УДК 519.854

С.Ю. Шабанов-Кушнаренко[1], А.И. Коваленко[2], Д.С. Булаенко[1]

[1] *Харьковский национальный университет радиоэлектроники, Харьков*
[2] *Харьковская государственная академия культуры, Харьков*

ПОСТРОЕНИЕ ОНТОЛОГИИ СЕМАНТИЧЕСКОГО ПОИСКА ДОКУМЕНТОВ

Проанализированы методы семантического поиска документов с использованием моделей онтологий предметных областей и концепции Semantic Web. Разработана модель онтологии, способная отражать понятия и структуры, свойственные текстам естественного языка. Разработана математическая модель семантического поиска, использующая созданную онтологию. Предложен алгоритм работы системы семантического поиска.

Ключевые слова: *онтология, информационный поиск, семантический поиск, поисковая система, Semantic Web, предметная область.*

Вступление

Основной задачей, возникающей при работе с полнотекстовыми базами данных, является поиск документов по их содержанию. Однако традиционные средства контекстного поиска зачастую не обеспечивают адекватного выбора информации по запросу пользователя. В настоящее время в поисковых системах используется релевантная модель оценки соответствия исследуемого документа поисковому запросу [1]. Основная проблема заключается в сложности точной формулировки запроса – подбора ключевых слов, которые предстоит искать в телах документов. Это может быть связано с рядом причин, как недостаточным знанием пользователем терминологии предметной области, наличием в языке многозначных и синонимичных слов, и даже орфографическими ошибками в написании искомых слов, которые могут встречаться как в текстах, так и в самом запросе. Одно из перспективных направлений развития информационно-поисковых систем – построение моделей семантического поиска ресурсов, использовать в модели семантического поиска онтологии предметных областей [2 – 5].

Постановка задачи

Онтологии включают доступные для компьютерной обработки определения основных понятий и объектов предметной области, свойства объектов и связи между ними, при этом онтологии обычно формируются экспертами в данной предметной области, преимущественно вручную.

Целью работы является исследование и усовершенствование методов построения онтологии для задач библиографического поиска документов.

Анализ методов семантического поиска

Для успешного поиска необходимо составить поисковый образ запроса. Чем конкретнее будет сформулирован запрос, тем точнее будет найденная информация. Правила составления поисковых образов являются правилами перевода текстов с естественного языка на информационно-поисковые языки. Для рассмотрения особенностей реализации поиска информации важно понимать, что поиск – это процесс, сводящийся к отбору через соотнесение отыскиваемого с каждым. Причем определяющими для понимания методологической основы автоматизации информационного поиска являются два фактора:

- сравниваются не сами объекты, а их описания – «поисковые образы запросов» (ПОЗ);
- сам процесс обычно реализуется последовательностью разнотипных операций.

На рис. 1 приведена обобщенная схема процессов в абстрактной АИПС.

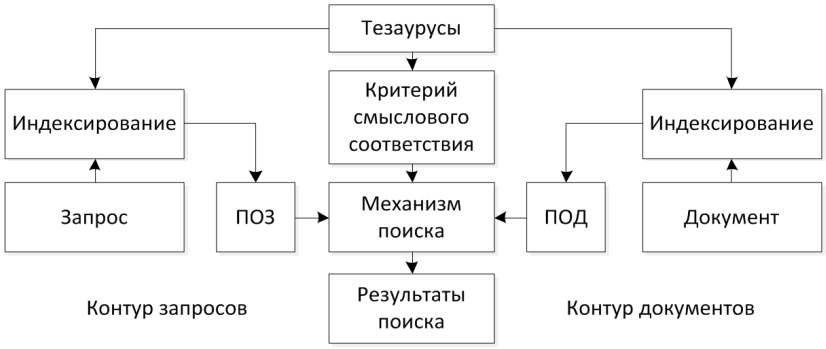

Рис. 1. Обобщенная схема процессов в абстрактной АИПС

Предполагаемый этой схемой алгоритм поиска включает процедуру отбора документов (ПОД):

- выборку очередного объекта из массива системы;
- сравнение выбранного объекта с образцом, выполняемое механизмом поиска;
- принятие решения, соответствует ли объект образцу (определение степени соответствия и применение некоторого критерия для принятия решения на уровне двузначной логики «соответствует»/«не соответствует»);
- переход к выборке следующего объекта или завершение процесса поиска.

Обеспечение возможности использования знаний предметной области стало одной из движущих сил недавнего всплеска в изучении онтологий. В данной работе предлагается применить онтологический подход для реализации семантической поисковой системы.

Семантический поиск позволяет найти документы, не содержащие слов из поискового запроса, но имеющие к ней отношение (например, по запросу «низшие формы жизни» могут быть найдены документы, содержащие слова «бактерии» и «вирусы»).

Основная задача семантического поиска заключается в анализе текста, т.е. извлечения смысла из текста и отображения его в формальную модель, которая позволяет находить смысловую близость двух текстов. Применительно к задаче поиска - близость запроса и документа.

Итеративный процесс разработки онтологии состоит из семи шагов:

- определение области и масштаба онтологии;
- рассмотрение вариантов повторного использования существующих онтологий;
- составление списка наиболее важных терминов в онтологии;
- определение классов и иерархии классов;
- определение свойств классов – слотов;
- определение ограничений (фацетов) свойств классов;
- создание экземпляров.

Semantic Web для построения модели онтологии семантического поиска

Semantic Web – многодисциплинарная тема, которая объединяет теории и методы трех областей:

- логика – формальные структуры и правила логического вывода;
- онтологии – описание типов сущностей, которые относятся к предметной области;
- теория моделей.

Для поддержки концепции Semantic Web W3C создал стандарты, понятия, технологии и форматы. В них входят URI (Uniform Resource Identifier), RDF (Resource Description Framework), OWL (Web Ontology Language), SPARQL (Protocol and RDF Query Language).

В рамках Semantic Web онтологии занимают центральное положение. Они задают отношения между понятиями и определяют логические правила для рассуждений о них.

Таким образом, компьютер может понимать смысл данных, обращаясь к онтологиям за требующейся информацией.

Математическая модель онтологии семантического поиска

Формально определим онтологию как множество

$$O = \{L, C, F_l, F_c, R_h\},$$

где:

$L = \{(w_i, x_i)\}_{i=1,n}$ - словарь терминов предметной области,

w_i - термин, возможно более одного слова,

x_i - его рейтинг относительно других терминов в концепции,

C – набор понятий (концепций), $C = \{c_i\}_{i=1,m}$,

$F_l(L) \to C$ - функция интерпретации терминов, сопоставляет набору терминов из словаря подмножество концепций,

$F_c(C_i) \to L$ - функция интерпретации концепций, сопоставляет концепции набор терминов из словаря,

R_h - отношения иерархии между концепциями.

Ведем следующие обозначения:

$w_i \in L$ - один термин из словаря,

$u = \bigcup_m w_m$ - запрос представляется в виде множества терминов из L, построенных на основе слов из этого запроса.

$P(c_i \mid u)$ - вероятность выбора концепции c_i при условии запроса u.

Итоговая формула для $P(c_i \mid u)$ выглядит следующим образом:

$$P(c_i \mid u) =$$

$$= \sum_{w \in L} \left(\frac{P(w \mid c_i)}{\sum_{c' \in C} P(w \mid c')} \cdot \frac{count(w, L)}{\sum_{w' \in L} count(w', L)} \right),$$

где $P(w \mid c_i)$ - вероятность вхождения термина w в концепцию c_i. Эта вероятность известна из мо-

дели нашей онтологии и имеет значение x_w^f (вес данного термина в данной концепции);

$count(w, L)$ - отношение количества вхождения термина w к общей сумме вхождений всех терминов из запроса в словарь.

Приведем алгоритм работы системы семантического поиска:

1. Формирование онтологии.
2. Ввод пользователем поискового запроса

$$u = \bigcup_m w_m ,$$

состоящий из множества терминов.

3. Модуль интерпретации запроса использует онтологию для выявления множества понятий (концепций), семантически эквивалентных запросу. Для этого используется функция интерпретации терминов

$$F_i(u) = \{c_i, p(c_i \mid u) = \max_{c_j \in C}(p(c_i \mid u))\}_{i=1,n} ,$$

определяющая вероятность для каждой концепции и возвращающая множество концепций с максимальной вероятностью.

4. Модуль уточнения запроса предоставляет пользователю интерфейс для выбора из полученного множества вероятных концепций той, которую он считает наиболее соответствующей теме запроса.

5. Модуль расширения запроса работает только для одной, указанной пользователем концепции. Для расширения запроса применяется функция интерпретации концепций

$$F_c(c_i) = \{w_j\}_{j=1,n},$$

формирующая для указанной концепции список наиболее релевантных, семантически связанных с данным понятием терминов.

6. Множество полученных терминов передается поисковой системе. ИПС ищет документы, содержащие все или часть терминов из расширенного запроса.

Найденные ссылки передаются модулю вывода результатов, который предоставляет пользователю возможность просмотреть найденные документы.

Выводы

Проведенные исследования показали актуальность построения онтологических моделей семантического поиска документов. В работе усовершенствована математическая модель семантического поиска, использующая онтологию предметной области. Разработана математическая модель онтологии, ориентированной на задачи информационного поиска, определены формальные функции интерпретации концепций и терминов. Предложен метод для автоматического создания онтологии на основе распределенных информационных ресурсов, имеющихся в сети Интернет.

Список литературы

1. Zuccon G. The Quantum Probability Ranking Principle for Information Retrieval [Текст] / Guido Zuccon, Leif Azzopardi, Keith van Rijsbergen // ICTIR 2009. – P. 232-240.

2. Ushold M. Ontologies: Principles, Methods and Applications [Текст] / Mike Ushold, Michael Gruninger // Knowledge Engineering Review, Volume 11, Issue 2, 1996. – P. 93-136.

3. Heflin J. Applying Ontology to the Web: A Case Study [Текст] / Jeff Heflin, James A. Hendler, Sean Luke // IWANN (2), 1999. P. 715-724.

4. Воскресенский А.Л. Средства семантического поиска [Текст] / А.Л. Воскресенский, Г.К. Хахалин // Материалы международной конференции «Диалог 2006», Москва. – С. 100-105.

5. Гусев В.С. Яндекс: эффективный поиск информации в Интернет. Краткое руководство [Текст] / В.С. Гусев. –М.: «Диалектика», 2007. – 224 с.

Поступила в редколлегию 10.07.2015

Рецензент: д-р техн. наук, проф. С.Ф. Чалый, Харьковский национальный университет радиоэлектроники, Харьков.

ПОБУДОВА ОНТОЛОГІЇ СЕМАНТИЧНОГО ПОШУКУ ДОКУМЕНТІВ

С.Ю. Шабанов-Кушнаренко, А.І. Коваленко, Д.С. Булаєнко

Проаналізовано методи семантичного пошуку документів з використанням моделей онтологій предметних областей і концепції Semantic Web. Розроблено модель онтології, здатну відбивати поняття і структури, властиві текстам природної мови. Розроблено математичну модель семантичного пошуку, що використовує створену онтологію. Запропоновано алгоритм роботи системи семантичного пошуку.

Ключові слова: *онтологія, інформаційний пошук, семантичний пошук, пошукова система, Semantic Web, предметна область.*

BUILDING OF ONTOLOGY FOR DOCUMENTS SEMANTIC SEARCH

S.Yu. Shabanov-Kushnarenko, A.I. Kovalenko, D.S. Bulaenko

The methods of semantic search for documents using models of domain ontologies and concept of Semantic Web are analyzed. Developed a model of ontology that can reflect the concepts and structure inherent in natural language texts. A mathematical model of semantic search using the created ontology is developed. The algorithm of the semantic search system is offered.

Keywords: *ontology, information retrieval, semantic search, search engine, Semantic Web, subject area.*