УДК 004.822

Интеграция онтологических ресурсов в документы открытых форматов для решения задачи семантического индексирования

С. А. Шумаков
Пермский государственный национальный исследовательский университет
Россия, 614990, Пермь, ул. Букирева, 15
shumakov.sergey.a@gmail.com

В. В. Ланин
Национальный исследовательский университет "Высшая школа экономики"
Россия, 614070, Пермь, Россия, ул. Студенческая, д. 38
vlanin@hse.ru

Описывается разработка программной библиотеки для включения онтологических метаданных в документы современных офисных форматов. Рассматривается модель документа, используемая для индексации контента документа понятиями включенной в него онтологии. Описываются существующие проекты, направленные на решение сходных задач.

Ключевые слова: *онтология; семантическое индексирование; форматы документов.*

Введение

В современных информационных системах наблюдается переход от обработки структурированных данных к оперированию неструктурированными данными, под которыми для определенности будем понимать традиционные электронные документы в различных форматах. Данная тенденция заметна как в корпоративном сегменте, так и среди частных пользователей. На протяжении всей истории развития информационных технологий для обработки документов разрабатывались и применялись специализированные программные средства и форматы хранения документов.

В настоящий момент исключительно важной частью информационного пространства стали новые классы систем (социальные сети, корпоративные порталы, wiki-ресурсы и т.д.), ключевым элементом которых является понятие "контента", которое можно обобщить до "электронный документ".

Повсеместно применяемая технология WYSYWIG стала "бомбой замедленного действия" для электронных документов. Большинство современных технологий подготовки и работы с документами (текстовые редакторы, язык HTML) ориентированы на организацию удобной работы с информацией для человека, так как зачастую методы работы с электронной информацией просто копировали методы работы с "бумажной" информацией. Отметим, что в текстовом редакторе присутствуют широкие возможности форматирования текста (представления в удобном для человека виде), но практически отсутствуют возможности для передачи смыслового содержания текста, т.е. отсутствуют средства для семантического индексирования. Автоматическая интеллектуальная обработка текста чрезвычайно затруднена, так как обычно мы имеем дело с "документом для человека", а не с "документом для человека и системы".

Современный подход к определению электронного документа кроме представления собственно содержимого требует наличия ме-

© Шумаков С. А., Ланин В. В., 2015

Исследование выполнено при финансовой поддержке РФФИ в рамках проекта № 14-07-31273-мол_а.

таданных, описывающих структуру и семантику представленных в документе данных. Благодаря такому подходу обработка электронных документов может быть организована на качественно ином уровне, так как становится возможным автоматический интеллектуальный анализ информации. Эта концепция заложена в проект Semantic Web, однако состояние проекта "семантической паутины" в силу целого ряда причин еще далеко от практической реализации. Однако идеи, заложенные в Semantic Web [1], могут быть реализованы в рамках отдельно взятой информационной системы благодаря меньшему масштабу ее предметной области [2]. В настоящее время данные, необходимые для обработки документов, рассредоточены (хранятся как в самом документе, так и в базах данных ИС, обрабатывающих документы) и специфичны для каждой из задач, решаемых в течение жизненного цикла документа в ИС. Очевидна необходимость использования единого механизма представления информации о документе. Решением может стать онтологический ресурс, описывающий различные аспекты электронного документа, рассматриваемые в течение всего его жизненного цикла. Этот ресурс может стать основой для решения широкого спектра задач, связанных с обработкой электронных документов в ИС.

Для комплексного решения поставленных задач необходимо разработать модель электронного документа, позволяющую включить в него метаинформацию, и онтологический ресурс, являющийся базой для семантического индексирования содержимого документа, разработать технологию внедрения метаданных в документ и предложить механизм обработки документов. Решению одной из указанных выше подзадач, а именно включению онтологического ресурса в документы, посвящена настоящая статья.

1. Модель документа

Электронный документ представляет собой *набор структурных элементов*, называемых в данной работе *фрагментами*. Примерами фрагментов могут служить таблица, заголовок, реквизиты углового бланка и т.д.

Таким образом, документ может быть представлен четверкой вида:

$$d = (S(F, R), C, o, M).$$

Здесь $S(F, R)$ – ориентированный гиперграф, вершинам которого сопоставлены элементы множества F (множество F – это множество фрагментов документа, а R – это множество ребер графа, соответствующее связям между фрагментами); элементы множества C представляют информационное содержание документа (его контент); o – онтология документа, M – отображение множества F на концепты онтологии o. Рассмотрим подробнее описанные компоненты.

Гиперграф $S(F, R)$ задает *отношение между фрагментами документа*. Ориентированность графа необходима, например, для отслеживания связей "часть-целое" между фрагментами. Вершины, входящие в ребро, пронумерованы, что позволяет установить порядок следования фрагментов в тексте документа. Очевидно, что ребро, включающее все вершины, соответствует документу целиком.

Фрагменты могут быть двух видов: *элементарные* фрагменты представляют простейшие неделимые элементы, такие как заголовок или дата составления документа, а *составные* содержат в себе другие фрагменты.

Определим формально *фрагмент* как пару вида:

$$f = (stat, inf),\ inf = \begin{cases} F^*, F^* \subseteq F; \\ c, c \in C. \end{cases}$$

Здесь *stat* – это статическая часть фрагмента, она может быть представлена текстом, изображением, ссылкой, каким-либо специальным символом, кроме того, здесь может содержаться и информация для представления фрагмента; *inf* – это часть фрагмента, которая либо указывает место для размещения элемента содержания c ($c \in C$), либо содержит множество фрагментов F^*.

Традиционно для представления документа используются обычные графы, чаще всего деревья (например, формат XML). Древовидная структура описания значительно упрощает работу с документом, но, вместе с тем, вносит и существенные ограничения. Выбор гиперграфа для представления структуры документа обосновывается возможностями гиперграфов представлять произвольные связи между фрагментами документа и их множествами.

В описанных выше обозначениях *шаблон документа* можно определить как $t = (S(F, R), C_0)$, где C_0 – *первичный контент* (например, стандартные заголовки, включенные в шаблон, и т.д.).

Учитывая специфику решаемых в данной работе задач, конкретизируем понятие *онтологии*:

$$o = (C, R, A),$$

где C – множество *понятий* (*концептов*) онтологии, R – множество *отношений* между концептами, A – множество *аксиом*, заданных на онтологии. Концептами могут быть как классы, так и экземпляры этих классов, а аксиомы используются для задания ограничений и правил, которые не могут быть выражены через отношения.

Для обработки документов необходимо реализовать операцию выделения произвольной части документа (назовем ее *операцией получения диапазона*), входным параметром которой является произвольное множество вершин графа, а результатом – подграф, порожденный этим множеством вершин. *Операция расшифровки* – "наложение" структуры на фрагмент (вершину графа). Помимо структуры и содержания в большинстве приложений важную роль играют *визуальное оформление документа* и его представление в определенном формате, поэтому необходима и операция *представления документа в определенном формате*, представляющая функцию, задающую соответствие между фрагментами документа и некоторым множеством форматов, элементы которого задают правила отображения фрагментов. Операция *поиска* применима к различным составляющим документа: структуре, содержанию и представлению, а результатом операции будут фрагменты документа, удовлетворяющие заданным критериям поиска.

Представленная выше модель документа достаточно хорошо согласуется с концепцией новых форматов офисных электронных документов, основанных на XML, таких как Open Document Format и Open XML.

2. Существующие решения и подходы

2.1. Проект Semantic Assistants

Semantic Assistants – это исследовательский Open Source проект [3, 4] разрабатываемый Канадской лабораторией Semantic Software Lab. Semantic Assistants помогает пользователям в извлечении, анализе и разработке контента, предоставляя контекстные услуги NLP (Natural Language Processing – обработка естественного языка), напрямую связываясь с настольными приложениями (текстовыми процессорами, почтовыми клиентами, браузерами), веб-информационными системами (например, wiki) и мобильными приложениями на базе Android. Semantic Assistants имеет открытую сервис-ориентированную архитектуру и использует OWL-онтологии.

Архитектурно Semantic Assistants состоит из четырех уровней. На первом уровне находятся клиентские приложения. На втором уровне находятся веб-сервисы и NLP Service Connector, который в настоящее время оборачивает GATE фреймворк для NLP, отвечает за связь с клиентами, чтение запросов и формирование ответов. На третьем уровне находится подсистема NLP, которая отвечает за извлечение, обобщение и индексацию информации, а также поиск. Четвертый уровень – ресурсный. Он содержит все необходимые внешние документы, к которым подсистема NLP должна иметь доступ.

На сегодняшний момент проект находится в стадии разработки и реализовано всего три клиента, из которых только один текстовой процессор – OpenOffice.org Writer. Для внедрения семантической информации в ODF документы Semantic Assistants не использует все возможности, предоставляемые спецификацией ODF 1.2, а использует механизм рецензирования, добавляя в документ примечания, тем самым сохраняя информацию в неструктурированном виде и доступном для редактирования пользователем, что не всегда является удобным.

2.2. Word Add-in for Ontology Recognition

Word Add-in for Ontology Recognition (Word Add-in) [5] – инструмент для ручного аннотирования документов в среде Microsoft Word. Word Add-in представляет собой надстройку уровня приложения (application level add-in) для Microsoft Office и создан на платформе .NET с использованием технологии VSTO. Word Add-in является Open Source проектом, что позволяет любому заинтересованному пользователю с легкостью адаптировать его для своих нужд.

Работа с Word Add-in начинается с выбора базы онтологий – электронного каталога, содержащего онтологии, относящиеся к одной предметной области. Затем пользователь может выбрать одну из онтологий выбранной базы, после чего при работе с Word Add-in начинает в фоновом режиме производить ана-

лиз вводимого текста. Если введенное слово совпадает с одним из концептов выбранной онтологии, оно специальным образом пометится (смарт-теги или custom actions). При активации смарт-тега или выбора в меню custom actions появляется специальное контекстное меню, с помощью которого можно просмотреть данный концепт в браузере онтологий.

Одной из основных проблем Word Add-in является синонимия и то, что одно слово может соответствовать нескольким концептам различных онтологий. В этом случае пользователь вынужден выбирать одну из онтологий, наиболее удовлетворяющую смыслу текста.

Несмотря на вышеперечисленные недостатки Word Add-in представляет собой вполне законченный продукт. Основным его преимуществом является высокий уровень интеграции с одним из самых популярных офисных пакетов – Microsoft Office, что позволяет использовать его широкому кругу пользователей, и при этом не предъявляется особых требований к их подготовке.

2.3. An Infrastructure for Managing Semantic Documents

Infrastructure for Managing Semantic Documents (ISDM) – специализированный промышленный инструментарий. В качестве основных функций Infrastructure for Managing Semantic Documents (ISDM) можно указать следующие:

- полуавтоматическое аннотирование электронных документов на основе онтологий с использованием размеченных шаблонов;
- контроль версий электронных документов;
- семантический поиск;
- уведомление об изменениях.

ISDM состоит из двух основных модулей: репозитория семантических документов (Semantic Document Repository – SDR), предназначенного для хранения электронных документов, и так называемого "основного модуля". Основной модуль, в свою очередь, имеет сложную структуру и может быть разделен на несколько модулей: модуль семантической разметки (Semantic Annotaion Module – SAM), модуль извлечения данных и контроля версий (Data Extraction and Versioning Module – DEVM), поисковый модуль (Search and Traceability Interface Module – STIM).

Модуль семантической разметки (SAM) позволяет добавлять метаданные, соответствующие предметной онтологии, в электронный документ. Версия ISDM, описанная в [6], поддерживает только один формат электронного документа – ODF 1.0. В данной версии формата еще отсутствовала удобная и гибкая модель метаданных, и поэтому авторы были вынуждены использовать наиболее подходящие для этого средства, предоставляемые форматом. Вместо использования ручного аннотирования документов в проекте предложен подход, основанный на использовании шаблонов, что позволяет повторно использовать метаданные.

Для представления метаданных используются так называемые "инструкции": instance и property. Для указания онтологий, применяемых при аннотировании, используется скрытое поле, имеющее название "Ontologies", в значении которого указываются URL онтологий.

Хотя данный проект актуален и по сей день, его основная часть, а именно механизм семантической разметки, значительно устарела, так как спроектирована в соответствии с ODF 1.0, в то время как новая спецификация ODF 1.2 предоставляет средства для добавления метаданных в ODF-документы.

3. Архитектура компонента

Перейдем к описанию разработанной авторами библиотеки OfficeMetadataLib, предназначенной для включения метаданных в документы формата Office Open XML и Open Document.

3.1. Предъявляемые требования

Программная библиотека должна предоставить следующие основные функции.

- Создание новых и открытие существующих текстовых электронных документов форматов Office Open XML и Open Document.
- Предоставление доступа (управления) к текстовым содержимым документов.
- Предоставление доступа (управления) предустановленными и пользовательскими метаданными документов.
- Внедрение онтологий в формате OWL в метаданные документа и предоставление к ним доступа (управления).

- Автоматизированный поиск и связывание фрагментов текста документа с концептами онтологий.
- Возможность расширения алгоритма и замены реализованных базовых алгоритмов поиска и лемматизации.

3.2. Описание архитектуры

Для обеспечения унифицированного доступа к электронным документам форматов Office Open XML и OpenDocument, и возможности расширения базовых алгоритмов поиска и лемматизации, программная библиотека OfficeMetadataLib имеет модульную архитектуру, схематично изображенную на рисунке.

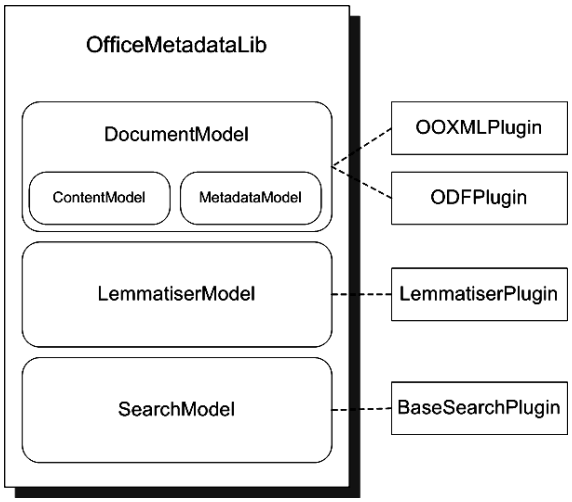

Схема модульной архитектуры

Модуль DocumentModel описывает обобщенную модель текстового офисного документа, состоящую из двух уровней:

1. *ContentModel* – модель содержимого документа;
2. *MetadataModel* – модель метаданных документа.

Данные модели спроектированы с учетом стандарта ISO/IEC 29500 [7, 8] и спецификации OASIS ODF 1.2 [9, 10].

Модуль *LemmatiserModel* описывает обобщенную модель лемматизатора. Модуль *SearchModel* описывает обобщенную модель поискового движка. Модули *OOXMLPlugin* и *ODFPlugin* реализуют обобщенную модель документа с учетом специфики каждого формата Office Open XML и OpenDocument соответственно.

Выделение реализации функций для каждого формата документа в отдельный плагин позволит дорабатывать и модифицировать код каждого плагина отдельно (например, в случае изменения спецификации формата документа), не меняя общей модели и не затрагивая исходный код основной библиотеки и других плагинов.

LemmatiserPlugin – конкретная реализация лемматизатора. Выделение реализации лемматизатора в виде плагина позволит подключать к библиотеке лемматизаторы сторонних разработчиков (реализующие соответствующие интерфейсы). Плагин ***BaseSearchPlugin*** реализует базовый поисковый движок. Аналогично лемматизатору может быть заменен сторонними разработками.

Заключение

В рамках описанного исследования была разработана программная библиотека, предоставляющая унифицированный доступ (управление) метаданными офисных документов форматов Office Open XML и OpenDocument Format. Основной составляющей данной библиотеки является OfficeMetadataLib.DocumentModel – модель электронного документа и его метаданных, основанная на моделях ISO/IEC 29500 (Office Open XML) и OASIS ODF 1.2. Данная модель адекватно отражает особенности обоих форматов и позволяет унифицированным образом работать с документами, использующими данные форматы. Также стоит отметить, что, несмотря на то, что OfficeMetadataLib.DocumentModel была изначально спроектирована для работы с документами в форматах Office Open XML и OpenDocument, благодаря ее гибкой структуре, присутствует теоретическая возможность применения данной библиотеки для работы с документами других форматов.

OfficeMetadataLib DocumentModel описывает модель документа и его метаданных, а программная реализация функций преобразования модели в документ конкретного формата содержится в специальных плагинах, которые используют для этого специализированные API (например, Open XML SDK, OpenOffice.org SDK). Использование подхода на основе плагинов позволяет избежать необходимости самостоятельной реализации всех особенностей работы с перечисленными выше форматами и позволяет использовать для этого уже существующие программные решения. Также стоит отметить, что использование

плагинов обеспечивает высокую степень гибкости и расширяемости. В случае устаревания какой-либо библиотеки или появления новой, более удобной, возможно просто заменить или добавить плагин, не изменяя уже существующий код модели.

Использование унифицированного подхода к разработке модели работы с электронными документами привело к отсутствию возможности использовать особенности конкретных форматов. Однако этот недостаток компенсируется возможностью подключения пользовательских плагинов.

В дальнейшем планируется реализовать интерфейс для выполнения SPAQL запросов к метаданным документа. Разработанная библиотека станет основой для интеграции офисных пакетов с *многоаспектной онтологией электронных документов* [11].

Список литературы

1. *Berners-Lee T., Hendler J., Lassila O.* The semantic web // Scientific American (May 2001). P. 28–37.
2. *Ланин В.* Онтологии как основа функционирования систем обработки электронных документов: матер. всерос. конф. с междунар. участием "Знания–Онтологии–Теории". Новосибирск, 2009, Т. 2. С. 173–177.
3. *Bakalov F., Sateli B., Witte R., Meurs M.-J., Komg-Ries B.* Natural Language Processing for Semantic Assistance in Web Portals // IEEE Sixth International Conference on Semantic Computing (ICSC 2012), 2012. P. 67–74.
4. *Witte R., Gitzinger T.* Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients // in 3rd Asian Semantic Web Conference (ASWC 2008), ser. LNCS, vol. 5367. Bangkok, Thailand: Springer, 2008. P. 360–374. Available: URL: http://rene-witte.net/semantic-assistants-aswc08.
5. *Fink JL, Fernicola P, Chandran R., et al.* Word add-in for ontology recognition: semantic enrichment of scientific literature. BMC Bioinformatics. 2010;11:103. doi:10.1186/1471-2105-11-103.
6. *Lucas de Oliveira Arantes, Ricardo de Almeida Falbo*. An Infrastructure for Managing Documents // 2010 14th IEEE International Enterprise Distributed Object Computing Conference Workshops.
7. *ISO/IEC* 29500-1 Third edition, 2012-09-01. Information technology – Document description and processing languages – Office Open XML File Formats. Part 1: Fundamentals and Markup Language Reference. 5030 p.
8. *ISO/IEC* 29500-2 Third edition, 2012-09-01. Information technology – Document description and processing languages – Office Open XML File Formats. Part 2: Open Packaging Conventions. 138 p.
9. *Open* Document Format for Office Applications (OpenDocument) Version 1.2 Part 1: OpenDocument Schema 29 September 2011. 846 p.
10. *Open* Document Format for Office Applications (OpenDocument) Version 1.2 Part 3: Packages 29 September 2011. 35 p.
11. *Lanin V., Sokolov G.* Using multidimensional ontology of electronic document for solving semantic indexing problem, in: Proceedings of the 8th Spring/Summer Young Researchers' Colloquium on Software Engineering (SYRCoSE 2014). M., 2014. P. 166–169.

Integration of Ontology Resources into Open Format Documents for Semantic Indexing

S. A. Shumakov[1], **V. V. Lanin**[2]

[1]Perm State National Research University, Russia, 614990, Perm, Bukirev st., 15
shumakov.sergey.a@gmail.com

[2]National Research University Higher School of Economics, Russia, Perm, Studencheskaya st., 38
vlanin@hse.ru

The article describes the development of a software library for ontological metadata inclusion into modern office documents formats. The model of the document used for indexing ~~the~~ its content by ontology concepts is given. Existing projects addressed for similar problems are overviewed.

Key words: *ontology; semantic indexing, document formats.*