На правах рукописи

УДК 519.767.6; 81'322.2

СИДОРОВА
Елена Анатольевна

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА ДЛЯ АНАЛИЗА ДОКУМЕНТОВ НА ОСНОВЕ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискании ученой степени
кандидата физико-математических наук

Новосибирск 2006

Работа выполнена в Институте систем информатики
имени А.П. Ершова СО РАН

Научный руководитель: Загорулько Юрий Алексеевич,
кандидат технических наук

Официальные оппоненты: Загоруйко Николай Григорьевич,
доктор технических наук, профессор

Лукашевич Наталья Валентиновна,
кандидат физико−математических наук

Ведущая организация: Томский политехнический университет

Защита состоится 15 декабря 2006 г. в 17 ч. 00 мин. на заседании диссертационного совета К.003.032.01 в Институте систем информатики имени А.П. Ершова Сибирского отделения РАН по адресу:
630090, г. Новосибирск, пр. ак. Лаврентьева, 6.

С диссертацией можно ознакомиться в читальном зале ИСИ СО РАН
(г. Новосибирск, пр. ак. Лаврентьева, 6).

Автореферат разослан _____ ноября 2006 г.

Ученый секретарь
Диссертационного совета,

к.ф.−м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Большой объем накопленной информации и высокая скорость поступления новой предъявляют все более жесткие требования к современным информационным системам (ИнС). Поскольку большинство источников являются текстовыми документами, то для их хранения организуются электронные библиотеки, в которых возможен поиск по ключевым понятиям, и в некоторых случаях проводится тематическая рубрикация документов. Но, так или иначе, в процессе работы человек имеет доступ к тексту документов, а не к основным смысловым фрагментам, содержащимся в них. Этого оказывается недостаточно: во-первых, в постоянно разрастающемся архиве становится трудно (практически невозможно) найти нужную информацию; во-вторых, данные часто дублируются и противоречат друг другу. Проблема усугубляется еще и тем, что пользователи используют для общения с поисковыми системами, как свои специальные термины, так и термины, широко используемые другими сообществами в ином контексте. А так как современные системы применяют в основном механизмы поиска по ключевым словам, не учитывающие ни семантику слов, входящих в запрос, ни его контекст, результатом их работы являются данные, подавляющее большинство из которых не относятся к существу запроса.

Для решения этих проблем требуется переход на новый качественный уровень обработки информации – необходимо вести обработку на семантическом уровне, т.е. учитывать смысл или содержание документов. За последние несколько лет это направление в информационных технологиях получило широкое развитие. Созданные на основе семантических технологий ИнС отличаются от традиционных тем, что используют явно выраженные (в виде онтологии) знания о предметной области. Часто онтология является не только основой для представления информации пользователям, ее хранения и поиска, но и для автоматической обработки поступающей текстовой информации.

До сих пор задача анализа текста на естественном языке рассматривалась многими исследователями независимо от той обстановки, где ее результаты планировалось использовать. Применяемые подходы либо никак не учитывают способ и форму хранения полученных результатов анализа в ИнС (например, исследования в рамках подхода «Смысл-Текст», разработанного И.А. Мельчуком), либо наоборот, строго привязаны к узким целям поставленной задачи и форме результата (например, при извлечении информации о персонах и организациях, что часто встречается в задачах компьютерной разведки). Классические подходы к семантическому анализу текста используют формальную модель языка и с “большой неохотой” переходят к модели предметной области, что не позволяет им естественным образом внедряться в ИнС с заданной предметной областью и удовлетворять поставленным перед такими системами требованиям. В отличие от работ, связанных с задачей полного извлечения смысла или извлечения всей информации из текстов документа, для большинства ИнС нет необходимости делать полный семантический анализ всего связанного текста. ИнС, построенные на основе онтологий, естественным образом задают как формат содержания того, что требуется извлечь из текста документа (или любого текстового ресурса), так и формат хранения результата в базе данных системы в виде семантической сети объектов, являющихся экземплярами понятий и отношений, заданных моделью предметной области.

Применение онтологий является одним из наиболее перспективных направлений исследований, поскольку позволяет формализовать и унифицировать операции обработки информации для повышения качества различных информационных услуг и сервисов. В работе проводится исследование одной из наиболее востребованных услуг – информационного наполнения системы.

В связи с этим особую актуальность приобретает разработка технологии анализа текста в контексте ее применения в различных информационных системах (в частности, для корпоративных систем документооборота или специализированных порталов знаний).

Ориентация технологии на деловую и научную лексику является вполне обоснованным решением, позволяющим эффективно применить семантически-ориентированные методы к решению задачи анализа текста на естественном языке.

Создание инструментальных средств – средств настройки онтологии, тезауруса и схем фактов, описывающих способы естественно-языкового выражения понятий и отношений в тексте, – дает возможность обеспечивать содержательную обработку текста документов без специальных навыков программирования непосредственным носителям знаний – экспертам и лингвистам.

Цель работы

Целью диссертационной работы является разработка способов описания лингвистических знаний и представления содержания документов в информационных системах, а также методов и инструментальных средств содержательного анализа текста на естественном языке.

Работа выполняется в рамках проекта по созданию технологии конструирования ИнС и направлена на автоматизацию наполнения различных ИнС данными, полученными в результате анализа содержания документов, поступающими в систему, либо в виде коллекции архивных документов, либо при регулярном оперативном поиске в сети Интернет. Технология должна включать средства описания предметной области и настройки лингвистической базы знаний.

Для достижения поставленной цели в диссертации последовательно решены следующие задачи исследования.

1. Определены требования, предъявляемые к технологиям автоматической обработки текста на естественном языке в современных информационных системах.
2. Создана лингвистическая модель описания фактов как способа естественно-языкового выражения понятий и отношений в тексте и как средства представления контента документа в информационной системе.
3. Разработана технология конструирования лингвистической базы знаний, реализующая предложенную модель.
4. Разработаны методы содержательного анализа документов, использующие предложенную модель.
5. Реализованы инструментальные средства, предназначенные для автоматического извлечения фактов из текста и формирования контента документа в информационной системе.
6. Апробированы компоненты предложенной технологии в реально функционирующих информационных системах.

Методы исследования

В диссертационном исследовании были использованы модели и методы искусственного интеллекта, компьютерной лингвистики, теории множеств, онтологический анализ, метод экспертных оценок, методы объектно-ориентированного проектирования и программирования.

Научная новизна

Проведенные исследования позволили разработать новый подход к содержательному анализу документов, позволяющий настраивать систему анализа на определенную предметную область.

В работе предложена комплексная модель представления знаний, включающая предметный словарь, модель предметной области и модель описания фактов.

Разработаны методы, которые по предложенной модели реализуют поэтапный анализ текста деловых документов ограниченной тематики: извлечение словарных объектов, поиск фактов и формирование контента документа в виде семантической сети объектов, являющихся экземплярами понятий и отношений, заданных моделью предметной области.

Реализованы проблемно-ориентированная оболочка, предназначенная для конструирования лингвистической базы знаний, и инструментальные средства, использующие построенную базу знаний для анализа текста.

Практическая ценность

Разработанная технология может быть применена как при создании новых информационных систем, так и при модернизации уже существующих.

Реализованы основные компоненты системы анализа текста документов, которые были апробированы при разработке ряда информационных систем.

Так при создании системы документооборота InDoc автором были разработаны и реализованы базовый алгоритм сборки фактов, использующий сегментацию, и алгоритм идентификации объектов, использующий глобальный контекст и позволяющий формировать контент анализируемого документа. Система InDoc прошла опытную эксплуатацию и была внедрена в производство.

При создании информационного портала по археологии и этнографии использовался словарный компонент в том виде, в котором он присутствует в технологии. Модуль индексирования археологических ресурсов использует при своей работе онтологию и создает контент ресурсов по тому же принципу, что и в предложенной технологии, однако, только для фиксированного набора схем фактов. Портал находится на стадии опытной эксплуатации.

Апробация работы

Основные выводы и научные результаты диссертационной работы докладывались на международных конференциях по компьютерной лингвистике и интеллектуальным технологиям «Диалог» в 2002, 2003, 2005 и 2006 гг., на международной конференции "Проблемы управления и моделирования в сложных системах" (г.Самара) в 2003г., на национальных конференциях по искусственному интеллекту КИИ в 2002 и 2004 гг., на международных конференциях по искусственному интеллекту ИАИ (г.Киев) в 2005 и 2006 гг.; печатались в журналах и сборниках: «Искусственный интеллект», №4, Киев, 2004; «Информационные технологии» № 11, 2004; «Молодая информатика: Сборник научных трудов аспирантов и молодых ученых» в 2005г., «Информационные технологии в гуманитарных исследованиях» в 2005.

По теме диссертации автором опубликовано 24 работы.

Структура и объем

Диссертационная работа состоит из 4 глав, введения, заключения, списка литературы содержащего 100 наименований. Общий объем работы 125 страниц, включая 2 приложения.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы исследования и формулируется задача диссертационной работы.

В **первой главе** рассмотрены типы информационных сервисов и их состояние на сегодняшний день. Выделяется объект исследования – деловая проза и описываются его свойства, формулируются требования к информационным системам, автоматизирующим работу с документами. Приводится классификация задач автоматической обработки текста, удовлетворяющих большей части потребностей ИнС. Дан обзор современных подходов и систем содержательного анализа текста на естественном языке.

В процессе развития системы информационных коммуникаций сформировались три вида информационного обслуживания – документальное, фактографическое и концептографическое.

Сущность документального обслуживания заключается в том, что информационные потребности удовлетворяются путем предоставления первичных документов, необходимые сведения из которых пользователи извлекают самостоятельно. В отличие от

документального обслуживания фактографическое предполагает удовлетворение информационных потребностей непосредственно, т. е. путем представления пользователям самих сведений (отдельных данных, фактов, концепций). Эти сведения предварительно извлекаются из первичных документов и после определенной их обработки предоставляются пользователям. При концептографическом обслуживании документы и полученные сведения подвергаются интерпретации, оценке, обобщению.

Возможности современных ИнС сводятся к фактографическому обслуживанию.

Рассматривая проблему извлечения сведений из документов, прежде всего, необходимо выделить основные свойства объекта анализа (текста первичного документа) и указать цели анализа – ограничить класс задач, что позволило бы формализовать данный процесс.

А.П. Ершов выделил деловую прозу как объект, с одной стороны, поддающийся автоматизации, с другой – остающийся естественным средством выражения мыслей для человека.

Для деловой прозы характерны следующие особенности.

1. Наличие строгой модельной ситуации, определяемой характером автоматизации или назначением создаваемой ИнС, для которой заданы правила распознавания и реакции на ее возникновение, хотя последовательность возникновения ситуаций может оставаться неопределенной. Это свойство приводит к тому, что деловая проза всегда внутренне формализована.
2. Ограниченность предметной области. Модель действительности определяется самой областью деловых отношений.
3. Ограничение естественного языка (т.е. используется концепция подъязыка как проекция общеупотребительного русского литературного языка на определенную предметную область и класс ситуаций общения). Потребность быстрого и точного взаимопонимания сделала язык деловой прозы четким, экономичным и жестким, а внешнее оформление текста документа – структурированным. Поэтому мы можем вводить в систему соответствующие ограничения и делать упор в большей степени на семантику текста, нежели на его синтаксическое представление.
4. Четкость функций каждого сообщения. Наличие цели, определяемой по заранее известным правилам, позволяет сконцентрировать анализ вокруг наиболее значимых понятий предметной области, к таким понятиям относятся, например, научный результат в научной статье или сообщение о какой-либо деятельности в деловом письме. Это свойство разительно отличает деловую прозу от других форм общения, например, стихов или пространного и эмоционального повествования.

Отмеченные выше свойства деловой прозы позволяют заменить расплывчатые категории смысла и понимания прагматическими концепциями адекватного восприятия языкового сообщения, взятом в четком контексте данной области деловых отношений.

К наиболее востребованным на сегодняшний день классам ИнС относятся различные информационные Интернет-ресурсы (например, специализированные порталы) и интеллектуальные системы документооборота, разрабатываемые для компаний и предприятий. Основным требованием к технологии автоматической обработки текста, удовлетворяющим большую часть потребностей таких ИнС, является поддержка содержательного анализа и поиска.

Во **второй главе** предложены основные формализмы знаний, используемые в разработанном подходе.

Совокупность всех знаний, хранимых системой и используемых для решения задач, преследуемых в данной работе, образуют *базу знаний технологии*, которая включает:

- знания, содержащиеся в самой информационной системе:
  - модель предметной области, средством описания которой выбрана онтология;
  - конкретные знания о предметной области, хранимые в базе данных системы;

- специализированные знания, используемые для решения задачи анализа текста на естественном языке; эти знания образуют *лингвистическую базу знаний*:
  - модель документов, которая описывает формальную структуру текста в зависимости от жанра документов;
  - словарь, который содержит ключевую лексику, используемую носителями языка и экспертами в данной предметной области для описания понятий и отношений, заданных в онтологии;
  - схемы фактов, описывающие языковые выражения, с помощью которых в тексте могут представляться понятия и отношения онтологии.

Онтология содержит понятия и отношения предметной области. Различные свойства понятий описываются с помощью атрибутов и ограничений, наложенных на области значений атрибутов. Структура понятия может варьироваться, но всегда имеет следующие характеристики:

- конечный набор атрибутов;
- наличие названий у атрибутов;
- закрепленный смысл каждого атрибута (трактовка значения атрибута);
- наличие типа у атрибутов;
- возможность присвоить одному атрибуту несколько значений;
- возможность указать атрибуту обязательность его заполнения;
- наличие набора ключевых атрибутов.

Структура понятия, обладающая данными характеристиками, достаточна для описания свойств объекта, существенных с точки зрения решаемой задачи.

Отношения в онтологии являются бинарными (имеют два аргумента) и могут иметь собственные атрибуты.

Определение. Онтология – это знаковая система

$O = \langle C, R, D, DV, TD, A, P, R_{AP}, R_{DV} \rangle$, в которой

$C = \{c_1, ..., c_n\}$ – конечное множество понятий,

$R = \{r_1, ..., r_m\}, R \subseteq C \times C, R = R_C \cup R_T \cup R_A$ – конечное множество бинарных отношений $\mathrm{r_i(c_x, c_y)}$ между понятиями,

- $R_C \subseteq C \times C$, $R_C \subseteq R$ – антисимметричное, транзитивное, нерефлексивное бинарное отношение наследования, являющееся отношением частичного порядка на множестве понятий C,
- $R_T \subseteq \{r_{T,i}(c_x, c_x)\} \subseteq C \times C$, $R_T \subseteq R$ – бинарное отношение «часть-целое»,
- $R_A \subseteq C \times C$, $R_A \subseteq R$ – конечное множество ассоциативных отношений,

$D = \{d_1, ..., d_q\}$ – конечное множество доменов,

$DV = \{dv_1, ..., dv_u\}$ – конечное множество конкретных значений стандартного типа string, включенных в некоторый домен,

$TD = D \cup \{string, boolean, \text{integer}\}$ – множество типов данных, включающее три стандартных типа string, boolean, integer и множество доменов,

$A = \{a_1, ..., a_w\}, A \subseteq \begin{cases} C \times TD, \\ R \times TD. \end{cases}$ – конечное множество атрибутов, т.е. бинарных отношений $a_i(c_x, td_y)$ или $a_i(r_x, td_y)$,

$P = \{p_1, ..., p_t\}$ – конечное множество конкретных свойств атрибута, включающее свойства {multiplicity, key, mandatory},

$P_{AP} \subseteq A \times P$ – бинарное отношение инцидентности между множествами атрибутов A и свойств атрибутов AP,

$P_{DV} \subseteq D \times DV$ – бинарное отношение инцидентности между множествами доменов D и доменных значений DV.

Онтология может использоваться в качестве схемы хранилища данных – информационного пространства системы. В этом случае наполнение БД системы образуют информационные объекты, являющиеся экземплярами понятий и отношений, заданных онтологией, и только эти объекты необходимо извлекать из текстов документов.

Информационный объект может быть рассмотрен в трех разных аспектах – структура, контекст и контент. Для ИнС значительный интерес представляет описание семантики объектов с точки зрения содержания или контента. Контент описывается в терминах онтологии, это означает, что любой информационный объект, которому соответствует некоторый документ (а в общем случае, это может быть носитель любого типа, например, звук, видео, рисунок и т.п.), связывается с набором других информационных объектов, присутствующих в БД системы и являющихся экземплярами понятий и отношений онтологии. Данный набор объектов отражает информационное содержание документа.

Определение. Информационное пространство системы, для которого задана онтология *O*, – это знаковая система:

$O_I = \langle I, R_I, V, A_I, TX, C_T, P_{IC}, P_{IR}, P_{IA}, P_{IT} \rangle$, в которой

$I = \{i_1, ..., i_n\}$ – конечное множество экземпляров понятий онтологии,

$R_I = \{ri_1, ..., ri_k\}$ – конечное множество конкретизированных отношений (экземпляров отношений), т.е. бинарных отношений $\mathrm{ri}_\mathrm{i}(\mathrm{i}_\mathrm{x}, \mathrm{i}_\mathrm{y})$ между экземплярами понятий,

$V = \{v_1, ..., v_q\}$ – конечное множество конкретных значений стандартного типа,

$A_I = \{ai_1, ..., ai_w\}$ – конечное множество конкретизированных атрибутов, т.е. бинарных отношений $ai_i(i_x, v_y)$ или $ai_i(ri_x, v_y)$ между экземпляром понятия или отношения и конкретными значениями,

$TX = \{tx_1, ..., tx_l\}$ – конечное множество носителей (текстов),

$C_T = \{ct_1, ..., ct_h\}$ – конечное множество контентных связей, т.е. бинарных отношений $ct_i(ti_x, ai_y)$ между текстом и конкретными атрибутивными отношениями, найденными в тексте в результате его анализа и составляющими контент документа,

$P_{IC} \subseteq I \times C$ – бинарное отношение инцидентности между множествами I и C,

$P_{IR} \subseteq R_I \times R$ – бинарное отношение инцидентности между множествами $\mathrm{R_I}$ и R,

$P_{IA} \subseteq A_I \times A$ – бинарное отношение инцидентности между множествами $\mathrm{A_I}$ и A,

$P_{IT} \subseteq I \times TX$ – бинарное отношение инцидентности между множествами I и TX.

В предлагаемом подходе документы являются информационными объектами и описываются в онтологии некоторым понятием(-ями). Текст, представляющий содержание таких объектов, анализируется с целью извлечения значимой информации и формирования контента. При анализе документа используется формальное представление структуры его текста, которая зависит от типа или жанра документа.

Текст в электронной форме имеет, по крайней мере, три уровня формальной структуры – физический, логический и жанровый. Первый представляет презентацию текста на странице, например, с помощью тегов или таблицы стилей. Ко второму уровню относятся такие элементы как абзац, строка, предложение и т.п. Третий уровень представлен разбиением текста на жанровые части, например, текст делового письма имеет следующие жанровые разделы: заголовок (отправитель, адресат, резюме и обращение), основной раздел (текст письма, примечания и приложения) и подпись.

Любую формальную структуру текста будем называть *сегментом*, а процесс извлечения сегментов из текста – *сегментацией*.

Жанровые разделы документа:

- характеризуются определенной лексикой, задаваемой в словаре,

- имеют определенную структурную организацию (состав и позиция относительно других жанровых разделов),
- реализуются в рамках определенных формальных сегментов.

Словарь включает термины следующего вида.

1. Лексема – слово во всей совокупности его форм и значений. В одну лексему объединяются разные парадигматические формы одного слова и разные смысловые варианты слова, зависящие от контекста, в котором оно употребляется.
2. Словокомплекс – это устойчивое терминологическое сочетание, характерное для выбранной предметной области.
3. Лексическая конструкция – несловарная единица, имеющая регулярную структуру, например, номер телефона, дата, инициалы и т.п. Для создания словаря лексических конструкций используется технология Alex, совмещающая в себе функции хранилища шаблонов, с помощью которых задаются лексические конструкции, и специализированного лингвистического процессора.

Любой термин описывается словарной статьей, которая включает наборы терминообразующих, статистических и семантических признаков. Термины словаря могут быть сгруппированы пользователем в синонимичные группы с выделенным главным термином, название которого автоматически становится названием всей группы.

Иерархии классов понятий и заданные на них семантические отношения позволяют представить структуру высказывания из предметной области в виде *факта*. Множество таких фактов составляет пропозициональное содержание документа.

Факт есть высказывание или языковое выражение, фиксирующее эмпирическое знание. Формализовав понятие факта, можно не только представить структуру высказывания, но и связать его с понятием или отношением, заданным в онтологии.

Для того чтобы извлечь факт из текста, его элементы должны удовлетворять определенным условиям или ограничениям. Выделяются семантические и структурные ограничения.

Семантические ограничения накладывают условия на семантические характеристики элементов факта. В предложенном подходе такие ограничения задаются таблично.

Таблица 1. Общая схема таблицы семантических ограничений для бинарных фактов.

Характеристики сочетания						Дополнительные характеристики								
1-ый аргумент			2-ой аргумент			1-ый аргумент			2-ой аргумент			Результат		
S_1	…			…	S_k	$S`_{k+1}$	…			…			…	$S`_n$

Характеристики сочетания содержат условия, которым должны удовлетворять параметры элементов (аргументов) факта. Дополнительные характеристики содержат значения, позволяющие либо уточнить объекты (аргументы), образующие факт, либо сформировать объект, соответствующий найденному факту (установить значения атрибутов данного объекта), либо уточнить значения атрибутов объекта документа (в тексте которого обнаружен факт).

Таблица семантических ограничений $\mathbf{Sem^F}$ задает n-арное отношение на k множествах семантических характеристик $S_1, S_2, ..., S_k$ и (n-k) множествах дополнительных характеристик $S`_{k+1}, S`_{k+2}, ..., S`_n$. Для каждого столбца i | $1 \le i \le k$ (характеристик сочетания) задается операция сравнения $\Theta_i: S_i \times S_i \rightarrow \{true, false\}$, позволяющая определить, соответствует ли значение характеристики, указанное в таблице, значению соответствующей характеристики аргумента факта. Эта информация, в частности, позволяет использовать иерархические отношения при сравнении таких характеристик как семантические классы.

Помимо семантических ограничений, необходимо учитывать ограничения других языковых уровней, которые в дальнейшем будут называться структурными. *Структурные*

ограничения накладывают условия на взаиморасположение элементов факта в тексте и их характер.

В предложенном подходе структурные ограничения **St** задаются списком значений фиксированных атрибутов. Все атрибуты **St** разделены на четыре практически независимые группы атрибутов:

- **St-seg** – условие на сегмент, т.е. в рамках сегмента какого типа должны располагаться аргументы;
- **St-pos** – взаиморасположение аргументов в тексте (контактность, пре- и постпозиция, приоритетность позиции при многовариантности выбора);
- **St-syn** – наличие синтаксических условий (валентности терминов, предложно-падежные сочетания и т.п.);
- **St-rul** – правила образования сочетаний (однородность, количество возможных связей, проективность, максимальная связность).

Таким образом, St = {St-seg, St-pos, St-syn, St-rul} определяется конечным множеством значений атрибутов $a_i(St, v_i)$, где значение $v_i \in d_i$, т.е. домену атрибута a_i. Количество атрибутов зависит только от количества элементов факта.

Для того чтобы находить факты, значимые для заданной предметной области, необходимо иметь механизм описания таких фактов. Декларативное описание структуры факта, условий его выявления и результат будем называть *схемой факта*.

Определение. Схема факта S_f – это тройка вида < A, Cs, Res >, где

$A = \{a_1,\ldots,a_n\}$ – конечное множество аргументов факта, где a_i задает класс объекта,

Cs = < Sem, St > – семантические и структурные ограничения,

Res = < t, op(t), P > – результат применения схемы факта, где

t – задает класс результирующего объекта,

op(t) – тип операции: создание или редактирование объекта,

$P = \{p_1,\ldots,p_m\}$ – конечное множество правил для формирования значений атрибутов результирующего объекта. Каждое правило ставит в соответствие атрибуту результирующего объекта один из следующих элементов: значение атрибута одного из аргументов, экспертное значение, заданное в таблице семантических сочетаний или значение по умолчанию.

S_f задает простую схему извлечения фактов из текста: если найдены аргументы из A, удовлетворяющие условиям Cs, то выполнить действия, задаваемые результатом Res.

Заданная таким образом схема фактов обладает двумя свойствами:

- моделируя промежуточные объекты (факты) можно обойтись только унарными и бинарными схемами фактов (т.е. схемами с одним и двумя аргументами),
- поскольку входными данными для схем являются термины Тезауруса, то возможно естественным образом упорядочить применение схем фактов к данным.

Рассмотрим упрощенный пример схем фактов для отношения Работает-в (Человек, Организация). Оно может быть выражено в тексте двояко:

1. Явным образом: «*Иванов работает в Организации N*».
2. Неявным образом, через должность «*Директор Организации N Иванов получил письмо*».

В первом случае схему факта можно выразить через промежуточный факт – предикат_место(Действие, Организация).

```
Работает-в_1 [
 A  : [Arg1: Человек; Arg2: предикат_место];
 St : [   St-seg: Предложение;
          St-pos: препозиция Arg1;
          St-syn: синтаксическая согласованность;
          St-rul: однородность ];
 Res: [t: Работает-в; ot: создать]
]
```

Во втором случае можно использовать промежуточный факт должность_организация(Должность, Организация).

```
Работает-в_2 [
 A  : [Arg1: Человек; Arg2: должность_организация];
 St : [   St-seg: Предложение;
          St-pos: контактность ];
 Res: [t: Работает-в; ot: создать; P: t.должность = Arg2.Arg1]
]
```

Предложенная модель знаний позволила разработать технологию настройки информационной системы на содержательный анализ текстов в ограниченной предметной области.

В **третьей главе** описана технология семантического анализа текста. Приведена архитектура системы, описаны методы и инструментальные средства, осуществляющие автоматическое извлечение фактов из текста и формирующие контент документа в ИнС.

Архитектура системы (Рис.1) включает четыре основных компонента: ядро, словарную подсистему, редакторы онтологии, схем фактов и формальных структур текста, подсистему взаимодействия с БД.

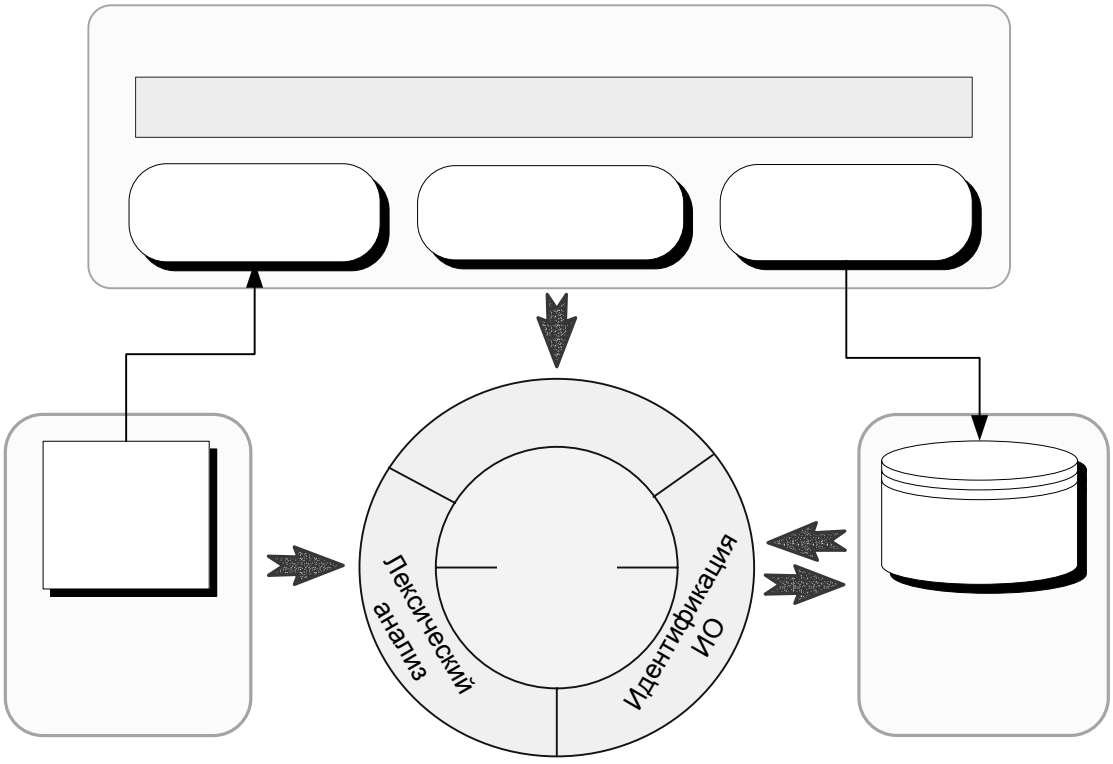

Рис. 1. Архитектура системы анализа текста на основе схем Фактов

Ядро системы обеспечивает сборку фактов по описаниям, созданным с помощью редактора схем фактов. Словарная подсистема обеспечивает создание словаря и предварительный этап обработки текста (сегментацию, лексический и морфологический анализ). В качестве редактора онтологии и модуля взаимодействия с БД используется компонент, реализованный в рамках проекта по созданию порталов знаний.

Предложенный подход к анализу текста документов включает следующие основные этапы: первичная сегментация, словарный поиск (обработка), жанровая сегментация, сборка фактов, формирование и добавление контента документа в информационное пространство системы.

Существуют два вида сегментации текста – первичная и жанровая. В процессе первичной сегментации осуществляется разбиение линейного представления текста на строковые объекты, оформленные как сегменты и упорядоченные в соответствии с порядком

их встречаемости в тексте. Жанровая сегментация осуществляется после лексического анализа на основе лексических объектов, маркирующих тот или иной жанровый сегмент. Механизм сегментации реализуется с помощью системы Алекс, входящей в качестве подсистемы в словарный компонент предлагаемой технологии.

Разбиение на сегменты используется в дальнейшем при сборке фактов, где, при наличии соответствующего структурного ограничения, на вход алгоритму подается не весь текст целиком, а только фрагмент текста. В этом случае алгоритм сборки фактов запускается столько раз, сколько найдено требуемых сегментов.

Лексический анализ осуществляется словарным компонентом системы.

Процесс создания словаря обычно очень трудоемкий процесс, требующий специалистов высокого уровня. Поэтому мы постарались разработать подходы и создать программные средства, облегчающие этот процесс, используя широко известные механизмы статистического сбора информации.

При разработке были выдвинуты следующие требования к словарному компоненту.

1. Наличие рабочего места лингвиста для конструирования словаря, поддерживающего классические функции редактирования, сортировки, фильтрации и просмотра конкорданса.
2. Поддержка автоматической наполняемости словарей на базе корпусов текстов.
3. Ориентированность на определенную предметную область, возможность настраивать и приписывать предметные характеристики элементам словаря.
4. Выполнение лексического анализа текста – извлечение из текста заданных в словаре терминов и их свойств.
5. Возможность накопления данных о статистико-комбинаторных свойствах лингвистических явлений.

К подключаемым и разработанным модулям автоматизированной настройки словаря и поддержки анализа относятся следующие модули.

- Модуль морфологического анализа, разработанный компанией Диалинг, подключаемый внешним образом.
- Подключаемый модуль сборки словокомплеков WordFinder. Этот модуль собирает именные и адъективные группы, некоторые типы групп наречия и групп глагола, учитывая наличие предлогов и союзов.
- Подключаемый модуль настраиваемой сборки сложных структур – лексических конструкций, на основе системы правил-шаблонов Alex.
- Модуль просмотра конкорданса, который позволяет в выбранном корпусе текстов просматривать встречаемость термина словаря.
- Модуль тематизации, обеспечивающий анализ текста в различных режимах: наполнение словаря, ведение статистики встречаемости терминов, классификация текста на основе статистики. Последовательный анализ текста в разных режимах позволяет поддерживать механизм расширения иерархии классов и «дообучения» словаря.
- Модуль выявления стоп-терминов, позволяющий отделить шумовую или общеупотребительную лексику от предметно-зависимой.

В ходе лексического анализа осуществляется извлечение словарных объектов из набора упорядоченных строковых объектов, полученного после первичной сегментации текста. В задачи данного этапа входит:

- применение лексических шаблонов и извлечение лексических конструкций;
- осуществление морфологического анализа и сборки словокомплексов;
- выделение жанровых сегментов.

Результатом лексического анализа является упорядоченный список объектов со следующим набором параметров: название (нормальная форма слова, словокомплекса или

имя шаблона – главная в списке альтернатив), позиция в тексте, значение (извлеченное числовое значение, подстрока и т.п.), грамматический класс и набор значений словоизменительных морфологических признаков для слов, набор семантических классов, статистические характеристики.

Механизм сборки фактов включает два этапа: планирование и исполнение. Причем если этап исполнения повторяется для каждого документа, то планирование осуществляется предварительно на основании заданных экспертом схем фактов.

После того, как эксперт описал весь набор схем фактов, необходимых для анализа документов, система проверяет корректность и сходимость набора схем и осуществляет планирование действий системы во время исполнения.

Задачами планирования являются:

1) Организация очереди схем фактов в порядке их исполнения или применения. При этом необходимо учитывать порядок создания объектов.

2) Генерация исполняемых правил на основе схем фактов. Такие правила мы будем называть – исполнителями. Исполнители включают набор методов, которые в зависимости от типа аргументов, типа требуемого результата и набора специфических условий по-разному реализуют сборку факта заданного типа.

Во время непосредственной обработки документа, менеджер системы осуществляет последовательный вызов исполнителей из очереди. Каждому исполнителю менеджер подает на вход данные, сгруппированные по сегментам (тип сегмента задается соответствующим условием в схеме факта) и вызывает функцию поиска для текущего исполнителя. Реализация данной функции зависит от типа исполнителя (т.е., в конечном счете, от заданных ограничений). Затем выполняется процедурная часть (создание или редактирование объекта) и проверка новых объектов (если они есть) на уникальность. Процесс исполнения завершается, когда очередь исполнителей становится пустой.

Исполнители осуществляют извлечение фактов из текста на основе четырех разработанных алгоритмов:

- алгоритм, использующий таблицу семантических сочетаний для поиска и формирования предварительного набора фактов;
- алгоритм, осуществляющий поиск и обработку однородных членов предложений (однородность, проективность, связность). Найденные объекты вначале объединяются в однородные группы (группа объектов одного класса, определенного аргументом схемы факта), затем проверяется сочетаемость (семантическая и/или синтаксическая) контактных групп;
- алгоритм, осуществляющий проверку синтаксической сочетаемости элементов факта с помощью модуля WordFinder;
- алгоритм, осуществляющий проверку остальных структурных ограничений.

Система спроектирована таким образом, чтобы в случае необходимости, можно было бы разработать и подключить дополнительные алгоритмы, которые бы осуществляли обработку (проверку) новых ограничений, включенных в схему фактов.

Дальнейшая обработка заключается в формировании контента документа. Для этого необходимо идентифицировать объекты, полученные в результате анализа, обеспечить корректность их добавления и, собственно, добавить их в информационное пространство системы.

Под идентификацией объекта понимается уточнение полученного объекта (уточнение атрибутов) и «склеивание» одинаковых объектов, на основе использования локального контекста, поиск объекта в информационном пространстве системы (глобальный контекст) и обеспечение корректности набора и значений его атрибутов.

Мы выделили три задачи, которые возникают при глобальном поиске объекта.

1. Идентификация объекта, найденного в тексте документа, по классу и набору атрибутов – поиск данного объекта в БД системы. При наличии нескольких объектов возникает *контекстная омонимия*, требующая однозначного разрешения.
2. Идентификация и уточнение класса объекта по иерархии классов (при этом может уточняться как объект найденный в тексте, так и объект из БД системы).
3. Идентификация и уточнение объектов по отношению «часть-целое» (иерархии вложенности), используемая, в частности, при разрешении контекстной омонимии.

Было предложено два метода разрешения омонимии.

Первый способ заключается в построении *фокусного множества ИО*, включающего все непосредственно связанные с данным ИО, и сопоставлении его с фокусным множеством объекта, найденного в тексте.

Второй способ заключается в использовании иерархии по отношению «часть-целое», в случае, когда объекты имеют сложную структуру, представленную линейными цепочками наименований, совокупность которых образует дерево (множество деревьев) информационных объектов. Для определения такого ИО требуется восстановить иерархию вложенности объектов документа данного типа путем сравнения с эталонной иерархией ИО из БД. Каждая пара объектов, удовлетворяющая определенным требованиям порядка слов, проверяется на предмет наличия между ними отношения вложенности (с учетом транзитивности). Результирующими являются те ИО, которые соответствуют листьям полученных древесных структур.

В общем виде механизм создания контента документа выглядит следующим образом.

1. Создается ИО документа и формируется его индекс.

2. Все найденные в тексте новые объекты и связи добавляются в БД системы.

3. Все атрибутивные отношения, добавленные в БД либо при создании нового объекта, либо при редактировании уже существующего, помечаются индексом ИО документа (для существующего ИО помечаются также старые атрибутивные отношения, обнаруженные в тексте и совпавшие с существующими).

Под термином «помечается» понимается добавление в БД системы специального отношения типа «объект – атрибутивное отношение», связывающего документ с контентом. Отметим, что этих связей оказывается достаточно, чтобы хранить информацию о содержании, т.к. любой объект идентифицируется набором своих ключевых атрибутов или ключевыми атрибутивными отношениями.

В **четвертой главе** рассмотрены применения разработанных подходов и методов в реальных информационных системах.

Так при создании в 2001-2003 гг. информационной системы документооборота InDoc был апробирован подход к анализу содержания документов путем извлечения фактов, при этом были разработаны и использованы алгоритм сборки фактов, использующий сегментацию, и алгоритм идентификации объектов, минимально использующий глобальный контекст – информационной наполнение системы, представленный иерархией вложенности объектов.

В системе InDoc рассматривался только один жанр документа – деловое письмо, как наиболее типичный для задачи интеллектуализации документооборота. Была разработана жанровая структура документа, что позволило ограничить возможную смысловую нагрузку той или иной части текста документа.

Для хранения контента документов была разработана структура семантического индекса, представленная в виде набора содержательных атрибутов, автоматически заполняемых системой анализа.

База знаний, необходимая для анализа содержания документов, в системе InDoc включает пять компонентов.

1. Метаонтология является ядром системы знаний и фиксирует базовые структуры, которые система должна использовать при автоматическом анализе и индексации текстов документов: компоненты семантического индекса, базовые классы понятий, жанровая структура документа, семантическая структура высказывания.

Общая структура извлекаемой информации представляется в виде некоторой пропозициональной структуры, выражающей связь предиката (процесс, действие, свойство) и множества его аргументов:

F = P(S, O, L), где *P* – действие (Работа), *S* – субъект (Организация), *O* – объект (Объект), над которым выполняется действие, *L* – место действия (Объект строительства).

Используются следующие схемы Фактов:

F1 = Работа + Объект,

F2 = F1 + Объект строительства,

F3 = F2|F1 + Организация.

2. Онтология предметной области включает иерархию классов понятий ПО и семантические отношения на этих классах:

- объектное отношение – это связь «Работа–Объект», на основе этого отношения формируется факты типа F1,
- агентивное отношение – это связь «Организация – F1», это отношение характеризует различные классы Организаций с точки зрения их деятельности,
- отношение вложенности – это отношение "часть–целое", заданное для Объектов строительства, которое позволяет идентифицировать объекты сложной структуры представляемые в тексте цепочкой понятий.

Знания о сочетаемости понятий представлены в виде трех таблиц, задающих ограничения на допустимые сочетания понятий указанных классов и их значений в фактах.

3. Предметные знания, т.е. знания о конкретных организациях и их типах, о видах деятельности, о типах строящихся объектов, об иерархии построенных объектов и т.п..

4. Знания о конкретном предприятии представлены списком сотрудников предприятия, структурой его подразделений, фильтрами адресации сотрудников.

5. Лингвистические знания об языке деловых писем представлены в словаре с помощью технологии Алекс.

Анализ текста в системе InDoc начинается с поиска ключевых понятий, выделяемых словарным компонентом системы. На этапе сегментации документа осуществляется жанровая декомпозиция текста документа, в результате которой определяются границы Основного текста при помощи служебных жанровых шаблонов *Обращение* и *Подпись*, выделяются все организации, упомянутые в шапке письма, и определяется организация-Отправитель.

Последующая обработка документа представляет собой процесс извлечения релевантной информации на основе ключевых понятий, выделенных в границах *Основного текста.*

На этапе идентификации объектов уточняются и идентифицируются все понятия, которые могут входить в состав фактов. На этом этапе определяются все возможные атрибуты понятия, позволяющие уточнить объект (например, для объекта строительства это может быть его номер, начальный и конечный километр участка и т.п.).

Последний этап – семантический анализ состоит в установлении семантических отношений между составляющими высказываний в Основном тексте, что позволяет представить содержание письма в виде совокупности упомянутых в нем фактов, выделить Тему письма (совокупность фактов типа F3, в которых фигурирует Отправитель) и определить соответствующие значения полей семантического индекса.

При создании информационного портала по археологии и этнографии использовался словарный компонент в том виде, в котором он присутствует в технологии. Модуль

индексирования археологических ресурсов использует при своей работе онтологию и создает контент ресурсов по тому же принципу, что и в предложенной технологии, однако, только для фиксированного набора схем фактов.

При разработке модуля индексирования были исследованы два жанра текстовых ресурсов: новостные сообщения и научные статьи по археологии и этнографии.

Из текста новостных сообщений извлекается информация о событиях и объектах, связанных с событиями. Часть онтологии портала – онтология научной деятельности, описывает все понятия и отношения, необходимые для анализа новостных сообщений.

При анализе научных статей в большей степени использовалась формальная структура документа. На основе описания жанровой структуры статьи извлекаются такие понятия как авторы статьи, организации, в которых работают авторы, название. Из основного текста статьи извлекается информация о научных результатах, описанных в данной статье, и связанных с ними объектах. Часть онтологии портала – онтология научного знания, описывает все понятия и отношения необходимые для анализа научных статей. Специфика понятий в данной предметной области, отражается онтологией предметной области (онтологией археологии).

Портал находится на стадии опытной эксплуатации.

В **заключении** сформулированы основные результаты, полученные в ходе диссертационной работы.

ЛИЧНЫЙ ВКЛАД АВТОРА

Результаты, которые выносятся на защиту в данной диссертационной работе, не были бы возможны без слаженной работы всего научного коллектива, в котором работал автор. Созданию технологии содержательного анализа текста предшествовала работа над проектом InDoc, а также дальнейшее развитие основных идей технологии в проектах, связанных с созданием технологии конструирования информационных систем. Наибольший вклад автором диссертации внесен при решении следующих задач.

- Теоретическая разработка подхода к анализу текста, изложенного в данной работе.
- Разработка и реализация основных алгоритмов сборки фактов.
- Разработка и реализация алгоритмов идентификации объектов, использующих отношение вложенности.
- Разработка архитектуры словарного компонента.
- Реализация ядра словарного компонента.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проанализированы существующие технологии представления и извлечения информации в электронном виде из текстовых документов. Сформулированы требования к компоненту информационных систем, отвечающему за извлечение данных из текстовых документов на основе онтологии и лингвистической базы знаний.
2. Предложена модель лингвистической базы знаний, включающая три составляющих: предметный словарь, модель документа и модель описания фактов как способа естественно-языкового выражения понятий в тексте и средства представления контента документа в информационной системе.
3. Разработана технология конструирования лингвистической базы знаний, реализующая предложенную модель.
4. Разработаны методы, которые по предложенной модели реализуют поэтапный анализ текста документов: извлечение словарных объектов, поиск фактов и формирование контента документа в виде семантической сети объектов, являющихся экземплярами понятий и отношений, заданных моделью предметной области;

5. Реализованы проблемно-ориентированная оболочка, предназначенная для конструирования лингвистической базы знаний, и инструментальные средства, использующие построенную базу знаний для анализа текста.
6. Разработаны приложения, в которых апробированы методы и компоненты предложенной технологии анализа текста.

ОПУБЛИКОВАННЫЕ РАБОТЫ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Кононенко И.С., Сидорова Е.А. Обработка делового письма в системе документооборота // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. –М.: Наука, 2002. –Т.2. –С.299–310.
2. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Представление знаний в интеллектуальной системе документооборота // Труды 8-й национальной конференции по искусственному интеллекту КИИ'2002. –М.: Физматлит, 2002. –Т.2. –С.867–875.
3. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Подход к разработке интеллектуальной системы документооборота инвестиционной компании // Труды IV-й международной конференции "Проблемы управления и моделирования в сложных системах". –Самара: Самарский Научный Центр РАН, 2002. –С.366–372.
4. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Классификация деловых писем в системе документооборота // Материалы международной научно-технической конференции «Информационные системы и технологии» (ИСТ'2003). –Новосибирск: Изд. НГТУ, 2003. –Т.3. –С.141–145.
5. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Проблемы организации электронного архива с семантическим индексированием документов // Труды международной конференции Диалог'2003 "Компьютерная лингвистика и интеллектуальные технологии". –Протвино, 2003. –С.724–731.
6. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Система InDoc: интеллектуальная обработка, распределение и поиск документов в электронном архиве. // Труды V-й международной конференции "Проблемы управления и моделирования в сложных системах". –Самара: Самарский Научный Центр РАН, 2003. –С.248–254.
7. Загорулько Ю.А., Кононенко И.С., Сидорова Е.А. Концепция интеллектуализации документооборота // Труды 9-й национальной конференции по искусственному интеллекту КИИ'2004. –М.: Физматлит, 2004. – Т.3. –С.986–993.
8. Сидорова Е.А. Интеллектуальная обработка документов // Искусственный интеллект, №4. –Киев, 2004. –С.738–747.
9. Марчук А.Г., Холюшкин Ю.П., Загорулько Ю.А., Воронин В.Т., Андреева О.А., Боровикова О.И., Булгаков С.В., Костин В.С., Нуртдинов А.Н., Сидорова Е.А. Разработка новых методов и информационных технологий представления и обработки археологических и этнографических данных // Информационные технологии в гуманитарных исследованиях Вып.7. – Новосибирск: Изд. НГУ, 2004. –С.10–22.
10. Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Холюшкин Ю.П. Разработка интеллектуального интернет-портала знаний для доступа к информационным ресурсам по археологии и этнографии // Информационные технологии в гуманитарных исследованиях. Вып.7. –Новосибирск: Изд. НГУ, 2004. –С.31–39.
11. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Подход к интеллектуализации документооборота // "Информационные технологии" №11, 2004. –С.2–11.
12. Сидорова Е.А. Методы интеллектуальной обработки документов, основанные на экспертных знаниях // Молодая информатика: Сборник научных трудов аспирантов и молодых ученых. –Новосибирск: Институт систем информатики им А.П. Ершова СО РАН, 2005. –С.95–104.
13. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Автоматизация сбора онтологической информации в Интернет-портале знаний // V международная конференция

«Интеллектуальный анализ информации ИАИ-2005». Сборник трудов под редакцией Т.А. Таран. –Киев: Просвита, 2005. – С.82–91.
14. Сидорова Е.А. Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов // Труды международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии". –М.: Наука, 2005. –С.443–449.
15. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний // Труды международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии". –М.: Наука, 2005. –С. 65–70.
16. Kononenko I.S., Sidorova E.A., Zagorulko Yu.A. A Knowledge-based Approach to Intelligent Document Management // Proceedings of the 6th International Workshop on Computer Science and Information Technologies CSIT'2005. –Ufa-Assy, Russia, 2005. –V1. –P. 33-38.
17. Андреева О.А., Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Циркин Б.Г. Организация содержательного доступа к систематизированным знаниям по археологии и этнографии через интернет-портал // Информационные технологии в гуманитарных исследованиях. Вып.9. –Новосибирск: Изд. НГУ, 2005. –С.25–32.
18. Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Холюшкин Ю.П., Система знаний информационного интернет-портала по археологии и этнографии // Информационные технологии в гуманитарных исследованиях. Вып. 9. – Новосибирск: Изд. НГУ, 2005. –С. 33–39.
19. Андреева О.А., Боровикова О.И., Загорулько Ю.А., Кононенко И.С., Сидорова Е.А. Коллекционер онтологической информации для портала знаний по археологии и этнографии // Информационные технологии в гуманитарных исследованиях. Вып. 9. – Новосибирск: Изд. НГУ, 2005. –С. 39–47.
20. Zagorulko Yu., Borovikova O., Bulgakov S., Sidorova E. Ontology-based approach to development of adjustable knowledge internet portal for support of research activity // Bull. of NCC. Ser.: Computer Science 2005. –Is. 23. –P.45-56.
21. Андреева О.А., Сидорова Е.А. Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов // Технологии Microsoft в теории и практике программирования. –Нововсибирск, 2006. –С.221–223.
22. Сидорова Е.А. Подход к описанию фактов для задачи фактографического анализа текста // VI международная конференция «Интеллектуальный анализ информации ИАИ-2006». Сборник трудов под редакцией Т.А. Таран – Киев: Просвита, 2006. –С.252–261.
23. Загорулько Ю.А., Боровикова О.И., Кононенко И.С., Сидорова Е.А. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике // Труды международной конференции Диалог'2006 "Компьютерная лингвистика и интеллектуальные технологии". –М.: Изд. РГГУ, 2006. – С.148–151.
24. Загорулько Ю.А., Кононенко И.С., Сидорова Е.А. Семантический подход к анализу документов на основе онтологии предметной области // Труды международной конференции Диалог'2006 "Компьютерная лингвистика и интеллектуальные технологии". – М.: Изд. РГГУ, 2006. – С.468–473.

/ Сидорова Е.А. /

Сидорова Е.А.

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА
ДЛЯ АНАЛИЗА ДОКУМЕНТОВ НА ОСНОВЕ
МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ

Автореферат

Подписано в печать Объем 1,1 уч.-изд. л.
Формат бумаги 60 × 90 1/16 Тираж 100 экз.
Отпечатано на ризографе "AL Group"
630090, г. Новосибирск, пр. акад. Лаврентьева, 6