УДК 004.5+004.82

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА

Бажанова А. И., Мартыненко Т. В.
Донецкий национальный технический университет
кафедра автоматизированных систем управления
E-mail: A_Bazh@rambler.ru

Аннотация

Бажанова А. И., Мартыненко Т. В. Исследование применения онтологических моделей для семантического поиска. *Рассмотрена схема работы семантического поиска информации, место онтологической модели в нем. Проанализированы основные средства построения онтологий. Проведен сравнительный анализ основных моделей представления данных в онтологиях, а также основных языков описания онтологий и редакторов для работы с ними.*

Общая постановка проблемы

Современные средства поиска, каталогизации, описания текстов не удовлетворяют нарастающим потребностям пользователей. Требуется их развитие в направлении повышения эффективности поиска информации и упрощения взаимодействия с пользователем.

В современных поисковых системах тексты автоматически индексируются по набору составляющих эти тексты слов. Такое представление текстов как простого набора слов имеет ряд очевидных недостатков:

- избыточность - в пословном индексе используются слова-синонимы, выражающие одни и те же понятия;
- слова текста считаются независимыми друг от друга, т. е. смысловая составляющая слова;
- многозначность слов - поскольку многозначные слова могут иметь два или более понятия, выражающих различные значения многозначного слова, то маловероятно, что все они интересуют пользователя.

Поэтому предлагается использовать семантическую модель информации, которая лишена этих недостатков, за счет использования концептуального индексирования, т. е. индексирование не по словам, а по понятиям. При такой технологии

- все синонимы сведены к одному и тому же понятию,
- многозначные слова отнесены к разным понятиям
- связи между понятиями и соответствующими словами описаны и могут быть использованы при анализе текста [1].

На рис. 1 показана схема поиска информации. Пользователь вводит запрос, который подвергается лингвистическому анализу, расширяется за счет использования синонимов, затем преобразовывается в и отправляется поисковой машине. Поисковая машина возвращает найденные документы, они также подвергаются лингвистическому разбору и формируются семантические образы документов. Образы документов сравниваются с образом запроса, делается вывод о релевантности каждого из документов и результаты анализа (документы, которые были признаны релевантными) предоставляются пользователю.

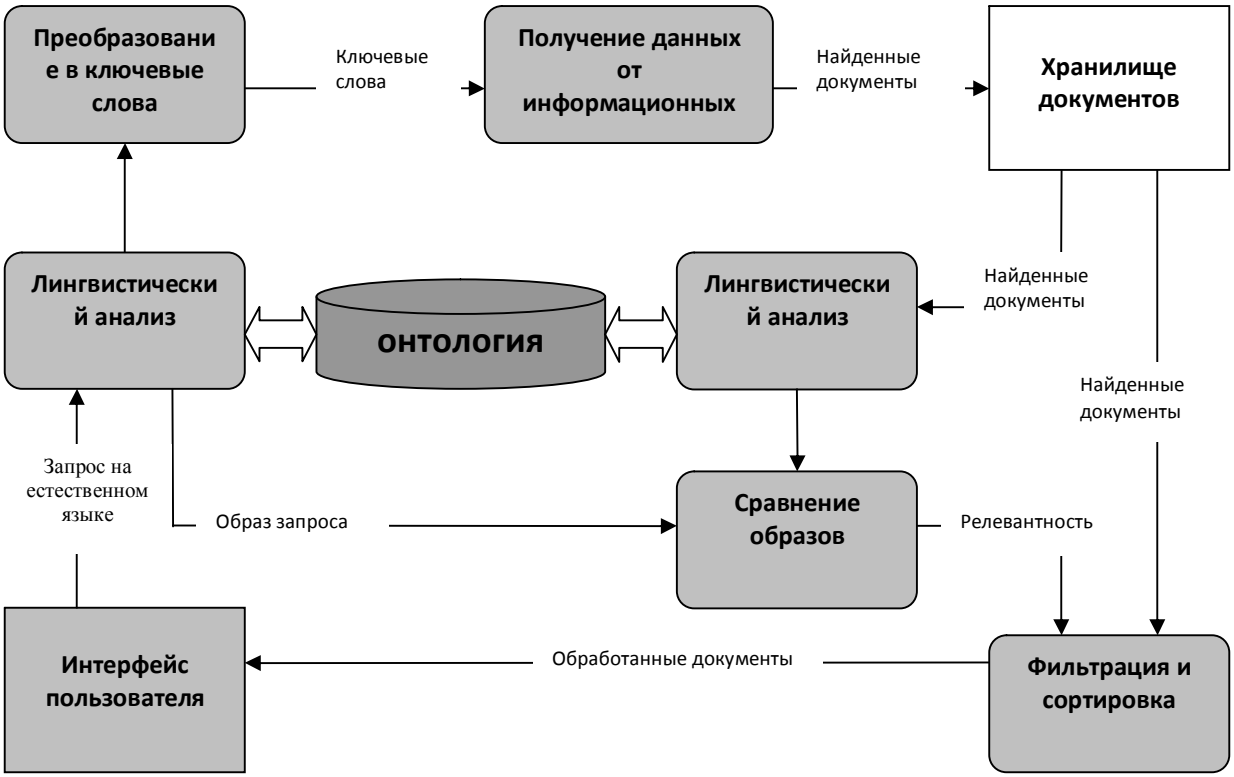

Рис. 1 – Диаграмма потоков данных при поиске.

Одной из важнейших стадий в разработке семантических поисковых систем является построение онтологических моделей, на основе которых можно будет делать вывод о том, какую именно информацию хочет получить пользователь, вводя запрос.

Под онтологией можно понимать:

- надежный семантический базис в определении содержания;
- общую логическую теорию, которая состоит из словаря и набора утверждений на некотором языке логики;
- основу для коммуникации между людьми и компьютерными агентами.

Информационные онтологии состоят из экземпляров, понятий, атрибутов и отношений и должны иметь формат, который компьютер сможет обработать.

Компоненты, из которых состоят онтологии, зависят от парадигмы представления. Но практически все модели онтологий в той или иной степени содержат *концепты* (понятия, классы, сущности, категории), *свойства* концептов (слоты, атрибуты, роли), *отношения* между концептами (связи, зависимости, функции) и дополнительные *ограничения* (определяются аксиомами, в некоторых парадигмах фасетами).

Постановка задачи построения онтологии. Формализованную модель онтологии предметной области можно представить как знаковую систему, где О – онтология:

$O = \langle C, R, A, P, D \rangle$, где

$C = \{c_1,..., c_n\}$ – конечное множество понятий, при $n \in \overline{1...N}$ – количество понятий, присутствующих в онтологии.

$R = \{r_1,..., r_m\}$ – конечное множество отношений между понятиями $r_i(c_x, c_y)$, при $m \in \overline{1...M}$ – количество отношений между понятиями.

$A = \{a_1,..., a_w\}$ – конечное множество атрибутов, т.е. бинарных отношений, при $w \in \overline{1...W}$ - количество атрибутов.

$P = \{p_1,..., p_t\}$ – конечное множество конкретных свойств атрибута, при $t \in \overline{1...T}$ – количество свойств атрибута.

$D = \{d_1,..., d_k\}$ – конечное множество типов отношений, при $k \in \overline{1...K}$ - количество типов отношений.

Исходя из требований, предъявляемых к онтологии, следует, что общее количество понятий, используемых в онтологии, должно стремиться к максимальному числу понятий, используемых в данной предметной области.

$$n \to N_{\max}$$

Что достигается постепенно при последовательном расширении онтологии. Т. е. для разработки онтологической модели необходимо выполнить следующие этапы:

- определение классов в онтологии;
- расположение классов в таксономическую иерархию (Подкласс – надкласс);
- определение атрибутов и описание их допустимых значений;
- заполнение значений атрибутов экземпляров.

После этого создается база знаний, определяются отдельные экземпляры этих классов, вводятся в определенный атрибут значение и дополнительные ограничения для атрибута.

Исследование применения онтологических моделей для семантического поиска. Для построения семантических моделей, или онтологий, необходимо разработать языки их представления, имеющие достаточную выразительную мощность и позволяющие пользователю избежать «низкоуровневых» проблем. При этом могут быть использованы такие специализированные языки как Resource Description Framework (RDF), Web Ontology Language (OWL) и т. д. Онтологии могут использовать различные модели представления знаний, такие как логика предикатов (First order logics - FOL), дескриптивная логика, фреймовые модели (Frames), концептуальные графы и т.п. Для создания онтологий могут использоваться различные редакторы (Protege, Ontolingua, WebOnto и др.), которые в свою очередь могут поддерживать различные форматы представления данных (языки), основанные на различных формализмах (логиках, моделях представления данных). Ключевым моментом в проектировании онтологии является выбор соответствующего языка спецификации онтологий (Ontology specification language) и редактора для работы с ней.

Онтологические модели за время исследований в этой области претерпели значительное развитие. В настоящее время для создания и поддержки онтологий существует целый ряд инструментов, которые помимо общих функций редактирования и просмотра выполняют поддержку документирования онтологий, импорт и экспорт онтологий разных форматов и языков, поддержку графического редактирования, управление библиотеками онтологий и т.д [3].

Наиболее известные инструменты инженерии онтологий, их основные характеристики представлены в таблице 1 [2].

Как уже говорилось выше, инструменты инженерии онтологий используют специализированные языки. Целью таких языков является предоставление возможности задавать дополнительную машинно-интерпретированную семантику ресурсам, сделать машинное представление данных более приближенным к реальному миру, повысить возможности концептуального моделирования слабо структурированных Web-данных. Такой подход распространился и на разнообразные языки описания онтологии и на инструментальные средства, предназначенные для работы с ними.

Таблица 1 – Инструменты инженерии онтологий

Название параметра	OilEd	Onto Edit	Ontolingua	OntoSaurus	Protege	WebODE	WebOnto
Архитектура приложения	3–х уровневая	3–х уровневая	Клиент/сервер	Клиент/сервер	3–х уровневая	n-уровневая	Клиент/ сервер
Хранение онтологий	файлы	файлы	файлы	файлы	файлы, СУБД	СУБД	Файлы
Язык ПО	Java	Java	Lisp	Lisp	Java	Java	Java+ Lisp
Осн. язык представления знания	DAML+ OIL	OXML	Ontolingua	LOOM	OKBC	-	OCML
Интерфейс пользователя	Локк-ое приложение	Локк-ое приложение	HTML	HTML	Локк-ое приложение	HTML и апплеты	Апплеты
Графич. редакт. таксономии концептов	-	+	-	-	+	+	+
Редактор формальных аксиом	+	-	-	-	+	+	-

На сегодняшний момент выделяют три основных класса языков описания онтологии, что показано на рис. 2:

- традиционные языки спецификации онтологии: Ontolingua, CycL и языки, основанные на дескрипционной логике (такие как LOOM), также языки, основанные на фреймах (OKBC, OCML, Flogic);

- более поздние языки, основанные на Web -стандартах (XOL, SHOE, UPML);

- специальные языки для обмена онтологией через Web: RDF(S), DAML, OIL, OWL [1].

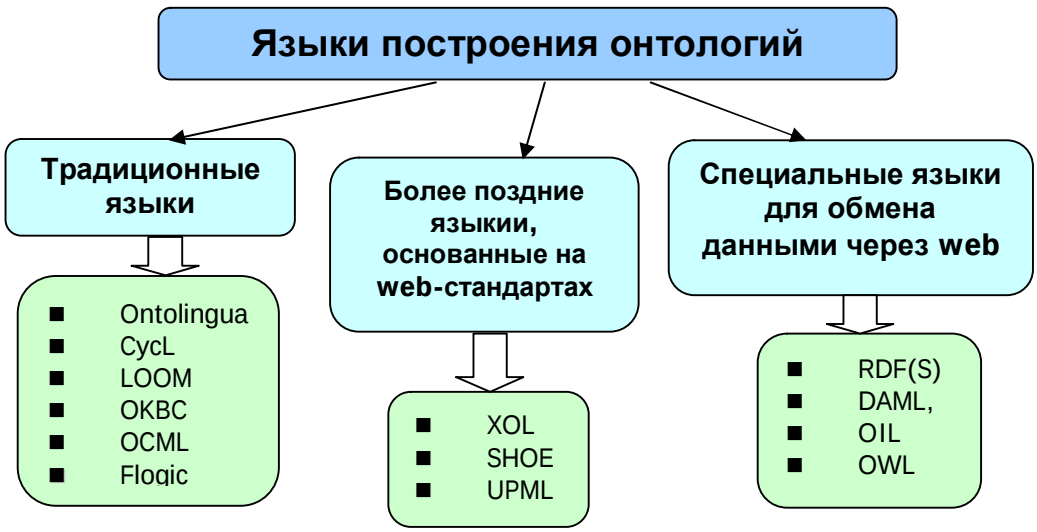

Рис. 2 – Классификация форматов представления данных

На сегодняшний день редакторы онтологий, кроме своего языка, поддерживают импорт и экспорт данных различных форматов, что показано на рис. 3.

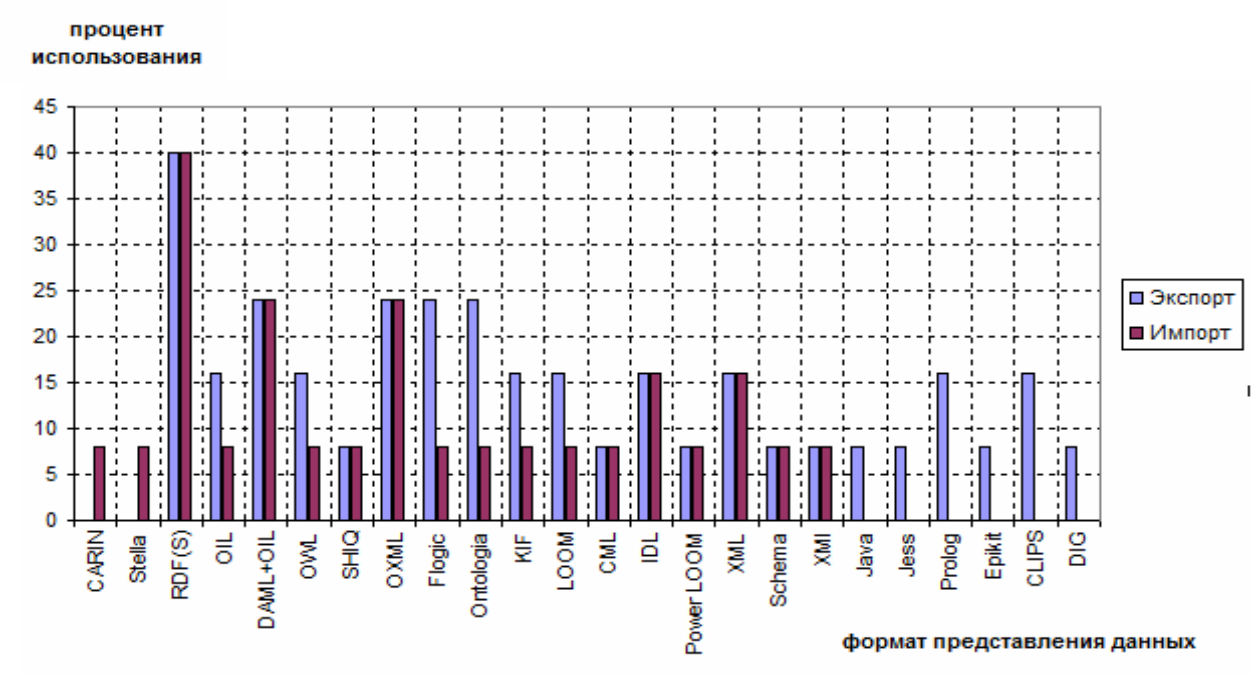

Рис. 3 – График применения различных форматов описания онтологий при импорте и экспорте данных

Исходя из графика, следует, что наиболее часто используемым форматом представления данных является RDF(S). Язык RDF обладает рядом преимуществ: представляет данные в виде rdf-триплетов (сущность-объект-предикат), а rdf-схема представляется в виде ориентированного графа, что является удобной для восприятия формой представления данных.

Выводы. В данной статье были рассмотрены основные недостатки поиска по ключевым словам, выделены преимущества семантической модели поиска информации. Рассмотрена схема потоков данных при семантическом поиске и место онтологической модели в нем.

Проведен обзор основных инструментов инженерии онтологий, форматов представления данных и языков для их описания. Исходя из анализа основных параметров различных редакторов онтологий, наиболее приемлемым является редактор Protege, именно он будет взят за основу в дальнейшей работе. Среди форматов представления данных, лидирующие позиции занял RDF(S), который будет использован для построения онтологии предметной области электронной библиотеки кафедры АСУ.

Литература

1. ОНТОЛОГИИ И ТЕЗАУРУСЫ: [Учебное пособие] / Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. – Москва: 2006. – 157с.

2. Обзор инструментов инженерии онтологий/ О.М. Овдей, Г.Ю. Проскудина // Журнал ЭБ. – 2004 – №4

3. Никоненко А.А. Обзор баз знаний онтологического типа// Искусственный интеллект.–2002.–№ 4. – С. 157–163