УДК 004.912

Д.Е. Семёнов

Модификация метода ручного формирования семантических метаданных корпоративных баз знаний

Предложен модифицированный метод ручного формирования семантических метаданных. Описан модуль расширения для текстового редактора Microsoft Word, позволяющий осуществлять аннотирование электронных документов для наполнения корпоративной базы знаний в ручном режиме. Описаны примеры использования.

Ключевые слова: онтология, метаданные, аннотирование, электронный документ, корпоративная база знаний.

Как показывает практика, в большинстве организаций уже сформировался набор программного обеспечения (ПО) для работы с электронными документами (например, Microsoft Office для текстовых документов или Microsoft Excel для электронных таблиц). С одной стороны, для построения корпоративной базы знаний из электронных документов компании необходимо использовать дополнительно ПО, что может быть связано с дополнительными расходами. С другой стороны, при неэффективной организации использования нового ПО может возникнуть отрицательный эффект (отторжение внедрения), что скажется на технологическом процессе работы. Исходя из этого, можно предположить, что **данную проблему рациональнее решить** через создание модулей расширения для существующего ПО. Такими примерами служат плагины для операционной системы, текстовых редакторов электронных документов и т.д.

Модули расширения дополняют существующее ПО новыми функциональными возможностями, не влияют на основную работу и не меняют внутренний формат хранения информации. Особенно это **актуально** при работе с существующими большими объемами электронных документов, построенных в виде иерархий информационных объектов (ИО). Принцип построения таких документов представлен в работе [1] и может быть рассмотрен в трех разных аспектах – структура, контент и контекст. Для решения задачи интеграции разнородных ресурсов информации и данных используется методология Semantic Web, предложенная консорциумом W3C [2], описанная в работе [3]. Построение моделей представления ИО достаточно широко применяется при разработке информационных и экспертных систем [1, 3, 6], основанных на корпоративных базах знаний (КБЗ) организации.

В результате исследований предложено использовать стандарт идентификации Universally Unique Identifier (UUID) [4] ИО, предназначенный для создания программного обеспечения. Все используемые идентификаторы (ID) ИО предлагается преобразовать в глобальные статистически уникальные 128-битные номера, позволяющие использовать специальные идентификаторы преднамеренно, для повторной идентификации той же самой сущности в различных контекстах. Такой подход предложен для построения модифицированной модели представления ИО с использованием ориентированных графов в реляционных базах данных [5]. Для успешной реализация этой модели были исследованы методы формирования КБЗ, направленные на аккумулирование информации, обработку и последующую выдачу результата.

Целью данной работы является анализ, выбор и модификация метода формирования семантических метаданных электронных документов и его реализация в виде модуля расширения для текстового редактора.

Онтологический подход для организации корпоративных баз

Для определения логической структуры взаимосвязей между элементами КБЗ необходимо выделить (абстрагировать) понятия из содержания знаний (документы, файлы, опыт сотрудников, записи в базах данных, ссылки и т.п.) и структурировать (организовать) их формальным способом, путем задания взаимоотношений между этими понятиями. Одним из распространенных способов описания знаний в виде множества понятий и взаимоотношений между ними являются онтологии

[1, 3]. Под онтологией предметной области будем понимать знаковую систему, которая может быть представлена в следующем формальном виде:

$$O_{\text{mod}} = \{X', R', F, F_{tag}\},$$

где *X'* – конечное множество понятий предметной области с уникальным идентификатором; *R'* – конечное множество отношений между понятиями, сгруппированное по концепциям; *F* – конечное множество функций интерпретации; F_{tag} – конечное множество функций расстановки ключевых слов к электронным документам.

Кроме элементов знаний в организации имеется множество потребителей данной информации (сотрудники). Порой время поиска необходимой информации много больше, чем время использования этой информации. Решение данной проблемы может быть найдено в использовании аннотированной КБЗ, построенной с использованием модифицированной модели представления ИО в виде ориентированных графов в реляционных базах данных [5]. Это обусловлено введением метаданных и ключевых слов для всех электронных документов, что обеспечивает минимальное время поиска нужной информации в конкретный промежуток времени. Метаданные – это данные о данных или структурированные данные, которые описывают характеристики объектов – носителей источников информации и способствуют идентификации, обнаружению, оценке и управлению этими объектами. На основе метаданных, накопленных в банке знаний, можно проводить анализ взаимосвязей, который позволит выявить зависимость между источником и приёмником данных.

Принимая во внимания тот факт, что данные для КБЗ представлены набором семантической информации о документах и непосредственно самими электронными документами объекта знаний, то составление КБЗ возлагается на группу экспертов, знания которых должны охватывать всю деятельность компании. Основная роль эксперта – указание семантических метаданных для предмета описания, т.е. создание (аннотирование документов) метаданных электронных документов. Эту функцию возможно выполнять во время работы с электронным документом напрямую из текстового редактора или из контекстного меню операционной системы.

Анализ методов формирования семантических метаданных корпоративных баз знаний

В работе [7] рассмотрены принципы ручного и полуавтоматического семантического аннотирования документов. В предложенном полуавтоматическом режиме создания документов [7] был использован подход, базирующийся на использовании лингвистических методов морфологического, синтаксического и поверхностного семантического (общеописательного) анализа документов на естественном языке. Реализация полуавтоматического режима создания аннотирования документов достаточно трудоемкая задача, для реализации которой требуется большое количество времени и наличие высококвалифицированной команды разработчиков. Однако в связи с тем, что задача понимания текстов на естественном языке до сих пор в полной мере не решена, не представляется возможным использовать полуавтоматический режим в полной мере, без вмешательства человека [6]. В случае, когда пользователь работает с существующими информационными системами документооборота, для формирования метаданных электронных документов необходимо прибегать к использованию «мастеров» заполнения ключевых слов, как это сделано в «1С: Документооборот» [9] и аналогичном ПО. Такой подход влечет за собой поочередное открытие большого количества дополнительных окон, что увеличивает время выполнения конкретной задачи. В ряде других систем работа только с целым документом, а не с конкретной его частью, как это сделано в «DocsVision» [10] и др., что не позволяет заполнить ключевые слова к произвольной части существующего электронного ресурса.

В результате исследований, развивая предложенный в работе [7] метод ручного семантического аннотирования документов, а также принимая во внимание труды, отмеченные в работе [6], и проведя анализ существующих программных продуктов [9, 10], автором предложено реализовать модуль расширения для текстового редактора Microsoft Word. Данный плагин позволит пользователю формировать метаданные для произвольного текста (контента), включая графику и объекты, не используя предустановленные шаблоны и не прибегая к дополнительным программным оболочкам, что обеспечит возможность выполнения простых операций с документами прямо в пакете Microsoft Office.

Модификация метода ручного формирования семантических метаданных заключается в отказе от использования предустановленных шаблонов [7] для текстового редактора, добавления возможности создания метаданных из произвольной выделенной области текста электронного документа,

включая графические элементы и другие встроенные объекты, основанного на использовании закладок в текстовом редакторе (закладки используются для позиционирования в электронном документе). Выделенные блоки текста могут быть сохранены в единую корпоративную базу, построенную с использованием модифицированной модели представления ИО [5]. «Мастер» по заполнению ключевых слов интегрирован прямо в текстовый редактор Microsoft Word.

В предложенном подходе выделенная область текста сохраняется с использованием закладок. Информация об областях помещается в служебный файл, расположенный в рабочей папке пользователя. При несанкционированном копировании электронного документа доступ к информации становится ограниченным, которая хранится в защищённом виде (обращение обеспечивается посредством служебного файла). Благодаря этому обеспечивается сохранность конфиденциальной информации (блокируется функция вывода на печать или копирования текста). При синхронизации с корпоративной базой служебный файл также становится элементом КБЗ.

Проектирование хранилища электронных документов корпоративной базы знаний

Модуль расширения для текстового редактора предоставляет пользователям единое информационное пространство. Это достигается за счет использования системы управления базой данных (СУБД) и архивом индексированных блоков текста из электронных файлов (аннотированных документов), расположенных на выделенном сервере или компьютере пользователя, под управлением операционной системы Windows, подключенного в глобальную сеть Интернет. Такая организация межсетевого взаимодействия используется в системах управления версий программного кода (Subversion [13], Team Foundation Server 2010 [14, 15] и др.). Основное отличие от аналогичных программных продуктов заключается в том, что метаданные ИО расположены в СУБД, по которым осуществляется полнотекстовый поиск и поиск по ключевым словам. На рис. 1 схематично представлено расположение файлов в КБЗ и размещение их метаданных (такую организацию ИО и метаданных в едином месте назовем репозитарием). В левой части рис. 1 представлены пользователи системы, которые работают с текстовым редактором Microsoft Word, со встроенным модулем расширения. Светлой стрелкой указана взаимосвязь электронных документов и индексированных файловых архивов, темной – метаданных и базы данных. Модуль расширения выступает посредником между репозитарием (сервер) и потребителем информации (клиент), сохраняя и выдавая необходимые ИО по запросу.

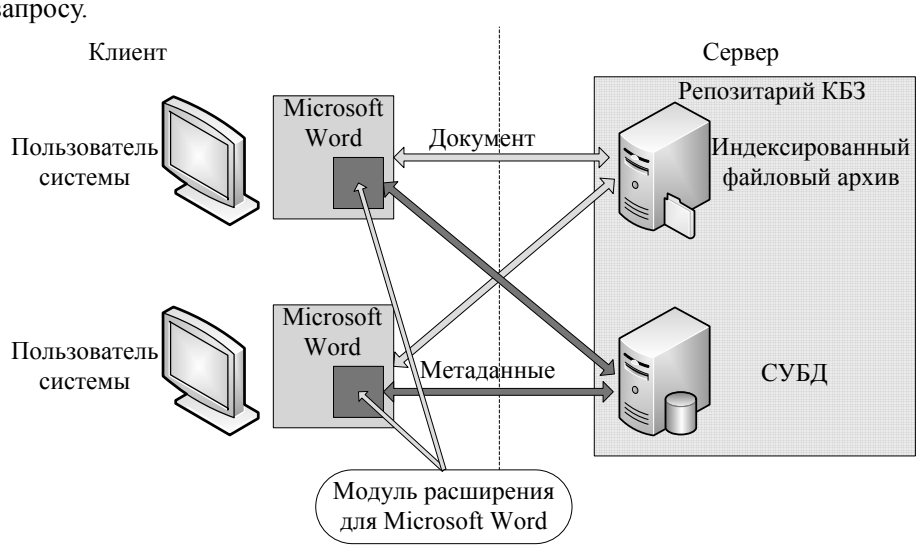

Рис. 1. Схема расположения модуля расширения для Microsoft Office

Реализация модуля расширения модифицированного метода ручного формирования семантических метаданных

Проведя анализ современных информационных технологий по созданию модулей расширений [11], была выбрана технология COM (Component Object Model – объектная модель компонентов) – это технологический стандарт от компании Microsoft, предназначенный для создания программного обеспечения на основе взаимодействующих распределённых компонентов, каждый из которых может использоваться во многих программах одновременно. Основой для выбора послужило то, что технология COM – это наиболее законченная, хорошо документированная, адаптированная к изме-

нениям стандартов программирования система. На базе этой технологии выполнены компоненты компании «Add-in express™» [12] для среды быстрой разработки приложений (Rapid Application Development Software) Embarcadero RAD Studio 2010, используемые при построении модуля расширения для текстового редактора Microsoft Office Word, выполненного в виде набора плагинов Microsoft Office COM Add-ins (поддерживаемые версии Microsoft Office 2000–2010).

Функциональные клавиши модуля расширения ручного формирования семантических метаданных документов расположены на дополнительных вкладках (для версии Microsoft Office 2007 и 2010) или дополнительных панелях (для версии Microsoft Office 2000 и 2003).

Перечислим основные функции модуля расширения:

1. Создание ИО из произвольной выделенной области текста в электронном документе (присутствует «мастер»).
2. Создание связей между созданными ИО в одном документе с указанием ключевых слов (присутствует «мастер»).
3. Навигация по ИО внутри документа.
4. Поиск схожих ИО по ключевым словам и описанию.
5. Просмотр ключевых слов и описания выделенного ИО.
6. Просмотр истории изменения ИО, с указанием даты и автора.
7. Экспортирование ИО во внешний файл (только текст в формате XML [16–17], один ИО в файл Word Document, экспортирование объектов в графическом виде с применением кадрирования, комбинирование нескольких ИО в отдельный файл Word Document).
8. Экспортирование ИО в другие программные продукты, поддерживающие протолок вызова удаленных процедур XML-RPC [16–17].
9. Шифрование ИО для защиты от несанкционированного копирования.
10. Синхронизация с хранилищем электронных документов.

Применение на практике

Рассмотрим типичные ситуации работы с электронными документами, демонстрирующие особенности предложенного модифицированного метода ручного формирования семантических метаданных.

Задача 1. Составить текст, состоящий из однотипных блоков, расположенных в других электронных документах.

Типовое решение. Поочередное открытие разных электронных документов и помещение необходимой информации посредством копирования и вставки. При составлении следующего аналогичного текста необходимо повторять эти операции снова.

Предлагаемое решение. Создание нового документа и добавление необходимых блоков текста из КБЗ через модуль расширения текстового редактора. Нет необходимости открывать большое количество файлов.

Задача 2. Экспортирование данных во внешнюю программу или файл.

Типовое решение. Открытие внешней программы и открытие большого количества однотипных файлов. Копирование выделенного текста в программу через буфер обмена операционной системы или сохранение в промежуточный файл.

Предлагаемое решение. Запуск мастера по экспорту данных из модуля расширения для текстового редактора, сохраняющий выделенные блоки текста в XML формат [16–17], подходящего для импорта во внешнюю программу. При необходимости возможен перенос данных напрямую во внешнюю программу посредством подключения дополнительных модулей (пример таких модулей не описан в данной статье).

Использование данного метода позволит отказаться от множественного повторного открытия электронных документов для однотипных действий с контентом.

Пример использования

Рассмотрим пример формирования билета для проведения компьютерного тестирования, состоящего из 10 заданий с четырьмя вариантами ответов.

В открытом электронном документе Microsoft Office Word, содержащем вопросы одинаковой сложности, которые могут быть использованы при тестировании студентов, выделяем поочерёдно вопрос за вопросом и варианты ответов, указывая атрибуты и взаимосвязь ответов с вопросами. После сохранения документа все выделенные области текста сохраняются в едином репозитарии в виде взаимосвязанных ИО. Модуль расширения автоматически сгенерирует необходимый XML-

документ, который может быть использован при тестировании в обучающей среде Moodle [18] или других аналогичных программных продуктах. В этом подходе нет необходимости создавать все вопросы в импортирующей программе и перемещать текст методом копирования и вставки.

Заключение

Проведенный анализ методов ручного аннотирования электронных документов показал, что данная задача весьма актуальна и находит обширное применение в большинстве организаций [1, 3, 6–8]. Новизной в данной работе является возможность добавлять метаданные к любому выделенному фрагменту электронного документа без использования предустановленных шаблонов, что особо актуально при работе с существующими файлами, с последующим индексированием в КБЗ и отражением метаданных в СУБД. Использование аннотированной КБЗ позволяет отслеживать динамику изменения электронных документов аналогично системам документооборота. Эта возможность позволяет вернуться к любой из версии документа и произвести сопоставление различий, что позволяет работать нескольким пользователям в едином информационном пространстве.

Разработанный модуль расширения добавляет функционал к текстовому редактору и операционной системе, что позволяет отказаться от внедрения новых сложных информационных или экспертных систем. Это позволяет использовать новые функциональные особенности быстрее, не обучая сотрудников организации новым программным продуктам, и работать прямо в пакете Microsoft Office.

В настоящее время предложенный в этой статье метод ручного формирования семантических метаданных КБЗ используется в основе интеллектуальной информационной системы экспертизы качества тестовых материалов [19, 20], что позволяет структурированно хранить используемую документацию и обеспечивать обратную связь между компьютерными программами тестирования и банком данных [21] экспертно-советующего контекстного помощника [22].

Работа выполнена при финансовой поддержке гранта «Индивидуальный грант молодого ученого Томского политехнического университета», приказ ректора от 21.05.2010 г. № 3401.

Литература

1. Сидорова Е.А. Семантический подход к анализу документов на основе онтологии предметной области / Е.А. Сидорова, Ю.А. Загорулько, И.С. Кононенко // Компьютерная лингвистика и интеллектуальные технологии: труды междунар. конф. «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2006. – С. 468 – 474.

2. Allemang D. Semantic Web for the Working Ontologist: Modeling in RDF, RDFS and OWL / D. Allemang, J. Hendler. – N.Y.: Morgan Kaufmann Publishers, 2008. – 350 p.

3. Черний А.В. Развитие информационной системы организации с использованием семантических технологий / А.В. Черний, А.Ф Тузовский // Матер. всерос. конф. с междунар. участием «Знания-Онтологии-Теория» (Новосибирск, 20–22 октября 2009 г.). – Новосибирск: ЗАО «РИЦ Прайс-Курьер», 2009. – Т. 2. – С. 52–59.

4. Электронная энциклопедия «Wiki» [Электронный ресурс]. – Режим доступа http://ru.wikipedia.org/wiki/, свободный (дата обращения 10.03.2010).

5. Семёнов Д.Е. Модификация модели представления информационных объектов с использованием ориентированных графов в реляционных базах данных // Доклады Томского государственного университета систем управления и радиоэлектроники. – № 1(21), ч. 2. – Томск: Изд-во ТУСУРа. 2010. – С. 142–148.

6. Системы управления знаниями (методы и технологии) / под ред. В.З. Ямпольского. – Томск: Изд-во НТЛ, 2005. – 260 с.

7. Тузовский А.Ф. Формирование семантических метаданных для объектов системы управления знаниями // Известия Томского политехнического университета. – 2007. – Т. 310, № 3. – С. 108–112.

8. Гаврилова Т. А. Интеллектуальные технологии в менеджменте: инструменты и системы: учеб. пособие. – 2-е изд. / Т.А. Гаврилова, Д. И. Муромцев. – СПб.: Высшая школа менеджмента, 2008. – 488 с.

9. «1С: Документооборот 8» [Электронный ресурс]. – Режим доступа http://v8.1c.ru/doc8/ свободный (дата обращения: 11.01.2011).

10. Продукты и решения «DocsVision». Клиент для Microsoft Office [Электронный ресурс]. – Режим доступа http://www.docsvision.com/catalog/produkti/dv_48.html свободный (дата обращения: 11.01.2011).

11. Бут И.А. Описание программного интерфейса для совместного использования распределённых компонентов приложения // И.А. Бут,Д.Е. Семенов, М.Е. Семенов // Сб. трудов VII Всерос. науч.-практ. конф. студентов, аспирантов и молодых ученых «Технологии Microsoft в теории и практике программирования». – Томск: Изд-во Том. политех. ун-та, 2010. – С. 145–146.

12. Библиотека VCL компонентов «Add-in express™» [Электронный ресурс]. – Режим доступа http://www.add-in-express.com (дата обращения: 10.09.2009).

13. Свободная централизованная система управления версиями. – URL: http://subversion.tigris.org (дата обращения: 04.03.2010).

14. Visual Studio Team Foundation Server 2010 [Электронный ресурс]. – Режим доступа http://www.microsoft.com/visualstudio/ru-ru/products/2010-editions/team-foundation-server свободный (дата обращения: 14.09.2010).

15. Работа с Visual Studio Team Foundation Server 2010 [Электронный ресурс]. – Режим доступа http://habrahabr.ru/blogs/vs/90911/ свободный (дата обращения: 15.09.2010).

16. Штайнер Г. HTML/XML/CSS : справочник / Г. Штайнер. – 2-е изд., перераб. – М.: БИНОМ. Лаборатория знаний, 2005. — 510 с.

17. Козлов С.В. Система удалённой работы с XML-документами / С.В. Козлов, А.Ф. Тузовский // Молодежь и современные информационные технологии: Сб. трудов рег. науч.-практ. конф. студентов, г. Томск, 25–26 февраля 2003 г. / Томский политехнический университет. – Томск, 2003. – С. 62–63.

18. Мясникова Т.С. Система дистанционного обучения MOODLE / Т.С. Мясникова, С.А. Мясников. – Харьков, 2008. – 232 с.

19. Муратова Е.А. Информационная система поддержки экспертизы качества тестовых материалов / Е.А. Муратова, Д.Е. Семёнов // «Интеллектуальные системы» (AIS-08) и «Интеллектуальные САПР» (CAD-2008): труды междунар. науч.-техн. конф.: в 4 т. (Дивноморское, 3–10 сентября 2008 г.). – М.: Физматлит, 2008. – Т. 1. – С. 182–190.

20. Семёнов Д.Е. Особенности программного обеспечения интеллектуальной информационной системы экспертизы качества тестовых материалов / Д.Е. Семёнов, Е.А. Муратова // Современные техника и технологии СТТ 2008: сб. докл. XIV Междунар. науч.-практ. конф. студентов и молодых ученых (Томск, 24–28 марта 2008 г.). – Томск, 2008. – С. 378–379.

21. Семенов Д.Е. Структуризация банка тестовых заданий // Электронные дидактические материалы в инженерном образовании: труды рег. науч.-метод. конф. ИДНО ТПУ (Томск, 11–12 октября, 2009 г.). – URL: http://www.lib.tpu.ru/fulltext/m/2009/m8/Repot/ Cemenov.html.

22. Семёнов Д.Е. Экспертно-советующий контекстный помощник (ЭСКП). Свидетельство о государственной регистрации программы для ЭВМ № 2011610676 от 11.01.2011 г.

Семёнов Дмитрий Евгеньевич
Аспирант каф. инженерной педагогики ИДНО ТПУ
Тел.: (8-382-2) 56-40-82
Эл. почта: dimomans@tpu.ru

Semenov D.E.

Modification of the method of manual format of semantic metadata in corporate knowledge bases

The paper describes the modified method of manual semantic metadata formation. The plug-in-module for Microsoft Word text editor is described, which allows annotation of electronic documents for corporate knowledge base in a manual mode. The examples of usage are given.

Keywords: ontology, metadata, annotation, electronic documents, corporate knowledge base.