Заключение

На основе проведённых исследований можно сделать следующие выводы:

1. Предложенный метод позволяет выполнять трёхстороннее слияние онтологий с учётом изменений формата онтологии, префиксов пространств имён, импортов и идентификаторов онтологии за приемлемое время.
2. Недостатком программной реализации является отсутствие возможности группировки изменений по сущностям, а также добавления новых изменений (не внесённых ни одним из пользователей). В дальнейшем предполагается усовершенствование программной реализации.

СПИСОК ЛИТЕРАТУРЫ

1. Motik B., Patel P.F., Parsia B. OWL 2 Web Ontology Language structural specification and functional style syntax // World Wide Web Consortium. 2009. URL: http://www.w3.org/TR/owl2-syntax/ (дата обращения: 23.10.2012).
2. Carroll J. Matching RDF Graphs. 2001. URL: http://www.hpl.hp.com/techreports/2001/HPL-2001–293.pdf (дата обращения: 23.10.2012).
3. DeltaView // ESW Wiki. 2007. URL: http://esw.w3.org/DeltaView (дата обращения: 23.10.2012).
4. Horridge M., Bechhofer S. The OWL API: A Java API for Working with OWL 2 Ontologies // OWL Experiences and Directions: Procs. of the VI Intern. Workshop. – Chantilly, 2009. – V. 529. – P. 53–62.

Поступила 31.10.2012 г.

УДК 004.02:004.82

СЕМАНТИЧЕСКОЕ АННОТИРОВАНИЕ ДОКУМЕНТОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

Ле Хоай, А.Ф. Тузовский

Томский политехнический университет
E-mail: lehotomsk@yahoo.com

Рассматривается задача выполнения аннотирования документов в электронных библиотеках, использующих семантические технологии. Описываются причины и преимущества использования таких аннотаций. Предложен новый метод выполнения полуавтоматического аннотирования, и показаны результаты тестирования его программной реализации.

Ключевые слова:
Аннотирование документов, семантические технологии, семантическая аннотация, электронная библиотека организации, методы аннотирования.

Key words:
Document annotation, semantic technology, semantic annotation, enterprise digital library, annotation method.

Введение

В настоящее время большинство электронных ресурсов, в особенности веб-страницы, создается только для их использования людьми. Форматы их описания практически не включают формальных описаний знаний, содержащихся в этих документах. Формальное описание основного смысла документа в удобном для программной обработки формате является целью использования активно разрабатываемых в последнее десятилетие семантических технологий, таких как RDF, RDFS, OWL и SPARQL.

Электронные библиотеки представляют собой специализированные информационные системы, которые выполняют управление коллекциями электронных ресурсов (например, таких как текстовые документы, изображения, мультимедиа файлы) с целью повышения эффективности использования содержащихся в них знаний некоторыми сообществами пользователей. Под семантическими электронными библиотеками (СЭБ) понимаются электронные библиотеки, использующие семантические технологии для организации всех процессов своей работы, таких как описание ресурсов, ведение каталогов, описание профилей пользователей, поиск и рекомендация ресурсов пользователям и т. п. Одной из важных функций семантических электронных библиотек является предоставление возможности аннотирования публикуемых ресурсов. В данной статье рассматривается проблема реализации такого подхода за счет перехода от человеко-читаемых к программно-обрабатываемым описаниям электронных ресурсов.

Аннотирование документов

В современной литературе существует разное понимание термина «аннотация ресурса». Обычно под аннотацией ресурса, понимаются некоторые метаданные (данные о данных) этого ресурса. Выделяют [1] следующие три вида аннотаций: неформальные, формальные и онтологические. Неформальная аннотация описывается на естественном языке и поэтому обычно не обрабатывается с помощью программ. Формальные аннотации составляются с использованием специальных языков (таких как XML и RDF), что позволяет выполнять их программную обработку. Если аннотация составляется на основе некоторой семантической модели (онтологии), описывающей основные понятия и отношения той предметной области, к которой относится описываемый ресурс, то она называется онтологической.

Формирование аннотаций ресурсов может выполняться с применением различных средств, таких как теги, наборы ключевых слов, наборы понятий и наборы триплетов. При использовании тегов аннотирование выполняется путем их добавления в тексты документов. Это делается для того, чтобы обрабатывающая программа могла определить формальный смысл выделенных с их помощью частей. Например, данный способ может использоваться для аннотирования веб-страниц, в которые, кроме HTML тегов, включаются и микроформаты RDFa [2]. Стандарт RDFa позволяет использовать URI идентификаторы для набора атрибутов, а также набор общепринятых словарей, таких как hCalendar, hCard или hAtom.

Другим способом аннотирования ресурсов является использование набора ключевых слов. Такое средство чаще применяется в электронных библиотеках для аннотирования научных статьей. Ключевые слова для конкретной статьи выбираются их авторами, а затем хранятся в базе данных для поддержки процесса поиска. Неоднозначность семантики ключевых слов приводит к использованию наборов понятий. Такие понятия могут выбираться из онтологий или тезаурусов, описываемых с помощью языка RDFS или OWL. При этом каждое понятие представляется в виде URI идентификатора [3].

Третьим способом аннотирования ресурса является его описание с помощью наборов утверждений, имеющих следующий формат: субъект–предикат–объект, которые также называются триплетами. Данный способ является самым новым, и его использование еще в достаточной степени не исследовано. В данной статье как раз и рассматривается подход к аннотированию ресурсов с использованием данного способа.

Аннотирование ресурсов представляет собой очень важную, но также и очень трудную и трудоемкую задачу. Следует отметить, что аннотирование является начальным этапом применения семантических технологий, а в настоящее время существует большое количество не аннотированных электронных ресурсов.

Онтологический подход к аннотированию документов

Основная часть информации (более 80 %) современных организаций содержится в виде текстов на естественных языках в бумажном и электронных форматах [4]. В связи с этим одной из наиболее сложных задач в разработке СЭБ является разработка методов и программных средств для составления достаточно точных метаданных для описания содержания текстовых документов.

Описание ресурсов в СЭБ организации выполняется с использованием набора специально разработанных онтологий, описанных на языке RDFS или OWL. В результате выполнения аннотирования для каждого ресурса создается семантическое метаописание, состоящее из набора триплетов, в состав которых могут входить контекстные и контентные семантические метаданные. Контекстные метаданные аннотируемого объекта представляют собой утверждения о его связи с другими объектами, понятиями онтологий, а контентные метаданные – утверждения о знаниях, содержащихся в самом объекте. Кроме этого, контентными метаданными ресурсов, как правило, является набор триплетов, созданный на основе онтологии, которая описывает ту предметную область, с которой связан данный ресурс. Именно контентные метаданные играют важную роль и представляют основную ценность для обработки и повторного использования, а также широко используются в среде организации [5].

Для составления контентных метаданных, содержащих наборы триплетов, используются онтологические модели.

Определение 1 (**онтологическая модель**). *Под онтологической моделью (онтологией) O понимается знаковая система <C,P,I,L,T>, где C – множество элементов, которые называются понятиями; P – множество элементов, называемых свойствами (двуместными предикатами); I – множество экземпляров понятий; L – множество текстовых меток или значений понятий и свойств; T – частичный порядок на множестве C и P.*

С помощью онтологий для ресурса могут создаваться их семантические аннотации.

Определение 2 (**семантическая аннотация**). *Семантической аннотацией ресурса является его формальное описание в виде набора кортежей вида (s,p,o,v), где s – URI идентификатор субъекта триплета; p – URI идентификатор предиката триплета; o – URI идентификатор некоторого объекта или конкретное значение некоторого типа, а v – весовой коэффициент значимости данного триплета. При этом субъекты, отношения и объекты выбираются из некоторого набора семантических моделей (онтологий). Начальные части таких кортежей, вида (s,p,o), называются триплетами.*

Онтологический подход к аннотированию заключается в формировании метаданных документа с использованием элементов онтологий [4]. На основе онтологической модели создаются контентные метаданные документа.

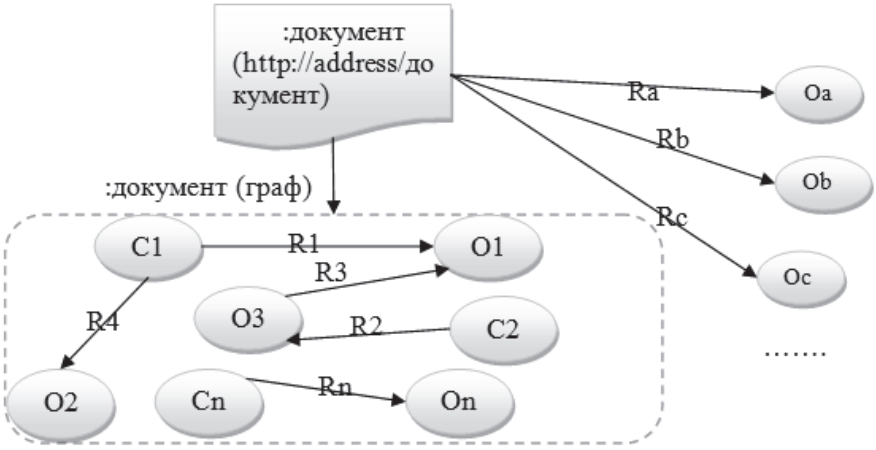

Рис. 1. *Пример семантических метаданных документа*

Определение 3 (**Контентные метаданные документа**). *Контентные метаданные документа – это набор семантических утверждений (триплетов) $M_c=\{t_1,t_2,...,t_m\}$, описывающих основные знания, содержащиеся в документе. Каждый триплет t_i имеет вид $t_i=(c,r,o,v)$, где c – это субъект утверждения ($c\subset C\cup I$); o – объект утверждения ($o\subset C\cup I$); r – отношение между субъектом и объектом ($r\subset P$), а v – весовой коэффициент, который оценивает значимость данного утверждения.*

Утверждения на основе онтологии O могут иметь следующий вид: $<C,P,C>$, $<I,P,I>$, $<I,P,C>$, $<I,P,L>$, где C, P, I – это уникальные URI-идентификаторы понятий, предикатов или экземпляров.

Как уже было отмечено, аннотирование может выполняться для создания контентных и контекстных метаданных. Но контекстные метаданные сохранятся в базе знаний как триплеты, субъектом которых является идентификатор аннотируемого документа, а контентные метаданные – как триплеты, составляющие некоторый подграф, в который входят идентификатор основных понятий, предикатов и экземпляров, являющихся важными для описания смысла данного документа (рис. 1).

Результаты аннотирования документа пополняют базу знаний не только новыми экземплярами, но и контентными метаописаниями документа. Это приводит к новым возможностям обработки документов, таким как: ***семантический поиск документа, формирование рекомендаций на основе профиля пользователя, автоматическая категоризация документов по заданной иерархии рубрик***.

Обзор методов семантического аннотирования документов

Аннотирование может выполняться разными способами: вручную, полуавтоматически или автоматически. При ручном способе аннотирования составлением метаописания документа занимается специалист. Однако программная система может оказывать ему помощь, например, показывая онтологии и имеющиеся экземпляры, выполняя проверку их правильности. При полуавтоматическом методе аннотирования система автоматически составляет начальный вариант аннотации, а затем специалист может проверять и выполнить ее корректировку. Ручной и полуавтоматический способы трудно использовать при большом количестве документов, для документов больших размеров, а также при недостатке специалистов, которые могут своевременного выполнять такую работу. В этих случаях возникает потребность в автоматическом методе аннотирования.

Ручное Аннотирование. В электронной библиотеке организации электронные ресурсы (в основном документы) создаются и формируются специалистами, которые могут создавать для них достаточно качественные аннотации. Для выполнения такой работы хорошо подходят инструменты ручного аннотирования.

При разработке инструментов или платформ ручного аннотирования рассматриваются следующие рекомендации [2]:

1. Поддержка ссылок к экземплярам и понятиям онтологии с помощью URI идентификаторов.
2. Создание инструментов на основе веб-технологий для обеспечения возможности аннотирования ресурсов различными группами пользователей.
3. Разрешение формировать аннотации различных типов (экземпляры, понятия и отношения).
4. Возможность работы с большим количеством онтологий и онтологиями, содержащими большое количество понятий и отношений.
5. Возможность аннотирования различных форматов документов, например таких, как HTML, PDF, XML, изображения и мультимедиа файлы.

Одной из известных комплексных платформ семантического аннотирования является CREAM [6]. Данная платформа работает как отдельная система со своей базой RDF-триплетов (созданной с помощью RDF-робота), которая пополняется информацией, содержащейся в веб-страницах. CREAM позволяет использовать созданные экземпляры для повторного применения, чтобы избе-

жать избыточности, а также предоставляет редактор содержания аннотируемого документа. Ручным редактором аннотаций платформы CREAM является инструмент OntoMat, работающий в веб-браузере. Данный редактор позволяет пользователям расширять онтологию новыми данными с помощью удобного интерфейса и обеспечивает три режима аннотирования: по написанию утверждения (typing statements), по разметке (markup), по управлению версиями (authoring).

Примером другого подхода к ручному аннотированию является система SMORE [7]. Она предоставляет такие же средства аннотирования, как и OntoMat, но работает совместно с редактором онтологий SWOOP [8].

Система Zemanta [9] представляет собой онлайн инструмент аннотирования для блогов и содержания электронной почты. Zemanta позволяет авторам использовать рекомендуемые лексические метки, чтобы пользователям становилась понятной связь между различными сообщениями блогов. Данный инструмент находит все страницы, связанные с заданными метками и предоставляет возможность выбирать ссылки для используемых меток, а также позволяет добавлять ссылки в виде изображений непосредственно в текст документа.

Полуавтоматическое и автоматическое аннотирование. В отличие от ручного метода, полуавтоматический метод создает частичную аннотацию документа с помощью анализа естественного языка документа. Извлечение информации является основной технологией для связывания неструктурированного текста с формальными описаниями, содержащимися в онтологиях. При извлечении информации часто используются такие компоненты обработки естественного языка, как разметчик частей речи (part-of-speech tagger), морфологический анализатор, сканеры именованных сущностей, полный (или поверхностный) синтаксический анализ и семантическая интерпретация.

Существует два подхода к извлечению информации:

1. На основе использования наборов правил, разработанных специалистами по лингвистике, которые создают словари и описывают правила извлечения требуемой информации.
2. На основе использования методов машинного обучения, позволяющих выполнять автоматическое обучение для решения различных задач извлечения информации.

Преимуществом первого подхода является то, что не требуется составлять обучающие выборки данных для создания правил и описания знаний предметной области. Благодаря созданным словарям и правилам системы аннотирования могут работать быстрее систем, разработанных с помощью машинного обучения. Для этого подхода обычно требуется создание коллекции проаннотированных человеком обучающих данных для достижения высокой точности. С одной стороны, на обучение такой системы требуется меньше затрат, чем на создание словарей и наборов правил извлечения информации из текста, но, с другой стороны, могут возникать проблемы, связанные с низким качеством составления аннотаций. В связи с этим выбор конкретного подхода зависит от характеристики системы электронной библиотеки организации. В [2] описаны и другие системы автоматического аннотирования, такие как AeroDAML, Amilcare и Melita.

Рассмотренные инструменты позволяют аннотировать документ понятиями и экземплярами онтологий, а также создавать новые экземпляры или понятия, содержащиеся в обрабатываемом документе. В отличие от описанных подходов, в данной работе предлагается метод аннотирования документов, который позволяет не только создавать новые экземпляры, но и составлять триплеты, описывающие основные знания, содержащиеся в аннотируемом документе.

Метод полуавтоматического семантического аннотирования

В общем виде для составления триплета аннотирования документа необходимо вручную выбрать субъект, определять его предикат (отношение), на основе его описания в онтологии, а затем выбрать связанный с ним объект. Созданный триплет сохраняется в базе знаний.

Выбор субъектов и объектов триплетов выполняется в ходе решения задач поиска кандидатов и преодоления многозначности.

С учетом изложенного выше, задача семантического аннотирования может быть структурирована следующим образом:

1. Выбор нужных понятий. Аннотирование документа выполняется в соответствии с некоторой онтологией предметной области, и при этом специалист имеет возможность ограничить некоторые категории понятий в онтологии предметной области (желаемые понятия) для преодоления многозначности.
2. Поиска кандидатов. Поиск в базе знаний основных кандидатов (понятий, предикатов или экземпляров). Для этого необходимо в документе находить термины (слова или словосочетания), которые совпадают с метками экземпляров или понятий онтологии или близки им в соответствии с некоторой оценкой семантической близости.
3. Преодоление многозначности. Данный шаг заключается в том, что из аннотации должны быть исключены все нерелевантные кандидаты. Эти кандидаты сходны с терминами документа по текстовым меткам, но в действительности не описывают содержание данного документа.

Решение задачи поиска кандидатов. Обозначим набор понятий и экземпляров онтологии, хранящихся в базе знаний как $M_{CE} - \{ce_1, ce_2, ..., ce_n\}$. Каждый элемент (*ce*) может иметь текстовые метки для их обозначения на разных естественных языках. Набор текстовых меток соответствующих элемен-

тов M_{CE} некоторой онтологий O обозначим как $M=\{m_1,m_2,...,m_n\}$, каждая метка может быть представлена в виде набора токенов в результате токенизации (tokenization). Под токенизацией понимается процесс разделения текста, содержащего метки экземпляров, понятий или документов, на последовательность токенов, при этом в качестве разделителя может использоваться знак пробела, а программа, выполняющая токенизацию, называется токенизатором.

Таким образом, аннотируемый документ (или набор документов D) может быть представлен в виде множества токенов. Для учета грамматики естественных языков возникает необходимость выполнять нормализацию токенов. Существуют два типа систем, выполняющих нормализацию токенов: лемматизатор (lemmatizer) и стеммер (stemmer). Как правило, лемматизатор токена возвращает его исходную форму, а стеммер – его корень с помощью правил отсечения. При нормализации могут быть удалены стоп-слова (шумовые слова), которые не несут никакой смысловой нагрузки.

Решение задачи поиска кандидатов может быть разделено на следующие шаги:

- преобразование D и M в наборы нормализированных токенов, обычно отсортированных в алфавитном порядке;
- поиск в D набора токенов для каждой метки $m_i \in M$.

В результате выполнения этих шагов будет получен набор кандидатов M_K ($M_K \subset M_{CE}$), лексические метки которых содержатся в документе [10].

Для эффективного решения задачи поиска кандидатов предлагается на основе множества D создать базу индексов *dbindx*, содержащих термины (токены), к которой можно обращаться с запросами для поиска описания любой метки m_i. Способ создания таких индексов, состоящих из терминов документов, можно найти в [11].

С учетом вышесказанного можно с помощью псевдокода составить следующий алгоритм функции поиска кандидатов:

поиск_кандидатов
Вход: M_{CE}, D.
Выход: M_K.
Начало
список M_K=[];*//пустое множество*
база_индексов *dbindx=создание_базы_индексов* (*D*);
Цикл для каждой $ce_i \in M_{CE}$ ***выполнять***
начало цикла
текстовые_метки m_i=***метка*** (ce_i);
//r – результат поиска – числовая оценка
оценка *r=поиск_метки* (m_i, *dbindx*);
Если *r*>0, *то* ***добавить*** ce_i ***к*** M_K;
конец цикла.
Конец.

Данный алгоритм создает базу индексов ***dbindx*** из аннотируемого документа с помощью функции ***создание_базы_индексов ()***. После этого для каждого экземпляра или понятия ce_i извлекаются метки m_i с помощью функции ***метка ()***. На следующем шаге выполняется поиск m_i в ***dbindx***, и если он будет найден (т. е. $r>0$), то понятие ce_i добавляется к списку кандидатов M_K.

При практической реализации для извлечения текста из аннотируемого документа требуется определить его формат (например, Microsoft Word, PDF или HTML). Также нужно указать естественный язык, на котором написан данный документ и используемые для его аннотирования метки (экземпляры, понятия и отношения). После этого создается база индексов, учитывающая грамматику используемого языка и использующая соответствующие ему токенизаторы.

Решение задачи преодоления многозначности. Существуют два подхода к автоматическому решению данной задачи: с помощью измерения семантической близости и с помощью измерения популярности [12].

Идея подхода измерения семантической близости для решения многозначности заключается в том, что для набора найденных кандидатов с использованием онтологий вычисляются их семантические близости с понятиями, подходящими для аннотирования документа. В результате этого система может выбирать тех кандидатов, близость которых с желаемыми понятиями больше некоторого порогового значения. В данном методе необходимо вычислять близость между понятиями и близость между понятиями и экземплярами. Один из возможных способов оценки такой близости описан в [3].

Допустим, что имеется конечный набор кандидатов M_K из онтологии O. Также имеется конечный набор желаемых понятий для аннотирования документа $M_C=\{c_i \in C \subset O\}$. Тогда релевантными будут считаться кандидаты из M_K, удовлетворяющие следующему условию:

$$\mathrm{Sem}(ce_i, M_C) = \max(\mathrm{Sem}_{c_j \in M_C}(ce_i, c_j)) > \varepsilon \forall \varepsilon > 0,$$

где Sem – семантическая близость и ε – установленное пороговое значение.

В результате объединения описанных выше задач получается следующий псевдокод полного алгоритма для функции поиска кандидатов для заданного документа:

поиск_кандидатов
Вход: M_{CE}, M_C, D.
Выход: M_K.
Начало
список M_K=[];*//пустое множество*
база_индексов *dbindx=создание_базы_индексов (D);*
Цикл для каждой $ce_i \in M_{CE}$ ***выполнять***
начало цикла
текстовые_метки m_i = ***метка*** (ce_i);
//r – результат поиска – числовая оценка
оценка *r=поиск_метки* (m_i, *dbindx*);
Если $r>0$ ***и*** *Sem* (ce_i,M_C)>***то добавить*** ce_i ***к*** M_K;
конец цикла.
Конец.

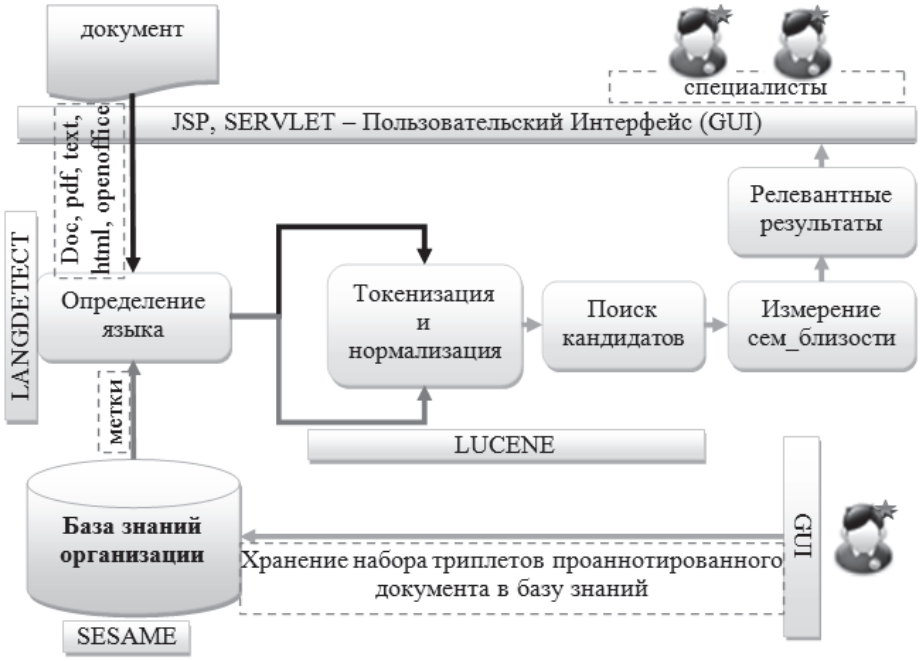

Рис. 2. *Структура системы – компонента полуавтоматического аннотирования*

Реализация системы полуавтоматического аннотирования

На рис. 2 показана обобщенная структура разработанной программной системы – полуавтоматического инструмента семантического аннотирования, а также другие использованные компоненты. Разработанная программа создана на основе технологии JSP и Servlet**.** Она может использоваться либо самостоятельно, либо как HTML-тег на странице JSP для аннотирования информационных объектов ЭБ.

В качестве системы управления базой знаний использовалась система **SESAME** [13], представляющая собой веб-сервис с открытым исходным кодом (на языке Java) для хранения триплетов описания всех понятий и их экземпляров онтологий организации.

Для определения языка, используемого в документе, и меток каждого понятия или экземпляра используется пакет **LangDetect** [14]. Данный пакет является библиотекой с открытым исходным кодом (на языке Java), которая позволяет идентифицировать большое количество естественных языков, в том числе и русский язык. Результат работы данного пакета используется для вызова конкретных компонентов анализа текста.

Самой важной частью данной системы является **Lucene** [15] – большая библиотека (на языке Java) полнотекстового поиска. Она позволяет выполнить индексацию документов различных форматов (с помощью специальных модулей). Одной из особенностей библиотеки **Lucene** является использование стеммера для нормализации токенов, написанных на русском языке, а также возможность хранения текста без стемминга.

Для поиска кандидатов и решения задачи многозначности выполняются следующие действия:

1. Специалист вначале загружает документ или набор документов, после чего система определяет формат документа для вызова соответствующего пакета с целью извлечения текста документа.
2. На основе извлеченного текста документа система определяет используемый в нем язык для вызова правильного анализатора текста из библиотеки **Lucene**.
3. Текст документа передается в модули **Lucene** для токенизации и нормализации. В результате этого из текста документа создается инвертированный файл индексов, хранящихся в оперативной памяти для повышения быстродействия алгоритма и используемых в качестве базы индексов для дальнейшего поиска.
4. Метки понятий онтологии и их экземпляров обрабатываются так же, как было описано выше (без создания базы индексов). После этого выполняется их поиск в базе индексов документа, созданной на шаге 3. Понятия и экземпляры, метки которых находятся в базе индексов, заносятся в список кандидатов для последующей обработки.
5. В результате поиска кандидатов получается список экземпляров и понятий, для которых выполняется оценка семантической близости со списком желаемых понятий, с целью исключения нерелевантных.
6. Релевантные кандидаты показываются специалисту для выполнения ручного аннотирования.

После этого специалист имеет возможность составлять триплеты, включающие найденных кан-

Рис. 3. *Интерфейс системы полуавтоматического аннотирования*

дидатов, для онтологического аннотирования загруженного документа.

Тестирование разработанного метода. Созданная система полуавтоматического аннотирования используется в составе разрабатываемой семантической электронной библиотеки (SemDL) для аннотирования электронных ресурсов, рубрик и профиля пользователя. Интерфейс данной системы показан в рис. 3.

Для проверки качества разработанного метода и программной системы было выполнено семантическое аннотирование данной статьи. В базе знаний хранятся экземпляры созданных понятий (Author и Concept). Система реализует поиск меток этих объектов в загруженном документе и выдает список рекомендуемых объектов, как показано на рис. 3.

С помощью данного инструмента возможно аннотировать документ набором триплетов с использованием значений (чисел, строк и дат, в качестве значения объекта триплета). Кандидаты экземпляров и понятий выводятся в отдельной области окна программы. Специалист может выбрать их для формирования триплетов.

Все созданные триплеты показываются в специальном окне программы, и специалист может выполнять их редактирование: удаление, изменение субъектов или объектов.

Выводы

В статье предложен новый подход к составлению семантических аннотаций электронных документов в виде набора триплетов и описаны разработанные алгоритмы для решения задачи поиска кандидатов и преодоления многозадачности при аннотировании. Для выполнения полуавтоматического аннотирования была разработана программная система, которая активно используется в разработке семантической электронной библиотеки. Использование данного подхода позволяет понизить трудоемкость составления семантических аннотаций и повысить их качество. Создание семантических аннотаций позволяет разрабатывать новые эффективные методы решения таких задач, как семантический поиск, автоматическая категоризация ресурсов и формирование рекомендаций.

СПИСОК ЛИТЕРАТУРЫ

1. Oren E., Hinnerk Möller K., Scerri S., Handschuh S., Sintek M. What are Semantic Annotations? URL: http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.97.7985&rep=rep1&type=pdf (дата обращения: 21.10.2012).
2. Domingue J., Fensel D., Hendler J.A. Handbook of Semantic web Technologies. – Heidelberg; Dordrecht; London; N.Y.: Springer, 2011. – 1077 p.
3. Saša Nešic, Mehdi Jazayeri, Fabio Crestani, Dragan Gaševic. Concept-Based Semantic Annotation, Indexing and Retrieval of Office-Like Document Units. 2010. URL: http://www.old.inf.usi.ch/file/pub/56/tech_rep.pdf (дата обращения: 21.10.2012).
4. Черний А.В., Тузовский А.Ф. Развитие информационной системы организации с использованием семантических технологий // Знания–Онтологии–Теория: Матер. Всеросс. конф. с междунар. участием. – Новосибирск, 20–22 октября 2009. – Новосибирск: ЗАО «РИЦ Прайс-Курьер», 2009. – Т. 2. – С. 52–59.
5. Тузовский А.Ф. Формирование семантических метаданных для объектов системы управления знаниями // Известия Томского политехнического университета. – 2007. – Т. 310. – № 3. – С. 108–112.
6. Handschuh S., Staab S. Authoring and Annotation of Web Pages in CREAM. 2002. URL: http://www2002.org/CDROM/refereed/506/ (дата обращения: 21.10.2012).
7. SMORE – Create OWL Markup for HTML Web Pages. 2005. URL: http://www.mindswap.org/2005/SMORE/ (дата обращения: 21.10.2012).
8. SWOOP – A Hypermedia-based Featherweight OWL Ontology Editor. 2004. URL: http://www.mindswap.org/2004/SWOOP/ (дата обращения: 10.21.2012).
9. Zemanta. URL: http://www.zemanta.com/ (дата обращения: 21.10.2012).
10. Ле Х.Х. Разработка электронных библиотек на основе семантических технологий // Научно-технический вестник Поволжья. – 2012. – № 3. – С. 138–145.
11. Manning C.D., Raghavan P., Shutze H. An Introduction to information retrieval. – Cambridge: Cambridge University Press, 2009. – 569 p.
12. Navigli R. Word sense disambiguation: a survey // ACM computing surveys. – 2009. – V. 41. – № 2. – P. 1–69.
13. Home of Sesame. URL: http://www.openrdf.org/index.jsp (дата обращения: 21.10.2012).
14. Language-detection. URL: http://code.google.com/p/language-detection/ (дата обращения: 21.10.2012).
15. Paul Th. The Lucene Search Engine. 2004. URL: http://www.javaranch.com/journal/2004/04/Lucene.html (дата обращения: 21.10.2012).

Поступила 16.01.2013 г.

УДК 004.942, 004.652.5

ФОРМАЛЬНАЯ МОДЕЛЬ СТРУКТУРЫ ВЗАИМОСВЯЗЕЙ РАЗНОТИПНЫХ ОБЪЕКТОВ ПРОЕКТИРОВАНИЯ

А.А. Вичугова, В.Н. Вичугов, Г.П. Цапко

Томский политехнический университет
E-mail: anya@aics.ru

В рамках развития существующих теоретических и практических положений технологий информационной поддержки жизненного цикла изделий поставлена задача разработки метода, позволяющего структурировать сущности, создаваемые при проектировании высокотехнологичной продукции на примере радиоэлектроники. Составлены теоретико-множественные модели, описывающие состав и взаимозависимости разнотипных объектов проектирования: изделие, его информационные модели, электронная структура и конструкторская документация. С использованием объектно-ориентированного подхода выполнено концептуальное проектирование базы данных: определен набор атрибутов, характеризующих стадии жизненного цикла рассматриваемых объектов проектирования, и предложена формальная информационная модель структуры их взаимосвязей в виде UML-диаграммы классов.

Ключевые слова:
Структурирование разнотипных взаимосвязанных объектов, объектно-ориентированный подход, теоретико-множественные модели.

Key words:
Structuring different related objects, object-oriented approach, set-theoretic models.

Введение

Проектирование высокотехнологичной продукции с длительным сроком активного существования (глубоководные и космические аппараты, искусственные человеческие органы и т. д.) представляет собой целый комплекс процессов со сложной структурой взаимосвязей и временной длительностью. Эта деятельность сопровождается большим количеством информации в виде информационных моделей изделия (ИМИ), конструкторской документации (КД) и электронной структуры изделия (ЭСИ). Готовность изделия к производству определяется состоянием КД на данное изделие. Под качеством КД понимается отсутствие ошибок в описании характеристик изделия по нормативно-техническим стандартам. Первоисточником данных для КД является совокупность элементов ЭСИ, формируемой на основе файлов