OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ИССЛЕДОВАТЕЛЬСКИЙ ПОРТАЛ ПО МОДЕЛИРОВАНИЮ ИНФОРМАЦИОННЫХ СИСТЕМ

Ланин В.В., Лядова Л.Н.

Пермский филиал федерального государственного автономного образовательного учреждения высшего профессионального образования "Национальный исследовательский университет "Высшая школа экономики", г. Пермь, Россия

lanin@perm.ru

lnlyadova@mail.ru

Статья посвящена описанию проекта разработки портала, ориентированного на поддержку работы исследователей (преподавателей, аспирантов, студентов), занимающихся вопросами моделирования информационных систем в различных предметных областях, создания и использования инструментальных средств разработки, основанных на (мета)моделировании (применении DSM, DSL и др.). В данной работе представлено описание архитектуры портала, средств информационного поиска и управления документами, создания единой системы документов, относящихся к данной области исследований.
Ключевые слова: портал; онтологии, моделирование информационных систем.

Введение

В настоящее время не существует единого информационного ресурса, ориентированного на поддержку работы исследователей, занимающихся вопросами моделирования информационных систем в различных предметных областях, хотя проекты по данному направлению выполняются в Санкт-Петербургском государственном университете (под руководством профессора Терехова А.Н.), Южном федеральном университете (руководитель – профессор Рогозов Ю.И.) и др. Информация об исследованиях размещается на сайтах отдельных организаций (кафедр университетов и пр.), фирм-разработчиков, имеются порталы, посвященные отдельным проектам, проводимым по данной тематике конференциям (http://www.metacase.com/, http://www.eclipse.org/, http://www.omg.org/, http://www.dsm-conference.org/ и др.). Поиск нужной информации (особенно для молодых исследователей) затруднён: с одной стороны, имеется множество ресурсов, с другой – многие из них оказываются бесполезными, не отвечающими поисковым запросам пользователей. Например, при поиске по слову «метамоделирование» этот термин появляется в сочетании с такими понятиями как «нейролингвистическое программирование», «психология», «гипноз» и т.д.

Цель создания портала в НИУ ВШЭ – создание «саморазвивающегося» ресурса, предоставляющего в распоряжение пользователей средства интеллектуального поиска и автоматизированной обработки полученных результатов (документов, источников), удобные средства навигации по найденным ресурсам. Реализация основана на использовании онтологий, описывающих как предметную область, так и обрабатываемые документы.

Пользователи портала должны получить возможность настройки этих средств в соответствии со своими потребностями, а также возможность оперативного взаимодействия, обмена информацией и публикации результатов своих исследований для обсуждения, коллективной работы над проектами.

Для работы с моделями должен быть разработан специализированный язык (DSL) с использованием созданных при выполнении проекта средств. При реализации проекта по созданию портала решаются задачи, связанные с организацией коллективной работы, поиском, сбором и анализом материалов и их публикацией.

Задачи подобного характера решались и ранее другими исследователями. Особенно хотелось бы отметить работы [Загорулько, 2004], [Загорулько, 2008], [Мальцева, 2008]. Новизна представленной работы заключается в комплексном подходе к разработке портала, интегрирующем возможности информационных технологий и систем различного назначения на основе знаний о предметной области системы.

При работе с порталом пользователи получают эффективные интеллектуальные средства поиска информации на основе семантической индексации, автоматической классификации и каталогизации найденных документов с построением семантических связей между ними и автоматического реферирования документов с использованием знаний. Эффективность работы с электронными документами предполагается значительно увеличить за счет их интеллектуального анализа, для которого применяются агентный и онтологический подходы [Ланин, 2009а].

1. Концепция портала

Ключевой идеей функционирования разрабатываемого портала является адаптируемость к потребностям пользователя. При создании портала необходимо учитывать не только потребности пользователей, но и особенности представления знаний в Интернет. Портал должен не только обеспечить более быстрый доступ к информационным ресурсам, но и предоставить в распоряжение пользователей дополнительные возможности по организации научных исследований и совместной работы над проектами. Таким образом, необходимо реализовать инструментарий, обеспечивающий автоматизацию трудоемких операций по поиску и анализу данных, разработке моделей инновационного развития и их апробации и пр. Многомерная классификация и удобная каталогизация ресурсов, наличие средств навигации, настраиваемых в соответствии с запросами пользователей, – еще одно требование к порталу.

В качестве основы исследовательского портала должно быть создано информационное ядро дисциплины, которую он представляет. На основе этого ядра, предоставляющего исследователям базовую информацию и набор сервисов, должны развиваться информационные ресурсы, создаваемые пользователями портала. Один из фундаментальных принципов – открытость в сочетании с защищенностью. Портал должен быть построен как «самоподдерживающийся» ресурс: развитие портала, расширение его ядра и наполнение новыми ресурсами, как и управление, ложится на самих пользователей.

При создании портала учитываются возможности современных информационных технологий, полезные с точки зрения решаемых задач. Перспективными тенденциями развития Интернет считаются концепция Web 2.0 и технология Semantic Web [O'Reilly, 2005], [Berners-Lee, 2001]. Вытесняющий традиционные Web-сервисы Web 2.0 не является технологией или специальным стилем Web-дизайна – его следует рассматривать как комплексный подход к организации, реализации и поддержке Web-ресурсов. Наиболее характерными технологиями, реализующими подход Web 2.0, являются wiki-ресурсы, блоги, технология FOAF, технологии синдикации новостей и пр.

Блоги (сетевые дневники) представляют собой один из самых ярких примеров использования принципов Web 2.0. Значительная часть Web-контента создается пользователями, а не владельцами ресурса. Для этого активно используют технологии RSS и FOAF, характерные для Web 2.0. Используются также тэги (tags, метки) для тематического структурирования контента. Технология FOAF (Friend Of A Friend) является одной из важнейших составляющих социальных Интернет-сетей. Пользователю предоставляется возможность подписаться на новости и материалы тех пользователей, которые находятся в так называемом «списке друзей». Этим самым поощряется общение пользователей Сети. Технология RSS (Really Simple Syndication) – это простая и эффективная технология экспорта гипертекста, используемая для создания новостных лент. Эта технология, как и другие технологии Web 2.0, основана на использовании языка разметки XML.

Все эти возможности используются при создании портала для организации эффективной работы пользователей, их оперативного взаимодействия в ходе исследований.

2. Архитектура портала

Архитектура портала основана на создании многоуровневых моделей, управляющих работой системы. Фактически программное обеспечение работает в режиме интерпретации этих моделей, что обеспечивает максимальные возможности адаптации системы через внесение изменений в модели в процессе эксплуатации системы [Лядова, 2008]. Упрощенная структура моделей портала показана на рис. 1. Предметно-зависимые модели – это модели онтологического уровня.

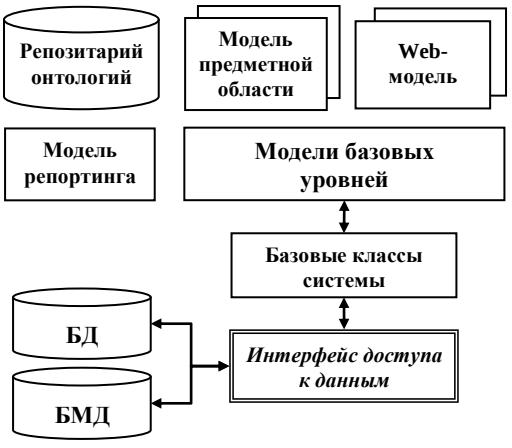

Рисунок 1 – Общая схема работы поисковой системы

Модели базовых уровней позволяют описать «ядро» портала, базовые понятия его предметной области, а также сгенерировать и настроить интерфейс пользователя и сформировать базу

данных портала. Портал функционирует в режиме интерпретации этих моделей. Базовые модели являются расширяемыми, на их основе могут быть созданы новые предметно-ориентированные (онтологические) модели. Для их создания разрабатываются специальные предметно-ориентированные языки. Пользователи могут вносить изменения в модели в процессе функционирования системы, настраивая ее на свои потребности.

Программные компоненты системы работают с моделями своих уровней (языковой инструментарий, средства реструктуризации данных, средства интеллектуального поиска, анализа и каталогизации документов, средства репортинга, средства анализа данных и моделей, подсистема безопасности и пр.). Система допускает расширение функциональности через подключение новых программных компонентов.

3. Описание документа с помощью онтологии

Процесс поиска информации с помощью поисковой системы может быть описан следующим образом [Ланин, 2009b]. У пользователя возникает информационная потребность (необходимость найти сведения по какому-либо вопросу). Затем пользователь некоторым образом формализует свою информационную потребность в виде запроса (в традиционных системах это выделенное множество ключевых слов с зафиксированными отношениями между ними). На следующем этапе через интерфейс поисковой системы вводится запрос. Система на множестве документов, являющемся информационно-поисковым пространством, осуществляет выборку документов, которые по внесенным в систему критериям соответствуют запросу пользователя, и формирует результат (отклик). Найденные документы по своему содержанию делятся на две группы: документы, соответствующие информационной потребности пользователя (релевантные), и документы, не соответствующие его информационной потребности, но соответствующие запросу пользователя с точки зрения информационно-поисковой системы (информационный шум).

Учитывая специфику решаемой задачи, процесс поиска информации может быть улучшен по двум направлениям: релевантности результата и представлению отклика. Обе задачи предлагается решать с помощью онтологического подхода, завоевывающего все большую популярность.

Модифицированная схема поиска, реализуемая при создании портала, представлена на рис. 2.

Основная особенность предлагаемого подхода – использование репозитария онтологий на этапах преобразования запроса и документа. Откликом является структурированный документ, т.е. документ, в котором выделены понятия онтологий.

Рисунок 2 – Общая схема работы поисковой системы

О структуре репозитария онтологий рассказано в следующем разделе.

4. Подсистема управления документами портала

Методы искусственного интеллекта, как правило, используются для решения трудно формализуемых задач, постановка которых проста и понятна для человека, но при разработке алгоритмов их решения возникают трудности. Одна из таких задач – работа с документами в информационных системах: их поиск и каталогизация, анализ и извлечение информации.

В настоящее время существуют различные подходы, модели и языки, ориентированные на интегрированное описание данных и знаний. Наиболее перспективным и универсальным, по мнению авторов, представляется онтологический подход.

Согласно общепринятому определению, под онтологией (в широком смысле) понимается база знаний специального типа, которая может «читаться» и пониматься, отчуждаться от разработчика и/или физически разделяться ее пользователями. Учитывая специфику решаемых в данной работе задач, можно конкретизировать понятие онтологии: онтология – это спецификация некоторой предметной области, которая включает в себя словарь терминов (понятий) предметной области и множество связей между ними, которые описывают, как эти термины соотносятся между собой [Ланин, 2009b].

Для построения иерархии понятий онтологии используются следующие базовые типы отношений: “is_a” («класс – подкласс», гипонимия); “part_of” («часть – целое», меронимия); “synonym_of” (синонимия). Следует учесть, что данные типы отношений являются базовыми и не зависят от онтологии, но необходимо предоставить

пользователю возможность добавления новых отношений, которые бы учитывали специфику описываемой предметной области.

В представленном подходе выделяются три типа онтологий:

– онтология предметной области конкретной информационной системы (ИС);

– онтология как база знаний (БЗ) интеллектуального агента;

– онтология как описание документа.

Рассмотрим назначение каждого из перечисленных типов онтологий.

Онтологии предметной области имеют наиболее типичное применение, они используются для описания понятий предметной области ИС. Например, школьное образование, социальная помощь гражданам или инновационное развитие регионов. В онтологии этого типа описывается связь понятий, языковые единицы для их выражения, аксиомы предметной области. Онтология предметной области используется для семантического индексирования и анализа всех документов системы.

Для анализа документов используется мультиагентный подход. Интеллектуальные агенты, руководствуясь онтологией как базой знаний (второй тип онтологий), производят поиск и анализ конкретных понятий документа. Каждая из вершин такой онтологии имеет определенный прототип, интерпретация которого известна агенту. Таким образом, агент использует онтологию как определенную программу своих действий. Вершинами онтологии данного типа могут являться понятия из онтологии предметной области.

Третий тип онтологий используется для описания структуры и содержания документов. Этот тип онтологий включает в себя два класса (плоскости) вершин. К первому классу относятся вершины, описывающие структуру документа. Например: таблица, дата, должность и т.д. (они представляют собой общие понятия, не зависящие от конкретной предметной области). Другим типом будут являться вершины, содержащие понятия документа. Первый тип вершин будем называть структурные вершины, второй тип – семантические вершины. Благодаря такому подходу из документа можно получить требуемые данные: известно, где искать данные и как они могут быть интерпретированы.

Если представлять документ с использованием онтологий, то задача сопоставления онтологии и анализируемого документа сводится к задаче поиска понятий онтологии в документе. Как следствие, системе необходимо ответить на вопрос: описывает ли данная онтология документ или нет. На последний вопрос можно ответить утвердительно, если в процессе сопоставления в документе были найдены все понятия, включенные в онтологию. Таким образом, исходная задача сводится к задаче поиска в тексте документа общих понятий на основе формальных описаний. На основе онтологии может быть получен фрейм, слоты которого заполняются в процессе анализа документа. В качестве слотов фрейма выступают понятия онтологии, а значения этих фреймов заполняются данными анализируемого документа. Таким образом из найденного неструктурированного документа может быть получен структурированный документ-фрейм.

Онтологии располагаются на трех уровнях репозитария. На первом уровне расположены онтологии, описывающие объекты, используемые в конкретной системе и учитывающие ее особенности. На втором уровне описываются объекты, инвариантные к предметной области. Объекты третьего уровня описывают наиболее общие понятия и аксиомы, с помощью которых описываются объекты нижележащих уровней.

Помимо описанных выше онтологий в процессе работы портала используются дополнительные онтологии: онтология источников информации и онтология форматов электронных документов. На данный момент в онтологии источников детально представлены ресурсы сайта http://www.dsmforum.org. Также в онтологии представлены конференции и другие мероприятия, проводимые по тематике моделирования информационных систем, блоги разработчиков инструментальных средств MetaCase и Microsoft Visual Studio. Онтология ресурсов портала базируется на понятиях Дублинского ядра метаданных. Онтология форматов документов используется для унификации обработки документов в различных форматах. Онтологии описаны на языке OWL 2.0 с помощью редактора Protégé 4.2.

5. Реализация подсистемы управления документами портала

Основными обрабатываемым на портале объектами будут являться электронные документы, поэтому подсистема управления документами является крайне важным компонентом. При проектировании портала принято решение использовать одну из существующих ECM-систем.

Согласно последнему опубликованному отчету агентства Gartner [Weintraub, 2011] по ECM системам в знаменитый магический квадрант попала единственная Open Source из всех представленных система Alfresco. Gartner отметили инновационный подход [Gilbert, 2012] – именно поэтому Afresco располагается в правом нижнем углу. Также Alfresco Software попала и в отчет агентства Forrester. Согласно Forrester, Alfresco может составить реальную альтернативу таким крупным игрокам рынка ECM систем, как IBM, Oracle, Open Text и EMC [Weintraub, 2011].

Пользователи могут начать использовать Alfresco в их повседневной работе сразу после установки,

однако это не позволит использовать преимущества, предоставляемые данной системой – одним из основных является гибкая система управления контентом.

Модель содержимого в Alfresco включает в себя 5 основных составляющих перечисленных ниже [Potts, 2008].

Пользовательские типы (custom types) предназначены для создания таксономии содержимого репозитория. С их помощью пользователь определяет типы документов, которые будут содержаться во внутреннем хранилище Alfresco. При этом пользовательский тип может обозначать не только конкретный тип документа, но и обозначать некоторую абстрактную категорию документов (например, «финансовый документ»), что позволяет гибко настраивать модель содержимого под нужды конкретной организации. Кроме того, для пользовательских типов определен механизм наследования, с помощью которого конкретные типы документов могут наследовать свои свойства от более абстрактного.

Однако обычно для создания полноценной модели содержимого конкретной организации зачастую бывает недостаточно использования только одних пользовательских типов. Часто пользователю требуется сопоставить с каждым из пользовательских типов некоторый набор метаданных, качественно характеризующий данный конкретный тип. Для этого в Alfresco используются *свойства (properties)*. При определении нового свойства пользователь должен указать его имя и используемый тип данных. Работа со свойствами в Alfresco напоминает работу с полями классов в ООП. Однако в последнем случае пользователю необходимо программировать всю бизнес-логику самостоятельно, в то время как Alfresco позволяет декларативно описывать ограничения, накладываемые на значения свойств.

Alfresco предоставляет возможность задания следующих *ограничений (constraints)*: ограничение на минимальное и максимальное значение свойства, ограничение на длину значения свойства, ограничение, контролирующее количество и вид элементов в случае использования типа данных «список», ограничение, задаваемое с помощью регулярных выражений.

Помимо определения таксономии документов организации и набора их метаданных, зачастую бывает необходимо определять связи между различными типами документов. Для этого в Alfresco используются *ассоциации (association)*, которые можно разделить на два класса.

1) Простые ассоциации, с помощью которых пользователь указывает, что два конкретных пользовательских типа связаны между собой.

2) Ассоциации, определяющие «родственные» связи между двумя пользовательскими типами (child associations). Ассоциации данного типа налагают дополнительное ограничение, предполагающее, что связанный документ зависит от документа, с которым он связан. Т.е., например, при удалении родительского документа, дочерний также должен быть удален. В Alfresco есть встроенная ассоциации данного типа – Contains, которая используется для определения связи между пользовательским типов, выступающим в качестве контейнера, например, «Каталог», и его содержимым.

С помощью использования *аспектов*, пользователь может расширить набор свойств отдельных пользовательских типов. Зачастую возникает необходимость, сделать доступными извне только некоторые документы. Для этого необходимо определить набор дополнительных свойств, с помощью которых Alfresco могла бы определять, должен ли быть доступен данный документ извне, или нет. Использование аспектов напоминает использование интерфейсов в объектно-ориентированном программировании. С их помощью пользователь может добавить ряд дополнительных свойств к отдельным типам, не меняя иерархию в целом.

Microsoft SharePoint – инструмент для создания сайтов, предоставляющих пользователям возможность для совместной работы. Создаваемые на платформе SharePoint сайты могут быть использованы в качестве хранилища информации, знаний и документов, а также использоваться для исполнения облегчающих взаимодействие веб-приложений, таких как вики и блоги.

Модель данных SharePoint 2010 состоит из перечисленных ниже понятий [Perran, 2010].

Списки (columns) представляет собой набор сведений, предназначенный для совместного использования различными пользователем.

Библиотеки (libraries) предназначены для хранения контента организации. Каждая библиотека отображает список документов и некоторый набор сведений о них, которые помогают пользователям в работе. Пользователи могут управлять отображением документов, а также контролировать доступ к ним, ограничивая круг пользователей, которым разрешается просматривать документы до их утверждения. Кроме того, библиотеки являются средством контроля версий: с их помощью возможно одновременно работать с различными версиями одного и того же документа. Содержимое библиотеки, также как и содержимое списка, определяется ее типом.

Столбец (lists) представляет собой атрибут метаданных, используемый для описания отдельных элементов в списке или библиотеке.

Тип контента (content types) используются для описания свойств элементов списков и библиотек. С помощью типов контента пользователь может, например, определить таксономию документов в библиотеке документов. Поддерживаются следующие типы контента: метаданные или

свойства, пользовательские формы, рабочие процессы, шаблоны документов.

В результате сравнения принято решение при разработке портала использовать Alfresco. Немаловажным факторами при выборе системы стали открытость исходного кода и использование открытых стандартов. Кроме того, система Alfresco реализована на языке Java, что значительно упрощает интеграцию с компонентами обработки онтологий и инструментами Semantic Web, так же реализованными на данном языке. SharePoint в свою очередь ориентирован на корпоративный сегмент и решения Microsoft, обладает ограниченными возможностями интеграции семантических технологий.

Заключение

Применение описанных подходов при построении портала существенно снизит трудоемкость поиска необходимой информации, ее анализа и возможности использования в исследованиях. Полученная в результате анализа документов информация может использоваться исследователями для усовершенствования моделей предметной области, построенных ими. Таким образом, появляется основа для создания интеллектуальной системы с высокой степенью обратной связи. Ориентация на знания является базовым механизмом функционирования портала, что позволяет комплексно решать поставленные задачи.

Работа выполнена при поддержке Программы «Научный фонд НИУ ВШЭ» софинансирования грантов РФФИ и РГНФ (проект № 12-09-0102).

Библиографический список

[Загорулько, 2004] Загорулько Ю.А., Булгаков С.В. Использование онтологий для построения инновационных цепочек в системе поддержки инновационной деятельности в регионе // Труды VI-й международной конференции «Проблемы управления и моделирования в сложных системах». Самара: Самарский Научный Центр РАН, 2004. С. 328–333.

[Загорулько, 2008] Загорулько Ю.А. Автоматизация сбора онтологической информации об Интернет-ресурсах для портала научных знаний // Известия Томского политехнического университета / Томск: Томский политехнический университет, 2008. Т. 312. № 5. С. 114-119.

[Ланин, 2009a] Ланин В.В. Методы и средства решения задач информационного поиска для системы поддержки научных исследований // Инновационное развитие регионов: методы оценки и поддержка исследований: межвуз. сб. науч. статей / Перм. гос. ун т. – Пермь, 2009. С. 80-88.

[Ланин, 2009b] Ланин В.В. Решение задач информационного поиска для исследовательского портала на основе агентного и онтологического подходов // Инновационное развитие регионов: методы оценки и поддержка исследований: межвуз. сб. науч. статей / Перм. гос. ун т. – Пермь, 2009. С. 89-96.

[Лядова, 2008] Лядова Л.Н. Метамоделирование и многоуровневые метаданные как основа технологии создания адаптируемых информационных систем // Advanced Studies in Software and Knowledge Engineering / International Book Series “Information Science & Computing”, Number 4. Volume 2, 2008. Institute of Information Theories and Applications FOI ITHEA, Sofia, 2008. P. 125-132.

[Лядова, 2009] Лядова Л.Н. О подходе к построению исследовательского портала на основе метамоделирования // Инновационное развитие регионов: методы оценки и поддержка исследований: межвуз. сб. науч. статей / Перм. гос. ун т. – Пермь, 2009. С. 74-79.

[Мальцева, 2008] Мальцева С.В., Проценко Д.С. Серверы отношений сетевых сообществ практики на основе онтологических моделей // Автоматизация и современные технологии. №3, 2008. Научно-техническое издательство «Машиностроение». С. 26-29.

[Мальцева, 2008] Мальцева С.В. Применение онтологических моделей для решения задач идентификации и мониторинга предметных областей // Бизнес-информатика, №3(05), 2008. С. 18-24.

[Berners-Lee, 2001] Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. Vol. 284, No. 5, 2001. P. 35-43.

[Potts, 2008] Potts J. Alfresco Developer Book. Customizing Alfresco with actions, web scripts, web forms, workflows, and more. Packt publishing. 2008

[Gilbert, 2012] Gilbert M. R., Shegda K. M., Chin K., Tay G., Koehler-Kruener Hanns Gartner’s Magic Quadrant for Enterprise Content Management. 18 October 2012

[Weintraub, 2011] Weintraub A. The Forrester Wave™: Enterprise Content Management, Q4 2011 Alan Weintraub November 1, 2011

[Perran, 2010] Perran A., Perran S., Mason J., Rogers L.- Beginning SharePoint 2010 - Building Business Solutions with SharePoint - 2010

[O'Reilly, 2005] O'Reilly T. What Is Web 2.0 [Электронный ресурс] [http://oreilly.com/web2/archive/what-is-web-20.html].

RESEARCH PORTAL OF MODELING INFORMATION SYSTEMS

Lanin V.V., Lyadova L.N.

National Research University Higher School of Economics, City of Perm

lanin@perm.ru

lnlyadova@mail.ru

The paper describes the development of a portal about development and use of tools based on the (meta) modeling (using DSM, DSL, etc.). The architecture of a portal, information retrieval subsystem and document management are described.

The purpose of the portal is the creation of "self-developing" resource, which provides intelligent search and automatic processing of the results (documents and sources), easy navigation on the found resources. Implementation is based on the ontologies approach.

The main feature of suggested methods is an integrated approach to development. The approach bases on a multi-level ontology repository. The portal allows searching and analyzing information, creating and researching model, publishing research results. Software gives an opportunity of a flexible customizing. The main topic of this paper is an intelligent information search means based on semantic indexation, automatic document classification, tracking of semantic links between documents and automatic summarization.