип – поиск структурных элементов по символьным обозначениям математических терминов. Пример – «найти определения, в которых встречается символ диэдральной группы D». Для выражения запросов данного типа используются словари языка разметки OpenMath, предназначенные для уточнения семантики математических формул:

```
SELECT ?d WHERE {?d a omdoc:Definition.
?d omdoc:hasProperty ?p.
?p omdoc:usesSymbol openmath:dihedral_group}
```

Третий тип – поиск структурных элементов и объектов математического знания. Пример – «найти теоремы, упоминающие термины из теории групп». В этом случае осуществляется поиск структурных элементов – определений, имеющих отношение к конкретной области математического знания:

```
SELECT ?t WHERE {?t a omdoc:Theorem.
?t hasMention ?s.
?s skos:subject dbpedia:Group_theory}
```

Формулирование запросов этого типа зависит от имеющегося в наличии терминологического ресурса. Кроме рассмотренного терминологического списка из Wikipedia также в качестве терминологических источников могут использоваться такие объекты, как иерархия с отношением КЛАСС – ПОДКЛАСС (пример – классификатор УДК); тезаурус с отношениями гипонимии, синонимии (подходящий объект для его получения – приложения учебников соответствующей прикладной области) и т. д. Логический вывод с использованием указанных отношений позволит выполнять запросы, весьма не тривиальные для обычных систем полнотекстового поиска.

6 Заключение

В данной работе описывается модель семантического поиска в коллекциях математических документов. Рассматриваются формат семантической разметки исходных документов, а также методы автоматического аннотирования документов и формального представления поисковых запросов. Поставлена задача извлечения отношений между структурными элементами из исходных документов на языке LaTeX. Требуют дальнейшего исследования вопросы, связанные с выполнением поисковых запросов, обработкой и представлением поисковых результатов.

Литература

[1] Altamimi M., Youssef A. An extensive Math Query Language // SEDE. – 2007. – P. 57-63.

[2] Bechhofer S., Miles A. SKOS core vocabulary specification // W3C Recommendation. – 2009. – http://www.w3.org/TR/2009/REC-skos-reference-20090818.

[3] CiteSeerX. – http://citeseerx.ist.psu.edu.

[4] DBPedia. – http://dbpedia.org.

[5] Google Scholar. – http://scholar.google.com.

[6] Hashimoto H., Hijikata Y., Nishida S. Search mathematical formulas by mathematical formulas // Human Interface and the Management of Information. Designing Information Environments. – LNCS. – 2009. – V. 5617. – P. 404-411.

[7] Kamareddine F.et al. Narrative structure of mathematical texts // Calculemus MKM / LNCS. – 2007. – V. 4573. – P. 296-312.

[8] Kohlhase M. OMDoc – an open markup format for mathematical documents. – Springer, 2006.

[9] Kohlhase M. sTeX: Semantic Markup in TeX/LaTeX. – 2005. – https://svn.kwarc.info/ repos/stex/trunk/sty/stex.pdf.

[10] Kohlhase M, Sucan I. A search engine for mathematical formulae // LNCS. – 2006. – V. 4120. – P. 241-253.

[11] Lange C. An extensible XML -> RDF extraction framework // CEUR Workshop Proceedings. – 2009. – V. 449.

[12] Lange C. SWiM – A Semantic Wiki for Mathematical Knowledge Management // ESWC. – LNCS. – LNCS. – 2008. – V. 5021. – P. 832-837.

[13] Mathematical thesaurus. – http://thesaurus.maths.org.

[14] Math-Net.Ru. – http://www.mathnet.ru.

[15] Miller B. LaTeXML: A LaTeX to XML converter, 2007. – http://dlmf.nist.gov/LaTeXML.

[16] Misutka J. Indexing mathematical content using full text search engine // WDS'08 Proc. of Contributed Papers. – 2008. – P. 240-244.

[17] Wikipedia. Словарь терминов по теории групп. – http://ru.wikipedia.org/wiki/Словарь_терминов_теории_групп.

[18] Youssef A. Roles of Math Search in mathematics //Mathematical Knowledge Management, 5th Int. Conf. – LNCS. – 2006. – P. 2-16.

[19] Zentralblatt MATH. – http://www.zentralblatt-math.org/zmath.

[20] arXiv. – http://arxiv.org.

[21] Аминева Н.Н., Антонов В.А. О группах с относительно большими централизаторами // Изв. высших учебных заведений. Математика. – 2003. – № 7. – С. 8-17.

[22] Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Физико-математические науки. – 2007. – Т. 149.– С. 49-72.

[23] Интеллектуальная поисковая система Нигма.РФ. – http://nigma.ru.

[24] Невзорова О.А. Онтолингвистические системы: технологии взаимодействия с прикладной онтологией // Ученые записки Казанского государственного университета. Серия физико-математические науки. – 2007. – Т. 149. – С. 105-115.

[25] Осипов Г.С., Тихомиров И.А., Смирнов И.В. Семантический поиск в сети Интернет средствами поисковой машины Exactus // Труды 11-ой национальной конф. по искусственному интеллекту КИИ-2008. – 2008. – С. 323-328.

Ontology-based semantic search model for the collections of mathematical documents

E.V. Birialtsev, A.M. Elizarov, N.G. Zhiltsov, V.V. Ivanov, O.A. Nevzorova, V.D. Solovyev

The paper proposes a semantic search model for the collections of mathematical documents. We consider the ontology-based representation of a mathematical document, classification and formalization of related search queries.

♣ Работа выполнена при финансовой поддержке РФФИ (проект 09-07-12059 офи-м)