УДК 004.8

Н. В. Рябова, канд. техн. наук,
А. А. Козополянская, аспирантка,
О. В. Шубкина, ассистент,
С. А. Гринев, младший научный сотрудник

Харьковский национальный университет радиоэлектроники,
г. Харьков

МОДЕЛЬ СЕМАНТИЧЕСКОГО РЕПОЗИТОРИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ ДЛЯ ОНТОЛОГИЧЕСКОГО ПОРТАЛА МОНУ

Данная работа является составной частью комплекса прикладных исследований, направленных на разработку онтологического портала менеджмента и оценки национальных ресурсов Украины в области образования и науки путем создания крупномасштабной онтологии национальных образовательных ресурсов и разработки онтологической Web-ориентированной распределенной информационной системы для взаимодействия с этой онтологией. Рассматривается проблема расширения функциональности онтологического портала МОНУ путем построения семантического репозитория текстовых документов (РТД), по сути информационного хранилища (Document Warehouse), которое должно учитывать интересы пользователей и их роли в организации.

Ключевые слова: *онтология, распределенная информационная система, семантическая аннотация, семантический репозиторий, текстовый документ.*

Введение. В современных Web-системах значительная часть информации, которую необходимо хранить и обрабатывать, представлена в виде текстовых документов различного типа. К таким системам можно отнести различные корпоративные системы, предусматривающие режим функционирования в Интернет-пространстве, информационные порталы и порталы знаний. В последнем случае ядром системы является онтологическая база знаний, в которой в эксплицитном виде представлена концептуальная модель предметной области (ПрО), основные понятия (концепты), и термины, релевантные ПрО, а также система отношений между ними. Корпоративный портал должен обеспечивать единый способ доступа к внутренней информации организации, позволяющий сотрудникам устанавливать взаимодействие друг с другом, связывать информацию с коллективным пониманием, системой ценностей и опытом. Корпоративный портал знаний можно рассматривать как интегрированный информационный репозиторий, доступный для оперативного обобщения и

анализа. Таким образом, особую актуальность при разработке порталов знаний приобретают методы и модели построения информационных репозиториев текстовых документов [1].

Анализ основных подходов к построению семантических репозиториев. Одним из наиболее часто используемых программных средств, которые могут выступать в качестве семантических репозиториев, является SESAME [2]. Это открытая среда и база данных для работы с RDF-графами. Основное понятие в структуре сервера SESAME — это репозиторий, который является контейнером для хранения RDF. Это могут быть простые объекты Java или реляционная база данных. Каждое действие, которое происходит на сервере SESAME проходит через репозиторий, несмотря на то, что каждая единица информации хранится отдельно.

Также SESAME поддерживает вывод в RDF Schema. Это означает, что при наличии RDF и/или RDF Schema на сервере, встроенные механизмы SESAME могут находить неявную информацию в данных. Также неявная информация добавляется в репозиторий при включении новых данных в систему. Следует отметить, что функция вывода в SESAME напрямую зависит от структуры репозитория.

SAIL (Storage And Inference Layer) API — это встроенный программный интерфейс SESAME, который обеспечиваю поддержку автоматического вывода. Выполнение SAIL обеспечивает поддержку кэширования и параллельного доступа к ресурсам репозитория. Каждый репозиторий SESAME имеет свой набор SAIL API, который его представляет.

Основной частью SAIL API являются модули SESAME SeRQL, RQL и RDQL машины вывода, административный модуль и RDF экспорт. Доступ к этим модулям обеспечивается через программные приложения SESAME, которые состоят из двух частей: API репозитория и API графа. API репозитория – это центральная точка доступа к репозиторию сервера, которая обеспечивает доступ высокого уровня — обработку запросов, хранение RDF-файлов, изъятие RDF, модернизацию локальных и отдаленных репозиториев. API репозитория общаются со всеми компонентами модели «клиент-сервер» как с локальными репозиториями. API графа поддерживают менее сложные процедуры обработки репозитория — добавление или удаление индивидуальных утверждений, создание небольших RDF-моделей непосредственно из кода. Также API графа поддерживает представление RDF-графа в виде Java-объектов. Эти программные приложения часто используются вместе. Также эти программные интерфейсы обеспечивают прямой доступ к функциональным модулям SESAME, каждой программе-клиенту и к компонентам следующего уровня архитектуры сервера. Эти компоненты обеспечивают HTML-подобный доступ к программным интерфейсам приложений сервера.

Boca — это многопользовательский RDF-репозиторий, который создан группой IBM Adtech в Кембридже [3]. Программный Boca разработан для реализации возможности создавать RDF-приложения, которые должны реализовывать функции, отсутствующие в других программных пакетах.

В центре системы Boca находится сервер, который может сохранять неограниченное количество RDF-триплетов в DB2 базе данных. Клиенты системы могут осуществлять запросы к репозиторию Boca и изменять/обновлять RDF-графы с помощью использования клиентского стека системы Boca. Клиентский стек обеспечивает широкие возможности по обработке данных системы, в том числе и роботе со структурой RDF-графа на стороне клиента, осуществляя управление всеми процедурами отдаленного сервера или выполнять эти процедуры на локальном клиенте с восстановленным на нем RDF-графе.

Клиентский стек Boca поддерживает совместимость с программным интерфейсом Jena, который является наиболее популярным программным интерфейсом для обработки RDF-структур. Также поддерживается язык запросов к RDF-структуры — SPARQL.

AllegroGraph RDFStore — современная, постоянная, скоростная база данных RDF-графов [4]. Система поддерживает дисковые хранилища, что дает возможность не ограничивать объем хранимой информации. Также AllegroGraph поддерживает вывод с помощью SPARQL, RDFS++ и Prolog, которые реализованы с помощью приложений Java. AllegroGraph обеспечивает широкие возможности относительно создания запросов и получению доступа к RDF-репозиторию. Также AllegroGraph поддерживает геопространственные и временные рассуждения, анализ социальных сетей, примитивные типы данных и эффективность запросов в пределах диапазона, индексацию любых текстов, именование графов для указания их веса, истинности и происхождения, кластеризацию и федерализацию RDF репозитория.

Онтологических подход к построению семантических репозиториев текстовых документов. Для повышения эффективности работы с репозиторием текстовых документов (РТД) в данной работе предлагается использовать онтологический подход, предполагающий выделение основной смысловой составляющей каждого текстового документа (Text Document — TD) в виде семантической аннотации (Semantic Annotation — SA). В настоящее время в системах интеллектуального анализа текстовой информации (Text Mining) широкое развитие получила разработка систем и методов семантического аннотирования текстовых документов, основная идея которого заключается в создании описания TD в машинно-понятной форме на основе онтологии предметной области (ПрО) для последующего использования интеллектуальными агентами [5].

Каждая SA при этом может быть представлена в виде RDF-файла, содержащего помимо стандартных тегов, характеризующих документ (автор, название, тип документа, год, источник документа, и т.п.), также и семантические теги, содержащие основные концепты и термины, выделенные предварительно из TD (например, методами Text Mining). Существует набор стандартных решений, которые разработаны для описания метаданных и формирования SA, как например, стандарт Dublin Core, проекты FOAF, SKOS. Однако набор заданных тегов для описания ТД не отражает информацию, которая является актуальной для текущей онтологии ПрО, а зачастую несет лишь общие сведения.

Семантическое аннотирование на основе онтологического подхода — один из многообещающих способов извлечения знаний из репозиториев текстовых документов, позволяющий избежать неоднозначности (неопределенности) информационного поиска (к примеру, страна или река Нигер), а также повысить возможность взаимодействия и интеграции информации из гетерогенных источников, накопленных организациями в процессе становления развития [6].

Большинство существующих методов для создания SA реализуют создание такого рода аннотаций на основе использования лингвистических шаблонов в качестве описания возможных конструкций предложений (Hearst patterns), а также путем формирования шаблонов на основе выделения именованных сущностей или разбора предложений по частям речи для последующего сопоставления с полученным образцом.

В предлагаемом подходе каждая SA может быть представлена в виде RDF-графа, что в свою очередь предоставляет возможность использовать онтологический подход как для концептуального индексирования самих семантических аннотаций, так и для их сравнения с целью поиска схожих документов или документов из одного смыслового поля. Таким образом, при решении комплекса задач, связанных с обработкой текстовых документов, появляется возможность работы с их семантическими аннотациями. Более подробно вопросы предварительной обработки TD, формирования SA и работы с ними рассмотрены также в работах [7; 8].

Примерами использования SA документов являются: построение профилей преподавателей, с возможностью просмотра, например, основных публикаций, направлений научных исследований; поиск экспертов в заданной ПрО; выявление основных научных приоритетов отдельных ВУЗов, кафедр, преподавателей.

Методы построения семантических репозиториев текстовых документов. Обобщенная архитектура семантического репозитория текстовых документов на основе предлагаемого подхода может быть представлена следующим образом (рис. 1):

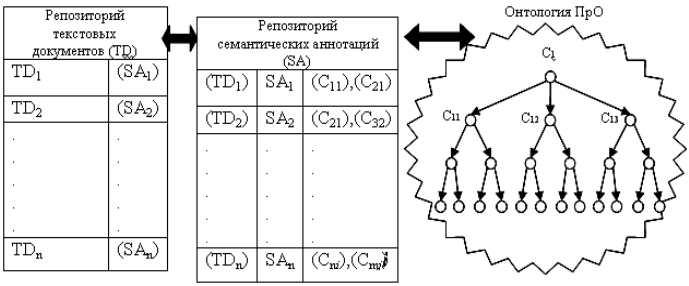

Рис. 1. *Обобщенная архитектура семантического репозитория TD*

Реализация семантического компонента репозитория осуществляется с помощью онтологии ПрО, к концептам которой «привязываются» термины из семантических тегов SA. Предполагается, что онтология ПрО, для которой создается РТД уже существует, по крайней мере ее основной прототип. Классы онтологии C_i соответствуют основным концептам ПрО. Предполагаем также, что онтология допускает графовое представление. Формирование самого репозитория TD возможно с использованием двух методов: организация централизованного хранилища документов и распределенное хранение TD. Отметим, что не исключены ситуации целесообразности гибридного метода создания семантического РТД, когда основные, наиболее важные или часто используемые документы, организуются в централизованное хранилище, а остальные — в распределенное хранилище.

В более простом случае, все документы физически хранятся в одном репозитории. При этом TD_i представляет собой «тело» документа, снабженное ссылкой на его семантическую аннотацию (SA_i), которая физически хранится в специально организованном репозитории семантических аннотаций. Тогда все документы TD связаны со своими семантическими аннотациями SA таким образом, что между ними существует взаимно однозначное соответствие. Каждая семантическая аннотация из репозитория SA имеет два типа ссылок: на местонахождение источника аннотации, т.е. самого документа (TD_i), а также ссылку на конкретные онтологические концепты (C_{ni}), (C_{mj}) и т.п. При этом связь конкретной SA с концептами (классами) онтологии осуществляется с помощью индексирования терминов в семантических тегах онтологическими концептами, которые предварительно систематизированы и пронумерованы.

Предложенные методы построения семантических репозиториев для текстовых документов позволяют осуществлять смысловой поиск различных типов документов по заданным семантическим критериям, смысловую обработку документов, хранящихся в репозитории, находить схожие документы и т.п.

Модель семантического репозитория текстовых документов для онтологического портала МОНУ. Архитектура портала регистра-

ции образовательных ресурсов МОН Украины основывается на принципах разработки распределенных многопользовательских систем, ориентированных на работу в Интернет. Портал представляет собой специализированную информационную систему, снабженную эргономичным Web-интерфейсом пользователя. Информационную основу портала составляет онтология и связанные с ней описания соответствующих ресурсов [9]. В случае, когда объемы текстовых документов весьма значительны и/или типология их обширна, имеет смысл использовать метод распределенного хранения отдельных типов документов (рис. 2). При этом документы *i*-ого типа (например, статьи в научных журналах) могут храниться в *i*-ом репозитории R^i. Каждый документ из R^i снабжен ссылкой на свою семантическую аннотацию $(SA)^i$, хранящуюся, как и в предыдущем случае, в репозитории SA. Вместо централизованного репозитория предлагается ввести в архитектуру системы R-медиатор, который является связующим звеном между локальными репозиториями отдельных типов документов и репозиторием семантических аннотаций. R-медиатор содержит не сами документы, а ссылки на адреса их физического хранения и/или локальные репозитории отдельных типов документов.

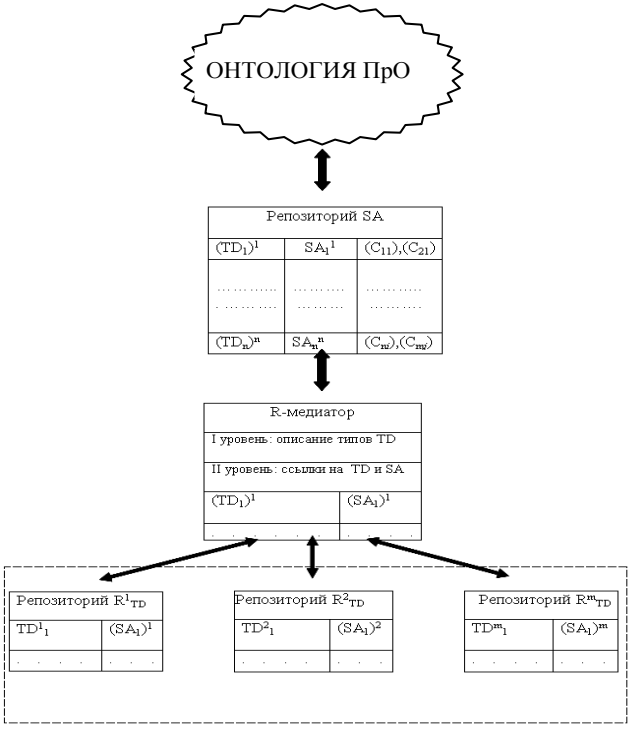

Рис. 2. *Распределенная архитектура семантического репозитория TD*

Выше были приведены различные архитектуры семантических репозиториев текстовых документов, в основе которых лежат семантические аннотации и онтология ресурсов предметной области (рис. 3).

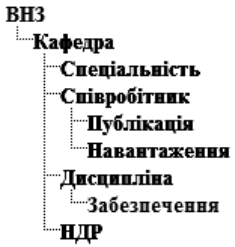

Рис. 3. *Дерево ресурсов онтологического портала*

Одними из основных составляющих семантического репозитория являются семантические аннотации, которые представляют собой метаданные текстовых документов, полученные путем обработки корпуса текстов и первоначальной онтологии методами интеллектуального анализа данных.

В результате каждый экземпляр концепта, например, «Преподаватель ВУЗа» в онтологии научно-образовательных ресурсов получает связь с публикациями, характеризующими области научных интересов сотрудников. Таким образом, формируется единое семантическое пространство, в котором возможно осуществлять мониторинг отдельных научных направлений и осуществлять семантический поиск экспертов.

Выводы. В данной работе рассматривается проблема расширения функциональности онтологического портала МОНУ путем построения семантического репозитория текстовых документов. Введены различные методы построения семантических репозиториев текстовых документов, а именно, централизованный, распределенный и гибридный, которые позволят осуществлять смысловой поиск различных типов документов по заданным семантическим критериям, смысловую обработку документов, хранящихся в репозитории, находить схожие документы. Предложена модель семантического репозитория текстовых документов для онтологического портала МОНУ, основанная на взаимодействии хранилищ текстовых документов, семантических аннотаций и онтологий предметных областей.

Также в данной работе рассматривается подход к представлению текстовых документов в машинно-понятной форме путем создания их семантических аннотаций на основе онтологии, отражающей знания о предметной области, что позволит избежать неоднозначности информационного поиска и повысит возможность взаимодействия и интеграции информации из различных источников.

Список использованной литературы:

1. Sullivan D. Document Warehousing and Text Mining. Techniques for Improving Business Operations, Marketing, and Sales. — Canada : John Wiley & Sons, Inc., 2001. — 542 p.
2. Официальный сайт проекта Sesame. — Режим доступу: http://www.openrdf.org.
3. Официальный сайт IBM Semantic Layered Research Platform. — Режим доступу: http://ibm-slrp.sourceforge.net.
4. Официальный сайт разработчика AllegroGraph RDFStore. — Режим доступу: http://www.franz.com/.
5. Рябова Н. В. Механизм формирования семантического описания текстовых документов / Н. В. Рябова, О. В. Шубкина // Материалы 9-й междунар. науч.-техн. конф. : «Проблемы информатики и моделирования». — Харьков, 2009. — С. 47—48.
6. Шубкина О. В. Методы разметки последовательностей для создания семантических аннотаций информационных ресурсов / О. В. Шубкина // Материалы междунар. науч.-практ. конф. : «Информационные технологии и информационная безопасность в науке, технике и образовании «ИНФОТЕХ-2009». — Севастополь : Изд-во СЕВНТУ, 2009. — С. 197—200.
7. Рябова Н. Застосування онтологічної семантики у зіставленні документів / Н. Рябова, Г. Козополянська, К. Дяденко // Матеріали III Міжн. конф. молодих вчених „Комп'ютерні науки та інженерія" CSE-2009, 14-16 травня, 2009, Львів. — Львів : НУ „Львівська політехніка", 2009. — С. 107—108.
8. Рябова Н. В. Методы и модели интеллектуальной обработки текстов в задачах онтологического инжиниринга / Н. В. Рябова // Математическое и программное обеспечение интеллектуальных систем (MPZIS-2008). Тез. докл. YI междунар. науч.-практ. конф., Днепропетровск, 12-14 нояб. 2008. — Днепропетровск, 2008. — С. 269—270.
9. Розробка і впровадження розподіленої архітектури онтологічного порталу МОНУ для надійного, безпечного та ефективного менеджменту та інтеграції освітніх ресурсів України: звіт про НДДКР № IT/543-2009 (пром.) / ХНУРЕ ; кер. В.Я. Терзіян. — Харків, 2009. — 97 с.

This work is part of the complex applied researches aimed to developing ontology-based portal management and evaluation of Ukrainian national resources in the field of education and science. This is achieved through the creation of large-scale ontology of national education-state resources and the development of ontological Web-oriented distributed information system to interact with this ontology. The problem was considered to extend the functionality of the ontological MESU portal by building a repository of semantic text documents (RTD), in fact an information storage (Document Warehouse), which should take into account the users' interests and their roles in the organization.

Key words: *ontology, distributed information system, semantic annotation, semantic repository, text document.*

Отримано: 24.04.2011