А.В.Сиренко

ЛИНГВОКУЛЬТУРНЫЙ ТЕЗАУРУС РУССКОГО ЯЗЫКА[1]

Общая характеристика работы

Актуальность темы. Одной из наиболее динамичных и востребованных разделов компьютерной лингвистики сегодня является семантический анализ текстов. Ему есть множество применений в сфере образования, интернет-технологиях и во многих других областях, где есть необходимость в автоматизации обработки больших массивов текстовых данных. Семантический анализ – процесс выявления смыслового содержания слов и словосочетаний в предложении. Он обеспечивает нормализацию синтаксической структуры предложений, распознавание терминов, классификацию терминов по семантическим признакам, выявление определений терминов.

Тема работы «Лингвокультурный тезаурус русского языка» непосредственно связана с задачей семантического анализа текста. Применяемая в работе пятикомпонентная структура единицы языкового сознания была разработана чл. корр. РАН Карауловым Юрием Николаевичем. Также в работе используются результаты ассоциативного эксперимента в виде ассоциативной сети.

Дипломная работа выполнена в рамках проекта РФФИ 05-06-80284 «Языковое сознание нашего современника: когнитивная структура и лингвокультурное содержание».

Цель работы. Целью работы является разработка лингвокультурного тезауруса русского языка на основе баз данных фигур знания и ассоциативного эксперимента, а также моделирование пассивного режима ра-

---

1 Статья представляет собой автореферат дипломного проекта по специальности 230102 – «Автоматизированные системы обработки информации и управления», выполненного на кафедре «Системы обработки информации и управления» Московского государственного технического университета им. Н.Э. Баумана в 2007 г. Руководитель проекта к.т.н., доц. Ю.Н.Филиппович; консультанты: Г.Б. Полаева, к.т.н., доцент И.В. Переездчиков; рецензент к.т.н., доцент В.Н. Поляков.

боты когнайзера через нахождение путей между знаками в ассоциативной сети.

Задачи. Для достижения поставленной цели решаются следующие задачи:

1. Анализ предметной области работы.
2. Постановка задач, решаемых системой.
3. Разработка архитектуры системы
4. Проектирование базы данных лингвокультурного тезауруса.
5. Разработка алгоритма функционирования пассивного режима когнайзера
6. Проектирование алгоритмов взаимодействия с пользователем и экранных форм.
7. Реализация программного продукта.

Практическая ценность. В работе был разработан программный продукт, осуществляющий работу с базой данных фигур знания, выборку фигур по заданным критериям, добавление данных в тезаурус из внешнего источника, а также реализующий пассивный режим работы когнайзера. Работа выполнена в рамках гранта Российского фонда фундаментальных исследований.

Объем работы. Дипломная работа содержит 120 страниц, 92 рисунков и таблиц, 11 источников, 19 страниц приложений.

Основное содержание работы

Во введении производится общая постановка задач, указывается на возможности их решения.

Глава 1 содержит описание предметной области.

Язык рассматривается в качестве отражения представлений об окружающем мире внутри сознания субъекта, так называемого Языкового Сознания.

Языковое Сознание можно представить в виде структуры:

<Языковое Сознание> = <Единица Знаний о Мире> + <Языковая Единица>

В этой зависимости языковое сознание представляет собой когнайзер, в котором происходит постоянное преобразование Единиц Знаний о Мире в Языковые Единицы и наоборот. Работу когнайзера по переходу от слова (знака) к знанию будем называть активным режимом работы, а от знания к знаку – пассивным.

Разработке лингвокультурного тезауруса предшествовал ассоциативный эксперимент. В нем респонденту называлось некое слово (стимул) и его задачей было назвать слово, ассоциирующееся со стимулом.

В результате формируется пара слов «стимул-реакция», которые можно рассматривать в качестве отражения Языкового Сознания, где стимул выступает в роли Языковой Единицы, а реакция в роли Единицы Знаний о Мире.

Знания об окружающем мире можно представить в виде так называемых Фигур Знания. Это элементарные когнитивные единицы, которыми мы оперируем в повседневной жизни. Они содержат как интенсиональные характеристики, а именно Знак, Способ, Формула смысла, так и экстенсиональные, отражающие положение Фигуры знания в общем пространстве знаний. Такими параметрами являются когнитивная область и функция. Схематичное представление фигуры знания можно увидеть на рисунке ниже.

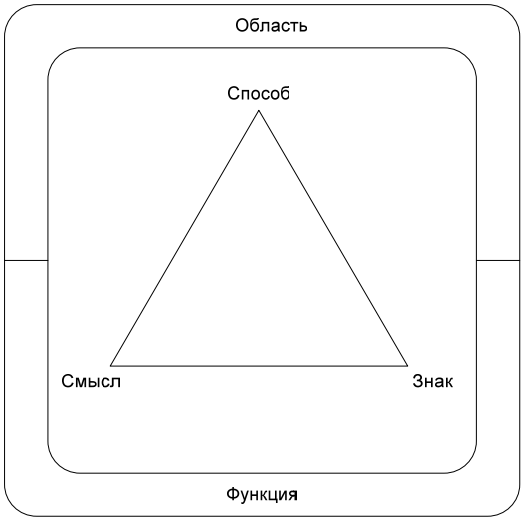

Рисунок 1 Фигура знания

В пассивном режиме работы когнайзера изначально имеется Формула смысла и, определенный через нее, Способ. Имея эти два параметра, когнайзер переходит к Знаку.

Фигуры знания в системе представлены материалами кроссвордов различной тематики.

Проводится обзор аналогов и прототипов. В качестве таковых приводятся системы:

- RSO Semantic Network SDK – инструментарий разработчика.

- www.lexfn.com – интернет-ресурс, работающий с англоязычной ассоциативной сетью.
- TextAnalyst – персональная система автоматического анализа текста.

Lexfn.com (интернет-ресурс компании «Datamuse Corporation», США, посвященный ассоциативным сетям) наиболее близок из всех представленных в обзоре систем к разработанной системе с точки зрения использования ассоциативной сети. Сайт предоставляет широкие возможности по поиску путей между знаками в ассоциативной сети. Система различает множество типов связей между словами, выполняет поиск не только непосредственно достижимых из начальной вершин сети, но и находит пути с ограничением по длине. На рисунке ниже изображен пример результата запроса.

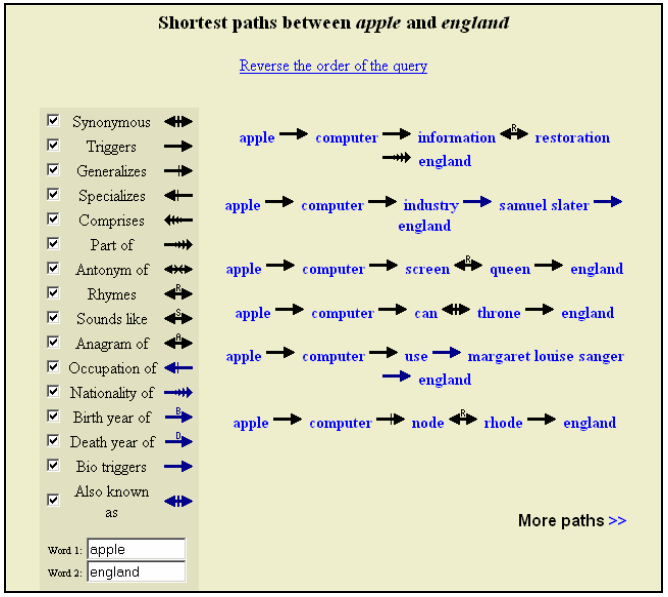

Рисунок 2 Результаты запроса

TextAnalyst разработан компанией “Microsystems, Ltd” в качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов. В результате анализа текста формируется семантическая сеть - основная структура, характеризующая смысл текста, в которой понятия (слова и словосочетания) объ-

единяются ассоциативными связями в соответствии с их совместной встречаемостью. Затем создается так называемое тематическое древо - представление структуры текста в виде многоуровневой иерархии тем и раскрывающих их подтем.

Библиотека *RCO Semantic Network*, разработанная компанией "Гарант-Парк-Интернет" позволяет автоматически анализировать содержание текстовых документов, представляя его в форме ассоциативной семантической сети. Ассоциативная семантическая сеть в программе представляет собой ориентированный граф, вершинами которого служат значимые темы, выделенные в анализируемом тексте, а дугами – связи между ними. С каждой вершиной связаны вес (значимость) и частота упоминания темы, а с каждой дугой – вес (сила) связи и частота подкрепления связи в тексте.

В главе 2 осуществляется разработка алгоритма пассивного режима когнайзера.

Для моделирования пассивного режима работы когнайзера необходимо осуществлять переход от формулы смысла к знаку, используя базу данных ассоциативного эксперимента. Для этого формула смысла разделяется на словоформы, для которых впоследствии ведется поиск соответствующих знаков ассоциативной сети.

Поиск путей от формулы смысла к знаку разделяется на поиск путей от каждого знака, входящего в формулу смысла к знаку фигуры знания.

К алгоритму поиска предъявляются требования:

- Определять длину пути.
- Восстанавливать сам путь при его нахождении.
- Должен быть не ресурсоемким.

Примененный в ЛКТ алгоритм поиска основан на алгоритме Дейкстра с единичной длинной пути (поиск в глубину).

Для каждой вершины в процессе поиска заполняется структура следующего вида:

Таблица 1

Наименование поля	Тип	Описание
Name	строка	наименование вершины
Level	целое число	Удаленность вершины от начальной
Old_elements	массив строк	Вершины, присутствующие в пути к данной вершине

Алгоритм поиска выглядит следующим образом:

1. Для начальной вершины заполняется структура.

Уровень = 0.
Список Old_elements пуст.
Текущей вершиной является начальная.

2. Если Level текущей вершины = Maxlevel тогда переход к пункту 8.

Иначе переход к пункту 3.

3. Рассматриваем список вершин, в которые можно попасть из текущей и которые не присутствуют в списке Old_elements текущей вершины.
4. Если в списке присутствует вершина, путь к которой мы ищем, выводим путь до текущей вершины с конечной вершиной как искомый путь. Обработка текущей вершины завершена. Переход к пункту 7.
5. Если список пуст, переход к пункту 8.

Иначе переход к пункту 6.

6. Для каждой вершины в списке заполняем структуру

Level = Уровень текущей вершины+1;
Old_element = Old_element текущей вершины + Name текущей вершины.
Для каждой вершины начинаем работу алгоритма в пункте 2.

7. Если Level < Minlevel переход к пункту 8.

Иначе вывод списка Old_elements в качестве пути до текущей вершины.
Переход к пункту 8.

8. Выход.

Функционирование алгоритма представлено на примере участка ассоциативной сети и поиске путей от вершины 1 к вершине 7.

Каждая вершина графа может быть точкой ветвления вычислений на n-1. Таких вершин может быть n-1. В результате сложность алгоритма вырастает до $\Theta(n^4)$.

Для оценки сложности алгоритма с учетом особенностей ассоциативной сети был проведен ее анализ.

Анализ ассоциативной сети, а также данные по аналогичным англоязычным сетям показал, что в рассмотрении путей длинной более 5 нет необходимости в силу высокой степени связности сети. Если вершина не достигается за 5 шагов, она считается недостижимой.

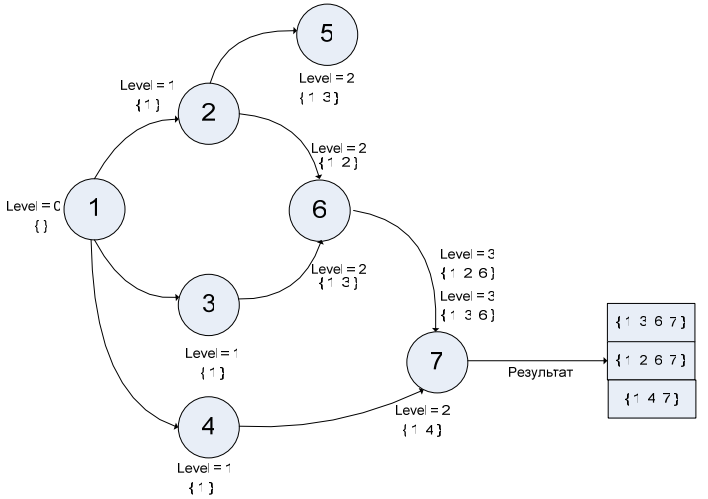

Рисунок 3

Для изучения функционирования поиска в ассоциативной сети было проведено изучение свойств сети в двух случаях:

- Изменение количества стимулов при неизменном числе реакций каждого стимула.
- Изменение количества связей сети, в том числе и с добавлением новых стимулов.

Для оценки характеристик сети при нарастании числа стимулов были произведены сокращения числа стимулов и расчет параметров:

- Количество связей
- Количество нелистьевых связей
- Процентное соотношение нелистьевых связей

Листьевыми будем называть связи, реакция в которых не является стимулов в каких-либо связях.

Таблица 2 Изменение числа стимулов

Процент стимулов	Количество стимулов	Количество связей	Количество нелистьевых связей	Процент нелистьевых связей
10	658	11005	701	6,36
20	1316	23788	3172	13,33
30	1974	36490	7328	20,08

40	2632	48560	12533	25,80
50	3290	60918	19649	32,25
60	3948	74305	28746	38,68
70	4606	86720	38630	44,54
80	5263	97857	49792	50,88
90	5920	109044	62322	57,15
100	6577	122059	78086	63,97

Ниже показана графическая интерпретация данной таблицы.

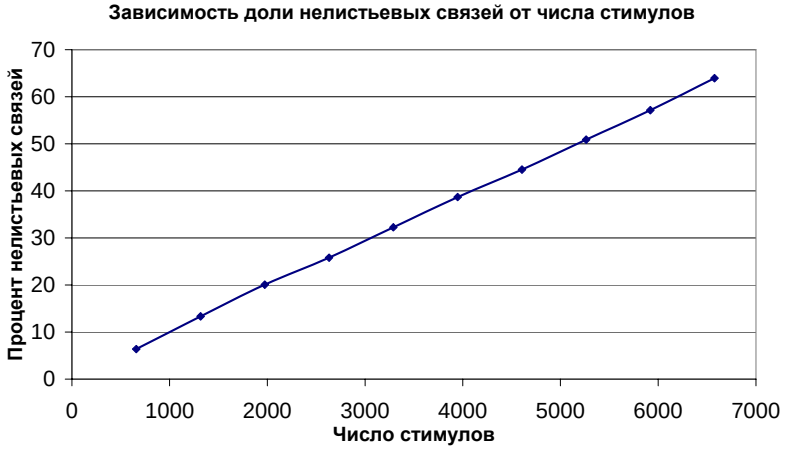

Рисунок 4. Изменение числа стимулов

На графике мы видим почти линейную зависимость процента нелистьевых связей от числа стимулов в диапазоне от 20 до 100 процентов стимулов исходной базы данных. Предложив респондентам новые стимулы, мы можем уменьшить число листьевых элементов сети.

Для оценки характеристик сети при нарастании числа связей были произведены сокращения числа связей и расчет параметров:

- Количество связей
- Количество стимулов
- Количество нелистьевых связей
- Процентное соотношение нелистьевых связей

На таблице ниже представлены результаты расчетов.

Таблица 3 Изменение числа связей

Процент связей	Количество стимулов	Количество связей	Нелистьевые элементы	Процент нелистьевых связей
1	932	1221	366	0,2997543
2	1521	2442	1013	0,414823915
3	1973	3663	1717	0,468741469
4	2343	4884	2491	0,51003276
5	2658	6105	3259	0,533824734
6	2929	7326	4009	0,547229047
7	3176	8547	4772	0,558324558
10	3838	12206	6934	0,568081272
20	5117	24412	14777	0,605317057
30	5740	36618	22706	0,620077557
40	6073	48824	30547	0,625655415
100	6577	122059	78086	0,639739798

На рисунках ниже приведены графические интерпретации содержимого таблицы.

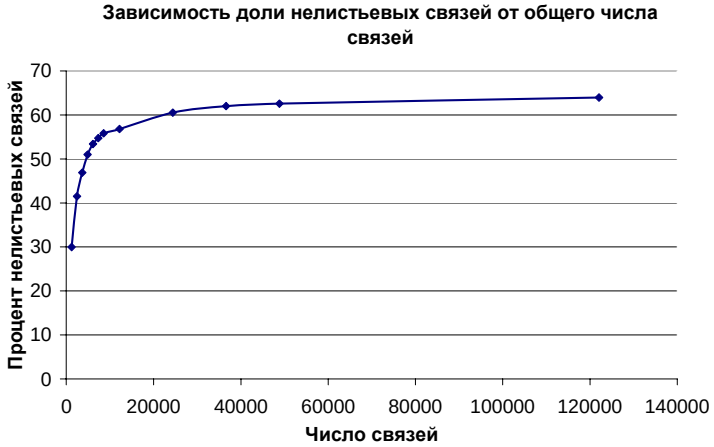

Рисунок 5 Зависимость доли нелистьевых связей

Судя по графику, при данном наборе стимулов, дальнейший опрос респондентов не приведет к значительному снижению доли листьевых элементов сети.

На графике мы видим заполнение ассоциативной сети во время опроса респондентов с точки зрения появления в ней новых стимулов.

Учитывая, что пути длиной более 5 в системе не рассматриваются, а также ограниченность количества реакций на стимул (что позволяет нам считать линейной зависимость между количеством вершин сети и ее связей с точки зрения расчета сложности алгоритма), результирующая сложность составляет $\Theta(k*n^2)$, где k - константа, $k<<n$.

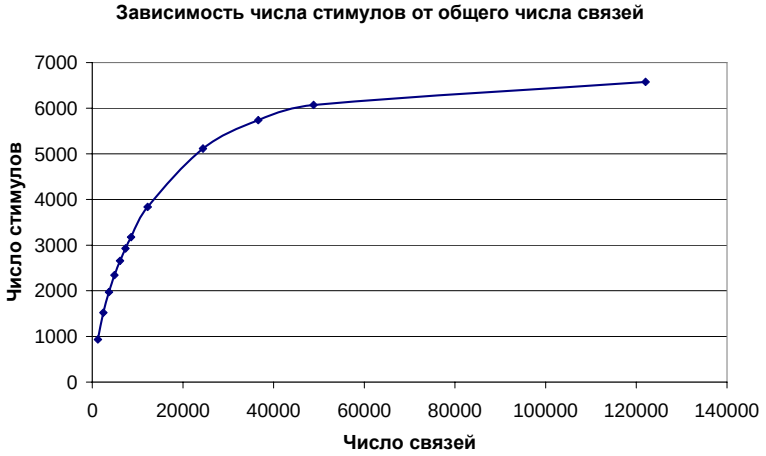

Рисунок 6 Зависимость числа стимулов от числа связей

Приведенный алгоритм обладает низкими требованиями к емкости используемой памяти, так как:

- для обработки вершин используется вызов рекурсивной функции
- поиск осуществляется в глубину
- после завершения работы экземпляра функции происходит освобождение памяти, занимаемой ею.

Приводится алгоритм доступа к базе данных Microsoft Access в случае изменения данных, а также поиска записей и их обработки.

Показан порядок формирования SQL-запроса, осуществляющего выборку фигур знания по критериям, установленным пользователем.

В главе 3 рассматриваются требования программному изделию, и в соответствии с ними обосновывается выбор Microsoft Access как используемой СУБД, а также Borland Delphi в качестве среды разработки.

Выполняется разработка схемы базы данных. Определяются функциональные зависимости схемы базы данных. Затем на их основе осуществляется доказательство оптимальности полученной схемы.

Производится выбор способа доступа к базе данных Access. В качестве альтернатив рассматриваются:

- Доступ через ODBC
- Компоненты DAO
- Технология ADO
- Компоненты KADao

ODBC (Open Database connectivity) – открытое соединение с базами данных – комплекс элементов, обеспечивающих стандартных средств взаимодействия с источниками данных при помощи синтаксиса SQL.

DAO (Data Access Objects) — объекты для доступа к данным — технология доступа к данным компании Microsoft, разработанная для работы с СУБД Microsoft Jet и позволяющая обращаться к базе данных Access.

ActiveX Data Objects — интерфейс программирования приложений для доступа к данным, разработанный компанией Microsoft и основанный на технологии компонентов ActiveX

KADao – компоненты доступа с базе данных Microsoft Access, использующие технологию DAO. Имеет простую систему классов, в значительной степени соответствующих таковой в ADO.

В результате выбор делается в пользу компонент KADao, имеющих простой для разработчика механизм работы, а множество функциональных возможностей.

Показан механизм импорта фигур знания в тезаурус на примере переноса данных их файла Microsoft Excel. Описана процедура обработки импортируемой таблицы, содержащей данные о добавляемых фигурах знания, для внесения их в схему базы данных тезауруса в виде фигур.

Приведены этапы осуществления пассивного режима работы когнайзера с точки зрения пользователя.

Дано описание справочной системы тезауруса и способа обращения к ней.

Глава 4 начинается с описания процесса установки тезауруса на компьютер пользователя.

Затем приводится граф диалога пользователя, отражающий возможные действия и доступ к основным функциям системы.

В главе присутствует описание методологии редактирования фигур знания. Это возможно двумя путями:

- Используя интерфейс главной формы, основанный на пятикомпонентной структуре фигуры знания
- Используя редактирование в табличном режиме

Табличный режим редактирования предназначен для пользователей, которым удобнее работать с данными в виде таблицы, чем со структурированным представлением фигур главной формы. Табличный режим редактирования не позволяет редактировать таблицы областей и способов.

Описан процесс выборки фигур знания по критерию. Необходимые действия для запуска данного режима работы тезауруса, настройки выборки, а также форма представления результата.

Показана работа пассивного режима когнайзера на примере фигуры знания

<Абориген – Коренной житель страны>

Производится поиск знаков ассоциативной сети, представленных в формуле смысла:

- Коренной
- Житель
- Страна

На поиск накладываются ограничения по длине путей: от 1 до 3 дуг.

В результате расчета найдены пути:

1. коренной-> абориген
2. житель-> абориген
3. страна->нищий->студент->абориген

В главе 5 рассмотрены вопросы эргономики и охраны труда оператора ПЭВМ**.** Показаны основные факторы, влияющие на утомляемость оператора. Приведены значения времени регламентированных перерывов в работе согласно СанПиН 2.2.2/2.4-1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы».

Описаны эргономические характеристики рабочего места оператора, схема расположения отдельных элементов, а также психологические требования к рабочему месту.

Произведен расчет освещения рабочего помещения, определено необходимое осветительное оборудование и места его расположения.

Глава 6 содержит организационно-экономическое обоснование разработки.

Разработка системы «Лингвокультурный тезаурус русского языка» состоит из семи стадий:

1) Формирование требований
2) Разработка концепции
3) Техническое задание
4) Эскизный проект
5) Технический проект
6) Рабочая документация
7) Ввод в эксплуатацию

Производится расчет стоимости разработки системы

Таблица 4. Затраты на разработку проекта

Статья затрат	**Стоимость, руб.**
1. Оплата машинного времени	2480
2. Затраты на материалы	2550
3. Затраты на оплату труда – всего	54518
3.1. Основная заработная плата разработчика	36000
3.2. Дополнительная заработная плата	7200
3.3. Единый социальный налог (26%)	11232
3.4. Отчисления в фонд обязательного социального страхования от несчастных случаев на производстве (0,2%)	86
4. Накладные расходы	17280
5. Услуги сторонних организаций	1100
Общая сумма затрат на разработку проекта	77928

В качестве обоснования разработки приводится актуальность работы и создание программного продукта в рамках гранта РФФИ.

Заключение содержит выводы по результатам работы системы, а также пути возможных дальнейших исследований.

Заключение

Основные выводы и результаты работы:

В рамках дипломного проектирования была разработана схема базы данных тезауруса и доказана ее оптимальность

1. Был разработан алгоритм функционирования пассивного режима работы когнайзера, а так же произведена оценка его сложности.
2. Произведен анализ структуры ассоциативной сети тезауруса, моделирующий процесс ее заполнения и сделаны выводы, касающиеся достаточной полноты опроса респондентов.
3. Разработана методология работы с базой данных фигур знания, включающая:
   - Поиск фигур по критерию
   - Редактирование фигур
   - Добавление фигур знания из внешнего источника
4. Произведен расчет эргономических параметров рабочего места оператора.
5. Определены затраты на разработку системы.