ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

Р.Р. Батыршина, А.М. Елизаров,
Е.К. Липачев

Организация коллекций цифровой математической библиотеки методами семантического анализа

Рекомендуемая форма библиографической ссылки

Батыршина Р.Р., Елизаров А.М., Липачев Е.К. Организация коллекций цифровой математической библиотеки методами семантического анализа // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 85-90. — URL: http://keldysh.ru/abrau/2019/theses/97.pdf doi:10.20948/abrau-2019-97

Размещена также презентация к докладу

Организация коллекций цифровой математической библиотеки методами семантического анализа

Р.Р. Батыршина[1], А.М. Елизаров[1,2], Е.К. Липачев[1,2]

[1]Институт математики и механики им. Н.И. Лобачевского
[2]Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Аннотация. Предложены методы формирования цифровых коллекций из набора документов – научных статей, монографий, докладов, представленных в различных форматах хранения. На основе анализа структуры документов и стилевых особенностей их оформления разработан алгоритм экстракции их метаданных. Представлен программный инструмент разделения сборников статей на отдельные документы и формирования их семантического представления. На примере набора сборников «Трудов Математического центра им. Н.И. Лобачевского», имеющих различные формат и структуру, описан алгоритм создания цифровой коллекции и ее включения в цифровую математическую библиотеку Lobachevskii-DML.

Ключевые слова: цифровая коллекция, цифровая математическая библиотека, метаданные, семантическая связь, семантический метод, Lobachevskii-DML

Organization of Digital Mathematics Library Collections by Semantic Analysis Methods

R.R. Batyrshina,, A.M. Elizarov[1,2], E.K. Lipachev[1,2]

[1] N. I. Lobachevskii Institute of Mathematics and Mechanics,
[2]Higher School of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University

Abstract. We offer methods for the formation of digital collections from a set of documents (scientific articles, monographs, collections of reports), which are presented in various storage formats. Based on the analysis of the structure of documents and the stylistic features of their design, we have developed an algorithm for extracting the metadata of these documents. We present a software tool for dividing collections of articles into separate documents and the formation of their semantic presentation. On the example of a collection “Proceedings of the Mathematical Center. N.I. Lobachevskii”, which have a different format and

structure, we describe the algorithm for creating a digital collection and its inclusion in the Lobachevskii-DML.

Keywords: digital collection, Digital Mathematics Library, metadata, semantic relation, semantic method, Lobachevskii-DML.

1. Введение

Как известно, управление информацией в Сети основано на использовании метаданных [1]–[3]. Создание цифровой коллекции из набора документов предполагает не только расширение набора метаданных, но и их нормализацию в соответствии с установленными схемами данных, например, Journal Archiving and Interchange Tag Suite (NISO JATS, https://jats.nlm.nih.gov/archiving/). При организации цифровых библиотек предъявляются дополнительные требования к составу и формату метаданных [4]. Так, при формировании метаданных цифровых математических библиотек учитывается специфика математических документов [5]. Нашей задачей было создание методов, позволяющих программным способом извлекать необходимые метаданные из цифровых математических документов и устанавливать семантические связи между объектами.

Настоящая работа посвящена методам создания цифровых научных коллекций из массива разнородных оцифрованных документов в соответствии с подходами, разработанными в рамках проекта создания Всемирной цифровой математической библиотеки (см. [6, 7]). На примере обработки набора файлов, содержащих тома “Трудов Математического центра им. Н.И. Лобачевского” за 1998–2018 гг., описан процесс формирования соответствующей цифровой коллекции и включения её в цифровую математическую библиотеку Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/), формируемую в настоящее время в Казанском (Приволжском) федеральном университете.

2. Алгоритм экстракции метаданных и семантических отношений

В состав Lobachevskii-DML входят различные коллекции журнальных статей, а также труды семинаров и конференций. Ниже представлены методы включения в эту цифровую библиотеку коллекции “Трудов Математического центра им. Н.И. Лобачевского”, издаваемую с 1998 года.

Основным назначением названного издания является публикация материалов математических конференций. Как следствие, большинство томов “Трудов” содержит по несколько десятков статей с ограниченным (с современной точки зрения) составом метаданных. С 1998 года (момента выпуска первого тома) использовалось несколько стилевых правил подготовки материалов, что отразилось на оформлении статей и форматах файлов сформированных сборников. Необходимыми условиями создания цифровой коллекции из файлового массива “Трудов” были разделение томов на отдельные статьи, выделение метаданных, описывающих каждую статью, генерация дополнительных метадан-

ных, содержащих, в частности, библиографическое описание статьи, ссылку на файл статьи в цифровой коллекции, а также связи с профилями авторов статьи на академических порталах и наукометрических базах (kpfu.ru, MathNet.ru, Scopus и др.).

Приведем основные этапы программной обработки набора файлов сборников статей, выделения метаданных, создания цифровой коллекции и включения её в библиотеку Lobachevskii-DML.

Прежде всего, была произведена кластеризация, в результате которой соответствующие тома были разделены на классы по сходству их структуры и оформления. Для каждого класса разработан набор паттернов регулярных выражений, задающих правила поиска информационных блоков. Основой этого алгоритма является подход, предложенный в [8]. Алгоритм реализован в виде программ на C#, позволяющих обрабатывать файлы в форматах TeX, OpenXML (.docx) и .pdf. TeX-файлы формировались с помощью стандартных функций, реализующих операции с текстовыми строками. Для работы с pdf-файлами использовались функции библиотек PDFLib (https://www.pdflib.com) и iTextSharp (https://www.nuget.org/packages/iTextSharp/). Для документов, представленных в виде docx-файлов, производился разбор файла “word/document.xml”, выделенного из .docx-архива в соответствии с форматом Office OpenXML (см., например, [9]).

Далее производилась обработка массива файлов “Трудов” с целью выделения метаданных, описывающих как том в целом, так и статьи, входящие в него. В частности, определялись номера страниц всех статей каждого тома. Для этого разработан алгоритм, использующий структурную однородность каждого тома и стилевую однозначность в оформлении статей в нем для поиска страниц с названиями статей. Это позволило выделить не только названия статей, но и списки авторов, блоки библиографии и другие метаданные (например, e-mail, ключевые слова), в случае их наличия в тексте.

Предложен XML-язык описания цифровых математических коллекций, состоящий из набора тегов и XML-схем, основанных на Journal Archiving and Interchange Tag Suite (https://jats.nlm.nih.gov/1.2d2/). В нотации этого языка на основании данных, полученных на этапе обработки массива файлов, проведено описание коллекции “Трудов”.

С помощью методов текстового анализа [10, 11] из документов цифровой коллекции выделены термины, из которых образованы наборы ключевых слов для включения в состав метаданных. Алгоритм извлечения терминов является развитием подхода, предложенного в [12].

Следующий этап создания цифровой коллекции состоял из процедур разделения каждого тома “Трудов” на отдельные статьи. Для этого из XML-файлов, в которых приведены метаописания томов, считываются теги, атрибуты которых указывают на начальные и конечные страницы статей. После этого производится разделение файлов на отдельные документы, к которым присваивается имена в соответствии с правилами цифровой коллекции. Процесс выде-

ления статей производился с помощью программы, разработанной на языке Python с использованием функций библиотеки PyPDF2 (http://pybrary.net/pyPdf/).

Ряд метаданных, таких, как адреса электронной почты авторов, их аффилиация, были импортированы и уточнены из профилей авторов на академических сайтах и базах. В этой процедуре были применены семантические связи, установленные в процессе формирования цифровой коллекции. Алгоритм назначения связей в цифровой коллекции основан на методе работы [13].

Система метаданных, подготовленных в процессе работы приведенного алгоритма, позволила сформировать цифровую коллекцию “Трудов Математического центра им. Н.И. Лобачевского” и включить ее в состав цифровой библиотеки Lobachevskii-DML (https://lobachevskii-dml.ru/) [14], [15].

Заключение

Для включения в международное научное пространство цифровых математических коллекций Казанского университета предложены методы их формирования из набора документов, представленных в различных форматах хранения. На основе анализа структуры документов и стилевых особенностей их оформления разработан алгоритм экстракции их метаданных, реализованный на примере «Трудов Математического центра им. Н.И. Лобачевского».

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 1.2368.2017/ПЧ, и при частичной финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160012.

Литература

1. Gartner R. Metadata. Shaping Knowledge from Antiquity to the Semantic Web. — Springer, 2016. — 114 p.
2. Sicilia M.-A. (Ed.) Handbook of Metadata, Semantics and Ontologies. — World Scientific Publishing Co. Pte. Ltd., 2014. — 570 p.
3. Lubas R., Jackson A., Schneider I. The Metadata Manual. — Chandos Publishing, 2013. — 216 p.
4. Alemu G., Stevens B. An Emergent Theory of Digital Library Metadata. — Elsevier Ltd., 2015. — 122 p.
5. Elizarov A.M., Lipachev E.K., Zuev D.S. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. — 2017. — Vol. 2022. — P. 317–325.
6. Developing a 21st Century Global Library for Mathematics Research. Washington: The National Academies Press. — 2014. — 131 p.

7. Ion P.D.F., Watt S.M. The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. Springer. — 2017. — Vol. 10383. — P. 56–69. — https://doi.org/10.1007/978-3-319-62075-6_5.
8. Elizarov A. M., Khaydarov Sh. M., Lipachev E. K. Scientific documents ontologies for semantic representation of digital libraries // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC). Vladivostok, 2017. — P. 1–5. — DOI: 10.1109/RPC.2017.8168064.
9. Standard ECMA-376 Office Open XML File Formats. URL: http://www.ecmainternational.org/publications/standards/Ecma-376.htm.
10. Ингерсолл Г.С., Мортон Т. С., Фэррис Э. Л. Обработка неструктурированных текстов. Поиск, организация и манипулирование. — М.: ДМК Пресс, 2015. — 414 с.
11. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. EMC // Education Services (Ed), Wiley, 2015. — 432 p.
12. Батыршина Р.Р. Метод извлечения терминов в цифровых математических коллекциях // Тр. Матем. центра им. Н. И. Лобачевского. — Казань: Изд-во Казан. матем. об-ва, 2017. — Т. 55. — С. 24–26.
13. Сабитова Э. М. Алгоритм извлечения связей в научных цифровых коллекциях // Тр. Матем. центра им. Н. И. Лобачевского. — Казань: Изд-во Казан. матем. об-ва, 2017. — Т. 55. — С. 123–126.
14. Елизаров А.М., Липачёв Е.К. Семантические методы и инструменты электронной математической библиотеки Lobachevskii-DML // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции. Москва: ИПМ им. М. В. Келдыша, 2017. — С. 130–136. — https://doi.org/10.20948/abrau-2017-73.
15. Elizarov A. M., Lipachev E. K. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. — 2017. — Vol. 2022. — P. 326–333.

Reference

1. Gartner R. Metadata. Shaping Knowledge from Antiquity to the Semantic Web. — Springer, 2016. — 114 p.
2. Sicilia M.-A. (Ed.) Handbook of Metadata, Semantics and Ontologies. — World Scientific Publishing Co. Pte. Ltd., 2014. — 570 p.
3. Lubas R., Jackson A., Schneider I. The Metadata Manual. — Chandos Publishing, 2013. — 216 p.
4. Alemu G., Stevens B. An Emergent Theory of Digital Library Metadata. — Elsevier Ltd., 2015. — 122 p.

5. Elizarov A.M., Lipachev E.K., Zuev D.S. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. — 2017. — Vol. 2022. — P. 317–325.
6. Developing a 21st Century Global Library for Mathematics Research. Washington: The National Academies Press. — 2014. — 131 p.
7. Ion P.D.F., Watt S.M. The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. Springer. — 2017. — Vol. 10383. — P. 56–69. — https://doi.org/10.1007/978-3-319-62075-6_5.
8. Elizarov A. M., Khaydarov Sh. M., Lipachev E. K. Scientific documents ontologies for semantic representation of digital libraries // 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC). Vladivostok, 2017. – P. 1–5. — DOI: 10.1109/RPC.2017.8168064.
9. Standard ECMA-376 Office Open XML File Formats. URL: http://www.ecmainternational.org/publications/standards/Ecma-376.htm.
10. Ingersoll G. S., Morton T. S., Farris A. L. Taming Text. How to Find, Organize, and Manipulate It. — Manning Publications Co., 2013. — 320 p.
11. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. EMC // Education Services (Ed), Wiley, 2015. — 432 p.
12. Batyrshina R.R. The Method of Extracting Terms in the Mathematical Digital Collections // Tr. Matem. centra im. N. I. Lobachevskogo. — Kazan: Izd-vo Kazan. Matem. Ob-va, 2017. — T. 55. — S. 24–26.
13. Sabitova E. M. Algorithm of Relation Extraction in Scientific Digital Collections // Tr. Matem. centra im. N. I. Lobachevskogo. — Kazan: Izd-vo Kazan. Matem. Ob-va, 2017. — T. 55. — S. 123–126.
14. Elizarov A.M., Lipachev E.K. Semanticheskie metody i instrumenty jelektronnoj matematicheskoj biblioteki Lobachevskii-DML // Nauchnyj servis v seti Internet: trudy XIX Vserossijskoj nauchnoj konferencii. Moskva: IPM im. M. V. Keldysha, 2017. — S. 130-136. — https://doi.org/10.20948/abrau-2017-73.
15. Elizarov A. M., Lipachev E. K. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. — 2017. — Vol. 2022. — P. 326–333.