На правах рукописи

ФАРРОХБАХТ ФУМАНИ МЕХДИ

ИССЛЕДОВАНИЕ И РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ СМЫСЛОВОЙ ОБРАБОТКИ ТЕКСТОВ В СИСТЕМЕ УПРАВЛЕНИЯ ЭЛЕКТРОННЫМИ АРХИВАМИ

Специальность:
05.13.01- Системный анализ управление и обработка информации (технические системы)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург 2013

Работа выполнена на кафедре автоматизированных систем обработки информации и управления Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В.И.Ульянова (Ленина)

Научный руководитель: кандидат технических наук, доцент
Шеховцов Олег Иванович

Официальные оппоненты: Копыльцов Алесандр Васильевич, доктор технических наук, профессор, Российский государственный педагогический университет, заведующий кафедрой «Информатика»

Назаренко Николай Александрович, кандидат технических наук, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ», доцент кафедры «Биотехнические системы»

Ведущая организация: Санкт-Петербургский университет телекоммуникаций им. проф. М.А. Бонч-Бруевича

Защита состоится "11" февраля 2013 г. в 14 час. на заседании диссертационного совета Д 212.238.07 в Санкт-Петербургском государственном электротехническом университете по адресу: 197376, г. Санкт-Петербург, ул. Профессора Попова, д. 5, корпус 1.

С диссертацией можно ознакомиться в научной библиотеке университета.

Автореферат разослан «____» декабря 2012 г.

Ученый секретарь
диссертационного совета
Д 212.238.07 Цехановский В.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Несмотря на широкое использование мультимедиа, текст остается одним из основных видов информации в большинстве электронных хранилищ. Огромное количество информации скапливается в многочисленных текстовых базах, хранящихся в информационных агентствах, библиотеках, корпорациях, в личных ПК и во всемирной глобальной сети. Объем информации увеличивается с поражающей скоростью и люди не в состоянии решать проблемы, связанные с этим ростом. Ввиду большого роста объемов текстовой информации и сложной структурированности естественно-языковых (ЕЯ) текстов, анализ текстов представляет собой актуальную проблему. Человечество нуждается в интеллектуальных электронных помощниках, которые могут справиться со смысловым анализом текста. Разработка эффективных подходов к обработке текстов с целью фильтрации, формирования смыслового портрета, навигации по базе текстов и т.д. является одним из наиболее актуальных направлений современных информационных технологий. В связи же с практическими потребностями быстрой переработки и поиска информации все более актуальной становится проблема смыслового преобразования текстов. Под преобразованием понимается такой процесс переработки текстов, результатом которого является создание некоторых вторичных текстов, близких по смыслу к исходным, но не заменяющих их полностью. В практическом плане эта проблема заключается в разработке конкретных методов автоматического аннотирования, реферирования, индексирования и др.

В настоящее время в мире существуют и активно развиваются системы смыслового поиска в полнотекстовых базах данных, которые поддерживаются ведущими фирмами - производителями серверов баз данных, например, Oracle, Microsoft, IBM и др. Такие системы строятся на основе многомерных хранилищ, из которых данные извлекаются и обрабатываются с помощью алгоритмов для заранее определенных субъект-объектных отношений между ними. Крупные поисковые серверы в Интернете (например, Google, Yahoo, Yandex) поддерживают алгоритмы поиска текстов “схожих” с данным и расчета релевантности найденных документов исходному запросу. Специализированные системы полнотекстового анализа (например, в России это “Следопыт”, ”ТекстАналист”) позволяют проводить автоматическую классификацию и реферирование текстов.

Классически в основе задачи обработки ЕЯ-текстов лежат морфологический и морфемный анализ, синтаксический и семантический анализ, результатами которых являются модели текста, адекватно отражающие его словообразовательные, грамматические и смысловые конструкции. В этом смысле, основные методы анализа текста достаточно подробно изложены в трудах Апресяна Ю.Д., Виноградова Д.В., Гладкого А.В., Клосса Б.М., Кожуновой О.С., Мельчука И.Д., Солтона Г., Н.С., Филмора Ч., Финна В.К., Шведовой Н.Ю., Дж.Дж.Катца, Дж.А.Фодора**,** Б. Патти, А. Вежбицкой и др. Ряд современных усовершенствованных методов представлены в статьях Ермакова А.Е., Леонтьевой Н.Н., Мозгового М.В., Плешко В.В., Сокирко А., Толпегина П.В., Тузова В.А. и др.

В настоящее время успешно решена задача морфологического анализа текстов, результаты которого применяются в поисковых Интернет-машинах, текстовых редакторах, подсистемах проверки орфографии и пр. Задачи синтаксического и, в особенности, семантического анализа не решены в полной мере. Синтаксический анализ (анализ грамматики) можно встретить в системах перевода, в подсистемах проверки грамматики. Несмотря на богатую теорию в области семантического анализа, применение находят лишь методы анализа основанные на статистических (факторных) характеристиках слов и словосочетаний анализируемого текста. Следует отметить, что подсистемы, реализующие указанные методы анализа текста, не предоставляют средств настройки процесса анализа, средств пополнения баз правил грамматики языка и часто эти подсистемы дают грубые результаты.

Семантические модели (СМ) текста, являющиеся результатом комплексного анализа, позволяют оценить корректность текста, в наглядной форме, визуально представить структуру сюжета, взаимосвязь объектов и процессов текста, их атрибуты. Последовательность моделей простых предложений текста и результирующая визуальная модель текста позволяют

реализовать обратную связь "воздействие на модель - реакция в тексте", благодаря чему можно в интерактивном режиме отлаживать процессы анализа текстов и доказательства объективности (однозначности) истолкования текстов на естественных языках.

Применение семантических моделей актуально в автоматизированных обучающих системах, при решении задач извлечения знаний из текстов, информационного поиска, реферирования, контроля корректности словарей терминов и определений, автоматической генерации ассоциативных связей в гипертекстовых базах данных (ГБД) и пр.

Учитывая вышеизложенное, а также то, что проблема смыслового анализа ЕЯ-текстов до настоящего времени не решена в полной мере, считаем, что совершенствование методов анализа ЕЯ-текстов и повышение степени их достоверности является актуальной задачей.

Разработанность проблемы. Исследования в области автоматической обработки текстов в Европе и США привлекают внимание крупнейших частных фирм и государственных организаций самого высокого уровня. Европейский Союз уже несколько лет координирует различные программы в области автоматической обработки текстов (например, проект IST, 1998-2001 гг.). В США с 1991 по 1998 гг. существовал проект TIPSTER, организованный Департаментом обороны совместно с Национальным институтом стандартов и технологий и Центром военно-воздушных и военно-морских вооружений. В работе консультативного совета этого проекта участвовали также ФБР, Национальный научный фонд и некоторые другие организации. Основной целью проекта было сравнение и оценка результатов работы различных поисковых систем и систем реферирования. По результатам проекта был опубликован подробный обзор и даны рекомендации по использованию этих систем. В США среди систем подобного рода наиболее известной является электронная архивная система “Excalibur RetrievalWare” производства компании Excalibur Technologies. Программные продукты этой компании используются Госдепартаментом, Библиотекой Конгресса, ЦРУ, компаниями Ford Motors, Lockheed, Reynold Electrical & Engineering, Maine Yankee Atomic Power.

Современные системы смыслового анализа текстов, особенностью которых являются: предпочтение скорости обработки текстов, точности семантического и морфологического анализа, выявление смысла текста, реферирование, автоматическое индексирование, эффективная навигация по текстовой базе, статистический частотный анализ словоупотреблений, автоматическая классификация и кластеризация текстов, смысловой поиск и расчет релевантности текстов поисковому запросу.

- OLAP-технологии. OLAP использует многомерное представление совокупных данных, чтобы обеспечить быстрый доступ к стратегической информации для дальнейшего анализа.

  Недостатки: a) функциональность систем ограничивается возможностями SQL, так как аналитические запросы пользователя транслируются в SQL-операторы выборки; b) сложно пересчитывать агрегированные значения при изменениях начальных данных; c) сложно поддерживать таблицы агрегатов; d) сложно изменять измерения без повторной агрегации; e) снижение скорости обработки из-за вычислений по требованию; f) ограничение на объем данных;
- система автоматического анализа текста TextAnalyst разработана в качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов.

  Недостатки: a) не имеет готового словаря русского языка; b) не применяет сколько-нибудь развитых лингвистических средств, например синтаксического и морфологического анализа;
- Oracle InterMedia Text. Одним из наиболее мощных продуктов, позволяющих реализовать поддержку полнотекстовых баз данных с доступом через интернет, является система InterMedia Text в составе СУБД Oracle8i. В InterMedia Text интеллектуальная обработка текста (тематическая классификация, аннотирование) сочетается с поисковыми возможностями, доступными при работе с реляционными базами данных.

Недостатки: a) большинство возможностей InterMedia оказывается доступно в полной мере лишь для английского языка и, в меньшей мере, еще для ряда европейских и восточно-азиатских языков; b) не задействует лингвистические технологии, которые зависят от лексики, грамматики и семантики языка; c) не устанавливает смысловые связи между темами;

- Russian Context Optimizer (RCO). Адаптацией технологий Oracle к русскоязычным базам данных занимаются специалисты компании «Гарант-Парк-Интернет», которая выпускает продукт под названием Russian Context Optimizer (RCO), предназначенный для совместного использования с системой InterMedia Text.
  Основной недостаток - функциональность системы ограничивается возможностями SQL, так как аналитические запросы пользователя транслируются в SQL-операторы выборки;
- Система “Ключи от Текста” - смысловой поиск и индексирование текстовой информации в электронных библиотеках.
  Недостатки: a) большие затраты интеллектуальной работы как при обработке первоисточника, так и при наполнении БД; b) в ней не учитывается коллективный характер использования Сети, а именно то обстоятельство, что ресурсы разделяемы;
- Интеллектуальная система “СЛЕДОПЫТ” помогает быстро находить текстовые фрагменты документов, и предназначена для тех, кто в результате своей деятельности имеет дело с большим объемом информации.
  Недостатки: a) ограничение на объем данных; b) зависит от сторонних программных продуктов, например, MS Office;

Большинство возможностей этих известных систем оказывается доступно в полной мере лишь для английского языка и, в меньшей мере, еще для ряда европейских и азиатских языков. Практически не поддерживают персидского языка.

В настоящее время в России и не только сложилась ситуация, что системы автоматизации управления корпоративными электронными архивами не поддерживают технологии автоматизированного смыслового анализа текстов, а современные системы анализа текстов не адаптированы к работе с электронными текстовыми архивами корпорации. Необходима разработка алгоритмов и методики автоматизированной смысловой обработки текстов и реализация программно-технического комплекса для внедрения смыслового полнотекстового анализа в технологию обработки электронных архивов. Данный комплекс также должен поддерживать персидский и другие азиатские языки.

Исходя из всего, что сказано выше, в данном диссертационном исследовании были сформулированы:

Объект исследования работы - математическое, информационное и программное обеспечение человеко-машинного взаимодействия на естественном языке.

Предмет исследования - модели, методы и алгоритмы смыслового анализа естественно-языкового текста.

Цель работы - исследование, разработка и научно-практическое обоснование алгоритмов и методики автоматизированной смысловой обработки текстов и внедрение их в технологию обработки текстов в системе управления электронными архивами.

Для достижения поставленной цели требуется решение следующих **основных научных и практических задач:**

1. Аналитический обзор существующих методов и систем анализа ЕЯ-текстов.
2. Исследование и разработка архитектуры автоматизированной системы смысловой обработки текстов, а также принципов смыслового анализа текстов.
3. Исследование и разработка онтологии предметной области **«смысловая обработка текстов на естественном языке»** и правил логического вывода как информационной основы построения системы с целью хранения и извлечения знаний о грамматиках естественных языков и о предметной области текста, а также выявления основных направлений снижения трудоемкости при проектировании алгоритмов смыслового анализа текстовой информации.

4. Разработка методов (статистических методов предварительного смыслового анализа текста, методики построения пересечения онтологий) и алгоритмов смыслового анализа текстов (алгоритм поиска, классификации, кластеризации, реферирования и т.д.), базирующихся на онтологиях ЕЯ.

5. Программная реализация автоматизированной системы комплексного смыслового анализа текстов и экспериментальное исследование предложенных методов и алгоритмов.

Методы исследования. Теоретические исследования выполнены с использованием моделей и методов системного анализа, статистического анализа, онтологического инжиниринга, теории множеств, семантических сетей, математической логики, теории проектирования баз данных. При разработке программного обеспечения использовались технологии объектно-ориентированного программирования и семантического web.

Достоверность и обоснованность полученных в работе результатов и выводов подтверждается корректным использованием математического аппарата и положительными результатами проведенных экспериментальных исследований.

Научная новизна.

1. Предложена архитектура автоматизированной системы смысловой обработки текстов.
2. Разработана онтология предметной области **«смысловая обработка текстов на естественном языке»**, включающая декларативные и императивные знания о грамматиках естественных языков и правила вывода с применением языка логики предикатов первого порядка.
3. Разработаны методы (взвешивания термов, взвешивания предложений, взвешивания абзацев, взвешивания разделов текста, взвешивания отношений между понятиями, оценки степени смысловой близости текстов) и алгоритмы (определения пересечения онтологий текстов, классификации текстов, кластеризации текстов, поиска по ключевым словам, смыслового поиска, реферирования текста) смыслового анализа ЕЯ-текстов.

Степень новизны полученных результатов.

1. Архитектура отличается от известных автору тем, что ее ядро основано на уникальной впервые созданной онтологии естественного языка, и способах извлечения из заданных текстов, соответствующих им онтологий; а также на уникальной методике определения пересечения онтологий текстов.
2. Онтология предметной области **«смысловая обработка текстов на естественном языке»** предложена впервые и не имеет известных автору аналогов
3. Методика определения пересечения онтологий текстов также не имеет известных автору аналогов. Все реализованные алгоритмы смыслового анализа ЕЯ-текстов основаны на данной методике, поэтому они в свою очередь также являются уникальными.

Практическая полезность. Проведение смысловой обработки ЕЯ-текстов по предложенной технологии позволит облегчить процесс их обработки, повысить доверие к результатам обработки, снизить издержки на обработку, обеспечить дальнейшее развитие систем смысловой обработки ЕЯ-текстов. Кроме того, результаты, полученные в работе, окажут положительное влияние на конгломерацию частных систем смысловой обработки ЕЯ-текстов в общую систему смысловой обработки ЕЯ-текстов. Также практическая значимость исследования заключается в возможности использования предложенных методов и алгоритмов смысловой обработки ЕЯ-текстов для повышения эффективности систем управления электронными архивами.

На защиту выносятся:

1. Архитектура автоматизированной системы смыслового анализа текстов.
2. Онтология предметной области **«смысловая обработка текстов на естественном языке»**, включающая декларативные и императивные знания о грамматиках естественных языков и правила вывода.
3. Методы и алгоритмы смыслового анализа ЕЯ-текстов.

Реализация результатов работы. Результаты работы использованы на кафедре «САПР» в преподавании дисциплины «Онтологический инжиниринг» для магистрантов направления «Информатика и вычислительная техника». Получено 2 акта о внедрении (использовании) результатов диссертационной работы.

Апробация работы. Основные результаты диссертационной работы докладывались и обсуждались на следующих конференциях и семинарах:

Материалы 63-й научно-технической конференции профессорско-преподавательского состава СПбГЭТУ. 2011.

Публикации. Основные теоретические и практические результаты диссертации опубликованы в 7 публикациях, включая 3 в изданиях, рекомендуемых ВАК, 3 статьи в международных журналах, 1 – материалы научно-технической конференции.

Структура и объем работы. Диссертационная работа состоит из введения, пяти глав, заключения и приложений. Основной текст изложен на 126 машинописных страницах с иллюстрациями. Список литературы включает 34 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность работы, сформулированы цель работы и перечень решаемых задач. Кратко изложено содержание работы, сформулированы научная новизна и практическая полезность.

В первой главе «Автоматизированная система смысловой обработки текстов» проведен анализ существующих программных продуктов обработки ЕЯ текстов. Указаны их достоинства и недостатки. Описано общее понимание смысловой обработки ЕЯ текстов. Сделан вывод о необходимости расширять возможности существующих систем автоматизации управления электронными архивами за счет включения в них средств смысловой обработки текстов. Рассмотрены:

а) базовые понятия смысловой обработки ЕЯ текстов (индексация, структуризация информации, формализация представления данных, классификация текстов, кластеризация текстов, смысловой поиск, реферирование, фрагментация текстов, формирование гипертекста, семантическая сеть текста, таксономия текста, онтология текста);

б) основные этапы смысловой обработки ЕЯ текстов (синтаксический анализ, семантический анализ, статистический анализ, выделение классов, определение отношений, формирование семантических сетей, формирование таксономий, получение онтологии);

в) основные методы (законы Ципфа; взвешивание термов, предложений и отношений) смыслового анализа текстов в полнотекстовых базах данных.

Введены и определены понятия: онтология естественного языка, онтология текста на естественном языке.

Во второй главе предложены структурная Рис. 1 и функциональная Рис. 2 схемы автоматизированной системы смысловой обработки текстов (SemTextProcessor).

В состав структурной схемы входят следующие подсистемы: **Управляющая подсистема** - предназначена для управления процессом смысловой обработки текстов; **Полнотекстовая база данных** – ориентирована на хранение текстовых документов в UTF-8 кодировке; **Подсистема ведения полнотекстовой базы данных** - предназначена для занесения, удаления и обновления информации о текстовых документах; **Подсистема синтаксического анализа** текстов - предназначена для выделения основных форм слов, составляющих текст, и их принадлежность к частям речи; **Подсистема семантического анализа** текстов - предназначена для обнаружения связи между словами, обусловленные конструкцией предложений; **Подсистема статистического анализа** текстов - предназначена для подготовки численных 2-мерных таблиц для методов анализа с целью классификации текстов; статистический анализ также используется для формализации задачи смысловой обработки текстов; Эта подсис-

тема дополняет лингвистические подсистемы и влияет на формирования семантической сети. **Подсистема формирования семантической сети** - предназначена для формирования множества понятий текста - слов и словосочетаний, связанных между собой по смыслу; **Подсистема формирования таксономии и онтологии** - предназначена для создания иерархии связанных тем и подтем, раскрывающих содержание тем; здесь формируется онтология текста. **Подсистема визуализации** - предназначена для отображения в удобном пользователю виде всей полученной из текста (текстов) информации - взаимосвязь и степень близости слов и групп текстов; **Подсистемы реализации алгоритмов обработки** – предназначены для реализации алгоритмов извлечений знаний и других манипуляции ими. Это алгоритмы классификации, реферирования и т.д.

Рис. 1

Рис. 2

Основой функционирования всей системы является полнотекстовая база данных, источниками информации для которой выступают корпоративные электронные архивы.

Система SemTextProcessor обеспечивает следующие функциональные возможности: **Выявление смысла текста** - формирование и экспорт точной семантической сети текста или текстовой базы. Р**еферирование** - качество краткого изложения текста обеспечивается сбалансированным сочетанием методов исследования лингвистической сети, статистических параметров и онтологии. **Исследование текстов по заданной тематике. Эффективная навигация по текстовой базе. Пояснение структуры основы текста** - создание древовидной структуры тематик и онтологии, представляющих семантику исследуемого текста. **Кластеризация текстов. Семантический поиск информации** - анализ запросов естественного языка на наличие важных слов и извлечение релевантных предложений из базы данных текстовых документов. В дополнение, формируется поддерево понятий, что детализирует поиск. **Машинный перевод текстов** - процесс перевода текстов с одного естественного языка на другой. **Автоматическое индексирование** - индексирование, технология которого предусматривает использование только формальных процедур обработки текста, осуществляемых с помощью вычислительной техники.

Глава также посвящена формализации представления данных. Рассматриваются два вида формализации:

а) Статистическая формализация представления данных заключается в построении матрицы объект-атрибут, где объектами будут исходные тексты, атрибутами - слова. Элементом матрицы является число словоупотреблений (или его логарифм, причем, при отсутствии слова логарифм считается равным - 1).

Рассчитываются следующие 2-мерные числовые матрицы:

1. Матрица текст/слово, элементы которой это число повторений данного слова в данном тексте.
2. Первая матрица слово/слово, элементы которой это число повторений данной пары слов в данном тексте.

3. Вторая матрица слово/слово, элементы которой это число текстов, содержащих данную пару слов.
4. Матрица текст/текст, элементы которой это число слов, встречающихся в данной паре текстов.

Дальше структуризация этой информации о распределении слов в текстах в числовом виде выполняется с помощью алгоритмов.

b) Онтологическая формализация представления данных. Она подробно излагается в главе 3.

Нужно отметить, что статистические оценки дополняются лингвистическими данными, которые хранятся в онтологии предметной области системы.

Матрицы, указанные выше рассчитываются по аналоги с латентным семантическим анализом (ЛСА).

В этой главе также приводятся общий процесс смыслового анализа текста рис. 3 и методика смыслового анализа текстов, сочетающая статистические, лингвистические и онтологические способы анализа текстов и которая состоит из следующих этапов:

Этап 1. Построение словаря терминов - обозначений «концептов» предметной области. Данный этап выполняется двумя подходами:

1. Лингвистический подход (рис. 4).
2. Статистический подход (рис. 5).

Этап 2 (рис. 6). Расширение словаря терминов именами ситуаций и свойств объектов предметной области. Формирование онтологии.

Этап 3 (рис. 7). Описание способов выражения отношений из онтологии в языке – типовых лексико-грамматических конструкций.

Этап 4 (рис. 8). Формальное описание онтологии на языке OWL-DL и генерирование запросов для извлечения знаний из онтологии.

Процесс смыслового анализа текста выполняется для каждого текста.

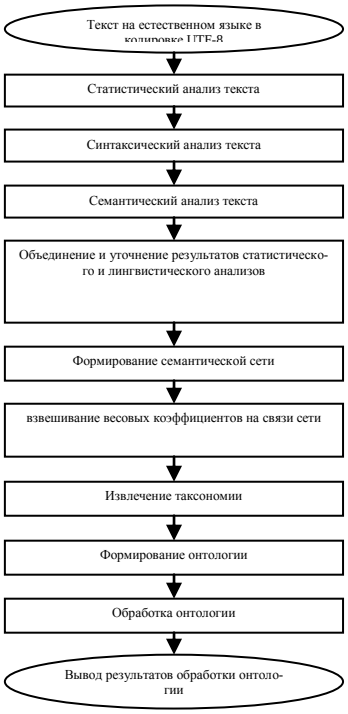

Рис. 3

Рис. 4

Для каждого предложения текста производится синтаксический анализ с получением дерева синтаксических зависимостей между составляющими предложения. Дерево зависимостей преобразуется в сеть синтактико-семантических отношений. На основе обхода сети синтактико-семантических отношений производится синтез термино-подобных словосочетаний. Для каждого термина словаря производится подсчет его полной и независимой частоты встречаемости. Отношение полной и независимой частот встречаемости позволяет учесть

иерархию смыслов, которая выражается в уровне синтаксической зависимости одних элементов словосочетаний от других. В итоге, те слова и словосочетания, для которых отношение величин «частота независимой встречаемости» (не в составе других словосочетаний) и «полная частота встречаемости» оказывается близко к нулю, могут быть отброшены как неполные части устойчивых терминов. Далее производится иерархическая группировка элементов словаря на основе лексической вложенности слов и словосочетаний. В конце проверяем и уточняем полученный словарь терминов, в том числе фиксация синонимичных обозначений одних и тех же объектов.

Алгоритмы смысловой обработки базируются на числовом анализе частотного распределения ключевых слов, выбранных из заданного массива текстов. Известно, что это распределение описывается эмпирическим законом Зипфа.

Модель представления данных заключается в построении матрицы объект-атрибут, где объектами будут исходные тексты, атрибутами - слова. Элементом матрицы является число словоупотреблений (или его логарифм, причем, при отсутствии слова логарифм считается равным - 1).

Рис. 5

Рис. 6

Рис.7

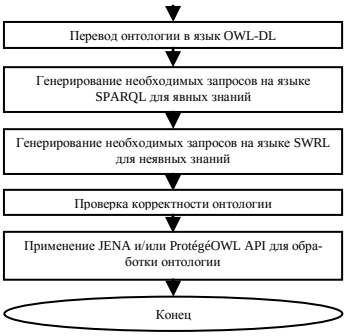

Рис. 8

Теперь объединяем плюсы статистического и лингвистического подходов. Для каждого ранее зафиксированного термина-объекта предметной области - поиск слов (словосочетаний), связанных связями типа «объект-атрибут» и «объект-ситуация», на основании шаблонов, задающих соответствующие конфигурации синтаксических связей. Потом формируем общий словарь терминов – объекты, их атрибуты и ситуации с их участием. Результирующий словарь представляет собой семантическую сеть взаимосвязанных сущностей трех классов, вход в которую возможен от частотного словаря имен объектов, атрибутов или ситуаций, а переход по связям между сущностями сопровождается возможностью просмотра текста, в котором связь раскрывается. Исследуем семантическую сеть и окончательное формирование концептуальной составляющей онтологии (шаг 1 процесса онтологического инжиниринга) - определение абстрактных понятий (классов объектов, их свойств и ситуаций) с определением типизированных отношений между сущностями этих классов; окончательное формирование фактического наполнения онтологии (шаг 2 процесса онтологического инжиниринга) - соотнесение всех терминов словаря с понятиями в схеме онтологии, в том числе фиксация синонимичных обозначений свойств и ситуаций, определение возможных иерархических отношений между сущностями одного класса.

Выявление множества ассоциативно-статистических связей между всеми терминами текста, для которых существует связь в онтологии. Ассоциативно-статистическая связь уста-

навливается между терминами, совместно упоминавшимися в предложениях текста не менее заданного числа раз. Далее построение списков цитат из текста для каждого типа связей из онтологии, с предварительным отсевом статистически малодостоверных связей и тех связей, которые выражаются уже известными способами и могут быть выделены на основании синтаксических шаблонов. Для выполнения алгоритмов смысловой обработки текстов, онтологию нужно представить в виде пригодном для машинной обработки. Такой вид выбран язык дескриптивной логики OWL-DL.

Третья глава посвящена инженерией знаний предметной области **«смысловая обработка текстов на естественном языке»**, результатом которой получение онтологическую базу знаний. Работа по конструированию базы знаний производится на нескольких этапах с применением различных формализмов представления знаний.

На первом этапе разрабатываются семантические сети, отражающие основные понятия предметной области и отношения между ними.

1. Основная семантическая сеть системы

Основная семантическая сеть представлена на рис. 10.

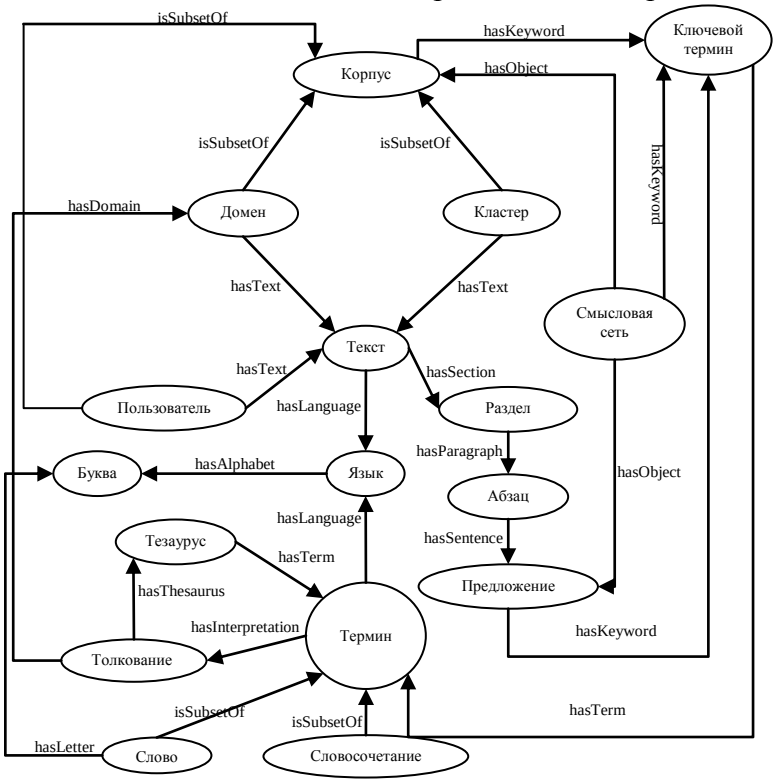

Рис. 10

2. Семантическая сеть понятия «слово»

Семантическая сеть «слово» представлена на рис. 11.

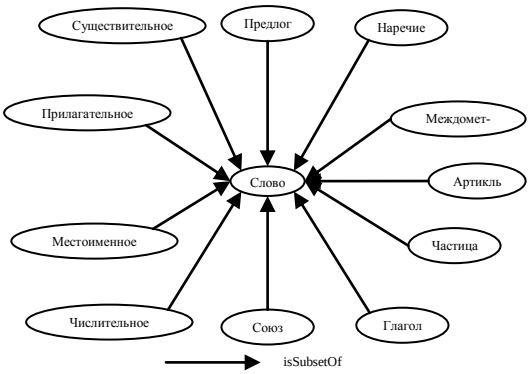

Рис. 11

3. Семантическая сеть понятия «прилагательное»

Семантическая сеть «прилагательное» представлена на рис. 12.

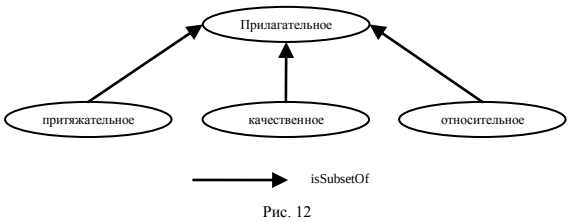

Рис. 12

4. Семантическая сеть понятия «местоименное»

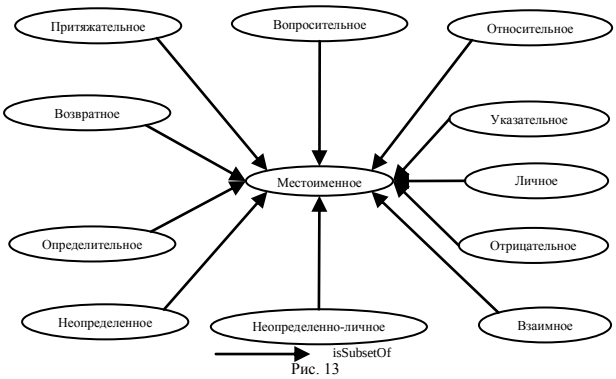

Рис. 13

Семантическая сеть «местоименное» представлена на рис. 13.

На втором этапе, исходя из анализа семантических сетей, мы построили древовидные иерархические структуры (таксономии) важных понятий (древовидные иерархии терминов) предметной области. В таксономии основным отношением является отношение подчиненности (наследования), т.е. класс-подкласс. Основная таксономия представлена на рис. 14.

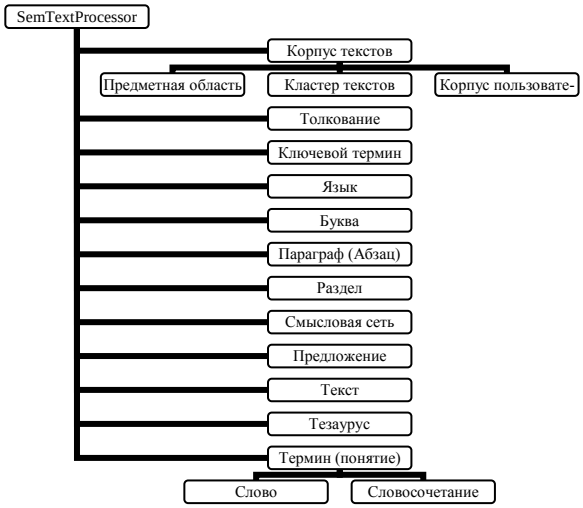

Рис 14.

На рис. 15. представлена таксономия понятия «слово».

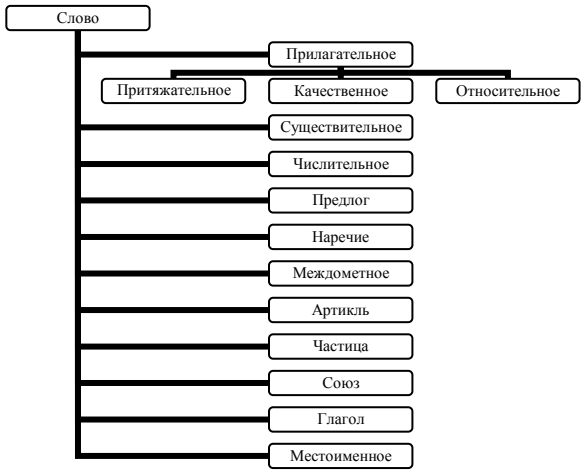

Рис 15.

На рис. 16. представлена таксономия понятия «местоимение».

Рис 16.

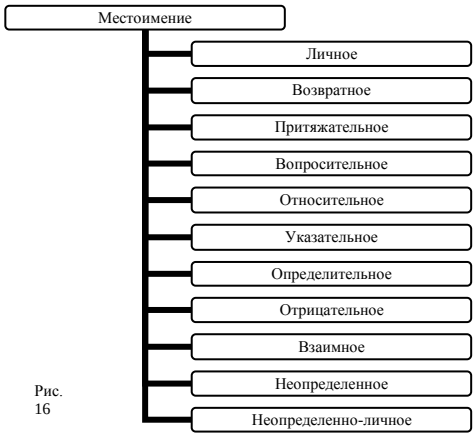

Рис. 16

На следующем этапе были разработаны фреймовые модели предметной области (см. главу 3 в диссертации).

На последнем этапе для выполнения рассуждения над базой знаний (онтологией) и логического вывода были разработаны с применением аппарата логики первого порядка и применены следующие правила логического вывода:

Правило смыслового поиска по запросу

Вычисление степени пересечения семантической сети запроса с семантическими сетями текстов дает возможность *отранжировать тексты по степени близости (релевантность и пертинентность) к запросу.*

Прежде всего, нужно указать, что любой текст может содержаться или не содержаться в результате поиска. Логически это пишется так:

$$\forall t, Text(t) \wedge (output(t) = TRUE \vee output(t) = FALSE),$$

$$TRUE \neq FALSE$$

Теперь определим собственно логическое правило смыслового поискового вывода текстов по заданному пользователем запросу.

$$\forall t, q, th, Text(t) \wedge Query(q) \wedge Threshold(th) \wedge$$

$$(\text{int}\, ersec tionLevel(SemanticNet(q), SemanticNet(t)) \geq th) \Rightarrow output(t) = TRUE$$

Правило классификации текстов по предметной области

Вычисление степени пересечения семантической сети текста с семантическими сетями рубрик (доменов) позволяет *автоматически отнести входной текст к одной или нескольким рубрикам*, то есть – отклассифицировать его.

Логическое правило смысловой классификации текстов по домену (предметной области) с учетом порога степени пересечения семантических сетей, заданного пользователем.

$$\forall t, d, th, Text(t) \wedge Domen(d) \wedge Threshold(th) \wedge$$

$$(\text{int}\, ersec tionLevel(SemanticNet(d), SemanticNet(t)) \geq th) \Rightarrow t \in d$$

Вероятность больше 0 и меньше или равна 1.

$$\forall p, \Pr obability(p) \wedge (p > 0) \wedge (p \leq 1),$$

$$1 \neq 0$$

Логическое правило смысловой классификации текстов по домену (предметной области) без порога степени пересечения семантических сетей. В этом случае текст классифицируется по доменам с определенной вероятности.

$$\forall t, d, p, Text(t) \wedge Domen(d) \wedge \Pr obability(p) \wedge$$
$$(p = get\Pr obability(\text{int } er\sec tionLevel(SemanticNet(d), SemanticNet(t)))) \wedge$$
$$(p > 0) \Rightarrow (t \in d) \wedge textDomen\Pr obability(t, d) = p)$$

Правило смысловой кластеризации текстов

Вычисление степени пересечения семантической сети одного текста с семантическими сетями других текстов дает возможность *отранжировать тексты по степени их близости.*

Логическое правило смысловой кластеризации текстов.

$$\forall t_2, c, p, \exists t_1, Text(t_1) \wedge Text(t_2) \wedge Cluster(c) \wedge empty(c) \wedge \Pr obability(p) \wedge (t_1 \in c) \wedge$$
$$(p = get\Pr obability(\text{int } er\sec tionLevel(SemanticNet(t_1), SemanticNet(t_2)))) \wedge$$
$$\Rightarrow ((p > 0) \wedge (t_2 \in c) \wedge textDomen\Pr obability(t_2, c) = p)) \vee$$
$$((p = 0) \wedge (\exists c_2, Cluster(c_2) \wedge empty(c_2) \wedge (c_2 \neq c) \wedge (t_2 \in c_2)))$$

Правило формирования реферата

SemTextProcessor может автоматически создавать реферат, который, в составе наиболее значимых предложений текста, позволяет осуществить первичное и быстрое знакомство с текстом.

Логическое правило формирования реферата текста.

$$\forall t, s, w, Text(t) \wedge Sentence(s) \wedge (s \in t) \wedge Number(w) \wedge$$
$$(hasWeight(s) > w) \Rightarrow addToSummary(s)$$

,

где **w** – среднеарифметическое значение весов предложений текста, т.е.

$$w = \frac{1}{N} \sum_{i=1}^{N} hasWeight(s_i)$$

$$hasWeight(s) > w$$

В четвертой главе вводятся понятия онтология естественного языка и онтология текста на естественном языке. Описываются разработанные статистические методы и алгоритмы смысловой обработки текстов. Для взвешивания термов ЕЯ текста используется метод tf (term frequency, частота терма) - вес определяется как функция от количества вхождений терма в документе; Вес конкретного предложения текста определяется следующей формулой:

$$sw = \frac{ns}{NS} + \frac{nk}{NK} + \sum_{i=1}^{nk} n_i * w_i ,$$

где

- sw - вес предложения,
- ns - число вхождений данного предложения в текст,
- NS - общее число предложений в данном тексте,
- nk - число ключевых термов в данном предложении,
- NK - общее число ключевых термов в тексте,
- n_i - число вхождений i-ого ключевого терма в данное предложение,
- w_i - вес i-ого ключевого терма.

Вес абзаца в тексте определяется следующей формулой:

$$pw = \frac{np}{NP} + \frac{nsk}{NSK} + \sum_{i=1}^{nsk} m_i * sw_i ,$$

где

- pw – вес абзаца,
- np - число вхождений данного абзаца в текст,
- NP - общее число абзацев в данном тексте,
- nsk – число ключевых предложений в данном абзаце,
- NSK – общее число ключевых (важных) предложений текста,
- m_i - число вхождений i-ого ключевого предложения в данный абзац,
- sw_i – вес i-ого ключевого предложения.

Вес конкретного раздела (подраздела) текста определяется следующей формулой:

$$\sec w = \frac{npk}{NPK} + \sum_{i=1}^{npk} l_i * pw_i ,$$

где

- secw – вес раздела (подраздела),
- npk – число ключевых абзацев в данном разделе (подразделе),
- NPK – общее число ключевых (важных) абзацев текста,
- l_i - число вхождений i-ого ключевого абзаца в данный раздел (подраздел),
- pw_i – вес i-ого ключевого абзаца.

Вес связи между двумя понятиями определяется следующей формулой:

$$r = \frac{nr * (w_1 + w_2)}{NR} + \frac{NR}{MR} + \frac{F}{F_1} + \frac{F}{F_2},$$

где

- r - вес отношения для данных двух понятий,
- nr - число встречаемости в тексте данной пары понятий с данным отношением,
- NR - общее число встречаемости данного отношения в тексте,
- MR - общее число отношений в тексте,
- w_1 - вес первого понятия,
- w_2 - вес второго понятия,
- F - частота совместной встречаемости этих двух понятий по любому отношению в тексте,
- F_1 - частота встречаемости первого понятия в тексте,
- F_2 - частота встречаемости второго понятия в тексте.

Для предварительной статистической обработки текста, определяются следующие двухмерные матрицы:

Для каждого слова подсчитывается число словоупотреблений в каждом тексте. Эти данные организуют матрицу ***текст/слово*** (***TW***).

$$TW = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}$$

Так как может быть представлено значительное число текстов, работать с матрицей текст/слово может быть затруднительно. Поэтому матрица текст/слово служит только для расчета матрицы ***слово/слово*** (***WW1***).

$$WW1 = \begin{bmatrix} y_{1,1} & \cdots & y_{1,n} \\ \vdots & \ddots & \vdots \\ y_{m,1} & \cdots & y_{m,n} \end{bmatrix}$$

Группы слов, организованные темами в матрице текст/слово, проявятся в матрице слово/слово как блоки, симметричные относительно главной диагонали.

Матрица слово/слово положительная, симметричная и имеет по диагонали 1. Смысл недиагональных элементов заключается в том, сколько раз встретилась данная пара слов во всех текстах. Также рассчитывается ***вторая форма матрицы слово/слово*** (***WW2***), элементы которой это число повторений данной пары слов в данном тексте.

$$WW2 = \begin{bmatrix} z_{1,1} & \cdots & z_{1,n} \\ \vdots & \ddots & \vdots \\ z_{m,1} & \cdots & z_{m,n} \end{bmatrix}$$

Рассчитывается матрица ***текст/язык*** (***TL***), элементом которой служит процент слов данного ЕЯ в данном тексте. Эта матрица не симметричная и имеет существенно меньшую размерность, чем матрица текст/слово. Группы текстов в данной матрице должны совпадать с группами текстов в матрице текст/слово.

$$TL = \begin{bmatrix} u_{1,1} & \cdots & u_{1,n} \\ \vdots & \ddots & \vdots \\ u_{m,1} & \cdots & u_{m,n} \end{bmatrix}$$

Для статистической группировки текстов рассчитывается также матрица ***текст/текст*** (***TT***). Матрица текст/текст симметричная положительная и имеет по диагонали 1, а недиагональные элементы меньше 1, смысл которых в том, какой процент из слов одного текста встретился в другом тексте. Так как матрицы текст/текст и слово/слово являются производными от матрицы текст/слово их поведение в целом совпадают.

$$TT = \begin{bmatrix} v_{1,1} & \cdots & v_{1,n} \\ \vdots & \ddots & \vdots \\ v_{m,1} & \cdots & v_{m,n} \end{bmatrix}$$

В главе предложен следующий алгоритм определения пересечения онтологий текстов:

1. Строится пересечения терминов (с учетом знаний об этих терминах из онтологии ЕЯ) двух онтологий $T(O) = T(O_1) \cap T(O_2)$.
2. Если пересечение T(O) не пусто, то уточним множество отношений R(O) между терминами из пересечения T(O) с использованием знаний из онтологии ЕЯ, O_1 и O_2.
3. Если пересечение T(O) не пусто, то для каждого термина t из этого пересечения строятся два множества T_{t1} и T_{t2} - термины, которые связанные с ним в каждой онтологии любыми отношениями. $T_{t1} = \{x \mid x \in T(O_1) \wedge (\exists r \in R(O_1)) \wedge xrt\}$, $T_{t2} = \{x \mid x \in T(O_2) \wedge (\exists r \in R(O_2)) \wedge xrt\}$.
4. Для каждого термина t из пересечения T(O) строится пересечение I_t множеств T_{t1} и T_{t2} (с учетом знаний о терминах в T_{t1} и T_{t2} из онтологии ЕЯ). $I_t = T_{t1} \cap T_{t2}$

5. Анализ и установка типов отношений между терминами из T(O) и I_t (отношения могут быть иерархические, синонимические, атрибутивные, производные и т.д.) с учетом знаний из онтологии ЕЯ.

Степень пересечения (близости) онтологий текстов определяется формулой:

$$l = \begin{cases} \dfrac{\left(\dfrac{N}{N1}+\dfrac{R}{R1}\right)}{\left(\dfrac{N}{N2}+\dfrac{R}{R2}\right)} * 100\%, при \left(\dfrac{N}{N1}+\dfrac{R}{R1}\right) \le \left(\left(\dfrac{N}{N2}+\dfrac{R}{R2}\right) \ne 0\right) \\ \dfrac{\left(\dfrac{N}{N2}+\dfrac{R}{R2}\right)}{\left(\dfrac{N}{N1}+\dfrac{R}{R1}\right)} * 100\%, при \left(\dfrac{N}{N2}+\dfrac{R}{R2}\right) < \left(\left(\dfrac{N}{N1}+\dfrac{R}{R1}\right) \ne 0\right) \end{cases},$$

где

- l - степень пересечения;
- N1 - общее число терминов первой онтологии;
- N2 - общее число терминов второй онтологии;
- N - общее число терминов в пересечении онтологий;
- R1 - общее число отношений первой онтологии;
- R2 - общее число отношений второй онтологии;
- R - общее число отношений в пересечении онтологий;

Если полученный коэффициент выше определенного пользователем коэффициента доверия (по умолчанию пресечение должно содержать не менее 50% терминов одной из онтологий), то считается, что эти онтологии по смыслу близки.

Также предложены следующие алгоритмы:

Алгоритм смыслового поиска по запросу

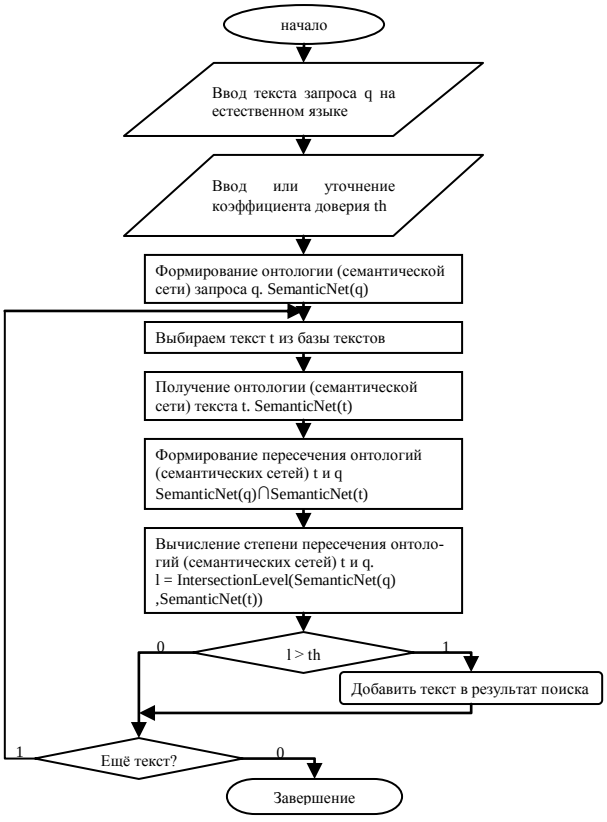

Рис. 17.

Алгоритмы классификации текстов по предметным областям

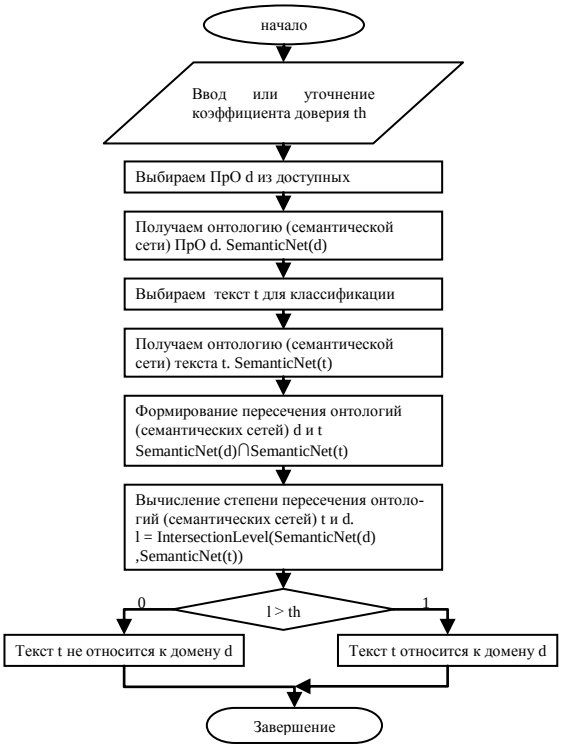

Рис. 18.

На рис. 19 представлен алгоритм классификации текста по доменам (ПрО) с определенными вероятностями.

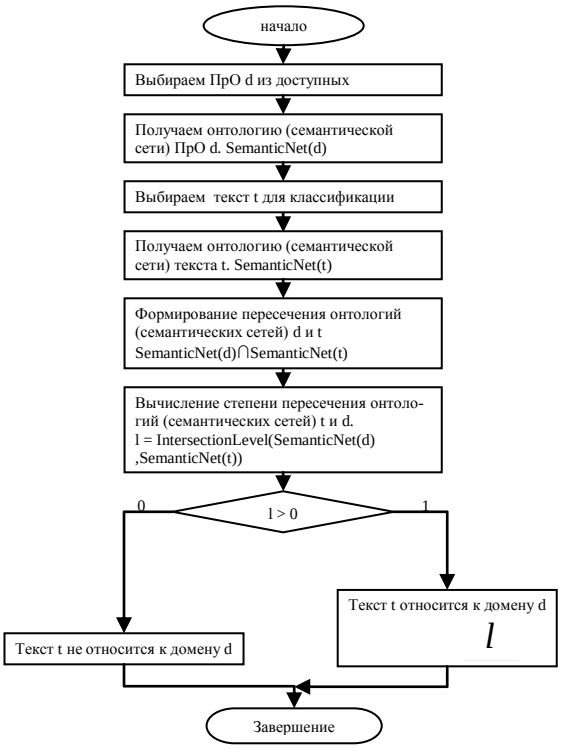

Рис. 19.

Алгоритм кластеризации текстов

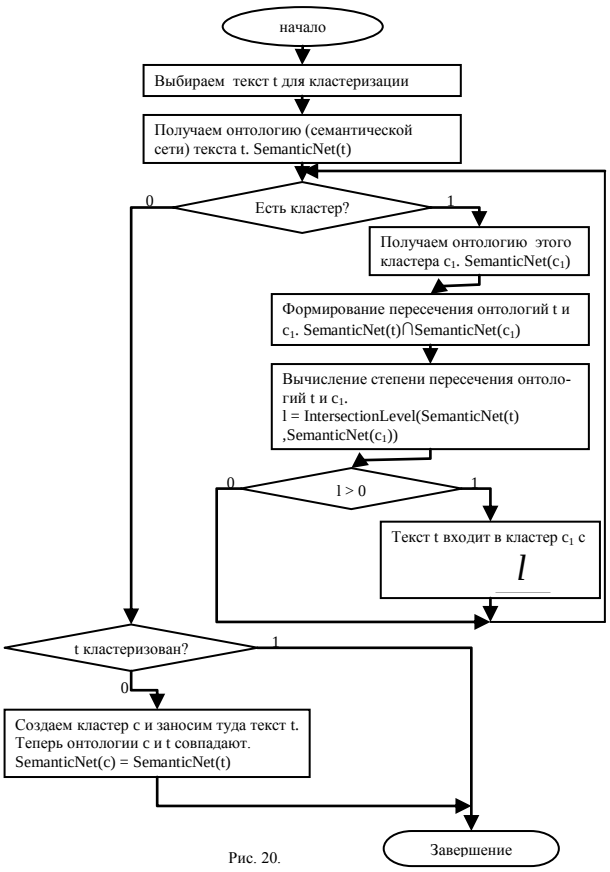

Рис. 20.

Алгоритм реферирования текста

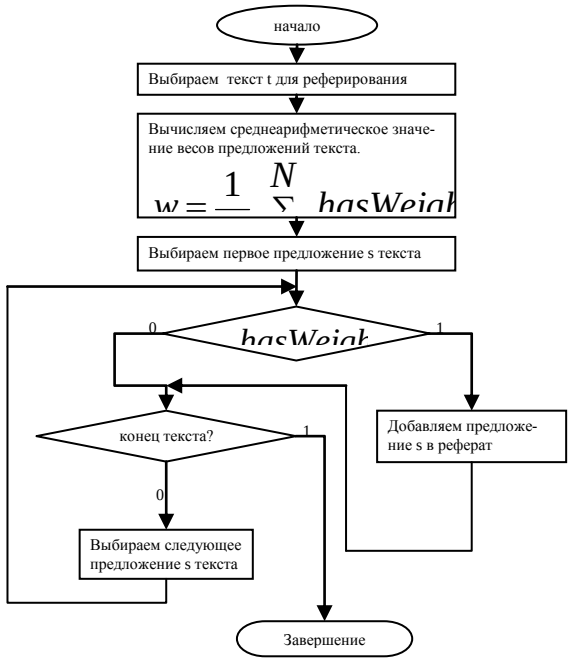

Рис. 21.

В пятой главе описываются проектирование и управление онтологией в среде Protégé, программная реализация и внедрение разработанных структур, методов и алгоритмов построения автоматизированной системы смысловой обработки текстов на естественном языке. Программа полностью реализована на языке Java.

Система реализует объектную модель системы управления онтологией, описанной на языке OWL-DL, и позволяет:

- получать доступ к онтологиям с использованием технологий Jena и ProtégéOWL API;
- хранить файлы с описаниями онтологий на языке OWL-DL в файловой системе;
- выполнять запросы к онтологии ЕЯ для извлечения явных знаний;
- выполнять запросы к онтологии ЕЯ для извлечения неявных знаний, т.е. осуществлять логический вывод.

Главная форма автоматизированной системы смысловой обработки текстов на естественном языке представлена на рис. 22. Она позволяет: полностью моделировать предметную область; управлять онтологией (посмотреть, создать, удалить и изменить классы или концепты,

свойства или роли, фасеты или ограничения); управлять таксономиями; посмотреть, создать, удалить и изменить экземпляры; сохранить изменения в онтологию; и т.д.

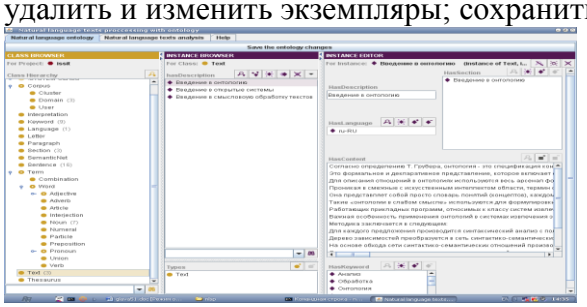

Рис. 23.

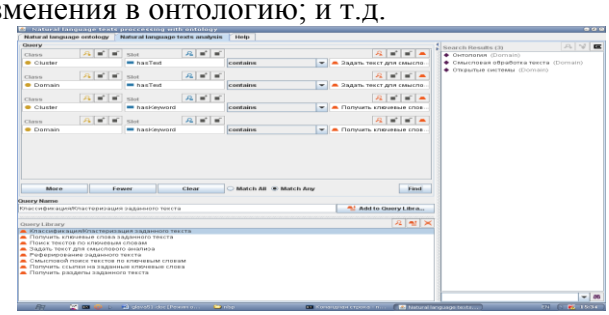

Рис. 22.

На рис. 23 представлена форма выбора и запуска вида смысловой обработки ЕЯ-текстов. Она позволяет осуществить: обычный поиск по ключевым словам; смысловой поиск текстов по ключевым словам; классификацию выбранных текстов; кластеризацию выбранных текстов; реферирование заданного текста; и другие промежуточные задачи.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Основной результат работы заключается в совершенствовании технологий смыслового анализа естественно-языкового текста. Полученные результаты относятся к направлению исследований «Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации». В работе получены следующие выводы и результаты:

1. Архитектура автоматизированной системы смыслового анализа текстов.
2. Онтология предметной области **«смысловая обработка текстов на естественном языке»**, включающая декларативные и императивные знания о грамматиках естественных языков и правила вывода.
3. Методы и алгоритмы смыслового анализа ЕЯ-текста, основанные на онтологии естественного языка и онтологическом описании предметов и процессов предметной области текста.
4. Реализация автоматизированной системы комплексной смысловой обработки ЕЯ-текстов «*SemTextProcessor*».

Публикации в журналах, входящих в перечень ВАК

1. Фаррохбахт Фумани Мехди, Автоматизированная система смысловой обработки текстов в системе управления электронными архивами// Известия СПбГЭТУ «ЛЭТИ» № 3. 2011. С. 40-44.
2. Фаррохбахт Фумани Мехди, методика автоматической смысловой обработки текстов в системе управления электронными архивами // Известия СПбГЭТУ «ЛЭТИ». № 4. 2011. С. 40-44.
3. Фаррохбахт Фумани Мехди, архитектура web-ориентированных подсистем оптимизации электронных схем //Перспективы науки №1(03). 2010. С. 90-94.

Статьи, опубликованные в других изданиях:

4. Farrokhbakht Foumani Mehdi, Automated semantic text processing in the management of electronic archives // The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, June 2012. http://springer.com
5. Farrokhbakht Foumani Mehdi, The technique of automatic semantic text processing in the management of electronic archives // The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, July 2012. http://springer.com
6. Farrokhbakht Foumani Mehdi, An ontological approach to semantic processing of natural language texts // The International Journal of Computer Science and Network Security, October 2012. http://www.IJCSNS.org

- Материалы конференций:

7. Фаррохбахт Фумани Мехди. Смысловой анализ текстов на основе алгоритма определения пересечения онтологий этих текстов // Материалы 63-й научно-технической конференции профессорско-преподавательского состава СПбГЭТУ. 2011.