В.В. Ланин[1]

Национальный исследовательский университет «Высшая школа экономики» (Пермский филиал)

lanin@perm.ru

ОРГАНИЗАЦИЯ ОБРАБОТКИ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА УЧЕБНО-ИССЛЕДОВАТЕЛЬСКОМ ПОРТАЛЕ С ИСПОЛЬЗОВАНИЕМ ОНТОЛОГИЙ

Введение

При реализации проекта создания портала «Моделирование сложных проблемно-ориентированных информационных систем» решаются задачи, связанные с организацией коллективной работы, поиском, сбором и анализом материалов и их публикацией.

Задачи подобного характера решались ранее другими исследователями. Особенно хотелось бы отметить работы [1, 2, 6, 7]. Новизна представленной работы заключается в комплексном подходе к разработке портала, интегрирующем возможности информационных технологий и систем различного назначения на основе знаний об информационных ресурсах, о предметной области системы и её пользователях. При разработке портала предполагается также использование технологий, основанных на предметно-ориентированном моделировании [5] и применении DSL (Domain Specific Languages).

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, также автоматического реферирования документов с использованием знаний. Эффективность работы пользователей с электронными документами предполагается значительно увеличить за счет их интеллектуального анализа, для которого применяются агентный и онтологический подходы [3].

[1] Работа выполнена при поддержке Программы «Научный фонд НИУ ВШЭ» (проект № 12-09-0102)

В данной работе представлено описание архитектуры поисковой системы учебно-исследовательского портала, средств поиска и управления документами, создания единой системы взаимосвязанных документов, относящихся к заявленной области исследований.

Существующие модели метаданных

Дублинское ядро (*Dublin core*) – набор элементов метаданных, предназначенный для описания контента документов различной природы (публикаций, аудиозаписей, видеозаписей и т.п.). Спецификация этого набора имеет статус официального международного стандарта (ISO:15836-2003). Текущая версия Дублинского ядра, в соответствии с указанными стандартами, включает 15 элементов метаданных. Развитие Дублинского ядра осуществляется под эгидой некоммерческой организации «Директорат Дублинского ядра». Семантика Дублинского ядра была разработана международной междисциплинарной группой специалистов библиотечного дела, компьютерных наук, кодирования текстов, музейного дела и других смежных групп.

Стандарт разделён на два уровня:

- *простой* (неквалифицированный, *simple*), состоящий из 15 элементов;
- *компетентный* (квалифицированный, *qualified*), состоящий из 18 элементов и группы так называемых тонкостей (или квалификаторов), которые уточняют семантику элементов для повышения релевантности поиска ресурсов.

Простой набор элементов метаданных Дублинского ядра (*Dublin Core Metadata Element Set*; DCMES) состоит из 15 элементов метаданных:

- Title – название;
- Creator – создатель;
- Subject – тема;
- Description – описание;
- Publisher – издатель;
- Contributor – внесший вклад;
- Date – дата;
- Type – тип;
- Format – формат документа;
- Identifier – идентификатор;
- Source – источник;
- Language – язык;
- Relation – отношения;

- Coverage – покрытие;
- Rights – авторские права.

Квалифицированный (компетентный) *набор элементов метаданных* Дублинского ядра, помимо 15 вышеперечисленных, может включать следующие элементы:

- Audience – аудитория (зрители);
- Provenance – происхождение;
- RightsHolder – правообладатель.

Каждый элемент опционален и может повторяться.

В документе «Инициатива метаданных Дублинского ядра» (Dublin Core Metadata Initiative; DCMI) описаны стандартные пути определения элементов и поощряется использование схем кодирования и словарей. Не существует заранее заданного порядка перечисления этих элементов. DCMI поддерживает также небольшой общий словарь, который рекомендуется использовать с элементом *Type* (Тип), который состоит из 12 слов. Полная информация по определениям элементов и отношениям между ними описана в реестре метаданных Дублинского ядра (Dublin Core Metadata Registry). Понятия Дублинского ядра метаданных были включены в онтологию, описывающую ресурсы разрабатываемого портала.

Схема поиска информации

Процесс поиска информации с помощью поисковой системы портала в общем случае был представлен в [3, 4]. Кратко его можно описать следующим образом. 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. 1.

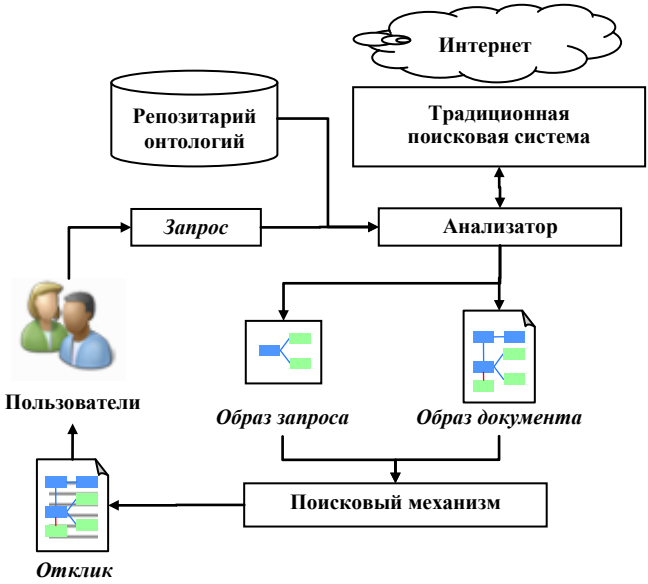

Рис.1. Общая схема работы поисковой системы

Основная особенность предлагаемого подхода – использование репозитария онтологий на этапах преобразования запроса и документа. О структуре репозитария онтологий рассказано в следующем разделе.

Откликом является структурированный документ, т.е. документ, в котором выделены понятия онтологий.

Описание ресурсов с помощью онтологии

Методы искусственного интеллекта, как правило, используются для решения трудно формализуемых задач, постановка которых проста и понятна для человека, но при разработке традиционных алгоритмов их решения возникают трудности. Одна из таких задач – работа с до-

кументами 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 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 [4]. Кроме того, для реализации системы предлагается создать онтологию ресурсов предметной области (спецификация документов, а также источников информации).

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 и решаемых пользователем задач.

В представленном подходе к созданию поисковой системы 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.

Онтологии портала

Помимо описанных выше онтологий в процессе работы портала используются дополнительные онтологии: *онтология источников* информации и *онтология форматов* электронных документов.

Онтологии описаны на языке OWL 2.0 с помощью редактора Protégé 4.2.

На данный момент в онтологии источников детально представлены ресурсы сайта http://www.dsmforum.org. В онтологии представлены также конференции и другие мероприятия, проводимые по тематике моделирования информационных систем, блоги разработчиков инструментальных средств MetaCase и Microsoft Visual Studio.

Онтология форматов документов используется для унификации обработки документов в различных форматах.

Заключение

Применение описанных подходов существенно снизит трудоёмкость поиска необходимой информации, её анализа и расширит возможности использования источников Internet в исследованиях. Полученная в результате анализа найденных 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.

Библиографический список

1. *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.
2. *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.
3. *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. нац. исслед. ун-т. – Пермь, 2009. С. 80-88.
4. *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. нац. исслед. ун-т. – Пермь, 2009. С. 89-96.
5. *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. нац. исслед. ун-т. – Пермь, 2009. С. 74-79.
6. *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.
7. *Мальцева С.В*. Применение онтологических моделей для решения задач идентификации и мониторинга предметных областей // Бизнес-информатика, №3(05), 2008. С. 18-24.