Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. John K M. DotNetNuke 5.4 Cookbook. Gardners Books, 2010. 432 p.

2. Ахметов Д.Ю., Грачев А.О., Герасимов А.Н., Елизаров А.М., Липачёв Е.К. Облачная платформа поддержки электронных научных изданий // Учёные записки Института социальных и гуманитарных знаний. 2014. № 1 (12), ч. 1. С. 13–19.

3. Stranack K. Getting found, staying found, increasing impact. Enhancing readership and preserving content for OJS journals // Public Knowledge Project. 2006. 40 p.

4. Суэринг С., Конверс Т., Парк Д. PHP и MySQL. Библия программиста. М.: Изд-во «Диалектика», 2010. 912 с.

5. Buyya R., Broberg J., Goscinski A. Cloud computing: principles and paradigms. John Wiley & Sons Inc., 2011. 674 p.

6. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачёв Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов// Докл. РАН. 2016. Т. 467, №4. С. 392–395.

УДК 004.91

СЕМАНТИЧЕСКИЙ АНАЛИЗ БОЛЬШИХ КОЛЛЕКЦИЙ НАУЧНЫХ ДОКУМЕНТОВ

А.М. Елизаров[1], Е.К. Липачёв[2], Ш.М. Хайдаров[3]
Казанский (Приволжский) федеральный университет
1 – amelizarov@gmail.com, 2 – elipachev@gmail.com,
3 – 15jkeee@gmail.com

Предложен метод автоматической обработки больших коллекций физико-математических документов, хранящихся в формате OpenXML, включающий валидацию документов и их преобразование в соответствии с правилами формирования коллекций, семантический анализ документов, извлечение метаданных и др. Описан алгоритм метода, приведен пример успешной его реализации при организации XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20 – 24 августа 2015 г.).

Ключевые слова: *Big Data, семантический анализ документов, структурный анализ текстов, метаданные, сервисы автоматической обработки больших коллекций*

Как известно (см., например, [1, 2]), большинство современных электронных коллекций научных документов (научные журналы,

сборники научных трудов, диссертации, научные отчеты, архивы и др.) представляет собой наборы неструктурированных документов, на базе которых трудно организовать семантический поиск, извлечение метаинформации и различные информационные сервисы. Кроме того, в настоящее время наблюдается значительное увеличение объема данных, включаемых в коллекции, что в свою очередь создает дополнительные трудности при обработке информации. Поэтому в условиях непрерывного роста объёмов, а также многообразия информации сейчас активно развиваются новые подходы, инструменты и методы обработки огромных объёмов данных, обозначаемых термином «большие данные» (Big Data). При управлении электронными научными коллекциями больших данных в полной мере остаются актуальными, а также появляются новые задачи, в их числе: семантическая разметка, организация поиска, выделение метаданных, формирование тематических кластеров документов, сбор наукометрической информации, подготовка сборников материалов и др. Насущными становятся проблемы анализа и управления данными в различных областях с интенсивным использованием данных (см., например, материалы конференции [3]).

К большим массивам научных документов сегодня можно отнести и материалы, поступающие на конференции. Их ручная обработка чаще всего не эффективна или даже невозможна. Именно такая ситуация возникла при подготовке проведения XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20–24 августа 2015 г.). В частности, при подготовке к печати материалов Съезда потребовалось решить задачу автоматизированной подготовки метаданных этих публикаций (в соответствии с правилами баз научного цитирования) общим объемом более 1500 статей в формате .docx. Естественно, что традиционными методами оперативно выполнить эту работу было невозможно. Основной задачей, решенной при формировании коллекции материалов Съезда, было приведение поступивших материалов к единому стилевому оформлению:

- единообразное представление названий статей и списка авторов докладов, структура аффилиации авторов, формат аннотации;
- приведение списков литературы к выбранному формату библиографического описания;
- единообразное шрифтовое оформление разделов текста статей;
- выбор форматов рисунков, схем, диаграмм;
- набор математических формул и системы ссылок на них;
- оформление ссылок на поддержку исследований грантами, благодарности.

Основным технологическим инструментом решения названных задач был структурный анализ рассматриваемых документов, проведенный с использованием техники регулярных выражений, а также различных эвристических методов: информация, извлекаемая из документа, содержит название статьи, список авторов с выделением для каждого аффилиации и адреса электронной почты, аннотацию и благодарности, список ключевых слов, основные разделы статьи, библиографический список. Результаты структурного анализа позволили сформировать семантическое представление формируемой коллекции. Опишем подробнее конкретные шаги проведенного структурного анализа.

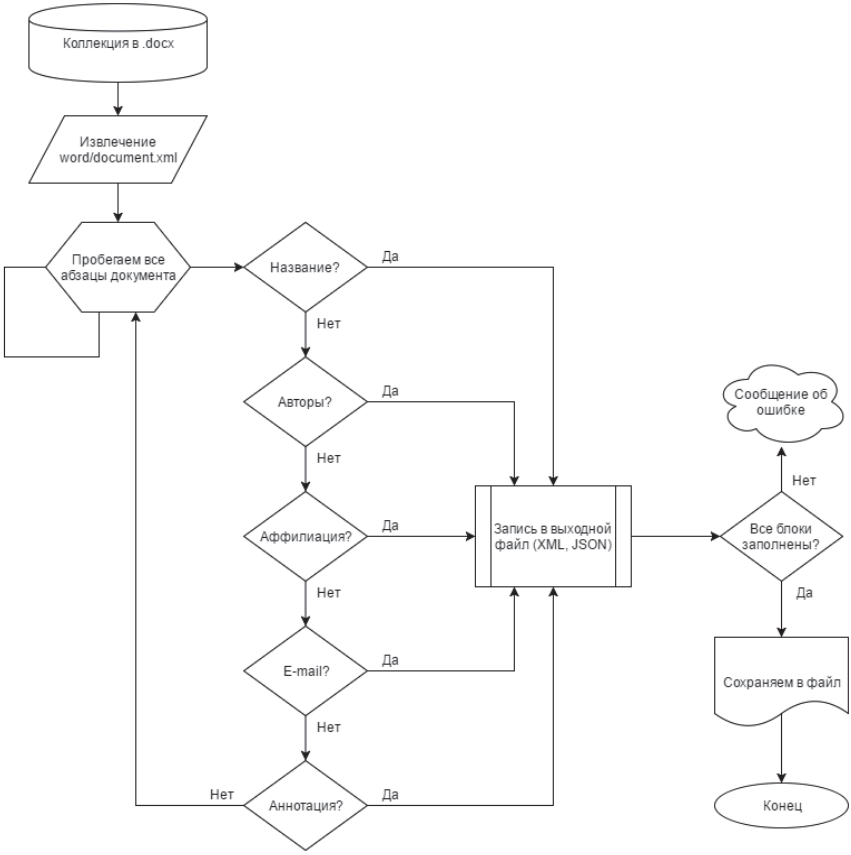

Рис. Алгоритм структурного анализа коллекции

Все материалы, поступившие на Съезд, были преобразованы в формат .docx, который основан на языке разметки XML и допускает семантическую обработку документов (см. [4]). Затем методом, предложенным в [5, 6], из каждого документа извлекался соответствующий ему файл «word/document.xml», который содержит информацию о шрифтовом оформлении и расположении основных блоков документа (название работы, перечень авторов, их аффилиация и др.). Для выделения метаданных использовалась техника регулярных выражений. Например, для выделения списка авторов статей использовались регулярное выражение

```
$fio="/([А-ЯA-Z]\.(?:[А-ЯA-Z]\.)?\s[А-ЯA-Z][a-za-я]+)(,\s)?(?1)?(,\s)?(?1)?/u";
```

и соответствующий скрипт

```
while(($w_ps->item($k+1)->nodeValue)!=""){
if(preg_match($fio,$w_ps->item($k+1)->nodeValue))break;
$articlename.=$w_ps->item($k+1)->nodeValue."%%";
$k++; }
```

Подчеркнем, что существенным условием применения описанного метода является единообразное стилевое оформление документов, что позволяет проводить структурный анализ документов коллекции в автоматическом режиме. Общий алгоритм метода представлен на рисунке выше.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

1. Афонин С.А., Бахтин А.В., Бухонов В.Ю., Васенин В.А., Ганкин Г.М., Гаспарянц А.Э., Голомазов Д.Д., Иткес А.А., Козицын А.С., Тумайкин И.Н., Шапченко К.А. Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА). Под ред. акад. В.А. Садовничего. М.: Изд-во Московского университета, 2014. 262 с.

2. Елизаров А.М., Липачев Е.К., Хохлов Ю.Е. Семантические методы структурирования математического контента, обеспечивающие расширенную поисковую функциональность // Информационное общество. 2013. № 1–2. С. 83–92.

3. Аналитика и управление данными в областях с интенсивным использованием данных: XVII Международная конференция DAMDID/RCDL’2015 (Обнинск, 13–16 октября

2015 года, Россия): Труды конференции/ под ред. Л.А. Калиниченко, С.О. Старкова. – Обнинск: ИАЕЭ НИЯУ МИФИ, 2015. 525 с.

4. Standard ECMA-376: Office Open XML File Formats. URL: http://www.ecmainternational.org/publications/ standards/Ecma-376.htm

5. Хайдаров Ш.М. Методы управления математическим контентом в информационных издательских системах // Тр. Матем. центра им. Н.И. Лобачевского. Материалы 14-й Всерос. Молодежной школы-конференции «Лобачевские чтения–2015 (Казань, 22–27 октября 2015 года). Казань. 2015. С. 162–165.

6. Хайдаров Ш.М. Семантический анализ документов в системе управления цифровыми научными коллекциями // Электронные библиотеки. 2015. Т. 18. № 1–2. С. 61–85.

УДК 004.91

ЭКОСИСТЕМА ONTOMATH И ПРОЕКТ ВСЕМИРНОЙ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ

А.М. Елизаров[1], Н.Г. Жильцов[2], А.В. Кириллович[3], Е.К. Липачёв[4], О.А. Невзорова[5],

Казанский (Приволжский) федеральный университет

1 – amelizarov@gmail.com,

2 – nikita.zhiltsov@gmail.com, 3 – alik.kirillovich@gmail.com,

4 – elipachev@gmail.com, 5 – onevzoro@gmail.com

Описаны возможности использования при проведении новых исследований всего корпуса накопленных научных знаний. Такое использование предполагает повсеместное внедрение информационно-коммуникационных технологий (ИКТ), обеспечивающих оптимальное управление имеющимися знаниями, организацию эффективного доступа к ним, а также совместное и многократное использование новых видов структур знаний. Наибольший эффект от внедрения современных ИКТ для дальнейшей организации научных знаний и повышения их понятности можно ожидать в области математики. Эти ожидания в полной мере подтверждены проектом создания Всемирной цифровой математической библиотеки (World Digital Mathematical Library – WDML). Представлены основные направления реализации проекта WDML и результаты по созданию экосистемы OntoMath как его составной части.

Ключевые слова: *WDML, Всемирная цифровая математическая библиотека, экосистема OntoMath, онтологии, семантический поиск*

В настоящее время благодаря повсеместному внедрению информационно-коммуникационных технологий (ИКТ) в научно-исследовательскую деятельность стало возможным при проведении новых