УДК 624.012:801.52

А.Т. АБДЫШОВА

E.mail. ksucta@elcat.kg

АНАЛИЗ МЕТОДОВ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ И ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ТЕКСТОВОЙ ИНФОРМАЦИИ

Бул макалада техникалык тексттердин гармониясын жаратуудагы иштердин эффетивдүүлүгүн жогорулатуунун жолдору каралды.

Целью данной работы является повышения эффективности поиска естественной текстовой информации по запросу пользователя на требуемом языке и уровня гармонии технического текста.

The aim of this articles is to improve the effectiveness of the natural search text information on the users request in the required language and the level of harmony technical text.

Для достижения поставленной цели необходимо решить следующие *задачи:*

Провести анализ применения онтологических моделей для семантического поиска информации и методов семантической обработки текста.

Развитие наукоемких отраслей человеческой деятельности в современном обществе сопровождается возрастанием роли компьютерных технологий. Появилась необходимость поиска новых способов ее хранения, представления, формализации и систематизации, а также автоматической обработки, способные без участия человека извлечь какие-либо сведения из текста (семантические связи). Как результат, на фоне вновь возникающих потребностей развиваются новые технологии, призванные решить заявленные проблемы. Неотъемлемым компонентом семантического веб является понятие онтологии, описывающей содержание семантической разметки.

Онтологии являются удобным средством представления и хранения знаний, поэтому развитие алгоритмической базы для создания, обновления и поддержки онтологий, является весьма актуальной задачей в настоящее время.

Разработать онтологическую модель для поиска информации в области компьютерной литературы. /1/

- Разработать алгоритм для автоматизированного расширения онтологий семантическими образами текстов и создать информационную систему.
- Предложена онтологическая модель для поиска естественной текстовой информации в компьютерной литературе, которая позволит получить полноценную базу знаний в предложенной предметной области.
- Разработана информационная подсистема с возможностью автоматизированного расширения онтологий .
- Семантическая обработка текста выполняется в три этапа: морфологический, синтаксический и собственно семантический анализ (рис. 1). Каждый этап выполняет отдельный анализатор со своими входными и выходными данными и собственными настройками. /1-3/

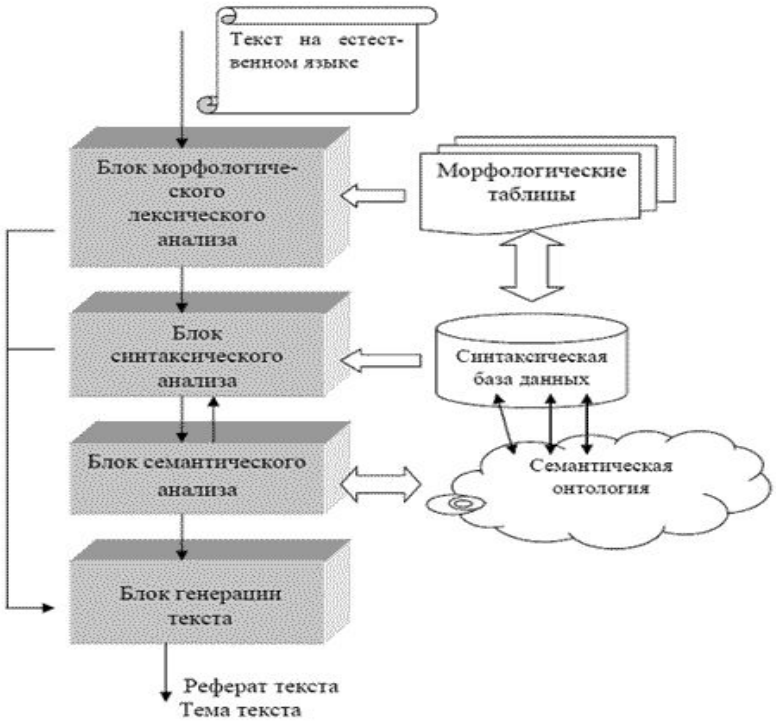

Рисунок 1 – Схема лингвистического анализа

Ввиду сложности выполнения всех этапов в работе рассматриваться будет только блок морфологического анализа. Среди методов морфологического анализа, использующихся в лингвистических процессорах, можно выделить методы с декларативной и с процедурной ориентацией. Основным недостатком декларативных методов является чрезмерно большой объем текста. Достоинствами метода является простота (и, как следствие, высокая скорость) анализа, а также термины аналитической механики.

Для процедурных методов время анализа одного слова может быть существенно выше, но объем используемых словарей в небольших системах позволяет загружать словари целиком в оперативную память. Существенным недостатком процедурных методов является отсутствие универсальности. Каждый из данных подходов имеет свои преимущества и недостатки, поэтому в дальнейшей работе будет использоваться комбинация этих методов для сочетания преимуществ каждого из них.

В общем виде схема морфологической обработки текста показана на рисунке 2.

Морфологический словарь

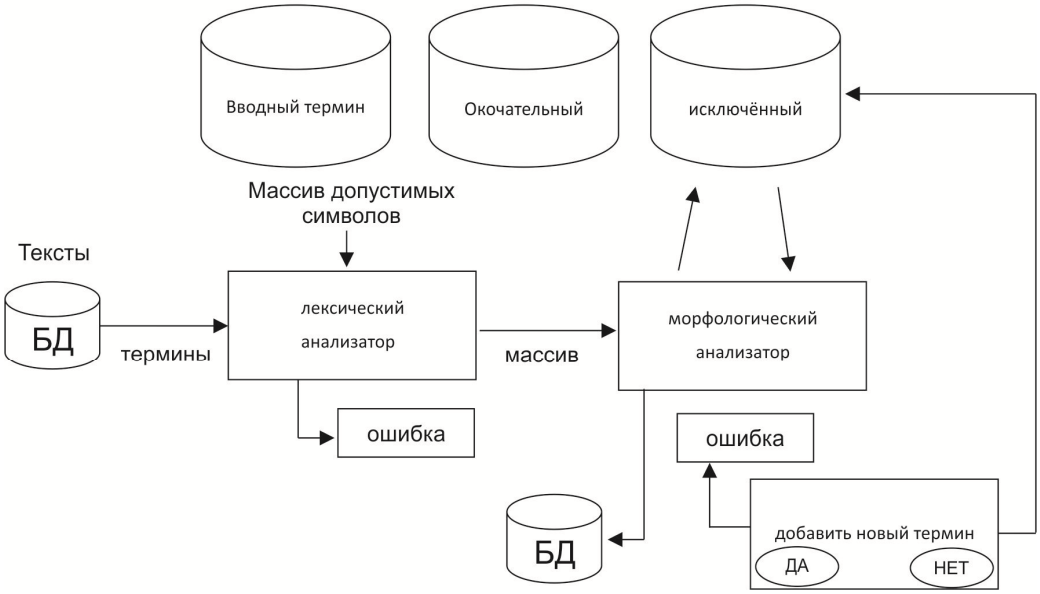

Рисунок 2 – Морфологический разбор текста.

Предварительно необходимо провести лексический анализ, т. е. проверить на допустимые символы. На вход лексического анализа подаются предложения из текста поочередно, а на выходе проверенный набор слов ./4-6/

Описание алгоритма работы морфологического анализатора:

1. На вход поступает массив “слов”, знаков препинания и чисел, выделенных из входного текста на этапе лексического анализа.
2. Для каждого “слова” анализатор выполняет процедуру поиска в словаре основ, загруженном в память. При этом ищутся все основы, с которых может начинаться анализируемое слово.
3. Если очередная основа удовлетворяет этому условию, то из словаря аффиксов извлекается строка, содержащая все возможные аффиксы для данной основы.
4. Каждый аффикс из этой строки поочередно присоединяется к основе, и результат сравнивается с анализируемым словом. В случае их точного совпадения формируется очередная запись в список результатов поиска: по порядковому номеру аффикса в строке аффиксов определяются переменные морфологические параметры слова (например, для существительного - число и падеж), а по словарной информации данной основы - его постоянные параметры (для существительного — род и одушевленность).
5. Если в результате такого поиска не найдено ни одного успешного варианта, то проводится поиск среди исключений. При поиске среди исключений приходится просматривать все словоформы всех присутствующих в словаре исключений. Это занимает много времени, поэтому поиск среди исключений проводится только в том случае, когда не найдено ни одного варианта среди обычных основ.
6. Если некоторая словоформа некоторого исключения точно совпадает с анализируемым словом, то по номеру словоформы определяются переменные морфологические параметры слова, а по словарной информации самого исключения — постоянные параметры слова.

Потребность в онтологиях связана с невозможностью адекватной автоматической обработки естественно-языковых текстов существующими средствами. Поэтому, для качественной обработки текстов и поиска релевантной информации, необходимо иметь детальное описание проблемной области, с множеством логических связей, которые показывают соотношения между терминами области. Использование онтологий позволяет представить естественно-языковый текст в таком виде, что он становится пригодным для автоматической обработки. /7-8/

При сравнении подходов семантического поиска с традиционными подходами поиска по ключевым словам можно отметить, что теоретически они имеют ряд преимуществ над традиционными подходами в смысле повышения релевантности получаемых результатов. Это связано с тем, что релевантными результатами являются документы, удовлетворяющие информационные потребности пользователей и релевантность оценивается по смыслу текстов.

Главным недостатком подходов семантического поиска в сравнении с традиционными подходами поиска является тот факт, что алгоритмы обработки смысла текстов зависят от особенностей конкретного анализируемого естественного языка, т.е. требуется создание специальных алгоритмов для разных естественных языков. При этом для каждого естественного языка должны учитываться его синтаксические и семантические особенно сти, отношения между словами и т.п. В связи с этим реализация подходов семантического поиска в многоязычных системах является очень сложной и трудоемкой работой.

В настоящее время основными используемыми системами информационного поиска являются представители традиционных подходов поиска по ключевым словам .Это говорит о несомненной эффективности таких подходов для решения данной задачи. Однако даже крупнейшим компаниям, предоставляющим сервисы информационного поиска трудно отказаться от огромных возможностей, которые могут дать семантические поисковые системы. Такие системы подтверждают тот факт, что семантический метод является перспективным направлением развития поиска информации.

Проведенные работы показали работоспособность созданного программного комплекса и его пригодность для широкого применения. Анализ полученных результатов позволяет сделать предварительный вывод о действенности метода выделения семантически близких понятий.

На основе информации о смысловой близости понятий удалось построить процедуру поиска текстовой информации. На качественном уровне удалось доказать повышенную надежность данной процедуры по сравнению с применяемыми в современных системах поиска. Построенная процедура поиска обладает рядом достоинств, которые делают ее востребованной в текстовых информационных системах большого объема.

В работе был проведен анализ существующих средств и методов построения онтологий. В ходе анализа было установлено, что существует множество инструментальных средств, для построения онтологий, однако не одно из них не позволяет автоматизировать этот процесс.

Для построения онтологий существуют различные специализированные языки, которые в свою очередь используют различные модели представления знаний и основаны на различных логиках. В результате проведенного анализа были сформулированы задачи для дальнейшей работы, выбраны методы и алгоритмы для их реализации в технических науках.

Список литературы

1. Исследование применения онтологических моделей для семантического поиска.
2. Никоненко А.А. Обзор баз знаний онтологического типа// Искусственный интеллект.–2002.–№ 4. – С. 157–163.
3. Семантический веб и микроформаты: Интуит. Лекция — Режим доступу http://www.intuit.ru/department/internet/mwebtech/20/
4. Королёв А.Н. Лингвистическое обеспечение информационно-поисковой системы Excalibur RetrievalWare: Аналитический аспект
5. Андреев А.М., Березкин Д.В., Брик А.В. Лингвистический процессор для информационно-поисковой системы – М: МГУ
6. Анисимов А.В., Марченко А.А. Система обработки текстов на естественном языке.// Искусственный интеллект.–2002.–№ 4. – С. 157–163.
7. Кутуев М.Д., Абдышова А.Т. Методы для семантического отображения технической информации в моделях.-2012,Б.,КГУСТА,Вестник №2(36)-С.7-12.

8. Кутуев М.Д., Абдышова А.Т. Методика автоматизации исследования технических текстов.-Б., МУИТ, Научный информационный журнал МАТЕРИАЛОВЕДЕНИЕ №1/2013(2)