Семантические сервисы для коллекций математических документов, представленных как Linked Data♣

© Н.Г. Жильцов

НИИММ им. Н.Г. Чеботарева Казанского (Приволжского) федерального университета
nikita.zhiltsov@gmail.com

Аннотация

Статья содержит обзор технологий Семантического Веба для представления научных математических документов. Обсуждаются вопросы формализации логической структуры математического документа и структуры объектов математического знания как ключевых характеристик исходных текстов. Также рассматриваются идеи семантических сервисов, расширяющих возможности электронных коллекций в области математики.

1 Введение

Современные специализированные электронные коллекции [13, 31, 32, 34] содержат большое число документов, представляющих интерес для отдельных групп математиков – исследователей, инженеров, преподавателей, студентов и т. д. Как правило, пользователям этих систем предоставляется доступ к исходным текстам публикаций, например, в форматах PDF или LaTeX, а также предлагается сервис полнотекстового поиска по ключевым словам с учетом метаданных – поиск по названию, автору, краткому описанию, году публикации. Расширенную функциональность реализуют системы поиска научных публикаций, которые кроме базового полнотекстового поиска предоставляют дополнительные возможности. Например, Google Scholar [5] позволяет находить как статьи, которые ссылаются на данную, так и статьи, сходные по тематике с данной. CiteSeerX [2] использует принцип общих цитат для поиска похожих статей. Scirus [24] позволяет фильтровать поисковые результаты, используя динамически генерируемые фасеты. Тем не менее, в рамках традиционного подхода к представлению и обработке математических документов игнорируется специфика исходных текстов: (i) наличие элементов математической нотации; (ii) структурированность математического документа; (iii) категоризация текстов по разделам математики. Очевидно, что использование этих особенностей требует развития альтернативных моделей математического документа. Наряду со стандартным полнотекстовым индексом такие модели должны специфировать дополнительные характеристики. Например, обработка элементов математической нотации находится в фокусе многочисленных систем поиска по формулам [11, 18, 30]. Как правило, такие системы используют особое формализованное представление формул, выраженное на языках OpenMath [21] и MathML [15]. На базе этих форматов решается более сложная задача – интерпретация семантики формул на языке LaTeX, которой, в частности, посвящены такие проекты, как Uniquation [33] и ArXMLiv [28].

Актуальность онлайновых сервисов для электронных коллекций и архивов научных публикаций широко обсуждается в отечественной литературе [35, 36]. В данном обзорном докладе делается акцент на математических научных публикациях и технологиях Семантического Веба, которые могут применяться при их интеллектуальной обработке.

2 Представление структуры математического документа

Специфика математических текстов позволяет выделить два типа структуры – логическую структуру математического документа и структуру объектов математического знания.

Логическая структура документа. Многие научные математические тексты имеют четкую логическую структуру. Даже языки, ориентированные на представление, имеют средства (в частности, пакет AMS-LaTeX) для разметки таких элементов, как теоремы, леммы, доказательства, определения, следствия и т. д. В последнее десятилетие разработаны различные методы для формализации логической структуры. Так, например, подход, описанный в [8, 20], выделяет элементы риторической структуры математического текста – главы, теоремы, доказательства – и отношения логического следования между ними. Авторы утверждают, что такое представление позволяет проводить частичную валидацию фактов, описанных в документе. Разработаны методы, направленные на улучшение навигации при чтении математического текста [19]. В проекте HELM [14] впервые была предпринята попытка представить структуру документа и объектов математического знания на языке RDF. Например, в онтологии HELM выделялись отношения между такими концептами, как Теория, Теорема, Доказательство, Заключение и т. д. Формат OMDoc [10], основанный на XML, позволяет выражать структурные элементы (утверждения, теоремы, леммы), объекты математического знания (теории и их морфизмы) и семантику математических формул с помощью языков OpenMath и MathML. Онтология OMDoc, реали-

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

зованная на языке OWL-DL, концептуально описывает формат OMDoc и выражает структурные элементы и отношения между ними. Например, формулируются такие утверждения, как «доказательство доказывает теорему», «теорема принадлежит теории», «пример относится к теореме» (см. рис. 1).

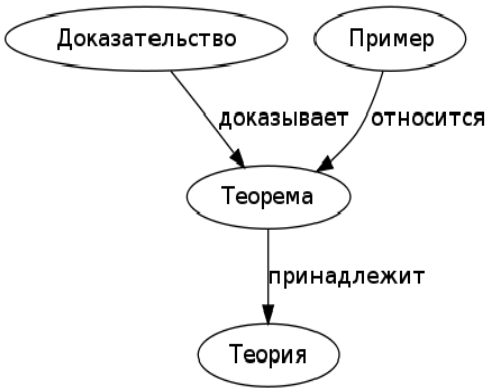

Рис. 1. Фрагмент онтологии OMDoc

Спецификация семантики математических формул также имеет важное значение. Во-первых, одни и те же элементы математической нотации могут обозначаться по-разному, например, биномиальный коэффициент записывается как C_n^k, C_k^n или $\binom{n}{k}$. Во-вторых, разные объекты могут быть отображены одинаково, как в случае обозначения счетчика при объявлении суммы и мнимой единицы в формуле Эйлера.

Структура объектов математического знания. Труды Н. Бурбаки были, по существу, первой попыткой построить онтологию математического знания из аксиоматики канторовской теории множеств. Они заложили общий фундамент для концептуального представления объектов математического знания и их отношений на уровне отдельных теорий. В работе [6] представлена формальная онтология математического моделирования для инженеров, покрывающая такие разделы, как абстрактная алгебра и метрология. Математический тезаурус Кембриджского университета [16] содержит список основных математических терминов с отношением гипоним-гипероним и отношением логической связи между терминами. Например, тезаурус содержит такой факт, что термин «моноид» является нижестоящим по отношению к термину «полугруппа» и определяется через понятие «нейтральный элемент». Широко известный тезаурус WordNet [4] содержит не только математические термины, но и указывает на отношение синонимии между некоторыми из них. В частности, WordNet приводит синоним понятия «абелева группа» – «коммутативная группа». Набор данных DBPedia [3] и онтология Yago [29] – примеры взаимосвязанных ресурсов, содержащих термины с отношением гипоним-гипероним и отношением принадлежности к некоторой категории. Например, DBPedia содержит тот факт, что Великая теорема Ферма относится к категориям «теория чисел» и «теория Галуа».

3 Представление математических документов как Linked Data

Термин Linked Data [1] – «связанные данные» – обозначает одну из центральных идей Семантического Веба, в основе которой – то, что первичными объектами веба являются описания сущностей с явным указанием их семантики и семантики ссылок (отношений) между ними. Технологически это обеспечивается представлением данных в виде триплетов «субъект – предикат – объект» на языке RDF, идентификацией данных с помощью URI, механизмом доступа по протоколу HTTP, спецификацией контролируемых словарей на языках RDFS и OWL. Также относительно недавно появился микроформат RDF [23], поддержанный поисковыми системами Yahoo и Google как расширение XHTML для аннотирования веб-страниц метаданными. Более высокая структурированность первичных объектов по сравнению с традиционным вебом документов позволяет предлагать более качественные сервисы, которые можно разделить на три группы: (i) браузеры (ii) семантические поисковые системы и (iii) мэшапы. Браузер Sparks O_3 [27] является примером приложения первого типа. Он отображает дополнительную информацию о факте, содержащимся в веб-документе. Например, в связи с упоминанием конференции браузер отображает информацию о месте проведения или о составе участников. Поисковая система Sindice [26] индексирует документы, представленные как Linked Data, и позволяет выполнять сложные семантические запросы. Например, Sindice позволяет находить документы, в которых встретилось упоминание о коллегах или знакомых пользователя. Sig.ma [25] – пример приложения-мэшапа. Мэшапы объединяют данные из нескольких источников в один интегрированный инструмент. Сервис Sig.ma, который можно рассматривать как проводник по Linked Data, агрегирует информацию по любому объекту Linked Data – конкретной личности, событию, предмету и т. д.

Технология, описанная в [22], предлагает оригинальный подход для публикации математических документов как Linked Data. Процесс преобразования документа выглядит следующим образом (рис. 2). Исходный математический документ на языке LaTeX аннотируется вручную с помощью пакета sTeX [9]. Остальные преобразования производятся в автоматическом режиме. С помощью утилиты LaTeXML [17] sTeX-документы конвертируются в формат OMDoc. На следующем этапе извлекаются данные в виде RDF с привлечением онтологии OMDoc и генерируются уникальные URI для структурных элементов [12]. Далее формируются доступные в вебе документы в форматах XHTML/MathML/RDF. Также авторами подхода разработан пример сервиса для интерактивного просмотра опубликованных математических документов [7]. Он позволяет просматривать определения терминов и выдавать объяснения элементов нотации.

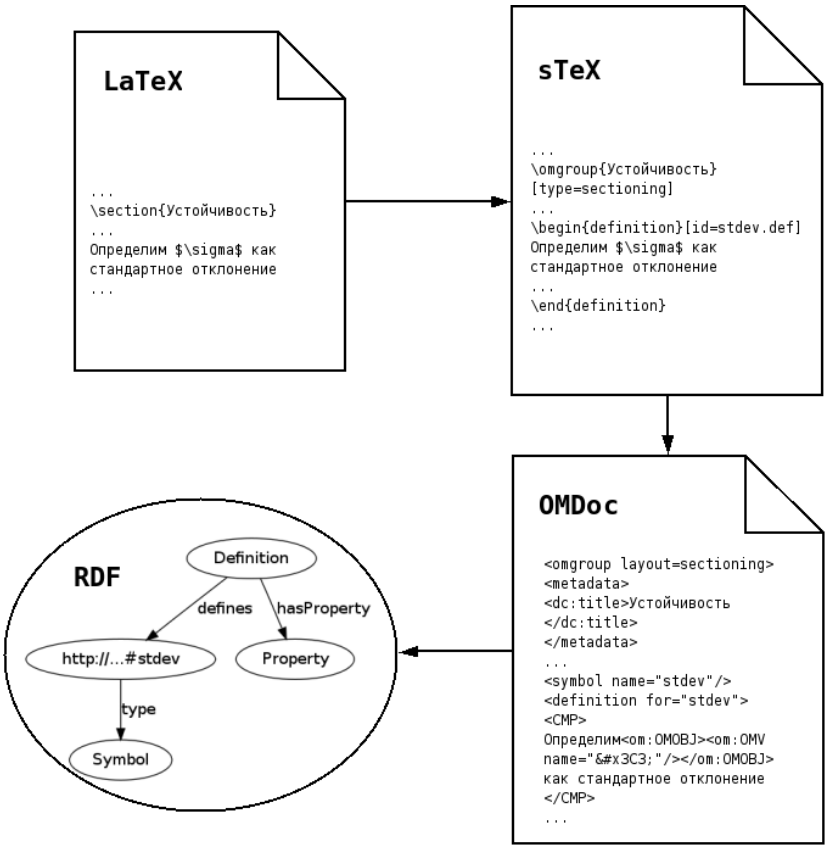

Рис. 2. Процесс преобразования математического документа

4 Семантические сервисы для коллекций математических документов

Проект Linked Data динамично развивается с 2007 года. По состоянию на ноябрь 2009 года размер «облака» Linked Data составляет 13.1 млрд. RDF-триплетов и 142 млн. RDF-ссылок. Важными источниками данных являются такие ресурсы, как DBPedia, Geonames, BBC, U.S. Government, Thomson Reuters и др. Тем не менее, до сих пор область математики не представлена в значительном объеме в данном проекте. В этой связи актуальна задача публикации математических документов в форматах Linked Data в автоматическом или полуавтоматическом режимах. Технология, кратко описанная в предыдущем разделе, является важным шагом в этом направлении. Далее обсуждаются возможные варианты использования математических текстов, представленных как Linked Data.

Для описания сервисов следует выделить различные группы людей, занимающихся математикой. Например, студентам могут быть интересны следующие сервисы:

- *расширенная навигация при просмотре текста;* такой сервис предлагает элементы навигации для перехода к определению встретившегося термина или тексту упоминающейся лекции;
- *объяснение элементов нотации и терминов*; сервис может выдавать дополнительную информацию для выделенного фрагмента текста или формулы, а также предлагать поиск связанных материалов в онлайн-ресурсах. Такие ресурсы, как Wikipedia и сайты онлайн-курсов известных университетов, представляют собой альтернативные источники знаний.

Лекторам полезны сервисы, упомянутые в [22]:

- *подбор материалов к очередной лекции;* подбор может осуществляться с учетом специфики аудитории, например, содержание курса математической логики варьируется для студентов математических, гуманитарных или физических факультетов в смысле уровня математической подготовки и примеров, специфичных для предметной области;
- *поиск проблем и задач для самостоятельных работ*; в этом случае удобно иметь возможность отмечать пропуски в доказательствах или ссылки на материалы для самостоятельного изучения.

Для профессиональных исследователей актуальны:

- *сбор и категоризация новых публикаций*; сбор и сортировка информации о публикациях могут проводиться по таким параметрам, как коды классификаторов, уровень доверия источника, прикладной/теоретический характер полученных результатов;
- *семантический поиск по документам с учетом структуры*; сервис может предлагать расширенные возможности поиска по ключевым словам, такие, как поиск теорем, доказательств, следст-

вий и других структурных элементов математических документов.

5 Заключение

В статье обсуждаются вопросы расширения функциональных возможностей современных информационных систем в области математики. Рассматриваются основные технологии Семантического Веба для представления математических документов, которые позволяют реализовывать сервисы с расширенной функциональностью. Предложены идеи возможных сервисов, которые могут представлять интерес для разных групп математиков.

Литература

[1] Berners-Lee T. Linked data – design issues. – 2006. – http://www.w3.org/DesignIssues/ LinkedData.html.

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

[3] DBPedia. – http://dbpedia.org.

[4] Fellbaum C. WordNet: An electronical lexical database. – The MIT Press. – 1998.

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

[6] Gruber T., Olsen G. An ontology for engineering mathematics // KR. – 1994. – P. 258-269.

[7] Kohlhase M., Giceva J., Lange C., Zholudev V. JOBAD – interactive mathematical documents. – AI Mashup Challenge. – 2009.

[8] Kamareddine F., Maarek M., Wells J.B. Toward an object-oriented structure for mathematical text // MKM. – LNCS. – 2005. – V. 3863.– P. 217-233.

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

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

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

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

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

[14] Asperti A. et al. Mathematical knowledge management in HELM // Ann. Math. Artif. Intell. – 2003. – V. 38, No 1-3. – P. 26-46.

[15] Mathematical Markup Language (MathML) Version 2.0 / Ausbrooks et al // W3C Recommendation. – http://www.w3.org/TR/MathML.

[16] Mathematical Thesaurus. – http://thesaurus.maths.org.

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

[18] Miner R., Munavalli R. An approach to mathematical search through query formulation and data normalization // Calculemus '07 / MKM '07: Proc. of the 14th Symposium on Towards Mechanized Mathematical Assistants. – 2007. – P. 342-355.

[19] Nakagawa K., Nomura A., Suzuki M. Extraction of logical structure from articles in mathematics // MKM. – LNCS. – 2004. – V. 3119. – P. 276-289.

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

[21] Buswell S. et al The OpenMath standard. – 2003. – http://www.openmath.org/standard.

[22] David C. et al. Publishing math lecture notes as Linked Data // ESWC. – LNCS. – 2010. – V. 6089. – P. 370-376.

[23] RDFa in XHTML: Syntax and Processing. – W3C Recommendation. – 2008. – http://www.w3.org/ TR/rdfa-syntax.

[24] Scirus. – http://www.scirus.com.

[25] Sig.ma – Semantic Information MAshup. – http://sig.ma.

[26] Sindice – the Semantic Web Index. – http:// sindice.com.

[27] Sparks O_3 Browser. – http://oak.dcs.shef.ac.uk/ sparks.

[28] Stamerjohanns H., Kohlhase M. Transforming the arXiv to XML // Proc. of the 9th AISC Int. Conf., the 15th Calculemas Symposium, and the 7th Int. MKM Conf. on Intelligent Computer Mathematics. – 2008. – P. 574-582.

[29] Suchanek F.M., Kasneci G., Weikum G. Yago: a core of semantic knowledge // WWW '07: Proc. of the 16th Int. Conf. on World Wide Web. – ACM. – 2007. – P. 697-706.

[30] The Wolfram functions site. –http://functions. wolfram.com.

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

[32] arXiv. – http://arxiv.org.

[33] (uni)quation. – http://uniquation.ru.

[34] Каталог ВИНИТИ. – http://catalog.viniti.ru.

[35] Паринов С.И., Когаловский М.Р. Технология поддержки электронных научных публикаций как «живых» документов // Труды RCDL'2009. – С. 53-58.

[36] Сальникова Е.Е., Сальников С., Кузнецов С.Д. Управление контентом в крупных научно-технических Internet-библиотеках // Труды RCDL'2009. – С. 193-199.

Semantic services for the collections of mathematical documents published as Linked Data

Nikita Zhiltsov

The paper gives short review of Semantic Web technologies for mathematical document representation. Key aspects of mathematical document formalization, such as logical structure specification and formalization of mathematical knowledge objects, are discussed. As a conlusion, some ideas of the semantic services, which could exploit the mentioned document models, are considered.

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