OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.8 + 004.9

АРХИТЕКТУРА СИСТЕМЫ ОБРАБОТКИ ПУБЛИКАЦИЙ ПО ТЕМАТИКЕ МОДЕЛИРОВАНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ

Ланин В.В.[*], Чугунов А.П.[**]

[*] *Национальный исследовательский университет «Высшая школа экономики», г. Пермь, Пермский край*

lanin@perm.ru

[**] *Пермский национальный исследовательский политехнический университет, г. Пермь, Пермский край*

chugunov@permedu.ru

В статье описывается реализация сервиса обработки научных публикаций. В рамках проекта планируется создание программных средств, предназначенных для решения задач поиска и обработки публикаций по тематике моделирования информационных систем на основе сочетания методов корпусной лингвистики и онтологического подхода.
Ключевые слова: интеллектуальная обработка документов, онтология.

Введение

Традиционные средства персонального, корпоративного и глобального поиска, имеющиеся в свободном доступе, предоставляют возможность поиска по вхождению строки в текст документа или его название или поиска на основе статистических методов. При этом не учитывается смысл текста. Однако, в большинстве случаев мы не знаем точных формулировок, а, зачастую, и слов, употребленных в целевом документе.

В настоящее время для решения задачи поиска применяются математико-статистические (латентный семантический поиск), графовые (набор документов представляется в виде ориентированного графа) и онтологические (поиск по готовым онтологиям) методы [Лапшин, 2010]. Каждый из перечисленных подходов имеет недостатки. Графовые методы неприменимы для поиска на локальном компьютере или в локальной сети из-за отсутствия явных ссылок между документами. Онтологические методы трудно применимы из-за отсутствия автоматических методов построения онтологий и значительных затрат на построение индексов и поддержание их актуального состояния. Латентный семантический поиск предполагает ведение поиска без учета семантики слов и выражений.

Несмотря на перечисленные недостатки, интеграция латентного семантического поиска и поиска на графовых структурах дает хорошие результаты: большинство поисковых систем в сети Internet использует сочетание этих подходов [Гасанов, 2002]. Однако, как говорилось, графовые методы неприменимы ни в локальной сети, ни на локальном компьютере и оба подхода не учитывают семантику запроса и документов. Как следствие, задача поиска «по смыслу» остается не решенной, а новейшие алгоритмы поисковых систем в сети Internet, дающие хорошие результаты, остаются неприменимыми на локальном компьютере и в локальной сети.

При дополнении процесса поиска третьим – онтологическим – методом обеспечивается возможность решения проблемы построения ориентированного графа документов и проблемы учета семантики. Построение полных онтологий при этом не обязательно, что делает второй недостаток онтологического подхода некритичным.

Предлагаемый подход сочетает в себе латентный семантический метод поиска и использование онтологий. Это позволяет использовать данный подход в локальной сети и на локальном компьютере (в отличие от алгоритмов ранжирования, показывающих высокие результаты работы при поиске в сети Internet и [Гасанов, 2002]). Предложенный подход позволяет, таким образом, устранить основной недостаток других подходов, основывающихся исключительно на онтологиях:

избежать высоких затрат на построение самих онтологий документов.

Задача агрегации информации из разных источников и ее структуризации является чрезвычайно актуальной. Кроме того, требуют решения задачи устранения дублирования информации и поиска противоречий в результатах поиска. Слабоструктурированный характер информации и гетерогенность еѐ источников предполагают применение средств и методов искусственного интеллекта для решения данной задачи (text mining, технологии Semantic Web и мультиагентные технологии).

1. Поиск информации

Модель, используемую для поиска документов, можно обобщѐнно представить в виде кортежа:

$$<D, Q, F, R(d, q)>,$$

где D – множество представлений документа; Q – множество представлений информационной потребности (запроса); F – средства моделирования представлений документа, запросов и их отношений; $R(d, q)$ – функция ранжирования, которая ставит в соответствие d из D и q из Q вещественные числа, а также определяет порядок на множестве документов относительно запроса q.

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.

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. О структуре репозитария онтологий рассказано в следующем разделе.

2. Описание документа с помощью онтологии

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 учитывали бы 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: моделирование систем и процессов,

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* (две «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 *структурными вершинами*, второй тип – *семантическими*. 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.

3. Архитектура сервиса

Общая архитектура разработанного сервиса представлена на рис. 1.

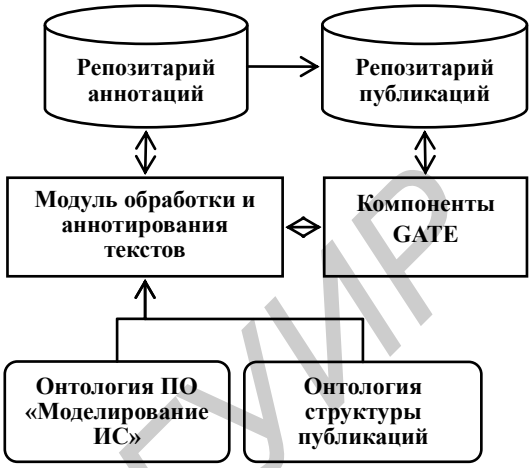

Рисунок 1 – Архитектура сервиса

При работе сервиса используются два онтологических ресурса: *онтология предметной области* «Моделирование ИС» и *онтология структуры публикаций*. Первая онтология содержит базовые понятия, относящиеся к различным аспектам моделирования информационных систем. Во втором онтологическом ресурсе представлены концепты, отражающие структурные элементы научной публикации и связанные с ними лексические маркеры.

Полученные в результате обработки документов аннотации хранятся в репозитарии аннотаций, а сами ресурсы хранятся в репозитарии публикаций. Модуль обработки и аннотирования текстов непосредственно выполняет обработку документов и записывает результаты в репозитарий аннотаций. Компоненты системы GATE [Cunningham, 2011] используются для выполнения базовых процедур по обработке текста.

4. Используемые программные и инструментальные средства

Программные компоненты разрабатываются в средах Microsoft Visual Studio 2012 и Eclipse на языках C# и Java соответственно. Онтологические ресурсы разрабатываются в редакторе онтологий Protégé 4.2.

Согласно 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 Alfresco располагается в правом нижнем углу. Также Alfresco Software попала и в отчет агентства Forrester.

Немаловажным факторами при выборе системы хранения документов 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, также реализованными на данном языке.

Заключение

Применение описанных подходов должно существенно снизить трудоёмкость поиска и анализа информации, обеспечить оперативность её использования в исследованиях, расширить доступный для обработки, изучения объем информации, извлекаемой из различных источников. Полученная в результате анализа документов информация, в свою очередь, может использоваться исследователями для усовершенствования моделей предметных областей, построенных ими. Таким образом, появляется основа для создания интеллектуальной системы поддержки научных исследований с высокой степенью обратной связи.

Ориентация на знания является базовым механизмом функционирования системы, что позволяет комплексно решать поставленные задачи.

Работа выполнена при поддержке Научного фонда НИУ ВШЭ по программе софинансирования грантов РФФИ (проект № 13-09-0143).

Библиографический список

[Segaran, 2009] Segaran T., Evans C., Taylor J. Programming the Semantic Web. – O'Reilly Media, 2009.

[Гасанов, 2002] Гасанов Э.Э. Теория Хранения и поиска информации / Э.Э. Гасанов, В.Б. Кудрявцев // М.: Физматлит, 2002.

[Лапшин, 2010] Лапшин В.А. Онтологии в компьютерных системах / В.А. Лапшин // СПб: Научный мир, 2010.

[Никоненко, 2009] Никоненко А.А. Обзор знаний онтологического типа / А.А. Никоненко // Научно–теоретический журнал «Искусственный интеллект».–2009.–№4.–С.208-219.

[**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.

[Cunningham, 2011] Cunningham H., Maynard D., Bontcheva K. Text Processing with GATE. – Gateway Press CA, 2011.

[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 H. 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.

AN ARCHITECTURE OF THE SCIENTIFIC PUBLICATIONS PROCESSING SYSTEM

Lanin V.V.[*], Chugunov A.P.[**]

** National Research University «Higher School of Economics» - Perm*

lanin@perm.ru

*** Perm National Research Polytechnic University, Perm, Russia*

chugunov@permedu.ru

The purpose of the project 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.