УДК 004.912
DOI: 10.15827/0236-235X.116.058-062
Дата подачи статьи: 08.08.16
2016. Т. 29. № 4. С. 58–62

ИЗВЛЕЧЕНИЕ МЕТАДАННЫХ ИЗ ПОЛНОТЕКСТОВЫХ ЭЛЕКТРОННЫХ РУССКОЯЗЫЧНЫХ ИЗДАНИЙ ПРИ ПОМОЩИ ТОМИТА-ПАРСЕРА

Р.С. Сулейманов, преподаватель, mail@ruslan.cc
(Московский педагогический государственный университет,
ул. Малая Пироговская, 1/1, г. Москва, 119991, Россия)

При публикации материалов в электронных библиотеках возникает необходимость извлечения метаданных после перевода печатного текста в электронный, что при обработке текста вручную является трудозатратным процессом.

В данной работе рассматривается возможность извлечения метаданных с помощью Томита-парсера, предназначенного для извлечения фактов из текста на естественном языке. Для обеспечения наиболее точного извлечения были разработаны грамматики для анализа полнотекстовых изданий на русском языке, сформирован список метаданных, являющихся обязательными при публикации издания. Разработанные грамматики были апробированы на 100 изданиях, после чего на основании анализа сформулирован ряд закономерностей. С учетом выведенных закономерностей алгоритм был оптимизирован, что позволило повысить эффективность автоматического извлечения данных. Определена необходимость программной обработки полученных данных, например, удаления повторяющейся информации и приведения данных к общему виду перед их публикацией.

С помощью оптимизированного алгоритма проведен масштабный эксперимент по автоматизированному извлечению метаданных из 10 000 изданий, выполнено сравнение его результатов с множеством метаданных, полученных вручную. Предложенный метод автоматического извлечения данных позволил корректно извлечь 86,7 % метаданных, и еще 4 % могут быть использованы после корректировки. Наибольшие проблемы (21 % данных извлечен неверно) возникли с наименованиями материалов вследствие отсутствия четкой структуры. Для четко структурированной информации, такой как ISBN и коды рубрикаторов, процент извлечения приближается к 100 %. Однако было установлено, что, несмотря на увеличение скорости и простоту нахождения метаданных, полностью исключить человека из процесса невозможно.

Ключевые слова: *метаданные, извлечение метаданных из электронных библиотек, извлечение метаданных из полнотекстовых изданий.*

Постоянное увеличение объемов информации с одновременным ростом требований к их доступности является одной из глобальных задач в эпоху формирования цифровой инфраструктуры информационного общества, в котором информация становится одним из базовых активов, необходимых для развития страны, общества и личности. Развитие информационных технологий привело к созданию и совершенствованию новых форм генерации информации, однако сохранение и передача накопленных знаний по-прежнему являются важнейшей задачей, которая решается и путем формирования электронных библиотек. Эти библиотеки позволяют обеспечить доступ к полнотекстовым изданиям российских и зарубежных авторов с любого устройства с доступом в Интернет, в то время как доступ к материалам в обычных библиотеках ограничен вследствие их местонахождения и правил ознакомления с материалами. По данным Министерства культуры РФ, охват населения России библиотечным обслуживанием составляет 34,4 %, при этом количество посещений снижается, в то время как аудитория российского Интернета составляет 82 млн человек, то есть 66 % населения России.

Оцифровка печатных материалов – трудоемкий процесс, обусловленный большим объемом накопленных фондов, их состоянием и возникающими попутными задачами. Если задачу наращивания производительности процесса получения цифровых копий изданий можно решать за счет установки более мощных сканирующих устройств и/или увеличения их количества, то задачу последующей обработки издания для его размещения в электронной библиотеке так просто не решить [1]. К числу наиболее трудоемких задач, сопровождающих формирование цифровых копий, является извлечение метаданных, использующихся в качестве атрибутивной информации при публикации материалов в электронных библиотеках и позволяющих осуществлять навигацию по ней [2].

При наличии материала в существующей электронной библиотеке метаданные можно получить из нее благодаря интерфейсу программирования приложений (API – application programming interface), наличию XML или JSON-сокетам, а также простым синтаксическим анализом HTML-кода страниц электронной библиотеки.

В случае, если материал не содержится в других электронных библиотеках и оцифровывается впервые, извлечение метаданных возможно двумя способами: вручную или при помощи анализа полных текстов материала. Очевидно, что извлечение метаданных вручную трудозатратно и неэффективно.

В данной работе исследуется процедура анализа полных текстов печатных материалов при переводе их в цифровую форму с целью определения

возможности сокращения времени автоматизированного извлечения метаданных из полнотекстовых материалов.

Построение алгоритма извлечения метаданных

Перед извлечением метаданных из материала необходимо преобразовать его из печатной в электронную форму (например, путем сканирования с распознаванием символа или перепечатки), в результате чего будут получены исходные данные для анализа текста [3]. Для анализа предложений в данном исследовании был использован Томита-парсер, созданный российской компанией «Яндекс» в 2014 году на основе GLP-парсера (Generalized Left-to-right Rightmost derivation parser) – обобщенного восходящего магазинного анализатора, расширяющего алгоритм LR-парсера и предназначенного для разбора по недетерминированным и неоднозначным грамматикам. Томита-парсер анализирует текст на естественном языке с учетом синтаксиса языка и морфологии обрабатываемого текста [4].

Томита-парсер является программным обеспечением, открытым для свободного использования, однако при этом для работы с ним необходимо изначально сформировать исходные файлы.

Для решения задачи извлечения метаданных на естественном языке при помощи Томита-парсера требуется создать КС-грамматики, газзетиры и файлы, описывающие факты [5].

КС-грамматика – это набор правил, описывающих синтаксическую структуру извлекаемых цепочек. Газзетир представляет собой словарь с ключевыми словами, которые используются при анализе КС-грамматиками. В файлах, описывающих факты, строится связь между грамматикой и настраивается способ интерпретации грамматики в факт.

Для оценки эффективности анализа полных текстов печатных материалов для извлечения метаданных была сформирована выборка из 100 случайных книжных материалов БД электронной библиотеки «Научное наследие России» [6]. Проверялась возможность извлечения следующих метаданных, являющихся обязательной информацией об издании и требующихся при публикации в электронной библиотеке для обеспечения каталогизации и доступности материалов:

- название материала;
- сведения об авторах;
- код ISBN (уникальный номер книжного издания);
- год публикации;
- место публикации;
- сведения об издателе;
- коды рубрикаторов (УДК, ББК, ГРНТИ).

Для анализа текстов были сформированы и использованы грамматики, приведенные в таблице 1.

Анализ результатов эксперимента

В 87 % обработанных материалов были обнаружены повторяющиеся паттерны, анализ которых позволил выявить закономерности, в дальнейшем учитываемые для оптимизации алгоритма извлечения метаданных.

1. Все рассматриваемые метаданные содержатся на первых или на последних трех страницах оцифрованного материала. Таким образом, для извлечения необходимых метаданных из материала достаточен анализ указанных страниц, анализ других страниц для решения поставленных задач не подходит, что снижает количество анализируемой информации для обнаружения метаданных.

2. Название материала встречается в аннотации в двух возможных сочетаниях:

- «Название» / «Автор»;
- («Издание» или «Публикация») «Название».

3. Авторы могут указываться как перед, так и после названия. Возможно различное написание Ф.И.О. автора: как с использованием инициалов, так и с полным именем.

4. Код ISBN обычно обозначается путем проставления слова ISBN перед цифровой последовательностью. В данном исследовании извлекались коды, соответствующие ISO 2108.

5. Год и место публикации указываются рядом. В качестве места публикации может выступать географический объект или наименование организации, например РАН, институт и пр.

6. Сведения об издательстве предваряются словом «ИЗДАТЕЛЬСТВО» или прилагательным «ИЗДАТЕЛЬСКИЙ» с существительными, например «ДОМ» или «ФИРМА».

7. Коды рубрикаторов предваряются наименованием рубрикатора, например «УДК», «КОДЫ».

Пример автоматического извлечения метаданных (фактов) приведен на рисунке.

Как видно из примера, полученные факты имеют ряд недостатков, в частности, данные повторяются в различных видах, как в случае с Ф.И.О. автора. Название книги также содержит информацию об авторе, в графе «место издательства» материала указаны и географические данные, и наименование организации. При обработке более широкого спектра материалов возможны другие сочетания дублирующейся и/или неточной информации.

Таким образом, возможно извлечение метаданных из полного текста материала, при этом на основании контрольной выборки из 100 книг были сформированы дополнительные уточняющие правила, позволяющие провести поиск метаданных с большей точностью и меньшими затратами вре-

Таблица 1

Грамматики для извлечения метаданных

Table 1

Grammars to extract metadata

Метаданные	Используемая грамматика
ISBN	S -> ('ISBN') (':') ('-') AnyWord<wfl="[0-9]{1,10}(-)?[0-9]{1,10}(-)?[0-9]{1,10}(-)?[0-9]{1,10}(-)?[0-9]{1}">; Isbn -> S interp (Material.Isbn);
Информация об издателе	PublisherDescr -> (Adj) 'издательство' | 'издательский' Noun; ForFact ->Word<h-reg1, gnc-agr[1], rt> (Word<gnc-agr[1]>*); CityOnly ->Word<gram="reo">; ForCity ->CityOnlyinterp (Material.PlaceOfPublish); S -> (ForCity) PublisherDescrForFactinterp (Material.Publisher::not_norm); S -> (ForCity) PublisherDescrForFact<quoted>interp (Material.Publisher::not_norm);
Коды рубрикаторов	UDKStart -> 'удк' (':') ('-'); UDKDeskr ->AnyWord<wff=/[0-9]{1,5}(\.|-)?([0-9]{1,5})?((\.|-)?)([0-9]{1,5})?(\.|-)?([0-9]{1,5})?/>interp (Material.RubricsUDK) (','); UDK ->UDKStartUDKDeskr+; BBKStart -> 'ббк' (':') ('-'); BBKDeskr ->AnyWord<wff=/[0-9]{1,5}(\.|-)?([0-9]{1,5})?((\.|-)?)([0-9]{1,5})?(\.|-)?([0-9]{1,5})?/>interp (Material.RubricsBBK); BBK ->BBKStartBBKDeskr+; GrntiStart -> 'грнти' (':') ('-'); GrntiDeskr ->AnyWord<wff=/[0-9]{1,5}(\.|-)?([0-9]{1,5})?((\.|-)?)([0-9]{1,5})?(\.|-)?([0-9]{1,5})?/>interp (Material.RubricsGrnti); Grnti ->GrntiStartGrntiDeskr+; S -> BBK | UDK | Grnti;
Дата и место публикации	CityOrOrg ->Word<gram="reo"> | "ран" interp (Material.PlaceOfPublish); S ->CityOrOrg (',') AnyWord<wfl="18[0-9]{2}|19[0-9]{2}|20[0-1][0-9]">interp (Material.YearOfPublish);
Автор и наименование	Initial ->Word<wff=/[А-Я]\./>; Initials ->Initial<h-reg1>Initial<h-reg1>; FullName ->InitialsWord<gram="фам"> | Word<gram="фам">Initials | Word<gram="фам"> (',') Word<gram="имя">Word<gram="отч">; Person ->FullNameinterp (Material.Person::not_norm); Year -> (',') AnyWord<wfl="18[0-9]{2}|19[0-9]{2}|20[0-1][0-9]">interp (Material.YearOfPublish) ('.') EOSent ; FromStart ->AnyWord<fw, h-reg1>AnyWord*; MaterialName ->FromStartinterp (Material.Name::not_norm) ('/') Person; NotFromStart ->AnyWord<h-reg1>AnyWord*; MaterialName -> 'научный' 'издание' NotFromStartinterp (Material.Name::not_norm);

мени и ресурсов. Однако полученные при этом данные не подходят в полной мере для незамедлительного использования и нуждаются в дополнительной корректировке.

После обработки полного текста с помощью Томита-парсера и извлечения данных возможно сохранение извлеченной информации в форматах Google Protobuf, обычного текста или XML, после чего может быть осуществлена корректировка. Дополнительная программная обработка требуется для удаления повторяющихся данных, приведения их к общему виду, распределения информации по верным рубрикам и иной корректировки извлеченных данных из выбранного формата и формирования итогового набора метаданных. После обработки полученных фактов возможно их использование для присвоения атрибутивной информации соответствующим изданиям.

Оценка эффективности разработанной методики извлечения метаданных

Для подтверждения корректности извлечения метаданных из полных текстов электронных вер-

Material							
Name	Person	Isbn	YearOfPublish	PlaseOfPublish	Publisher	RubricsUDK	RubricsBBK
						7.0	
							85
			2009	РАН			
				Москва	Восточная литература		
Николаева Н.С. Образы Японии: очерки и заметки	Н.С. Николаева						
		978-5-02-036405-9					
		978-5-02-036405-9					
	Николаева Н.С.						
		978-5-02-036405-9					
			2009	РАН			

Пример извлечения фактов из полного текста книги

The example of fact extraction from a full book text

сий печатных материалов был проведен ряд экспериментов. В качестве тестовой площадки для их проведения использован набор данных, содержащий 10 000 русскоязычных книг из БД электронной библиотеки «Научное наследие России» [7]. Для всех материалов, подлежащих тестированию, были доступны метаданные, с которыми было проведено сравнение извлеченных данных методом сравнения полей.

В результате экспериментов получены результаты, представленные в таблице 2.

Таблица 2

Корректность извлечения метаданных из тестовой выборки материалов

Table 2

Metadata extraction accuracy in a test data sample

Поле	**Извлечено верно (%)**	**Извлечено неверно (%)**	**Требуется уточнение (%)** *
Наименование материала	76	21	3
Сведения об авторах	91	7	2
Код ISBN	98	0	2
Год публикации	89	10	1
Место публикации	84	12	4
Сведения об издателе	79	14	7
Коды рубрикаторов	90	1	9
Результаты в среднем	**86,7**	**9,3**	**4**

* Выявлены ошибки при оптическом распознавании текста (OCR), данные извлечены не полностью либо извлечена лишняя информация.

Средний показатель корректно извлеченых метаданных составил 86,7 %, еще 4 % извлеченных фактов поддаются последующей корректировке и могут быть использованы после ее проведения. При этом наибольшие проблемы наблюдаются с извлечением наименований материалов, которые не имеют четко утвержденной структуры, могут содержать любое количество символов и знаков препинания.

Извлечение сведений об издательствах также проблематично, поскольку может не подчиняться вышеприведенным правилам и не быть обозначенным при печати, что затрудняет поиск данных по тексту, однако вместе с данными, поддающимися корректировке, процент доступного извлечения превышает 80.

Наиболее полной обработке поддаются код ISBN и коды рубрикаторов из-за четкой структуры, а также благодаря тому, что в большинстве случаев коды предваряются соответствующим названием, что значительно облегчает их поиск. При этом они поддаются корректировке, так что процент их извлечения приближается к 100.

Извлечение ISBN может помочь, если материал и данные о нем были ранее размещены кем-либо в сети Интернет. В таком случае метаданные можно уточнить путем поиска по кодам, например через Google Books ISBN API [8].

Таким образом, в данной работе предложены правила обработки книжных материалов для эффективности поиска метаданных. При проведении эксперимента было показано, что автоматизированное извлечение метаданных из полных текстов русскоязычных книг позволяет существенно сократить затраты на ввод метаданных в любую электронную библиотеку, однако полностью исключить участие человека в этом процессе невозможно. Для обеспечения максимально полной и достоверной информации о материале необходима редакторская проверка корректности полученных программным путем метаданных.

Процент успешного извлечения метаданных из полных текстов можно увеличить за счет улучшения качества оптического распознавания печатных материалов, а также улучшения КС-грамматик и газзетиров.

Литература

1. Кириллов С.А. Эволюция систем оцифровки печатных изданий на примере их использования в проекте ЭБ «Научное наследие России» // Информационное обеспечение науки. Новые технологии: сб. науч. тр. 2011. С. 227–237. URL: http://www.benran.ru/SEMINAR/SEM/Sb_11/sbornik/doc_227.pdf (дата обращения: 07.08.2016).

2. Антопольский А.Б. Системы метаданных в электронных библиотеках // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества: сб. тр. VIII Междунар. конф. 2001. URL: http://gpntb.ru/win/inter-events/crimea2001/tom/sec4/Doc5.html (дата обращения: 07.08.2016).

3. Васильев А., Козлов Д., Самусев С., Шамина О. Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей // RCDL2007: сб. тр. Всерос. науч. конф. Переславль: Изд-во Ун-та г. Переславля, 2007. Т. 1. С. 175–181.

4. Economopoulos G.R. Generalised LR parsing algorithms. PhD thesis, Univ. of London Royal Holloway, August 2006, 253 p.

5. Инструмент для извлечения структурированных данных из текста Томита-парсер. URL: https://tech.yandex.ru/tomita/ (дата обращения: 06.08.2016).

6. Электронная библиотека «Научное наследие России». URL: http://e-heritage.ru/index.html (дата обращения: 06.08.2016).

7. Каленов Н.Е., Савин Г.И., Серебряков В.А., Сотников А.Н. Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты и системы. № 4. 2012. С. 30–40.

8. Using the API. URL: https://developers.google.com/books/docs/v1/using (дата обращения: 06.08.2016).

Software & Systems
DOI: 10.15827/0236-235X.116.058-062

Received 08.08.16
2016, vol. 29, no. 4, pp. 58–62

EXTRACTION OF METADATA FROM THE FULL-TEXT ELECTRONIC MATERIALS WRITTEN IN RUSSIAN USING TOMITA-PARSER

R.S. Suleymanov[1], *Lecturer, mail@ruslan.cc*

[1] *Moscow State University of Education, M. Pirogovskaya St., 1/1, Moscow, 119991, Russian Federation*

Abstract. Publishing information in digital libraries requires metadata extraction after transforming initial material into e-text. This procedure is time-consuming in case of performing it manually. This paper considers metadata extraction using Tomita-parser method, which is software designed to extract facts from a natural language text. To ensure the most accurate extraction there were formulated spatial grammars for analyzing full-text books in Russian and a list of metadata for publication was made. Designed spatial grammars were tested on 100 editions, the analysis served as a base for observing a number of consistent patterns. The algorithm has been optimized with regard of derived patterns. This allowed improving the efficiency of automatic data extraction. The authors determined a need for manual data processing, such as removing repetitive information and data reduction to general view before publishing. The optimized algorithm helped to conduct a large-scale experiment of metadata automated extraction from 10,000 publications. Its results were compared to manually extracted data. The proposed method allows extracting correctly up to 86,7 % of meta-data with further 4% which can be used after adjustment. The biggest problem (21 % of data were extracted incorrectly) has been discovered in the names of the materials due to the lack of a clear structure. As for clearly structured information (such as ISBN and rubricator codes) the percentage of correct extraction approaches 100 %. However, despite the speed increase and easiness of metadata extracting, it was proved that it is impossible to completely eliminate a human from the process.

Keywords: metadata, metadata extraction from electronic libraries, metadata extraction from full-text materials.

References

1. Kirillov S.A. Print media digitizing system evolution on the example of their using in the EB project “Scientific heritage of Russia”. *Informatsionnoe obespechenie nauki. Novye tekhnologii: sb. nauch. tr.* [Proc. “Russian Information Support. New technologies”]. 2011, pp. 227–237 (in Russ.).

2. Antopolsky A.B. Metadata systems in electronic libraries. *Sb. statey konf. “Biblioteki i assotsiatsii v menyayushchemsya mire: novye tekhnologii i novye formy sotrudnichestva”* [Proc. Conf. “Libraries and Associations in a Changing World: New Technologies ad New Forms of Cooperation”]. Crimea, 2001. Available at: http://gpntb.ru/win/inter-events/crimea2001/tom/sec4/Doc5.html (accessed August 7, 2016).

3. Vasilev A., Kozlov D., Samusev S., Shamina O. Extracting metadata and references from Russian scientific papers. *Trudy konf. RCDL2007* [Proc. Conf. RCDL2007]. Pereslavl, Pereslavl Univ. Publ., 2007, vol. 1, pp. 175–181 (in Russ.).

4. Giorgios R. *Generalised LR Parsing Algorithms*. PhD thesis, Univ. of London Royal Holloway, 2006, 253 p.

5. *Instrument dlya izvlecheniya strukturirovannykh dannykh iz teksta Tomita-parser* [Tomita-parser tool to extract structured data from texts]. Available at: https://tech.yandex.ru/tomita/ (accessed August 6, 2016).

6. *Nauchnoe nasledie Rossii* [Scientific Heritage of Russia]. Available at: http://e-heritage.ru/index.html (accessed August 6, 2016).

7. Kalenov N.E., Savin G.I., Serebryakov V.A., Sotnikov A.N. Scientific Heritage of Russia Digital Library: constriction and sources aggregation philosophy. *Programmnye produkty i sistemy* [Software & Systems]. 2012, no. 4, pp. 30–40 (in Russ.).

8. *Using the API*. Available at: https://developers.google.com/books/docs/v1/using (accessed August 6, 2016).