ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

на правах рукописи

Бердник Владислав Леонидович

Модели и методы семантического сравнения строк символов в коллекции документов

05.13.01 – Системный анализ, управление
и обработка информации (промышленность)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Волгоград - 2008

Работа выполнена в Волгоградском государственном техническом университете.

Научный руководитель	доктор технических наук, профессор Заболеева-Зотова Алла Викторовна.
Официальные оппоненты:	доктор технических наук, профессор Ковалев Сергей Михайлович.
	кандидат технических наук, доцент Тарасов Валерий Борисович.
Ведущая организация	Брянский государственный технический университет

Защита состоится «31» октября 2008 г. в 15-00 часов на заседании диссертационного совета Д212.028.04 при Волгоградском государственном техническом университете по адресу: 400131, г.Волгоград, пр.Ленина, 28 (ауд. 209, Главный корпус).

С диссертацией можно ознакомиться в библиотеке Волгоградского государственного технического университета.

Автореферат разослан 30 сентября 2008г.

Ученый секретарь
диссертационного совета Водопьянов В. И.

СПИСОК СОКРАЩЕНИЙ

ПИС – Предикативное Имя Сущности

БД – База Данных

УКО – Условно Кодовое Обозначение

ДО – Дефиниционное Отношение

ЕЯ – Естественный Язык

ОЕЯ – Обработка Естественного Языка

ИПС – Информационно-Поисковая Система

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Анализ документов является одним из важнейших аспектов человеческой действительности. В настоящее время для этого активно используются информационные технологии обработки информации в базах данных.

При этом одной из глобальных проблем интеллектуальной обработки данных является нахождение эффективного способа именования объектов реального мира. В реляционных базах данных такая идентификация (референция) реализуется или как формализованная система атрибутов (например, кортеж <марка автомобиля, цвет, гос. номер>), или как концептуально нечленимая символьная строка, ориентированная на понимание человеком (например, «Кабель питания компьютера 3м»). Необходимость определения семантической эквивалентности[1] двух и более символьных имен сущности[2] возникает в таких задачах, как исключение семантически дублирующих записей таблиц БД (нормализации по 1NF), перенос сведений между не реплицированными БД в виде электронных или печатных документов, ведомостей или прайс-листов, а также, в системах электронной коммерции (например, http://www.price.ru).

В настоящее время **не существует эффективных способов семантического сравнения таких имен сущностей в таблицах БД.**

В данной работе исследуется подмножество символьных имен сущности с предикативным способом указания признаков – предикативное имя сущности (ПИС). В общем случае, ПИС не является элементом ЕЯ, не используется в устной речи, на письме выделяется особым образом (кавычками, шрифтом и т.д.), может содержать УКО, элементы сообщения в «телеграфном стиле», а также полностью состоять из них. По этой причине, **востребованная на практике задача недостаточно исследуется лингвистами.**

В системах сравнения и поиска ПИС, существующих в настоящее время (PRICE.RU, «Анализ прайсов TradesMan», система «АПЛ», «Анализ прайс-листов» компании b2b-soft и т.д.) используются методы информационного поиска, от простейших дескрипторных моделей, до методов, с использованием словарей синонимов и статистики встречаемости термов. В тоже время, **вопрос о кореферентности символьных строковых идентификаторов исследуется недостаточно.**

В существующих системах используются модели и методы ориентированные на поиск и сравнение по критерию релевантность. Вопрос об **адекватности использования такого критерия для рассматриваемой проблемы остается открытым.**

Следует также отметить близкие к данной задаче исследования в области поверхностно-семантического анализа (технологии Alex) Российский НИИ искусственного интеллекта под руководством Нариньяни А.С. по выделению на основе настраиваемых синтаксических шаблонов **отдельных параметров** идентифицируемых изделии.

---

[1] Под семантически эквивалентными будем понимать элементы, однозначные по отношению к общему денотату.

[2] Термин «имя сущности» впервые введен в работах по информационной алгебре, и, в настоящее время активно используется в теории баз данных.

В тоже время, предикативное имя сущности создается и воспринимается некоторой группой лиц, что выражается в значительной аналогии грамматики ПИС и ЕЯ. Предикативное имя сущности как явление активно используется в торговле и все чаще встречается в повседневной жизни там, где существует недостаточность естественно языковых средств для выделения явления или объекта: указание и название маршрутов общественного транспорта, адресов, книг и т.п. **Сочетание языковых и неязыковых свойств ПИС увеличивает сложность его исследования.**

Решение задачи такого рода возможно активно развивающимися в последнее время методами системного анализа.

В данной работе ПИС рассматривается как статическая символьная система, что подразумевает влияние некоторой организации элементов с устойчивыми связями на выход системы (денотат, референт). Главной функцией имени сущности в БД является выделение текущей записи среди других записей в таблице, т.е. *дистинкция*[3]. В случае идентификации предметной области с эволюционирующей онтологией[4], это достигается за счет добавления в исходное ПИС нового дифференцирующего признака – терма, или модификации условно-кодового обозначения. Предикативное имя сущности семантически замкнуто, что означает автономность идентификации сущности и отсутствие связи с другими системами по силе больше или равное связям внутри системы (отсутствие ссылок на другие идентификаторы, местоимения, и т.п.). Наличие связей внутри ПИС подразумевает наличие в составе более одного элемента.

Учитывая высокую практическую востребованность такого рода компьютерных систем и фактическое отсутствие достаточных теоретических исследований и результатов, данная задача является **актуальной научной проблемой**. Так как предикативное имя сущности является сложной символьной системой, а процесс их сравнения - это определение класса сходства систем, семантическое сравнение ПИС является задачей системного анализа.

Цель диссертационной работы

Целью диссертации является исследование возможностей методов системного анализа для повышения эффективности процесса семантического сравнения предикативных имен сущности.

Задачи исследования

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ существующих методов поиска и сравнения ПИС.
2. Построить модель предикативного имени сущности и провести анализ влияния составляющих параметров на выход системы.
3. Разработать методы определения семантической эквивалентности предикативных имен сущности. Сформировать алгоритмы автоматизации семантического сравнения ПИС.

[3] Дистинкция - процедура отличения и отграничения одних (определяемых) предметов от других.

[4] Часто, (например, в торговле) наряду с основным изделием, производятся различные его подвиды, имеющие небольшие, но важные для потребителя изменения, что свидетельствует об эволюционном развитии онтологии отдельных предметных областей.

4. Разработать и реализовать программную систему – инструментарий для анализа системных свойств и определения эквивалентности предикативных имен сущности.
5. Провести анализ эффективности автоматизации методов сравнения семантически эквивалентных предикативных имен сущности.

Методы исследования

Методы системного анализа, статистические и лингвистические методы обработки естественно-языковых текстов, методы принятия решений в условиях неопределенности.

Объект исследования

Подмножество символьных имен сущности с предикативным способом указания признаков (ПИС).

Предмет исследования

Автоматизация семантического сравнения ПИС.

Научная новизна

Впервые разработаны средства и методы семантического сравнения предикативных имен сущности:

1. Разработана модель предикативного имени сущности и проведен анализ влияния термов на выход системы.
2. Разработана информационная модель рода сущностей.
3. Разработан метод определения семантической эквивалентности предикативных имен сущности на основе информационной модели рода сущности.
4. Разработан метод ранжирования альтернативных вариантов семантически толерантных ПИС.
5. Разработаны алгоритмические операции для автоматизации сравнения семантически эквивалентных предикативных имен сущности.

Достоверность результатов диссертационной работы.

Достоверность научных положений, выводов и рекомендаций подтверждена результатами экспериментов, а также результатами использования материалов диссертации и разработанной системы в коммерческой организации ООО «Прайм» в соответствии с актом внедрения.

На защиту выносятся:

- Модель предикативного имени сущности как статической системы.
- Модель рода сущностей как эффективное средство семантического сравнения ПИС.
- Метод определения семантической эквивалентности предикативных имен сущности на основе информационной модели рода сущности.
- Метод ранжирования альтернативных вариантов семантически толерантных ПИС

Практическая значимость работы

Разработанные методы и алгоритмы позволяют повысить качество поиска информации о товаре в системах электронной коммерции по запросу

пользователя, маркетинговых службах и снабжении предприятий, а также сократить время, затрачиваемое на перенос сведений между не реплицированными базами данных за счет автоматизации процесса сравнения кореферативных ПИС.

Реализация результатов работы

Результаты диссертации и программная система использованы при автоматизации бизнес процессов на предприятии ООО «Прайм», что подтверждается актом внедрения.

Апробация работы

Основные положения и результаты диссертации докладывались, обсуждались и получили одобрение на конференциях:

- «Интеллектуальные системы» (AIS'08) и «Интеллектуальные САПР» (CAD-2008), международная научно-техническая конференция, Дивноморское, 2008;
- Информационные технологии в науке, образовании, телекоммуникациях и бизнесе (IT+SE`07).-XXXIV международная конференция и дискуссия научного клуба, Ялта-Гурзуф, 2007;
- Инновационные технологии в управлении, образовании, промышленности "АСТИНТЕХ-2007", Астраханский государственный университет, Астрахань, 2007;
- Системные проблемы надёжности, качества, математического моделирования, информационных и электронных технологий в инновационных проектах: Инноватика-2007;
- Информационные технологии в образовании, технике и медицине, международная научно-техническая конференция, Волгоград, 2000

Публикация результатов работы.

По теме диссертации опубликовано 12 работ, в том числе: 4 статьи опубликованы в изданиях, входящих в перечень ВАК; 8 статей в сборниках трудов; 4 материалов конференций.

Структура и объем диссертации

Диссертационная работа изложена на 131 страницах машинописного текста, содержит 6 рисунков и 8 таблиц, состоит из введения, пяти глав, заключения, списка литературы из 165 наименований и 2 приложений на 7 страницах.

Автор выражает глубокую **благодарность** научному консультанту доктору технических наук, профессору Андрейчиковой Ольге Николаевне за оказанное содействие и поддержку в ходе выполнения диссертационной работы.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы диссертации, определяются объект, предмет, цель, задачи диссертационного исследования, методы анализа, раскрываются научная новизна, теоретическая и практическая значимость работы, формулируются положения, выносимые на защиту.

В первой главе вводится определение предикативного имени сущности, рассматриваются вопросы применения методов и мер информационного поиска в

существующих коммерческих системах. В связи очевидной аналогией естественного языка и предикативных имен сущности, рассматриваются современные научные лингвистические подходы и методы их решения, в том числе с применением ЭВМ.

Определение 1.1 ***Предикативное имя сущности*** (ПИС) – это символьная строка конечной (лимитированной) длины, в которой отдельные слова (термы) или группы слов задают отдельные признаки, в совокупности и во взаимодействии обеспечивающие идентификацию и выделение сущности или вида сущности среди остальных в предметной области. Такая строка должна идентифицировать сущность, либо совокупность семантически близких сущностей, воспринимаемых, согласно предметной области, как атомарный идентифицируемый объем.

В настоящее время существует более десяти косвенных и прямых аналогов разработанной программной системы. Наиболее известная из них – PRICE.RU, которая в своей основе использует методы на основе локального и глобального взвешивания термов строк прайс-листов и запросов[5]. Такая система электронной коммерции отличается от обычных поисковых машин Интернет (например, www.yandex.ru) наличием дополнительных словарей, которые позволяют транслировать различные способы написания (и сокращения) одних и тех же признаков изделий (например, «Диалог» и «Dialog») в формализованное представление.

При проведении анализа других аналогов («Анализ прайсов TradesMan», система «АПЛ», «Анализ прайс-листов» компании b2b-soft и т.д.) обнаруживается, что в сравнительно недорогих системах используется дескриптивная модель поиска, т.е. для каждой итерации поиска пользователь вынужден выделять подстроку –дескриптор. В более дорогих коммерческих системах реализованы алгоритмы информационно-поисковых машин, как это сделано, например для PRICE.RU.[6]

В рамках информационного поиска изучаются вопросы поиска документов, обработки результатов поиска, разрабатываются и исследуются критерии, метрики и меры, а также ряд смежных вопросов: моделирования, классификации, кластеризации и фильтрации документов, проектирования архитектур поисковых систем и пользовательских интерфейсов, языки запросов, и т.д. Важнейшим критерием оценки информационно-поисковой системы является *релевантность* ее ответов, т.е. соответствие ответов системы информационным потребностям пользователя. Строго говоря, термин *релевантность* используется для ссылки на семейство разных критериев. Каркас этого семейства имеет три размерности: *Информационная потребность*, *информационные ресурсы (потенциальная возможность), контекст использования информации(область интересов пользователя, его знаний и т.д.)*

Сравнение предикативных имен сущности имеет ряд принципиальных отличий. Единственным критерием сравнения ПИС является их кореферентность. Референт является системным выходом ПИС и зависит не только от состава, но и от связей между элементами. Предикативное имя сущности выполняет функции

[5] Информация получена из E-mail переписки с разработчиками PRICE.RU
[6] Информация получена из документации к указанным программным продуктам.

дистинкции и дефиниции. Референт является значением дистинкции, а понимание ПИС человеком является значением дефиниции. Таким образом, не все элементы и связи между ними влияют на идентификацию сущности (дистинкцию референта).

Среди наиболее известных и влиятельных работ, посвящённых формальному описанию языков, можно выделить теорию формальных грамматик Н. Хомского[7] и модель «смысл – текст» И. Мельчука[8]. Теория Смысл – Текст, являясь практически первой в СССР «кибернетической» теорией в области лингвистики, ставящей перед собой прикладные цели — создать двунаправленный лингвистический процессор, использовала Толково-Комбинаторный Словарь для задания морфологических, синтаксических, семантических характеристик и толкований заглавного слова. Теория постулирует многоуровневую модель языка. Различают следующие уровни: фонологический, поверхностно-морфологический, глубинно-морфологический, поверхностно-синтаксический, глубинно-синтаксический, семантический. Каждый уровень характеризуется набором собственных единиц и правил представления, а также набором правил перехода от данного уровня к соседним.

Технология лексического анализа Alex (система создана в Российский НИИ Искусственного Интеллекта, под руководством А.С. Нариньяни) позволяет с помощью настраиваемых лексических шаблонов решать задачи:

- Поиск в текстовых массивах различной степени структуризации определенных фрагментов, извлечение знаний;
- Нормализация слабоструктурированных массивов данных как с точки зрения их структуры, так и с точки зрения качества их наполнения.

Технология лексического анализа Alex позволяет транслировать слабоструктурированные символьные строки в единицы поверхностно-синтаксического уровня. Для семантического сравнения ПИС требуется разработка моделей и методов идентификации сущности по признакам.

Моделирование ПИС значительно упрощается, если в предметной области выделить таксонометрические единицы. Предикативное имя сущности, как определяющее дефиниционного отношения содержит в своем составе классифицирующий родовой признак.

В самом общем виде дефиниционное отношение представлена схемой Dfd+Dfn, в которой «Dfd» - это дефиниендум, или определяемое, «+»- связка, «Dfn»-дефиниенс или определяющее. В классической логической дефиниции дефиниендум представляет собой видовое понятие, а дифиниенс –(ближайшее) родовое понятие и видовое отличие. Структурный анализ дефиниций предполагает выделение таких отношений как тождество, включение и аддиция (дистинкция). Предикативное имя сущности можно представить как определяющее (дефиниенс) ДО. Определяемое находится вне языкового представления.

Таким образом, вопрос семантического сравнения предикативных имен сущности остается открытым. Существующие методы информационного поиска

[7] Хомский Н. Аспекты теории синтаксиса. — М.: Изд-во БГК им. И.А.Бодуэна Де Куртенэ, 1999. — 235 с.

[8] Мельчук И. А. Опыт теории лингвистических моделей «Смысл - Текст». М.: Наука, 1974.

не имеют формального аппарата для определения кореферентности ПИС, отсутствуют адекватные задаче методы, критерии и меры. В настоящий момент не разработано подходов и методов для семантического анализа ПИС, поскольку при явной аналогии ПИС и естественного языка, в общем случае, предикативное имя сущности не является естественно- языковой конструкцией.

Во второй главе проведен структурно-функциональный анализ предикативного имени сущности, выполнен анализ компонента дистинкции предикативного имени сущности, описаны системные связи внутри предикативного имени сущности, и связи с надсистемой и средой.

С точки зрения семиотики, предикативное имя сущности можно рассматривать как сложный по внутренней структуре искусственный символьный знак, восполняющий недостаточность естественно-языковых средств идентификации элементов описываемой предметной области (выполняющий номинативную функцию). Атомарность объема понятия ПИС инвариантна во времени и пределах информационной системы, но зависит от выбранной при проектировании СУБД концептуальной модели и предметной области, может быть единичным экземпляром, видом объектов (явлений), или совокупностью (набором, комплектом) и т.д.

Введем следующие обозначения:

- U – универсум ПИС, в данном случае, специальный корпус текстов;
- S – множество идентифицируемых сущностей (видов сущностей);
- C – универсум признаков.

Областью наших исследований является случай, когда между множествами U и S существует сюръекция ref:U->S, где ref – отношение идентификации сущности (референция). Это достигается тем, что во множестве S некоторые группы семантически близких сущностей, согласно предметной области, представлены как единый элемент множества. Из этого следует, что универсум U состоит из подмножеств U_s *имен сущности- синонимов:*

$$U_s = \{u: (\exists s \in S)\ (\forall u \in U)\ [s=ref(u)]\} \qquad (2.\ 1)$$

Отношение ref задано в общем виде, объективно существует, и является одной из главных целей нашего исследования. Рассмотрим более подробно элементы множества S. Проводя аналогию с конструкциями естественного языка, сущность s является денотатом соответствующих ПИС.

Различные сущности должны иметь различный набор признаков. Пусть s_1, s_2 – сущности, C_1, C_2 –соответствующие множества признаков.

$$(\ \forall s_1=\gamma(C_1))(\forall s_2=\gamma\ (C_2))[\ s_1 \neq s_2\ \Rightarrow C_2 \neq C_1], \qquad (2.\ 2)$$

где

γ:C->S, где γ – отношение идентификации сущности по набору признаков.

В отличие от лингвистического описания сущности, введем более строгие ограничения. В подмножествах C зададим как наличие, так и отсутствие семантического признака. Для множеств C и S (множества признаков и сущностей), введем бинарное отношение M=S×C. С помощью характеристической функции (предиката) множества это можно представить в виде:

$$\mu_M<s,c>=\begin{cases}1, & если \quad s \quad обладает \quad c\\ 0, & если \quad s \quad не \quad обладает \quad c\end{cases} \qquad (2.3)$$

Определим минимальный набор признаков C_m для идентификации сущности s.

$$(\exists C_m \subset C)\ (\forall c_m \in C_m)\ [\ s=\gamma\ (C_m)\ \Rightarrow\ \gamma\ (C_m) \neq \gamma\ (C_m \setminus c_m)] \qquad (2.4)$$

$$(\forall s \in S)(\forall C_1 \subset C)\ (\exists C_m \subset C)\ [\ s=\gamma\ (C_1) = \gamma\ (C_m)\ \Rightarrow\ C_m \subseteq C_1] \qquad (2.5)$$

Предикативное имя сущности имеет две функции. Во-первых, как элемент программной системы, оно должно выделять идентифицируемую сущность среди множества других, т.е. выполнять функцию ***дистинкции***. Во-вторых, как средство информирования человека, оно дает определение и описывает наиболее значимые признаки сущности, выполняя функцию ***дефиниции.***

Логически верное ПИС должно идентифицировать единичный объем. Свойство идентификации единичного объема (единичного объекта), позволяет утверждать, что функция дистинкции предикативного имени сущности является первичной по отношению к функции дефиниции. Кроме того, не обладая необходимым качеством для использования в вычислительной среде, таким как выделение сущности среди прочих, ПИС будет нарушать детерминированную модель информационной системы (например, будут нарушаться условия нормализации БД по 1NF), и, следовательно, является ошибочным. Отсутствие дефиниционного компонента в ПИС, приведет к дополнительным трудностям понимания непосвященному кругу лиц, что не исключает возможности использования такого предикативного имени сущности (пример ПИС: «LG 80130N»). Компонент дистинкции может быть полностью включенным, частично включенным и не включенным в компонент дефиниции.

Введем множество $O \subset C$ родовых предикатов S, следующим образом.

$$(\forall s \in S)\ (\exists o \subset O)\ (\exists C_s \subset C)\ \ [s=\gamma(C_s)\ \Rightarrow o \in C_s] \qquad (2.6)$$

В ПИС, используемых в маркетинге и торговле, в качестве родового признака указывается классификационный признак продукта или изделия.

Множество S состоит из родов сущностей S_o по родовому признаку o.

$$(\forall o \in O)\ (\exists S_o \subset S)(\ \forall s_o \in S_o)(\ \exists C_s \subset C)\ [s_o=\gamma(C_s)\ \Rightarrow o \in C_s] \qquad (2.7)$$

где o– родовой признак (предикат),

C_s- предикаты видовой характеристики и дистинкции.

Роды сущностей имеют собственные подмножества признаков.

$$L=C\times O,\ \varepsilon_L<c,o>=\begin{cases}1, & если \quad \exists C_L, c \in C_L, o \in C_L, \gamma(C_L) \neq \varnothing\\ 0, & если \quad \forall C_L, c \in C_L, o \in C_L, \gamma(C_L) = \varnothing\end{cases} \qquad (2.8)$$

ПИС, в случае идентификации сущности, которая в своем составе содержит иные сущности, может принимать *сложносоставную или односоставную форму*.

Сложносоставная форма предикативного имени сущности применяется при отсутствии установившегося классифицирующего термина–прототипа в предметной области. В таком случае в качестве родового признака сущности используются слова «набор», «комплект», «комбайн» и т.п. Например, «Набор для рыбалки (спиннинг, фонарь, садок)». Таким образом, многосоставная ПИС представляет собой сложную систему, и состоит из предикативных имен сущности – подсистем идентифицирующих входящие в состав сущности.

Повторяемость структуры (грамматики) предикативного имени сущности также проявляется для определения понятий-признаков. Например, «Черный картридж для *принтера Hewlett Packard LJ 1600»* **.** Подстрока «принтер Hewlett Packard LJ 1600» содержит собственный родовой признак «принтер», видовую характеристику «Hewlett Packard» и компонент дистинкции «LJ1600». Однако, идентифицируемое изделие «картридж» не содержит в своем составе изделия «принтер», а, следовательно, ПИС по форме является односоставной. Таким образом, предикативное имя сущности, включенное в состав основного ПИС, несет функцию (имеет цель) дефиниции понятия-признака: «для принтера определенной модели». Такие предикативные имена будем называть: *подчиненные ПИС*. Подчиненное ПИС отличается от составной ПИС не только назначением, но и объемом идентифицируемого понятия (обычно, более одной сущности) и допустимыми способами дефиниций.

В общем случае, предикативное имя сущности может иметь следующие формы:

1. Простейшее ПИС – состоит из родового признака или компонента дистинкции.
2. Простое ПИС - не содержит в своем составе подчиненных или составных ПИС.
3. Составное ПИС – имя сущности входящей в состав более сложной сущности.
4. Подчиненное ПИС – указание параметра (признака) через номинацию некоторого непустого множества ассоциативно связанных сущностей.
5. Комплексное ПИС – предикативное имя сущности, содержащее в своем составе подчиненное ПИС.
6. Многосоставное ПИС – предикативное имя сущности, содержащее составные ПИС.
7. Управляющая ПИС – комплексное или многосоставное предикативное имя сущности за исключением подчиненных или составных ПИС.

Компонент дистинкции ПИС реализуется, преимущественно, как условно-кодовое обозначение или имя собственное. Условно-кодовое обозначение состоит из изменяемой части, но может содержать неизменяемую (постоянную) часть. Постоянная часть, как правило, содержит тип идентификатора, который соотносится с областью его применения, либо иные классифицирующие признаки (например, код региона на государственных автомобильных номерах). Имя собственное позволяет восстановить через пресупозицию внелингвистическую информацию, которая в дальнейшем может быть использована для анализа ПИС.

Предикативное имя сущности преимущественно используется в СУБД и, в частности, может быть обычной конкатенацией символьных атрибутов кортежа отношения БД. Существует сходство между элементом теории БД – домен атрибутов и семантическим полем. Содержимое домена атрибутов определяется проектировщиком БД, а также может пополняться «стихийно». Семантическое поле – более формализованное и исследованное понятие, чем домен атрибутов БД.

Связи между подсистемами на уровне компоновки символьной структуры ПИС преимущественно реализуются, как инкапсуляция (внедрение) подчиненных (нижележащего уровня) подсистем в управляющие (вышележащего уровня)

подсистемы. Таким образом, многосоставные и комплексные ПИС, как формат представления символьных данных, являются контейнером, который содержит атрибуты управляющий ПИС и ПИС нижележащего уровня.

Применяется, но сравнительно редко, организация ПИС с внутренними ссылками (например, местоимениями). Организация такого рода системных связей неэкономно расходует место в символьной строке. Пример, «Набор (гелевая ручка, ***карандаш*** STAFF, линейка), ***карандаш*** с ластиком». Такое комплексное или многосоставное ПИС не членимо на отдельные предикативные имена сущностей, так как подсистемы нижележащего уровня семантически не замкнуты, а связаны с другими частями системы. Предикативные имена сущностей с внутренними ссылками будем называть *сильносвязанными ПИС*.

Разрешение многозначности символьных единиц и их конструкций предикативного имени сущности возможно только на основании внелингвистической модели. В рамках родовой принадлежности возможно обобщенное моделирование подмножества идентифицируемых сущностей. Учитывая сложность такого моделирования, обобщенная модель рода сущностей не может претендовать на универсальность, однако, при условии небольшого размера множества рода и достаточной точности его модели, можно разработать высокоэффективные алгоритмы сравнения предикативных имен сущности на семантическую эквивалентность.

В третьей главе исследуется метод ранжирования семантически толерантных предикативных имен сущности, определяются критерии метрики и меры, разрабатывается модель родов сущностей, а также метод определение семантической эквивалентности ПИС на основе моделей родов сущности.

Под *документом сущности* D^s будем называть совокупность информации о сущности s.

$$D^s =< U^s, Y^s>, \qquad (3.\ 1)$$

где

U^s – множество семантически эквивалентных (кореферентных) ПИС;

Y^s – разнородная информация о сущности и способах сравнения на эквивалентность (рассмотрено в следующих разделах).

Соответственно, *коллекция документов* K:

$$K=\{D^s_1, D^s_2, \ldots D^s_n,\}, \qquad (3.\ 2)$$

где n-количество документов в коллекции.

В предыдущей главе введены множества U, S, U_s и отношение ref:U->S. Множество U является объединением двух подмножеств $U=U^+\cup U^-$. Соответственно, U^+ содержит предикативные имена сущности, для которых значение функции s=ref(u) установлено. Множество U^- содержит предикативные имена сущности, для которых значение функции ref(u) необходимо определить.

ПИС идентифицируют только одну общую для подмножества сущность.

$$(\forall u^s_i\in U^s)\ (\forall u^s_j \in U^s)\ [ref(u^s_i) \equiv ref(u^s_j)] \qquad (3.\ 3)$$

Для разбиения множества U выполняется следующее условие.

$$U= U^s_1\cup U^s_2\cup\ldots\cup U^s_k \text{ так, что } U^s_I \cap U^s_j =\varnothing \text{ для } i\neq j. \qquad (3.\ 4)$$

Элементы множества U^s семантически эквивалентны внутри подмножества, так как указывают на одну и ту же сущность.

Определим отношение (синтагму) эквивалентности R:

$$(\forall x,y\in U^s)\ [(U^s\subset U) \Rightarrow (xRy=True)], \quad (3.5)$$

$$(\forall x\in U^s{}_i,\ U^s{}_i\subset U)\ (\ \forall y\in U^s{}_j,\ U^s{}_j\subset U\ )\ [(U^s{}_i\cap U^s{}_j=\varnothing) \Rightarrow (xRy=False)] \quad (3.6)$$

Верны и обратные утверждения:

$$(\forall x,y\in U)\ [(xRy=True) \Rightarrow (\ \exists U^s\subset U\ \&\ x,y\in U^s)] \quad (3.7)$$

$$(\forall x\in U^s{}_i,\ U^s{}_i\subset U)\ (\ \forall y\in U^s{}_j,\ U^s{}_j\subset U)\ [(xRy=False) \Rightarrow (U^s{}_i\cap U^s{}_j=\varnothing)] \quad (3.8)$$

Из выражения (3.5), в случае тождественности x≡y выводится условие рефлексивности:

$$xRx=True \quad (3.9)$$

Заменив в выражении (3.7) xRy на yRx, мы получаем, что также $x,y\in U^s$, и из выражения (3.5) получаем xRy=True.

$$(\forall x,y\in U)\ (yRx=True) \Rightarrow (\ \exists U^s\subset U,\ x,y\in U^s) \Rightarrow xRy \quad (3.10)$$

Таким образом, соблюдается условие симметричности.

$$yRx \Rightarrow xRy \quad (3.11)$$

Заменив в выражении (3.5) xRy на yRz, мы получаем:

$$(\forall x,y\in U)\ (xRy=True) \Rightarrow (\ \exists U^s{}_i\subset U\ \&\ x,y\in U^s{}_i), \quad (3.12)$$

$$(\forall y,z\in U)\ (yRz=True) \Rightarrow (\ \exists U^s{}_j\subset U\ \&\ y,z\in U^s{}_j), \quad (3.13)$$

из чего следует:

$$(x,y\in U^s{}_i)\ \&(y,z\in U^s{}_j) \Rightarrow (\ U^s{}_i\cap U^s{}_j=y) \quad (3.14)$$

учитывая выражение (3.4):

$$U^s{}_i\cap U^s{}_j=\varnothing \text{ для } i\neq j, \quad (3.15)$$

получаем, что i=j.

$$(\forall x,y,z\in U)\ (xRy=True)\ (yRz=True) \Rightarrow (\ \exists U^s\subset U\ \&\ x,y,z\in U^s) \Rightarrow (xRz=True)$$

Таким образом, соблюдается условие транзитивности.

$$xRy\ \&\ yRz \Rightarrow xRz \quad (3.16)$$

Отношение R на множестве U является отношением эквивалентности, так как для него соблюдаются условия рефлексивности, симметричности и транзитивности.

Введем вектор соответствия X' документов коллекции K и предикативного имени сущности $u^-\in U^-$:

$$X'=\{x_1, x_2, \dots, x_n), \quad (3.17)$$

где, учитывая, что $D_s^i=<U_s{}^I, Y_s{}^i>$, $D_s^i\in K$; i=1..n; n=|K|

$$(\forall u^+\in U^s{}_i)\ \ x_i=R(u^-,u^+) \quad (3.18)$$

Коллекции K не всегда содержит искомый документ сущности анализируемого ПИС u^-. В этом случае необходимо создание и добавление документа в коллекцию. Введем в вектор X' элемент x_z, который свидетельствует о необходимости пополнения коллекции K .

В этом случае, X является единичным координатным вектором.

$$X=\{x_1, x_2, \dots, x_n, x_z), \quad (3.19)$$

где $x_z=1-\sum(x_i)$, $i=1..n$; $n=|K|$

Таким образом, задача поиска семантически эквивалентных предикативных единиц текста в коллекции документов сводима к нахождению единичного координатного вектора X.

Как упоминалось в главе 1, существующие коммерческие системы используют методы информационного поиска. Эти методы основаны на эвристической оценке релевантности:

- чем чаще терм встречается в документе, тем он более релевантен по отношению к документу;
- чем чаще терм встречается среди всех документов коллекции, тем хуже он отражает различие между документами;
- если некоторый терм не встречается в документах, то он указывает на неполноту коллекции.

Согласно векторной модели, близость документа D к запросу оценивается как корреляция между векторами их описаний[9]. Эта корреляция может быть вычислена как скалярное произведение соответствующих векторов описаний.

Сравнение предикативных имен сущности имеет ряд принципиальных отличий.

1. Единственным критерием сравнения ПИС является их кореферентность.
2. Референт является системным выходом ПИС и зависит не только от состава, но и от связей между элементами.
3. Предикативное имя сущности выполняет функции дистинкции и дефиниции. Референт является значением дистинкции, а понимание ПИС человеком является значением дефиниции. Таким образом, не все элементы и связи между ними влияют на идентификацию сущности (дистинкцию референта).

Руководствуясь вышеизложенным, мы приходим к выводу, что любой документ коллекции состоит из совокупности термов. Ранжируя термы по значимости для выражения предикативным именем сущности функции дистинкции следует выделить отдельные классы.

1. Условно-кодовые обозначения и имена собственные наиболее значимые термы для выделения сущности среди остальных в коллекции. Такие термы имеют наивысшую оценку.

2. Родовые (классификационные) признаки термы – указывают на состав и структуру ПИС. Они в большинстве случаев обладают высокой глобальной частотой и разделяют область поиска в коллекции документов, в том смысле, что сравниваемые ПИС должны принадлежать к одному роду. Термы рода управляют взвешиванием, но сами на значение меры не влияют.

3. Для термов –признаков вида следует использовать такую меру, которая позволит определять степень выражения функции дистинкции в ПИС. Реализующие функцию дистинкции признаки обязательны для идентификации сущности, хотя соответствующие им термы могут отсутствовать, и выражаться через УКО или имя собственное. Таким образом, такие термы наиболее частотны

[9] Сравнительный анализ различных метрических мер близости сделан в работе: Козачков Л. С., Патиоха А. А., Хоменко А. И., О моделировании некоторых метрических мер близости. /Сб. Информационный анализ и лингвистические проблемы информационных систем. К., ИК АН УССР, 1975, стр. 7-15.

по локальным и глобальным схемам взвешивания. Термы, не участвующие в реализации функции дистинкции, поясняют пользователю потребительские качества товара. Так как термы дефиниции (например, рекламы и пояснения) маловероятны для появления во всех ПИС документа, локальная частота таких термов ПИС невысокая. Таким образом, признаки вида необходимо сравнивать с учетом локальной частоты терма в документе, которая определяется по формуле.

$$l_{ij} = \frac{\left|U^s_{j,i}\right|}{\left|U^s_j\right|}, \tag{3.20}$$

где

$|U^s_j|$ - мощность множества семантически эквивалентных ПИС документа D^s_j коллекции K;

$U^s_{j,i}$ - подмножество $U^s_{j,i} \subset U^s_j$ документа D^s_j коллекции K содержащих терм p_i. $|U^s_{j,i}|$ - мощность этого подмножества.

Локальная частота терма в документе является косвенным признаком участия терма в реализации функции дистинкции. С другой стороны, чем больше количество видовых признаков ПИС совпадают при сравнении, тем более вероятно совпадения минимального необходимого набора видовых признаков для дистинкции. Малочастотные термы, наиболее вероятно, реализуют функцию дефиниции ПИС, а следовательно не должны участвовать в анализе. Таким образом, для высокочастотных термов признаков вида адекватна аддитивная мера сравнения, например, мера включения множества термов анализируемого ПИС в множество всех термов документа.

$$a_{ij} = \frac{\left|P^s \cap P^s_j\right|}{\left|P^s\right|}, \tag{3.21}$$

где

P^s - множество термов анализируемого ПИС;

P^s_j - множество термов документа D^s_j коллекции K.

Модель рода сущности можно представить в виде кортежа.

$$E=<R,G,P,W,K,M>, \tag{3.22}$$

где

R - Правила определения рода для анализируемого (неизвестного) предикативного имени сущности.

G – Типовая структура ПИС (число подчиненных ПИС и связи между ними).

P - Необходимый набор видовых признаков для идентификации.

W - Семантические поля термов для разрешения семантической неоднозначности (омонимии, полисемии) и метода семантических противоречий.

K - Значение признаков «по умолчанию».

M - Модели построения (интерпретации) компоненты дистинкции ПИС.

Правила определения рода анализируемого ПИС представляет собой набор шаблонов. При наложении или поиске в анализируемом предикативном имени сущности можно сделать предположение о соотнесении сущности с определенным родом. Шаблоны предназначены для выделение: термов – признаков рода, непосредственно указанных в ПИС; компоненты дистинкции с

характерных для рода способом построения; видовых признаков, свойственных только определенному роду или характерной совокупности признаков.

Зарезервированные (терминальные) символы шаблонов:

'[' и ']' – открывающая и закрывающая скобка символьного шаблона;

'{' и '}' – открывающая и закрывающая скобка шаблона, основанного на допустимости любого терма из указанного в скобках идентификатора семантического поля;

'?' – любой символ;

'&' – оператор дизьюнкции, указывает на необходимость реализации более одного шаблона.

Типовая структура предикативного имени сущности G – это дерево (граф) вершинами которого является двойка <'Родовой признак', 'Тип ПИС'>. Тип ПИС может иметь одно из значений {'Составное ПИС', 'Подчиненное ПИС'}. Дуги графа указывают на связи между ПИС.

Необходимый набор P видовых признаков для идентификации – это совокупность кортежей, элементами которых являются идентификаторы семантических полей применительно к каждой вершине графа.

Семантические поля термов W объединяют семантически связанные между собой термы, и полностью описывают возможные значение признака сущности применительно к роду сущности(вершине графа).

Модели построения M компоненты дистинкции ПИС это способ интерпретации содержимого УКО или имени собственного. Компонент дистинкции в этой модели представлен как вектор (массив). Для каждого элемента или последовательности элементов вектора назначен тип поля и способ его интерпретации.

Семантическое сравнение ПИС на основе модели рода сущности состоит из последовательности действий: определение рода сущности ПИС, или совокупности признаков рода; анализ границ и выделение простых ПИС в составе сложного; определение кореферентности простых ПИС; ранжирование альтернативных (толерантных) ПИС; выбор лицом принимающим решение верного варианта.

В четвертой главе описываются общие принципы построения программной системы ASTEND сравнения предикативных имен сущности, рассматриваются вопросы представления данных и интеграция с коммерческими системами.

Программная система ASTEND состоит из подсистемы интеграции с корпоративной информационной системой, хранилища коллекции документов, модуля поверхностно-синтаксического анализа, словаря синонимов, модуля выделения и обработки компоненты дистинкции, таблицы пресупозиций, таблицы признаков рода, хранилища моделей родов, алгоритмов семантического сравнения предикативных имен сущности.

Подсистема интеграции с корпоративной информационной системой должна реализовывать программный интерфейс (например, COM, OLE, .NET) и предоставлять сервис семантического сравнения предикативных единиц текста. Для наполнения системных таблиц, отладки и контроля работы алгоритмов сравнения необходим сервисный интерфейс пользователя. В качестве основного межпрограммного интерфейса была использована технология COM объектов с

описанием программных интерфейсов на языке IDL. Подобный подход универсален для большинства информационных систем, и, в частности, позволил интегрировать программу ASTEND с системой управленческого учета «1С Предприятие 7.7 Торговля и склад». Для репликации таблиц – справочников товаров 1С Предприятия и коллекции документов ASTEND используется прямой доступ к единой БД на сервере MS SQL Server 2000, но допускается более медленная репликация через методы интерфейса COM сервера.

Хранилище коллекции документов представляет собой таблицу БД со следующими полями:

|*Ключ сущности* |*тэг записи* |*значение записи* |*символьная строка* |,

где

ключ сущности – идентификатор принадлежности кортежа документу сущности;

тэг записи – указывает способ интерпретации полей *значение записи* и *символьная строка*. В частности, отдельное значение этого поля предназначено для репликации коллекции документов с корпоративной информационной системой.

Модуль поверхностно-синтаксического анализа с использованием словаря синонимов приводит различные способы написания и сокращения термов к единой форме. Устойчивые словосочетания предметной области, такие как «коврик для мыши» воспринимаются модулем как единый терм с трансляцией в служебное кодовое обозначение. На данном уровне производится морфологический анализ термов с приведением к именительному падежу единственного числа. Модуль производит преобразование символов верхнего регистра в нижний. Алгоритмы модуля используют упрощенную библиотеку настраиваемых пользователем шаблонов.

Модуль обработки и выделения компоненты дистинкции в своей работе использует эвристические приемы (например, если терм содержит цифры, то это условно-кодовое обозначение). Учитывая неконтролируемое качество содержимого предикативного имени сущности, допускаются различные способы написания УКО (например, «A18E4», «A-18E4», «A 18 E4»). Анализ такого рода компонентов дистинкции основывается на предположении (модели), что информативным является начертание символов и их последовательность. Знаки препинания, пробелы и прочие служебные символы УКО при сравнении можно игнорировать. Такая модель показала высокую эффективность при эксплуатации программы ASTEND на протяжении 4 лет.

Таблица пресупозиции позволяет восстанавливать внелингвистические признаки сущности, соответствующие имени собственному. Например, «Stylus» - струйный принтер Epson.

Таблица признаков рода содержит название, идентификатор рода и правила его выделения. Использование таблицы признаков рода позволяет в процессе семантического сравнения ПИС индексировать поиск необходимых записей, как в коллекции документов, так и в хранилище моделей родов.

Хранилище моделей родов сущностей содержит значения полей информационной модели рода сущности в соответствии с описанием в третьей главе.

В программе ASTEND реализованы алгоритмы: декомпозиции сложного ПИС на простые ПИС на основе определения границ, анализа подчиненных ПИС и семантического сравнения простых ПИС. Анализ сильносвязанных ПИС в программе не предусмотрен. Учитывая, что допускается наличие различных альтернативных вариантов анализа ПИС, необходимо их ранжирование как семантически толерантных ПИС. Соответствующие методы и метрики критериев задачи принятия решения описаны в третьей главе.

В соответствии с ранжированием, сравниваемые варианты предлагаются лицу принимающему решение.

В пятой главе рассмотрены примеры семантического сравнения предикативных имен сущности в процессе эксплуатации программной системы. Приведены данные по трудозатратам и эффективности использования программной системы при коммерческом применении.

Рассмотрим следующий содержательный пример: *«Набор, фотобумага LOMOND + картридж T036140 для Stylus C42Plus/ C42S/ C42SX/ C42UX (черн.), подарочный экземпляр»*. Эта многосоставная комплексная ПИС может быть представлена в виде следующего дерева.

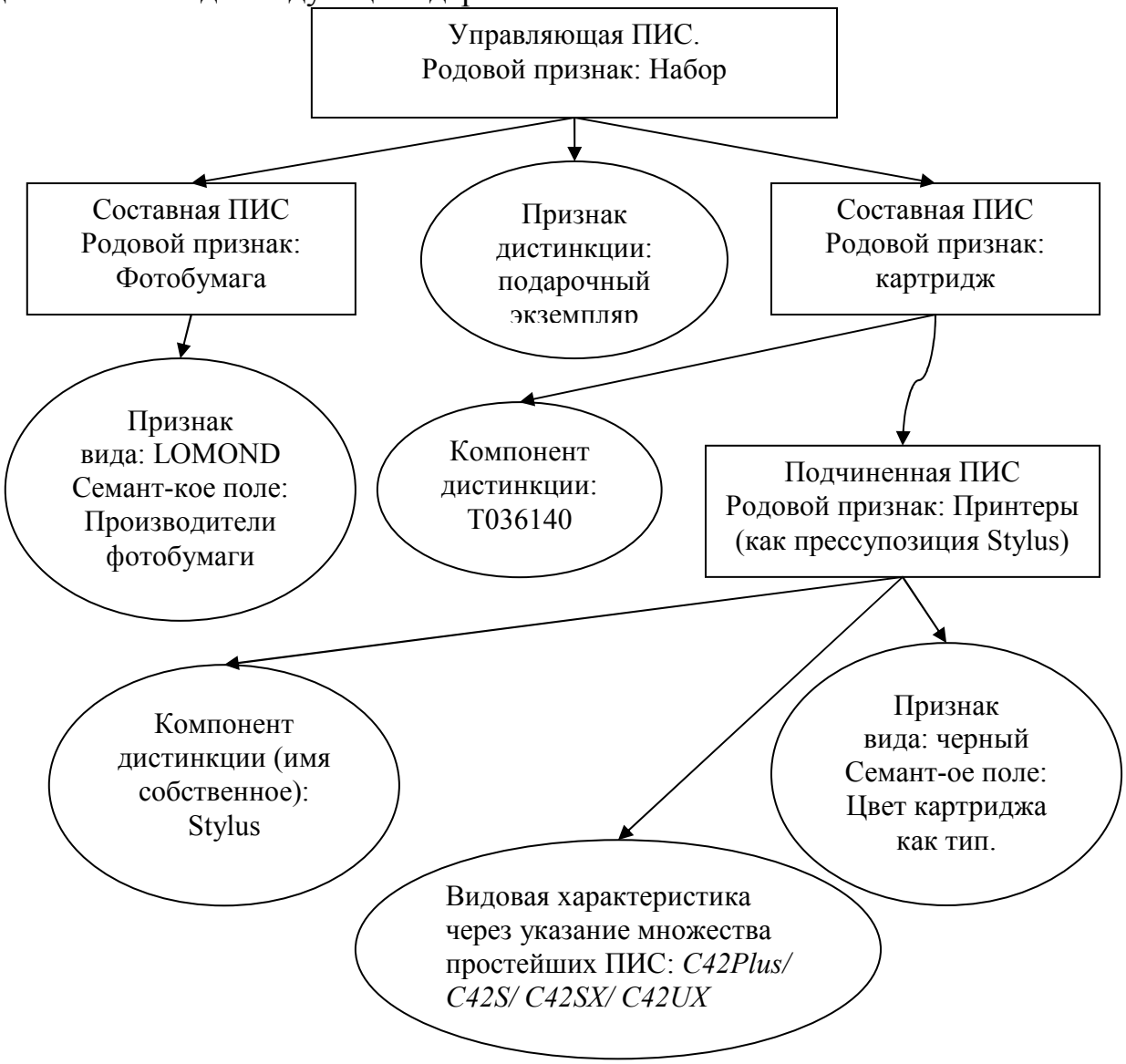

Рисунок 5.1 Иерархическая структура предикативного имени сущности.

Рассмотрим документ сущности $D^s =< U^s, Y^s >$.
U^s=**{**
'Видеокарта AGP 256МБ ASUS "AX1650Pro/HTD" (Radeon X1650 Pro, DDR2, D-Sub, DVI, TV) (ret)**',**
'AX1650PRO/HTD/256 SVGA <AGP8x> Asus 256Mb ATI Radeon X1650PRO 600MHz, DDR2 800MHz/128 bit, DVI, D-SUB**',**
'Видеоплата AGP 256M ATI Radeon AX1650Pro/HTD ASUS DDR2 128bit TV**'}**
Y^s ={**<**'УКО'**,'**ax1650pr0htd**'>,<**'Род '**,** 'Видеоплата'**>}**

Модель рода сущности E=<R,G,P,W,K,M>.

Шаблоны определения рода сущности: [видеокарта] и [видеоплата].

Типовая структура ПИС: одна вершина <Род: видеоплата, Тип: составная>

Необходимый набор видовых признаков для идентификации:{Производитель, Объем памяти, Графический чипсет, Стандарт системного разъема, Наличие tv-out}

Семантические поля: Производитель={ASUS, …, Palit}, Объем памяти={128MB, 256MB,…, 1024MB}, Графический чипсет= {X1650SE, …, X3140}, Стандарт системного разьема = {AGP, PCI-E}, Способы упаковки={OEM, RTL}.

Значение признаков «по умолчанию»: <Способ упаковки, OEM>

Модель компоненты дистинкции: отсутствует.

Анализируемое ПИС u=
'VCASEAX1650ProHTD256 Видеокарта PCI-E ASUS EAX1650Pro/HTD 256MB DDR2 <ATI Radeon X1650Pro,TV-Out, HD**'.**

Примерная последовательность действий для семантического сравнения ПИС:

1. Анализируемое ПИС идентифицирует сущность рода 'Видеокарта'.
2. Анализ структуры сложного ПИС не требуется.
3. Тип ПИС: составная. Следовательно кореферентность определяется минимальным набором признаков.
4. Не совпадение признаков (семантическое противоречие) по семантическому полю 'Стандарт системного разьема'.
5. Анализ закончен: документ и анализируемое ПИС не эквивалентно.

Применение программной системы ASTEND для семантического сравнения предикативных имен сущности требует наполнения данными для анализируемой предметной области. Начальные затраты ручного труда включают в себя: анализ предметной области и выделение родов сущностей, заполнение таблиц трансляции (словарей) сокращений термов, наполнение моделей родов сущностей, таблиц пресупозиции и компонентов дистинкции. Эффективность применения программы тем выше, чем больше выше потребность в сравнении ПИС. Экспериментальные данные показывают следующую зависимость.

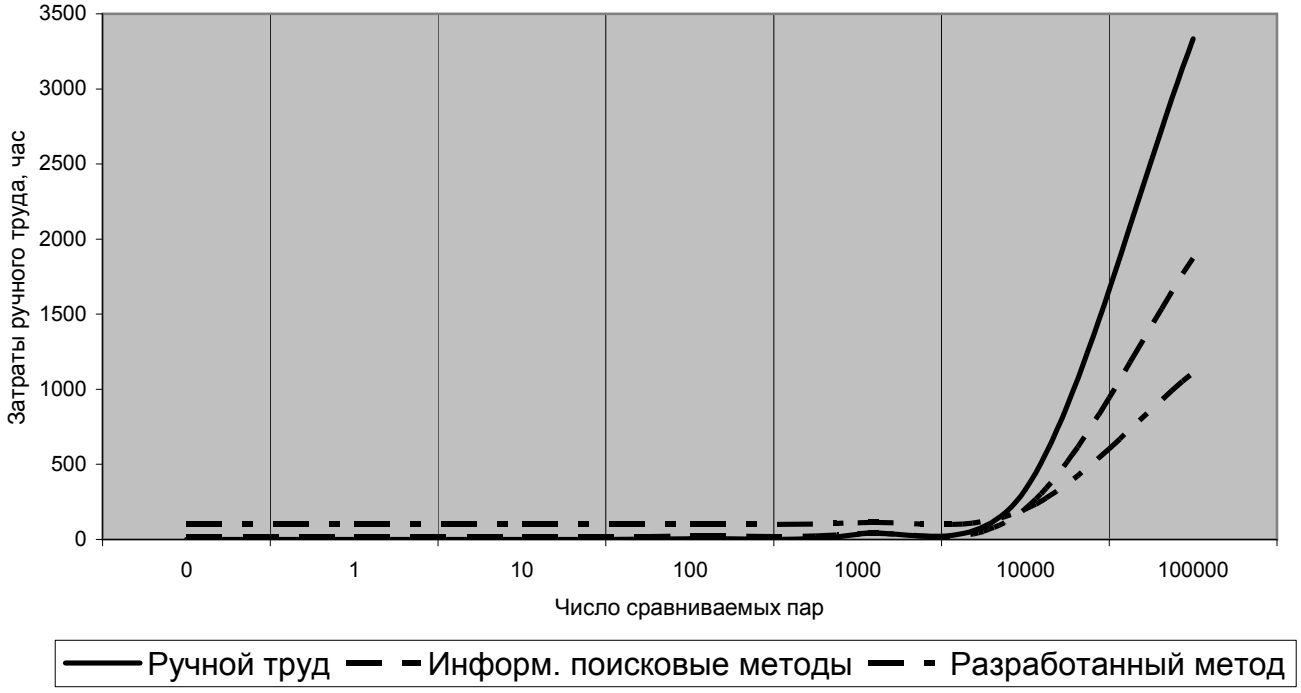

Рисунок 5.2 Зависимость затрат времени для решения задач различного объема.

В **заключении** сформулированы основные выводы и результаты диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Предлагаемые средства и методы семантического сравнения предикативных имен сущности обладают достаточной универсальностью и могут быть использованы для повышения эффективности большого класса систем от информационно-поисковых до интеграции БД систем управления ресурсами предприятий.

1. На основании исследований, проведенных в диссертации, установлено, что предикативное имя сущности является сложной иерархической системой. Референт является системным выходом ПИС. Каждая подсистема именует определенную сущность или их совокупность. Идентификация сущности в подсистемах осуществляется за счет указания не менее чем минимального набора параметров. В результате исследований, разработана обобщенная информационная модель предикативного имени сущности.

2. Разработана информационная модель для рода сущности как эффективное средство семантического сравнения предикативных имен сущности. Такой подход сокращает объем труда оператора ЭВМ по моделированию ПИС за счет описания обобщенной структуры рода сущностей.

3. На основе модели рода сущности предложены методы выделения параметров из символьной строки и их интерпретация, методы определения семантической эквивалентности, метод ранжирования альтернативных вариантов толерантных ПИС.

4. Разработанный программный комплекс позволяет решить задачу семантического сравнения предикативных имен сущности в коллекциях

документов. Комплекс может использоваться как инструмент интеграции БД систем управления ресурсами предприятий.

5. Предложенные методы, алгоритмы и разработанный программный комплекс использован в службе снабжения коммерческого предприятия ООО «Прайм».

Публикации по теме диссертации

1. Бердник В.Л. Интеллектуальные методов в системах проектирования топологии сети //Концептуальное проектирование в образовании, технике и технологии: Сб. науч. тр. /ВолгГТУ.- Волгоград, 2000.-С.35-36.
2. Данилов Д.А., Бердник В.Л. Безопасные экспертные системы на нейронных сетях //Концептуальное проектирование в образовании, технике и технологии: Межвуз. сбор. науч. трудов /ВолгГТУ.- Волгоград, 2001.- С.34-36.
3. Бердник В.Л., Борисенко С.Г., Лукьянов В.С. Автоматизированное рабочее место проектировщика топологии древовидной сети большой размерности //Концептуальное проектирование в образовании, технике и технологии: Сб. науч. тр. /ВолгГТУ.- Волгоград, 1997.-С.8-11.
4. Бердник В.Л. Сопоставление высказываний идентификации сущности: фактический стандарт корпоративных решений или технология будущего. // Конференция 10-ой научно-практическая конференция “Реинжиниринг бизнес-процессов на основе современных информационных технологий. Системы управления знаниями” (РБП-СУЗ-2007). 17-18.04.07/ МЭСИ – Москва, 2007.
5. Бердник В.Л., Заболеева-Зотова А.В. Система поиска и сопоставления предикативных имен сущностей идентификации сущности //Системные проблемы надёжности, качества, мат. моделирования, информ. и электронных технологий в инновационных проектах: (Инноватика-2007): матер. междунар. конф. и Рос. науч. школы /Рос. акад. надёжности [и др.].-М., 2007.- Ч.2, т.III.- С.316-320.
6. Бердник В.Л. Задача идентификации сущности и методы её решения //Открытое образование: прилож. к журн.: по матер. XXXIV междунар. конф. и дискусс. науч. клуба, Ялта-Гурзуф, 20-30.05.07: Инф. технол. в науке, образ., телеком. и бизнесе (IT+SE`07).- 2007.-[Б/н].-С.247-249.
7. Бердник В.Л., Заболеева-Зотова А.В. Поддержка решения задачи идентификации сущности методами информационного поиска //Инновационные технологии в управлении, образовании, промышленности "АСТИНТЕХ-2007": матер. всерос. науч. конф., 18-20 апреля 2007 г. /Астрахан. гос. ун-т и др.- Астрахань, 2007.- Ч.2.- С.100-103.
8. Бердник В.Л. Задача бинарного синтеза и метод ее решения на начальных этапах проектирования //Информационные технологии в образовании, технике и медицине: Сб.науч.тр. междунар. н.-техн. конф., Волгоград, 19-21.09.00 /ВолгГТУ и др..- Волгоград, 2000.- Ч.2.-С.15-17.

Публикации в ведущих рецензируемых научных журналах и изданиях РФ рекомендуемых ВАК по специальности 05.13.01

9. Бердник В.Л., Заболеева-Зотова А.В. Поддержка решения задачи идентификации сущности методами информационного поиска //Программные

продукты и системы: приложение к междунар. журналу "Проблемы теории и практики управления".- 2007.-№2.- С.32-35.
10. Бердник В.Л. Декомпозиция задачи идентификации сущности для учёта нелингвистических составляющих //Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ.- 2007.-Вып.3, №9.- С.39-43.
11. Бердник В.Л., Заболеева-Зотова А.В. Задача идентификации сущности заданной слабоструктурированным текстом //Изв. ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ.-2007.-Вып.2, №2.- С.26-28.
12. Бердник В.Л., Заболеева-Зотова А.В. Семантический анализ предикативных имен сущностей идентификации сущности //Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ.- 2007.-Вып.3, №9.- С.43-46.

Подписано в печать 25.09.2008 г. Формат 60х84/16
Усл.печ.л 2.0 Тираж 150 экз.
Заказ 1036 от 25.09.08 г.

Типография “Стигма”
400078, г. Волгоград, пр. Ленина, 67
Отпечатано с оригинал-макетов заказчика.