Федеральное государственное автономное образовательное учреждение
высшего образования
Санкт-Петербургский национальный исследовательский университет
информационных технологий механики и оптики

На правах рукописи

Диковицкий Владимир Витальевич

МЕТОДЫ ИНТЕРФЕЙСНОЙ НАВИГАЦИИ И ПОИСКА НОРМАТИВНО-СПРАВОЧНЫХ ДОКУМЕНТОВ В КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Специальность 05.13.11 – Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Диссертация
на соискание ученой степени кандидата технических наук

Научный руководитель
д.т.н.
Шишаев Максим Геннадьевич

Санкт-Петербург – 2016

Оглавление

ВВЕДЕНИЕ 5
ГЛАВА 1. Теоретические и практические основы повышения эффективности доступа к информации в корпоративных информационных системах, основанных на знаниях 11
1.1 Специфика проблемы поиска информации в корпоративных информационных системах, основанных на знаниях 11
1.2. Системы управления знаниями 13
1.3. Модели представления знаний 14
1.4. Получение знаний 19
1.5. Источники знаний организаций 21
1.6 Основные методы и технологии обеспечения унифицированного доступа к знаниям организаций 26
1.7. Классификация применяемых информационных систем 33
1.8. Примеры систем повышения эффективности поиска, основанных на формализованных знаниях. 40
Выводы по главе 1 48
ГЛАВА 2. Подход и модели построения корпоративных информационных систем, основанных на формализованных знаниях 49
2.1. Подход к построению мультипредметных информационных систем, основанных на знаниях 49
2.2. Сценарная модель мультипредметной информационной системы 55
2.3. Формальное определение мультипредметной информационной системы 61
2.4. Модель предпочтений пользователей мультипредметной информационной системы 63
2.5. Архитектура мультипредметной информационной системы 67
Выводы по главе 2 70

ГЛАВА 3. Методы формирования и функционирования мультипредметных информационных систем 71
3.1. Метод автоматизированного формирования семантической модели предметной области информационной системы на основе принципа «пользователь как эксперт» 71
3.2.1. Формальное описание семантической модели предметной области 72
3.2.2. Интеграция семантических образов документов организации 73
3.3. Метод формирования модели предпочтений пользователя мультипредметной информационной системы 78
3.4. Метод интерфейсной навигации в мультипредметных информационных системах 81
3.4.1. Модель навигационного интерфейса 82
3.4.2. Ограничения на структуру пользовательского интерфейса 86
3.4.3. Метод формирования навигационной структуры, адекватной модели предпочтений пользователей 93
3.4.4. Визуализация формализованных знаний на основе методов визуального анализа информации 97
3.5. Метод поиска на основе семантической сети с субтрактивными связями 102
Выводы по главе 3 109
ГЛАВА 4. Применение методов мультипредметных информационных систем в рамках документооборота организаций 110
4.1. Реализация сервиса поиска информации 111
4.2. Реализация сервиса семантического индексирования 117
4.3. Реализация сервиса интерфейсной навигации на основе СМПО и формировании модели предпочтений пользователей 120
4.4. Апробация мультипредметной информационной системы в рамках документооборота организаций. 124

Выводы по главе 4 129
ЗАКЛЮЧЕНИЕ 131
Определения 133
Список обозначений и сокращений 134
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 135
ПРИЛОЖЕНИЕ 1. Акты внедрения 152

ВВЕДЕНИЕ

Актуальность работы связана с возрастанием требований к эффективности информационного обеспечения организаций, деятельность которых требует аккумуляции и обновления знаний в различных предметных областях. Нормативно-справочные документы: классификаторы материалов и оборудования, регламенты – это стратегический актив компании, который предприятия используют в процессе постоянного информационного обмена. Многочисленность нормативных документов и их версий влечет дублирование и рост объема хранимой информации. Одни и те же процессы и объекты с разных сторон описываются в различных документах и могут рассматриваться специалистами с различных точек зрения, что усугубляет проблему поиска релевантных документов. Это обусловливает необходимость формирования единого информационного пространства предприятия в виде мультипредметной информационной системы организации. Создание и поддержка функционирования такой информационной системы (ИС) порождает ряд важных задач, требующих научно обоснованного подхода к их решению: формирование семантической модели предметной области мультипредметной информационной системы предприятия (МИСП) на основе интеграции существующих формализованных знаний, обеспечение актуальности контента единого информационного пространства в условиях динамики предметной области, обеспечение корректного и быстрого восприятия предоставляемой специалистам различных предметных областей информации, релевантности и пертинентности результатов информационного поиска.

Задача построения информационных систем, основанных на формализованных знаниях, известна довольно давно. По данной тематике опубликовано и выпущено множество статей и монографий, разработан целый ряд моделей и методов и спроектированных на их основе информационных

систем, которые находят широкое применение в различных областях. Вместе с тем, несмотря на высокий уровень исследований в этой области, создание информационных систем, делающих возможным эффективный (в плане уменьшения времени) доступ к нормативно-справочной информации (НСИ) предприятия различным специалистам, остаётся сложной, до конца не решённой проблемой. Предлагаемым в данной работе решением является разработка новых методов автоматического формирования семантической модели предметной области НСИ, методов интерфейсной навигации и поиска документов с целью повышения эффективности доступа специалистов к НСИ. Повысить эффективность доступа к требуемой информации представляется возможным за счет адаптации пользовательского интерфейса к различным категориям пользователей. Использование автоматически формируемой семантической модели предметной области (СМПО) позволяет разделить семантику и контент информационных баз, что обеспечивает семантическую интеграцию НСИ предприятия.

Цель работы состоит в совершенствовании процессов поиска нормативно-справочной информации путем повышения полноты и точности поиска информации за счет автоматизированного формирования интегрированной семантической модели предметной области и разработке в рамках этой модели методов интерфейсной навигации и адаптивного поиска документов.

Для достижения поставленной цели в работе решаются следующие **задачи**:

1. Анализ особенностей построения и использования информационных систем, основанных на знаниях;
2. Разработка архитектуры мультипредметной информационной системы предприятия, основанной на семантической модели предметной области и моделях предпочтений пользователей.
3. Разработка методов формирования семантической модели предметной области информационной системы, интерфейсной навигации и поиска

документов, реализующих уточнение семантической модели и адаптированное к различным пользователям представление информации.

4. Программная реализация и проверка эффективности разработанных методов.

Методы исследований. Для решения поставленных задач использованы методы системного анализа, теории графов, методы информационного поиска, математического моделирования, модульного и объектно-ориентированного программирования, искусственного интеллекта и инженерии знаний.

Положения, выносимые на защиту:

1. Метод автоматизированного формирования семантической модели предметной области, реализующий уточнение автоматически сформированных знаний.
2. Метод поиска документов, обеспечивающий автоматизированное расширение запроса и адаптивное ранжирование документов.
3. Метод интерфейсной навигации, обеспечивающий формирование навигационной структуры интерфейса, соответствующей модели предпочтений пользователя.
4. Комплекс программных средств, реализующий предложенные модели и методы для повышения эффективности доступа к ресурсам информационных систем.

Научная новизна работы состоит в создании моделей, алгоритмов и методов формирования и функционирования мультипредметных информационных систем. Применение данных методов позволяет повысить эффективность механизмов информационного поиска нормативно-справочной информации и человеко-машинного взаимодействия. Научной новизной обладают следующие результаты:

1. Разработан метод автоматизированного динамического формирования семантической модели предметной области мультипредметных информационных систем, использующий опыт пользователей для уточнения автоматически сформированных знаний, отличающийся интеграцией существующих формализованных знаний, результатов семантического анализа новых документов и моделей предпочтений пользователей.
2. Разработан метод поиска документов, обеспечивающий автоматизированное расширение запроса и оценку релевантности результатов поиска на основе совместного анализа модели предпочтений пользователя и семантической модели предметной области с учетом субтрактивных отношений.
3. Предложен метод интерфейсной навигации для формирования пользовательских интерфейсов мультипредметной информационной системы, адаптированных для различных категорий пользователей. Повышение эффективности человеко-машинного взаимодействия обеспечивается за счет отображения модели предпочтений пользователей на автоматически формируемую навигационную структуру интерфейса.
4. Создан комплекс программных средств для повышения эффективности доступа к документам организаций, отличающийся использованием методов, способных к автоматическому уточнению и адаптированному представлению информации организаций.

Аспекты научной новизны соответствуют областям исследований «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем», «Системы управления базами данных и знаний» и «Человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности,

мультимедийного общения» специальности «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Практическая значимость работы состоит в том, что для решения задач исследования создан комплекс программных средств организации эффективного доступа к ресурсам информационных систем, позволяющий наглядно и обозримо провести систематизацию НСИ предприятия. Применение разработанного комплекса программных средств позволяет усовершенствовать процессы обработки данных и знаний в компьютерных системах и сетях. Разработанный комплекс может успешно применяться в локальных и глобальных информационных системах.

Апробация работы. Научные положения и практические рекомендации диссертационной работы в целом, а также отдельные ее разделы докладывались и обсуждались на международных научно-технических конференциях «Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2012) » (г. Минск, 2012 г.), «Современные проблемы прикладной информатики» (г. Санкт-Петербург, 2011 г.), «Интеллектуальные системы и технологии: современное состояние и перспективы» (г. Тверь, 2011 г.), на всероссийских научных конференциях «Прикладные проблемы управления макросистемами» (г. Апатиты, 2010, 2014 гг.), «Теория и практика системной динамики» (г. Апатиты, 2011 г.), а также на семинарах лаборатории региональных информационных систем ИИММ КНЦ РАН.

Внедрение результатов. Результаты работы были использованы в информационной системе крупного промышленного предприятия региона (АО «Апатит»), в учебном процессе в ФГБУ ВПО КФ ПетрГУ.

Публикация результатов работы. По теме диссертационной работы опубликовано 24 статьи, в том числе 4 статьи в изданиях, рекомендованных ВАК, оформлено 3 свидетельства о регистрации программ для ЭВМ.

Структура и объем работы. Диссертация объемом 153 машинописные страницы, содержит введение, четыре главы и заключение, список литературы (135 наименований), 4 таблицы, 38 рисунков, одно приложение с копиями актов внедрения.

В первой главе предлагается общая характеристика решаемой в работе проблемы. Рассматриваются особенности предметной области и существенные проблемы повышения эффективности доступа к информации. Рассмотрены системы управления знаниями организаций, модели представления знаний, источники знаний, а также способы представления формализованных знаний в аспекте повышения эффективности доступа к информации. Выполнена оценка существующих решений, представлены основы формирования, интеграции, представления и использования формальных знаний для повышения эффективности доступа к информации и человеко-машинного взаимодействия. Рассмотрены подходы к интеграции гетерогенных источников информации.

Во второй главе представлен поход к построению мультипредметных информационных систем, основанных на знаниях. Представлены концептуальная и сценарная модель, формальное определение мультипредметной информационной системы. Представлена модель предпочтений пользователей мультипредметной информационной системы и ее архитектура.

В третьей главе представлены методы, обеспечивающие формирование и эффективное функционирование компонентов мультипредметных информационных систем: Метод автоматизированного формирования семантической модели предметной области, метод интерфейсной навигации, метод поиска информации с учетом субтрактивных отношений.

В четвертой главе представлено прикладное программное обеспечение, реализующее предложенные методы. Приводятся результаты вычислительных экспериментов. В заключении изложены основные результаты исследования.

ГЛАВА 1. Теоретические и практические основы повышения эффективности доступа к информации в корпоративных информационных системах, основанных на знаниях

В данной главе представлена общая характеристика решаемой в работе проблемы. Рассматриваются особенности предметной области и существенные проблемы повышения эффективности доступа к информации.

Рассмотрены существующие решения по управлению знаниями организаций, модели представления знаний, источники знаний, а также способы представления формализованных знаний в аспекте повышения эффективности доступа к информации.

Представлены общие методы интеграции корпоративной информации и их недостатки. Описаны основные подходы к семантической интеграция на основе онтологий. Рассмотрена проблематика использования онтологий в информационных системах организаций. Рассмотрены технологии и стандарты Semantic Web и их роль в семантической интеграции информации.

Также рассмотрены существующие, основанные на знаниях, информационные системы организаций в аспекте одной из основных задач – обеспечение эффективного доступа пользователей к нормативно-справочной информации (НСИ).

В заключении сформулированы выводы и основные требования к подходам повышения эффективности доступа пользователей корпоративных информационных систем к требуемой информации.

1.1 Специфика проблемы поиска информации в корпоративных информационных системах, основанных на знаниях

На сегодняшний день уже существуют средства, методы и технологии, которые в целом решают проблему доступа к корпоративной информации. Однако их применение возможно лишь в масштабах одной организации или предприятия, имеющей один административный центр, который вырабатывает

детальную программу процесса интеграции, осуществляет выбор средств и технологий, производит контроль полученного результата. Однако такой сценарий проведения типовой корпоративной интеграции сложно реализовать в виду наличия некоторых особенностей рассматриваемого класса информационных систем.

Среди таких особенностей можно выделить следующие:

1. Разнородность информационных ресурсов организаций. Количество объектов и процессов, содержащихся в нормативно-справочной информации, описывающей бизнес-процессы предприятия, затрудняют доступ пользователя к данной информации, и требуют от последнего временных и трудовых затрат на поиск и изучение множества соответствующих документов. Кроме того, затруднения вызывает семантическая разнородность, заключающаяся в том, что одни и те же процессы и объекты с разных сторон описываются в различных документах.

2. Мультипредметность информационных ресурсов организаций. Знания различных структурных подразделений организаций не являются автономным информационным активом одного подразделения. Это выражается в существовании общего информационного «пула» нормативно-справочной информации предприятия, различные фрагменты (документы) которого требуются различным структурным подразделениям.

3. Динамика информационных ресурсов организаций. Нормативно-справочные документы: классификаторы материалов, оборудования, регламенты – это стратегический актив компании, который предприятия используют в процессе постоянного информационного обмена. Различие в версиях данных документов приводит к их многочисленному дублированию в различных отделах организации. В существующих информационных системах необходимость синхронизации фрагментированных баз данных НСИ снижает эффективность внедрения и использования информационных технологий, возлагая возросшую нагрузку на пользователя.

Кроме перечисленных особенностей, следует так же отметить экономическую составляющую обеспечения функционирования информационных систем. Существующие решения подразумевают наличие дорогостоящих экспертов по знаниям, систематизирующим НСИ промышленного предприятия. Однако, динамика информационных активов и диверсификация производства современных промышленных предприятий делают невозможным привлечение экспертов для формирования баз знаний, предъявляя высокие требования к автоматизации данного процесса.

Учет данных особенностей определяет применимость какого-либо из существующих общих подходов к обработке информации в качестве основы, а также направление его последующей модификации. Существующие ИС реализуют большую часть функций, требуемых для обеспечения пользователей информацией. Однако анализ особенностей предметной области показал, что эффективность информационного обеспечения в корпоративных информационных системах, основанных на знаниях, может быть повышена за счет учета специфики предметной области при построении и обеспечении функционирования ИС.

1.2. Системы управления знаниями

Знание в широком смысле – форма существования и систематизации результатов познавательной деятельности человека. Знания в рамках организации - это закономерности предметной области (принципы, связи, законы), полученные в результате практической деятельности и профессионального опыта, позволяющие специалистам ставить и решать задачи в этой области [12]. Управление знаниями (УЗ) (Knowledge Management) – это процесс сбора, разработки, распространения и эффективного использования знаний организации [89]. Как отмечается в [12], информации в компаниях накоплено даже больше, чем она способна оперативно обработать, при этом часто одна часть предприятия дублирует работу другой просто

потому, что невозможно найти и использовать знания, находящиеся в соседних подразделениях. Различные организации пытаются решать этот вопрос по-своему, но при этом каждая компания стремится увеличить эффективность обработки знаний. Управление знаниями можно рассматривать и как новое направление в информатике для поддержки процессов создания, распространения, обработки и использования знаний внутри предприятия.

Концепция «управление знаниями» (УЗ) или Knowledge Management (КМ) призвана поменять взгляд на автоматизацию обработки информации. Концепция УЗ заключается в задаче – накопления не разрозненной информации, а знаний, т.е. закономерности и принципы, позволяющие решать реальные задачи. При этом в расчет берутся и те знания, которые «невидимы» – они хранятся в памяти специалистов, а не на материальных носителях. [12]

Далее рассмотрим основные модели представления знаний.

1.3. Модели представления знаний

Существуют множество моделей представления знаний для различных предметных областей. В работе [12] представлена следующая классификация:

1) Продукционные модели;
2) семантические сети;
3) фреймы;
4) формальные логические модели.

Продукционная модель – основанная на правилах модель, которая позволяет представить знания в виде условий «Если (условие), то (действие)». Под условием или антецедентом понимается некоторый образец, по которому осуществляется поиск в базе знаний. Под действием или консеквентом понимаются действия, выполняемые при успешном исходе поиска.

Вывод на основанной на продукционной модели базе знаний может быть прямой (от данных к поиску цели) или обратный (от цели для ее подтверждения). Данные представляют исходные факты, хранящиеся в базе

фактов, на основании которых запускается машина вывода или интерпретатор, перебирающий правила из продукционной базы знаний. Продукционная модель находит применение в промышленных экспертных системах. Ее отличают наглядность, модульность, простотой механизма логического вывода.

Семантические сети имеют вид ориентированного графа, вершины которого содержат понятия, а дуги отношения над понятиями. Понятиями выступают абстрактные или конкретные объекты предметной области, а отношения обозначают связи типа: «ВИД», «ЧАСТЬ - ЦЕЛОЕ», «Принадлежит». Особенностью семантических сетей является обязательное наличие трех типов отношений [14]:

1) Класс — элемент класса;
2) свойство — значение;
3) пример элемента класса.

В зависимости от типов отношений между понятиями, существует несколько классификаций семантических сетей:

По количеству типов отношений семантические сети могут быть однородные (с единственным типом отношений), и неоднородные (с различными типами отношений).

В зависимости от типов отношений сети могут быть бинарные (в которых отношения связывают два объекта), N-арные (в которых есть отношения, связывающие более двух понятий). Наиболее часто в семантических сетях используются следующие отношения: «часть — целое» ,«класс — подкласс», «элемент —множество», функциональные связи (определяемые глаголами действия), количественные (больше, меньше, равно...), пространственные, временные (раньше, позже), атрибутивные связи (иметь свойство, иметь значение), логические связи, лингвистические и др.

Как отмечается в [14], поиск решения в базе знаний, имеющей в основе модель семантической сети, сводится к задаче поиска фрагмента сети, соответствующего некоторой подсети, отражающей запрос.

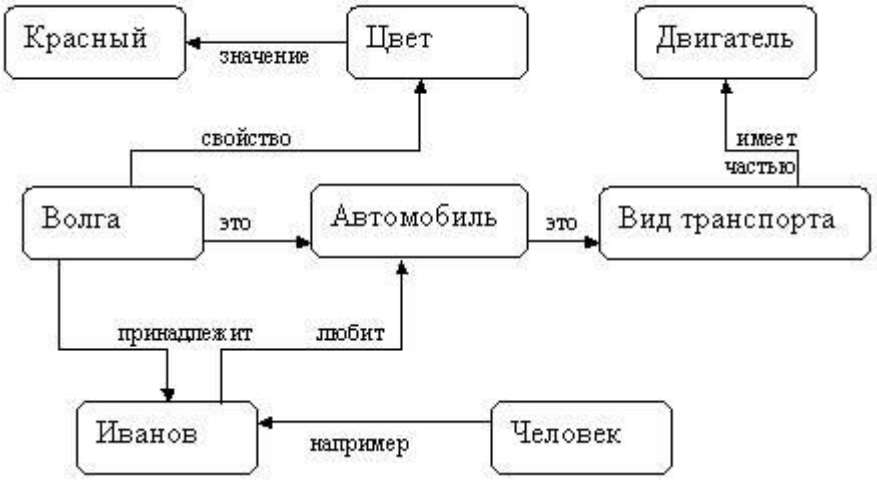


Рисунок 1 - Семантическая сеть

На рисунке 1 изображен пример семантической сети. Вершинами являются понятия предметной области.

Основным преимуществом семантической сети является то, что она наиболее соответствует представлениям об организации долговременной памяти человека. К недостатку модели можно отнести относительную сложность реализации процедур поиска и логического вывода. Существуют специальные сетевые языки, например NET, для реализации данной модели представления знаний. Известные экспертные системы, PROSPECTOR, CASNET TORUS[12], используют семантические сети в качестве языка представления знаний.

Фрейм - структуры знаний для восприятия пространственных сцен [11]. В психологии и философии существует понятие абстрактного образа[12]. Произнесение вслух слова “комната” вызывает у слушающих образ комнаты: жилое помещение с четырьмя стенами, полом, потолком, окнами и дверью. Если убрать из этого описания, например, окна, мы получим не комнату, а чулан, но в нем будут слоты - незаполненные значения некоторых атрибутов. В теории фреймов подобный образ называется фреймом. Формализованная модель для отображения образа также называется фреймом. Фреймы различают на образцы, хранящиеся в базе знаний, и экземпляры, отображающие реальные ситуации на основе поступающих данных. Модель фрейма является достаточно

универсальной, поскольку позволяет отобразить все многообразие знаний о мире посредством [11] :

- Фреймы структуры позволяют обозначать объекты и понятия;
- фреймы роли;
- фреймы сценарии;
- фреймы ситуации и др.

Структура фрейма может быть представлена как список свойств этого фрейма, например, (ИМЯ: (имя 1-го слота: значение 1-го слота),(имя 2-го слота: значение 2-го слота),(имя N-го слота: значение N-го слота)), а так же фот же фрейм можно представить в виде таблицы, дополнив ее столбцами:

Таблица 1 – Структура фрейма

Имя фрейма			
Имя слота	Значение слота	Способ получения значения	Присоединенная процедура

Дополнительные столбцы в таблице 1 предназначены для описания способа получения значения и возможного присоединения к тому или иному слоту специальных процедур, что допускается в теории фреймов. В качестве значения слота может выступать имя другого фрейма с образованием сети фреймов. Получения слотом значений во фрейме-экземпляре может осуществляться несколькими путями:

- По умолчанию;
- средствами наследование свойств;
- по формуле, указанной в слоте;
- через присоединенную процедуру;
- явно, например, пользователем;
- из базы данных.

Преимуществом фреймов перед другими моделями представления знаний является отражение концептуальной основы организации памяти человека, гибкость и наглядность.[11]

Формальные логические модели основаны на исчислении предикатов 1-го порядка. При этом предметная область или задача должна быть описана набором аксиом. Данная логическая модель редко применяется в силу высоких требований и ограничений к предметной области.

Как отмечается в [16], недостатки сетевым моделям, относятся не столько к самим моделям, сколько к сложностям их реализации на традиционных (фон-Неймановских) компьютерах. Преимущество моделей такого класса очевидно, т. к. их выразительные возможности намного превышают возможности других моделей.

В работах [41] рассмотрено ассоциативно-онтологического представление данных. Отмечается, что методы обработки текстов на естественном языке на основе ассоциативно – онтологического подхода подходят для выделения признаков текста для построения поисковых индексов, автоматического реферирования научных и технических документов, отнесения текста к предметной области, поиска в коллекции документов. Так же отмечается роль ассоциативного окружения ключевых слов при недостатке начальных данных, например в задаче поиска документов, недостающие ключевые слова добавляются постепенно при помощи их итеративного выбора из ассоциативного окружения.

Таким образом, целесообразно использовать в качестве основы для хранения знаний модель семантической сети, как наиболее гибкой и соответствующей современным представлениям об организации долговременной памяти человека. Далее рассмотрим вопросы, связанные с приобретением и формализацией знаний.

1.4. Получение знаний

Для повторного использования экспертных знаний предметной области знания необходимо получить и сохранить. При этом основным вопросом является процесс получения знаний – то есть перенос компетентности от экспертов к инженерам по знаниям. Процесс переноса компетенции в литературе получил несколько названий[12]:приобретение (анг. acquisition), извлечение (анг. elicitation), формирование, получение, добыча, выявление знаний.

Под извлечением знаний (англ. knowledge elicitation) понимается процесс взаимодействия эксперта и источника знаний, результатом которой являются формализация процесса рассуждений специалистов при принятии решения и структура их представлений о предметной области. Данная процедура является самым узким местом при построении экспертных систем. При этом разработчикам приходится самостоятельно разрабатывать методы извлечения знаний, сталкиваясь со следующими очевидными трудностями [12]:

1) Организационные неувязки;
2) неудачный метод извлечения, не совпадающий со структурой знаний в данной области;
3) неадекватная модель представления знаний.

К данному списку можно добавить трудность в наладке контакта с экспертом, терминологическое разнообразие, отсутствие целостной системы знаний в результате извлечения только части знаний и др.

Процесс извлечения знаний длителен, при этом без готовых сформированных знаний запуск информационной системы невозможен. Так же данный процесс необходимо повторять каждый раз при изменениях в предметной области.

Еще одной трудностью является необходимость воссоздания инженером по знаниям, обладающим компетенцией в психологии, системному анализу, математической логике и пр., модели предметной области, используемой

экспертами. Как отмечается в [12], на практике часто разработчики с целью упростить процедуру получения знаний подменяют инженера по знаниям экспертом, что приводит к неполноте полученных знаний, в силу, того, что большая часть знаний эксперта — это результат ступеней опыта, часто, зная, что из A следует B, эксперт не отдает себе отчета, что цепочка рассуждений была гораздо длиннее, например A —> D или A -> Q -> R -> B. В ходе объяснения инженеру по знаниям размытые ассоциативные образы эксперта приобретают четкие словесные ярлыки. Эксперту труднее создать модель предметной области вследствие объема информации, объекты реального мира связаны более чем 200 типами отношений, которые образуют сложную систему, для выделения структуры которой нужна соответствующая методология.

К альтернативным существующим подходам к извлечению знаний относится, например, обучение на примерах. Источником знаний является множество примеров предметной области, а процесс обучения заключается в задании алгоритма распознавания путем предъявлении примеров заранее заданной классификации. Отличие обучения на примерах от машинного обучения заключается в том, что результат обучения должен быть интерпретирован в модели предметной области, и преобразован в способ представления, допускающий использование результатов обучения для моделирования рассуждений.

Таким образом, получение знаний является сложным процессом в информационных системах, основанных на знаниях. Привлечение экспертов и инженеров по знаниям требует временных и трудовых затрат. Актуальным является разработка методов и средств автоматизации процессов получения и уточнения знаний, а так же автоматизации переноса компетенций пользователей информационной системы.

1.5. Источники знаний организаций

В рамках корпоративных информационных систем источниками знаний являются знания квалифицированных сотрудников, являющихся экспертами в отдельных предметных областях, а так же информационные ресурсы, содержащие нормативно-справочную информацию в виде документов ИС, представляющих стратегический актив компании, который организации используют в процессе постоянного информационного обмена.

В современных организациях большую часть НСИ составляют текстовые данные, однако следует отметить, все большее распространение приобретают мультиформатные ресурсы, использующие для повышения эффективности функционирования различные формы структуризации контента. Результатом структуризации становится деление информации на собственно данные, мета-данные, описывающие их структуру. Такие особенности контента, в явном или неявном виде, определяют подходы к получению знаний в рамках соответствующего набора ресурсов.

Еще одним важным обстоятельством, оказывающим существенное влияние на эффективность механизмов получения знаний, является распределенный и, как следствие, гетерогенный характер НСИ организаций. Ориентированные на использование в условиях однородных информационных систем и ресурсов механизмы извлечения знаний (например, на основе заранее заданной формальной структуры знаний) резко теряют свою эффективность в применении к НСИ организаций, где форматы представления данных и, соответственно, мета-данные отличаются от документа к документу или от подразделения к подразделению. Это обстоятельство заставляет искать пути создания универсальных методов и технологий получения знаний, адекватных требованиям современных корпоративных информационных систем.

Текстовая информация в различных форматах составляет значительную долю информационных ресурсов информационных систем. Поэтому создание и развитие технологий обработки текста привлекали большое внимание на всех

этапах развития информационных систем. Далее представлены основы методов обработки естественного языка с целью получения формальных знаний. Главное место в таких методах занимает обработка естественного языка (ЕЯ). Обработка ЕЯ (Natural Language Processing, NLP) - решение задач, связанных с пониманием, анализом, выполнением различных операций над текстами, а так же их генерацией [38]. Примеры подобных задач: классификация, кластеризация хранимых коллекций документов, анализ текстов, перевод документов с одного языка на другой и т.д.

Методы автоматической формализации основываются на обработке и анализе текста. Перечислим основные трудности, возникающие при обработке текстов на ЕЯ:

- Проблема синонимии;
- проблема омонимии;
- устойчивые сочетания слов;
- морфологические вариации.

Проблема синонимии. Одно понятие может быть выражено различными словами. В результате релевантные документы, в которых используются синонимы понятий, указанных пользователем в запросе, могут быть не обнаружены системой.

Проблема омонимии и явлений «смежных с омонимией». Грамматические омонимы - разные по значению слова, но совпадающие по написанию в отдельных грамматических формах. Это могут быть слова одной или разных частей речи. Лексические омонимы - слова одной части речи, одинаковые по звучанию и написанию, но разные по лексическому значению.

Устойчивые сочетания слов. Словосочетания могут иметь смысл отличный от смысла, который имеют слова по отдельности.

Морфологические вариации. Во многих естественных языках слова имеют несколько морфологических форм, различающихся по написанию.

В современных технологиях обработки текстов используется не только аппарат лингвистики для анализа, но и статистические методы, математическая логика и теория вероятностей, кластерный анализ, методы искусственного интеллекта, а так же технологии управления данными. Рассмотрим два основных подхода к обработке и анализу текстов ЕЯ с целью их формализации– статистический и лингвистический (рисунок 2).

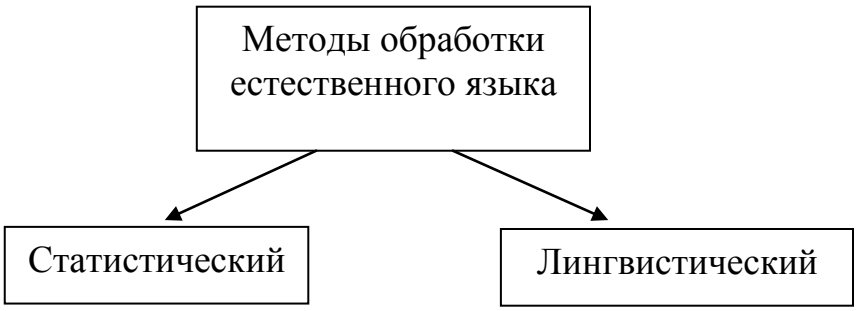


Рисунок 2 – Методы обработки естественного языка

В основе *статистического подхода* лежит предположение, что знания, отражающие содержание текста, можно выразить наиболее часто встречающимися словами. Суть статистического анализа заключается в подсчете количества вхождений слов в документ. Распространенным является сопоставление каждому терму *t* в документе некоторого неотрицательного весового коэффициента. Веса термов вычисляются множеством различных способов. Самый простой из них – положить «вес» равный количеству появлений терма *t* в документе *d*, обозначается *tft,d* (term frequency)[43]. Этот метод взвешивания не учитывает дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов по коллекции, лучше работает схема *tf-idf* вычисления весов, определяемая следующим образом:

$$tf - idf_{i,d} = tf_{i,d} \times idf_i , \quad (1)$$

где $idf_i = \log \frac{N}{df_i}$ - обратная документальная частота (inverse document frequency) терма t , dft - документальная частота (document frequency), определяемая как

количество документов в кол-лекции, содержащих терм *t, N* - общее количество документов в коллекции. Схема *tf-idf* и ее модификации широко используются на практике.

Эффективным подходом, основанным на статистическом анализе, является латентно-семантическое индексирование. Латентно-семантический анализ – это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных [48]. Латентно-семантический анализ основывается на идее, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые в значительной степени позволяют определить похожесть смысловых значений слов и множеств слов между собой.

Главный недостаток статистических методов состоит в невозможности учета связности текста, а представление текста как простого множества слов недостаточно для отражения его содержания как знаний. Текст представляет набор слов, выстроенных в определенной заданной последовательности. Преодолеть этот недостаток позволяет использование лингвистических методов анализа текста.

Существуют следующие уровни *лингвистического анализа*: графематический, морфологический, синтаксический, семантический. Результаты работы каждого уровня используются следующим уровнем анализа в качестве входных данных (рисунок 3).

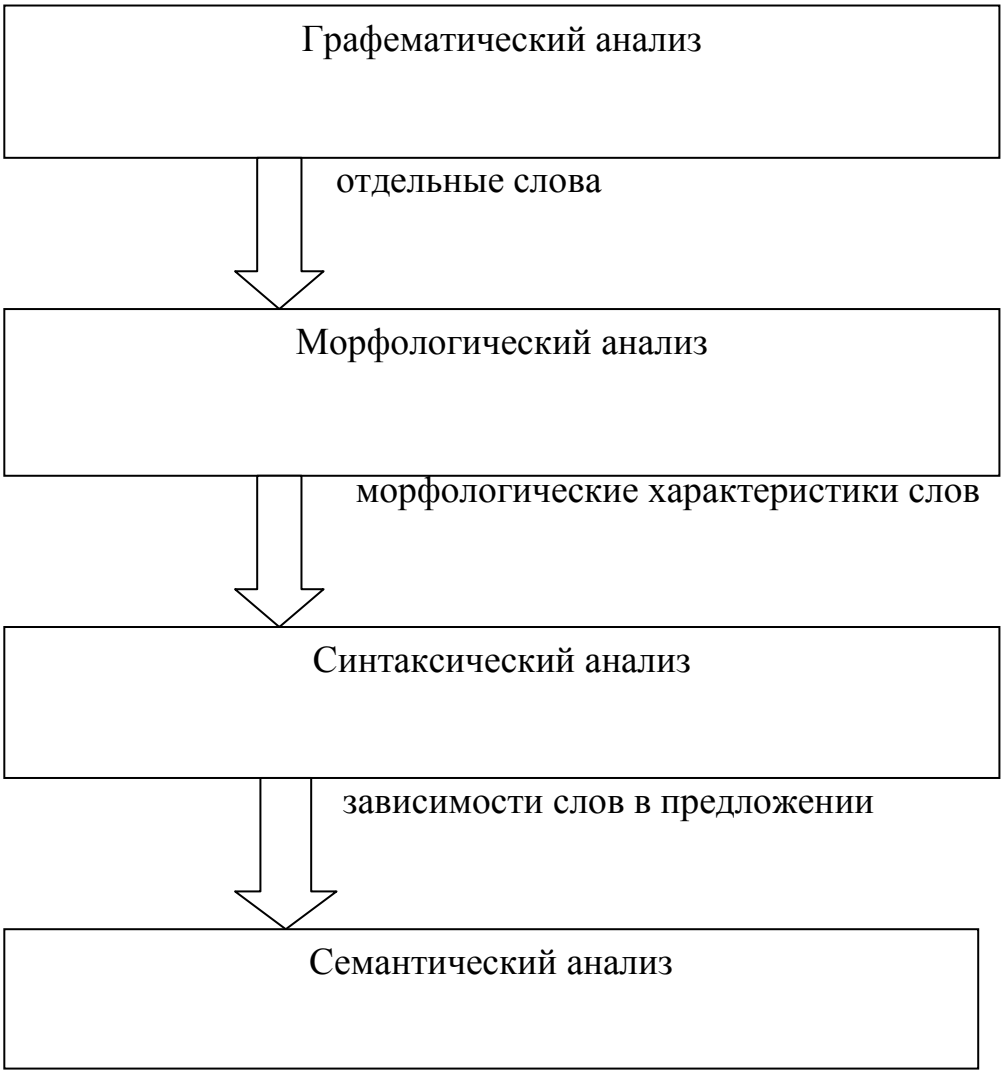


Рисунок 3 - Уровни лингвистического анализа

Целью графематического анализа является выделения элементов структуры текста: параграфов, абзацев, предложений, отдельных слов и т. д.

Целью морфологического анализа является определение морфологических характеристик слова и его основной словоформы. Особенности анализа сильно зависят от выбранного естественного языка.

Целью синтаксического анализа является определение синтаксической зависимости слов в предложении. В связи с присутствием в русском языке большого количества синтаксически омонимичных конструкций, наличием тесной связи между семантикой и синтаксисом, процедура

автоматизированного синтаксического анализа текста является трудоемкой. Сложность алгоритма увеличивается экспоненциально при увеличении количества слов в предложении и числа используемых правил.

Разработки в области семантического анализа текста связаны с областью искусственного интеллекта, делающей акцент на смысловом понимании текста. В настоящее время успехи в этом направлении достаточно ограничены. Разработанные семантические анализаторы обладают высокой вычислительной сложностью и неоднозначностью выдаваемых результатов [35]. В связи с чем, актуальным становится разработка технологий вовлечения квалифицированных пользователей информационной системы, обладающих знаниями предметной области, в процесс формализации текстовой информации.

Далее рассмотрим подходы к интеграции формализованных знаний и обеспечению унифицированного доступа к информации.

1.6 Основные методы и технологии обеспечения унифицированного доступа к знаниям организаций

1.6.1. Подходы к интеграции информации

Интеграция данных в информационных системах понимается как обеспечение единого унифицированного интерфейса для доступа к некоторой совокупности, неоднородных независимых источников данных [108]. Интеграция является механизмом повышения эффективности доступа к информации. Как отмечается в [47] интеграция характеризуется разнообразием постановок задач, подходов и методов, используемых для их решения. С точки зрения архитектуры можно выделить системы с материализованной и виртуальной интеграцией [37]. С точки зрения преодолеваемых видов гетерогенности информационных ресурсов, методы можно разделить на структурные или логические и семантические методы интеграции.

Структурные методы интеграции предполагают оперирование данными на основе структур данных, в которых они хранятся. Отсюда и следует

основной недостаток данных методов – высокая вероятность появления семантических конфликтов между фрагментами информации из разных источников. Это вызвано тем обстоятельством, что структурной семантики, выражаемой, например, в схеме реляционной базы данных, явно недостаточно для установления соответствия или различия между информационными фрагментами. [47] Структурные методы обычно применяются для интеграции данных в корпоративных информационных системах с жестко заданной логикой бизнес-процессов. Среди структурных методов интеграции выделяют консолидацию, федерализацию и распределение данных. Общим недостатком данных подходов к интеграции данных, как отмечается в [47] является оперирование данными на основе их структурной спецификации и без принятия во внимание семантики, что неизбежно влечет появление смысловых конфликтов и противоречий между информационными фрагментами. Образованное в результате использования данных методов интегрированное информационное пространство можно будет применять для решения ограниченного круга задач, сформулированных на этапе проектирования интеграционной системы. Это вполне допустимо, например, в рамках одного отдела или небольшой организации. Однако современные предприятия, особенно промышленной сферы, характеризуются динамикой предметной области и множеством субъектов, преследующих разные цели и имеющих различные точки зрения на одни и те же объекты, используя при этом одни данные. Ограниченная выразительность единого информационного пространства, полученного методами структурной интеграции, затрудняет его использование для информационно-аналитического обеспечения организаций.

Семантические методы интеграции лишены недостатков структурных методов. Семантическая интеграция [47] основывается на знании и учете природы данных. Хранение данных вместе с метаданными создает дополнительные сложности, но обеспечивает большее удобство работы. При их использовании интегрируемые источники рассматриваются не как

совокупности фрагментов информации определенной структуры, а как описания объектов, субъектов и процессов предметной области.

Широкое распространение в качестве средства описания концептуализации получили онтологии [47]. Их достоинствами являются большие выразительные возможности, наглядность, а также, что является особенно важным, возможность формального отражения семантики.

Онтология – спецификация концептуализации [96]. Данное определение позволяет называть онтологиями огромное множество моделей, используемых для описания понятийных систем предметных областей. Отличие онтологических подходов к интеграции данных от структурной интеграции заключается в том, что унификация проводиться в отношении заданной в онтологии формальной семантики, а не формата представления. Среди подходов к семантической интеграции выделяют [47] централизованный, децентрализованный и гибридный.

Централизованный подход заключается в использовании единой онтологии. Концепты онтологии и связи между ними являются понятийной системой предметной области, с элементами которой связываются фрагменты данных интегрируемых источников. Процесс разработки общей онтологии проводиться группой экспертов предметной области и инженеров по знаниям. Эксперты договариваются о значении терминов предметной области и отношениях между ними, а инженеры по знаниям задают в онтологии аксиомы, отражающие значение терминов. Основным преимуществом является скорость проведения процесса интеграции на начальном этапе при известном перечне ресурсов для интеграции. Однако с увеличением количества связей и ограничений онтологии затрудняется ее поддержка. Также проблематичной становиться повторное использование такой онтологии при интеграции данных в связанной предметной области, так как определения новых понятий могут приводить к большому числу различных противоречий. Их можно разрешить

путем упрощением существующих определений и тем самым потерей части формальной семантики, или достаточно сложной модификацией общей онтологии, выполненной после ее тщательного анализа.

Мульти-онтологический подход, в отличие от предыдущего, подразумевает описание каждого информационного ресурса отдельной онтологией. Вследствие этого нет необходимости в обобщающей онтологии, и каждая новая онтология может разрабатываться независимо от других, что облегчает подключение новых информационных ресурсов. Недостаток подхода заключается в необходимости определения соответствие между различными онтологиями. На практике реализация связи онтологий представляет собой очень сложную задачу, поскольку онтологии разнородны.

В гибридном подходе используется общий словарь, на основании которого строятся частные онтологические описания. Гибридный подход позволяет обеспечить выразительность при создании онтологий исходных информационных ресурсов и, как следствие, более точное отражение семантики понятий, а также, в отличие от децентрализованного, существенно облегчается задача установления различных отношений с терминами отдельных онтологий. Соответствие между концептами двух онтологий означает наличие этого отношения только между множествами их интерпретаций, а не их элементами. Иными словами это означает, что отсутствует возможность установить соответствие между экземплярами разных онтологий, интерпретации которых представляют один и тот же объект реального мира. В контексте семантической интеграции НСИ промышленного предприятия это является серьезным недостатком, так как информационные источники часто содержат данные, описывающие одну и ту же сущность. Их обработку необходимо вести совместно во избежание появления различного рода противоречий.

1.6.6. Технологии и стандарты Semantic Web в контексте семантической интеграции

Особый интерес в плане применения и разработки семантических методов интеграции информации представляет глобальная сеть интернет. Это обусловлено тем, что в глобальной сети Интернет представлено огромное количество разнородных информационных ресурсов. Их использование заключается в отборе ресурсов и, далее, содержащихся в них данных, наиболее релевантных решаемой задаче. Данный процесс можно рассматривать как проведение семантической интеграции информации человеком с целью образования информационного пространства, на основании которого он способен выработать определенное решение.[47] Существенной трудность при этом является гигантский объем данных в сети Интернет, требующий разработки новых подходов обработки данных.

Для решения данной проблемы используются поисковые машины. Они позволяют облегчить сбор информации путем предоставления ранжированного списка ресурсов, который можно также отнести к интегрированному представлению информации об объекте запроса. Однако такое преставление, как правило, включает большие фрагменты данных, в которых может находиться лишь небольшое количество нужной пользователю информации. Это не решает задачи в полной мере, т.к. приводит к необходимости последующей обработки представленного набора информационных ресурсов человеком. Отдельное усовершенствование механизмов индексации и поиска не оказывает должного влияния на эффективность информационного обеспечения пользователя. Например, использование в поисковых машинах методов компьютерной лингвистики позволяет повысить релевантность результатов. Что достигается за счет выявления семантики в индексируемых ресурсах в процессе их комплексного языкового анализа, но ценой этого является существенное уменьшение быстродействия поисковых систем и требует привлечения дополнительных вычислительных ресурсов[53].

Веб-ресурсы представлены в виде блоков текста в формате HTML, связанными друг с другом URL ссылками. Машинопонимаемой семантики информации такие ресурсы не несут. Данное обстоятельство является следствием того, что с появлением ЭВМ долгое время существовала точка зрения, что данные должна хранить ЭВМ, а их семантику – человек. Разумеется, в рамках решения одной задачи это вполне допустимо. Программист создает некоторую структуру с данными и программный код, в котором реализует свои знания о семантике данных. Это приводит к тому, что программа «знает» что значит тот или иной фрагмент информации, и каким образом он должен быть обработан. Однако Интернет предполагает общий доступ к информационным ресурсам, то есть их обработка может осуществляться впоследствии множеством программ и людей. Но если для людей, как правило, семантика информации представлена, в виду описания данных посредством естественного языка, то для программ такого сказать нельзя. Это не позволяет использовать ЭВМ для осуществления эффективной обработки большого объема данных и приводит к проблеме информационного хаоса в сети Интернет.[47]

На преодоление обозначенных трудностей направлен проект построения так называемой *семантической сети (Semantic Web)*[82]. Основной идеей данного проекта является представления у любой информации ее семантики в виде метаинформации в рамках одного информационного ресурса. Это позволит сделать данные *машинопонимаемыми (machine readable)* и соответственно обеспечить их обработку с помощью программных агентов.

В рамках проекта *Semantic Web* используются такие технологии, языки и стандарты как:

- *XML (Extensible Markup Language)* [108], является гибким текстовым форматом для описания документов произвольной структуры. XML обеспечивает возможность включения метаинформации, несущей машинопонимаемую семантику, в контент ресурса;

- *RDF (Resource Definition Framework)* [119], стандарт принятый в 1999 году консорциумом W3C и поддержанный ведущими производителями ПО. Он включает две части: способ описания ресурсов и способ задания схем, по которым ресурс описывается. Первая часть (RDF) определяет простую модель для описания информационного ресурса в виде троек или триплетов, состоящих из элементов: объект, атрибут и значение. Вторая часть (RDF Schema) определяет более сложную модель, позволяющую представить структуру предметной области в виде сходном с диаграммой классов UML;

- *OWL (Web Ontology Language)* [114] определяет модель и язык, расширяющие возможности RDF и RDFS. Язык OWL использует синтаксис XML, включает конструкции для представления таксономии классов их свойств и экземпляров. Основной целью языка OWL является описание онтологий в виде веб-ресурсов. Онтологии в этом случае используются для определения семантики метаданных, которыми в свою очередь аннотируются фрагменты данных в информационных ресурсах;

- *SPARQL (Protocol and RDF Query Language)* [122] - язык запросов к RDF ресурсам и, одновременно, протокол передачи информации в виде RDF троек.

Как отмечается в [45], использование технологий, стандартов и языков, применяемых в проекте Semantic Web в решении задачи интеграции данных позволяет преодолеть их синтаксическую и структурную гетерогенность за счет повсеместного использования языка XML и хранения данных в виде наборов триплетов RDF. Также это создает предпосылки для успешного осуществления семантической интеграции в виду наличия модели и языка OWL, позволяющего формально представит семантику ресурса в машинопонимаемом виде посредством описания его онтологии.

1.7. Классификация применяемых информационных систем

Далее кратко рассмотрим классификацию используемых информационных систем по способу их организации, масштабу и сфере применения. По степени распределенности среди информационных систем можно выделить 2 группы:

1) Локальные;
2) распределённые;

Локальные информационные системы выполняются на одном компьютере, все компоненты (БД, СУБД, приложения) также расположены на одном компьютере.

Среди распределённых информационных систем можно выделить файл-серверные и клиент-серверные.[39] Отличие последних в расположении базы данных. В файл-серверных ИС БД находится на сервере, а на рабочих станциях расположены СУБД и приложения. В информационных системах, имеющих клиент-серверную архитектуру, базы данных и система управления базами данных расположены на сервере, а приложения размещены на рабочих станциях.

Клиент-серверные ИС делятся на многозвенные и двухзвенные. В ИС двухзвенного типа два типа узлов: сервер с БД и СУБД, и рабочие станции, на которых выполняются приложения. В многозвенных ИС присутствуют промежуточные узлы - серверы приложений, при этом приложения не обращаются к СУБД напрямую, а взаимодействуют с сервером приложений.

Отдельного внимания заслуживают ИС на основе «облачных вычислений» [135]. Концепция облачных вычислений (англ. «Cloud computing») подразумевает обеспечение повсеместного и удобного сетевого доступа по требованию к общему пулу (англ. pool) конфигурируемых вычислительных ресурсов (например, сетям передачи данных, серверам, устройствам хранения данных, приложениям и сервисам — как вместе, так и по отдельности), которые

могут быть оперативно предоставлены и освобождены с минимальными эксплуатационными затратами или обращениями к провайдеру .

На базе облачных вычислений реализована модель обслуживания «программное обеспечение как услуга», или SaaS (англ. «software as a service»).

В рамках данной модели, пользователю предоставляется возможность использования программного обеспечения, работающего в облачной инфраструктуре, доступного посредством различных клиентских устройств или тонкого клиента, в качестве которого может выступать веб-браузер (например, веб-почта) или посредством интерфейса программы. Управление и контроль за виртуальной инфраструктурой: серверов, операционных систем, хранения, индивидуальных возможностей программного обеспечения осуществляется облачным провайдером. Основное преимущество модели SaaS для потребителя состоит в отсутствии затрат на установку, обновление и поддержку работоспособности оборудования и работающего на нём программного обеспечения.

По размеру можно произвести классификацию ИС на одиночные информационные системы, групповые и корпоративные информационные системы.

Одиночные ИС выполняют работу на персональном компьютере без использования сети. Одиночные ИС включают несколько приложений, и рассчитаны на работу одного пользователя. В основе подобных систем, как правило, локальные системы управления базами данных (СУБД).

Групповые информационные системы ориентированы на использование коллективом и формируются на базе локальной вычислительной сети организации и сервера баз данных.

Развитием коллективных ИС являются корпоративные информационные системы являются. Пользователями являются крупные организации с

территориально удаленными узлами. Для корпоративных информационных систем характера архитектура клиент-сервер или многоуровневая архитектура.

Можно выделить 3 наиболее важных тенденции корпоративных информационных систем:

1) Появление методик управления предприятием;
2) улучшение производительности ИС;
3) развитие методов реализации компонентов и служб информационных систем.

Прогресс в области технологий организации сетей и передачи данных обеспечивают рост функциональности и производительности современных информационных систем организаций. Наряду с ростом функциональных возможностей аппаратной части информационных систем наблюдается недостаток новых, эффективных и универсальных методов реализации программно-технологической части информационных систем. Согласно [34], в современном развитии информационных систем выделяются три наиболее существенных новшества.

1. Новые подходы к программированию. Модульное программирование фактически вытеснено объектно-ориентированным. Методы построения объектных моделей непрерывно совершенствуются. Внедрение технологий объектно-ориентированного программирования позволяет сократить сроки разработки сложных информационных систем, а также упростить поддержку и развитие.
2. Развитие сетевых технологий способствует замещению клиент-серверными и многоуровневыми локальных информационных систем.
3. Развитие глобальной сети Интернет позволяет реализовывать удаленную работу с подразделениями, использовать средства электронной коммерции. Также использование Веб-технологий во внутренних сетях

предприятий позволяет централизовать работу отделов. Следует также отметить, что использование глобальной сети Интернет в рамках корпоративных информационных систем повышаются требования к обеспечению безопасности и надежности данных.

Далее рассмотрим классификацию информационных систем по области решаемых задач. Можно выделить четыре группы [34]:

1. Системы обработки транзакций в реальном времени (англ. Online Transaction Processing);
2. системы поддержки принятия решений (англ. Decision Support System);
3. информационно-справочные системы;
4. информационные системы офисной автоматизации.

Системы обработки транзакций разделяются на пакетные и оперативные. В информационных системах управления преобладает режим оперативной обработки транзакции для отражения актуального состояния предметной области в любой момент времени, а пакетная обработка практически не задействована.

Системы поддержки принятия решений позволяют средствами сложных запросов произвести фильтрацию и анализ данных в различных разрезах: временных, географических и т.п. Большинство информационно - справочных систем осуществляю поиск по гипертекстовым документам и мультимедийному контенту.

Офисные информационные системы осуществляют преобразование документов на бумажном носителе в электронный вид, автоматизацию управление документооборотом и делопроизводства.

Следует отметить, что приведенная классификация весьма условна. Применяемые информационные системы часто обладают признаками всех перечисленных классов.

Так как ИС создаются для решения задач конкретной предметной области, то в каждой предметной области можно выделить тип ИС. Рассмотрим в качестве примера следующие типы ИС[38]:

1. Экономические информационные системы характеризуются набором организационных, программных и технических средств, объединённых в единую систему объединенную целью сбора, хранения, обработки и выдачи информации с целью осуществления функций управления.

   К данному классу относятся:

   – Информационные системы банков;
   – информационные системы фондового рынка;
   – информационные системы страховых компаний;
   – информационные системы налоговых служб;
   – информационные системы организаций (особое место по значимости и распространенности в них занимают бухгалтерские ИС);
   – статистические информационные системы.

2. Медицинские информационные системы - информационные системы, объединяющие функции хранения записей пациентах, данные медицинских исследований, мониторинга состояния пациента с медицинских приборов, средства общения, финансовая и административная информация.
3. Геоинформационные системы (также ГИС - географическая информационная система) – системы сбора, хранения, анализа и визуализации пространственных гео-данных и связанной информации о представленных в данной системе объектах. Это набор инстрементов для поиска, анализа и формирования карт и дополнительной информации об объектах. ГИС обычно состоят из СУБД, графических редакторов,

аналитических средств картографии, геологии, метеорологии, управлении, транспорте, экономике.

Информационные системы промышленных предприятий как правило включают в себя несколько подсистем разного уровня. Например, по данным MESA International [71] (Manufacturing Enterprise Solutions Association — Международная ассоциация производителей систем управления производством, www.mesa.org) структура информационной системы промышленного предприятия может быть представлена следующей схемой (рисунок 4):

Рисунок 4 - Информационно-управляющая структура промышленного предприятия

- На первом уровне находятся АСУТП;
- на втором уровне MES-системы;
- на третьем уровне — ERP-системы;
- на верхнем уровне находятся OLAP-системы.

АСУТП — автоматизированные системы управления технологическими процессами;

MES — (Manufacturing Execution System) — исполнительная система производства, автоматизированная система управления производства, информационно-вычислительная система. Системы такого класса решают задачи синхронизации, координируют, анализируют и оптимизируют выпуск продукции в рамках какого-либо производства в режиме реального времени.

ERP - (Enterprise Resource Planning) — Система планирования ресурсов предприятия. Основное назначение ERP — управление финансовой и хозяйственное деятельностью предприятия. ERP-система работает на самом верхнем уровне в иерархической лестнице систем управления, она затрагивает основные аспекты всех элементов производственной и торговой деятельности предприятия.

OLAP — (On-Line Analytic Processing) — Оперативный многомерный анализ данных. Аналитическая обработка в реальном времени, технология обработки информации, включающая составление и динамическую публикацию отчетов и документов. Используется аналитиками для быстрой обработки сложных запросов к базе данных. Служит для подготовки бизнес-отчетов по продажам, маркетингу, в целях управления, т.н. — data mining — добыча данных (способ анализа информации в базе данных целью отыскивания аномалий и трендов без выявления смыслового значения записей).

Передача информации осуществляется по всем ступеням иерархи системы (рисунок 4). Из производственной зоны (АСУТП) информация поступает к MES-системам, проходит стадию обработки, а затем уже обработанная информация поступает к MES-системам, проходит стадию обработки, а затем уже обработанная информация поступает в ERP-системы, и далее — на уровень высшего менеджмента предприятия (OLAP).

1.8. Примеры систем повышения эффективности поиска, основанных на формализованных знаниях.

На сегодняшний день можно выделить несколько проектов компаний по проектированию информационных систем, решающих, в том числе, и проблему повышения эффективности доступа к информации организаций. В основном, данные решения направлены на обеспечение информационного взаимодействия между подразделениями организации на основе внедрения системы электронного документооборота, интеграции информационных ресурсов и осуществления информационного поиска по документам предприятия.

В данном разделе рассмотрены примеры таких информационных систем с точки зрения их функционала, а также методов и технологий, применяемых для их реализации.

Архивариус 3000

Программа предназначена для поиска документов и почтовых сообщений в рамках одного компьютера, в локальной сети и в съёмных дисках. Поиск производится по содержимому документов, с учётом морфологии. Архивариус 3000 выполняет запросы на естественном языке. Документы могут быть найдены по ключевым словам или с использованием языка запросов, также как в традиционных поисковых системах. Во время поиска программа автоматически использует все грамматические формы слова и обеспечивает смысловой поиск на 18 языках[3].

В процессе индексирования документов и почтовых сообщений извлекает и сохраняет полную информацию. Во время поиска, даже если документ физически недоступен, программа найдёт его по содержимому и определит, на каком диске находится разыскиваемый файл. Возможность хранения текстов можно отключить. [49] Сервер удалённого поиска, встроенный в программу, позволяет через сеть искать и использовать документы с другого компьютера, используя web – интерфейс. Таким образом, программа позволяет осуществлять

поиск, получать удаленный доступ к документам и отсылать их по электронной почте.

Coredge Logik My Edition 2[69]

Пакет программ Logik 2.0 является средством организации неструктурированных данных в логические библиотеки. Пакет программ Logik извлекает ключевые фразы из документов автоматически, проводя систематизацию без привлечения пользователей. Правила классификации для обеспечения процесса формирования связей между документами, вырабатываются программой, а не человеком. В инженерном тестовом центре разработчика отмечается, что такой подход оправдан при работе системы в динамичной предметной области. [49]

Точность работы пакета программ Logik коррелирует с множеством начальных фраз, введенных в библиотеку. В случае многозначности лексических единиц, программа часто неверно выполняет классификацию документов. Для реализации поиска совместно с темами применяются фильтры. Список фильтров задается через интерфейс администратора. Библиотека визуализируется в пакете программ в виде дерева в левой части экрана. Папки отображаемой библиотеки являются взаимосвязанными темами и организованы в соответствии с введенными пользователями первоначальными фразами. Древовидную структуру можно изменять, например, после изменения расположения папки все связи между словами обновляются в автоматическом режиме. В процессе осуществления поиска документов, пакет программ Logik отображает содержание документов, полученное как множество первых 30 тематик документа. При просмотре документа программа выделяет список тем форматированием. Следует отметить существование возможности интеграции пакета программ Logik с программами Microsoft Word, Excel и PowerPoint, что облегчает процесс добавления документов в библиотеку. Также существует возможность интеграции с программами Microsoft Outlook и Internet Explorer. После осуществления поиска программа производит синтаксический анализ

документов с целью создания краткого содержания и ключевых тем документа. Пакет программ Logik также отслеживает экспертов по отдельным темам, позволяя установить авторство документа или группы документов. Программа также осуществляет перевод документов. [49]

IBM Lotus Discovery Server 2.0.1[70]

Пакет программ Discovery Server фирмы Lotus ставит своей целью автоматизировать процессы управления знаниями и документами. Программа отслеживает информацию, с которой работает пользователь. Обрабатываются данные из электронной почты, файл-серверов, баз данных и платформ коллективной работы. Данные сохраняются автоматически без привлечения пользователя (IBM данные пользователя названы «путеводными знаками»). На основе «путеводных знаков» пакет программ Discovery Server формирует сводку знаний и обязанностей пользователя. Также среди функциональных возможностей заявлено создание профиля пользователя на основе документов, с которыми они работают, также предпринята попытка отследить персональную информацию, а именно предпочтения, навыки и образование. На основе полученных данных формируется профиль пользователя. Данный профиль в дальнейшем используется в процессе поиска для формирования критериев. Существует возможность использования профилей сторонних программных продуктов, например, ERP-системы PeopleSoft или каталога Active Directory от Microsoft. В пакете программ Discovery Server существует механизм вычисления вероятности соответствия результатов на основе критериев поиска. Вывод результатов осуществляется в порядке убывания вероятности. Пакет программ также формирует краткое содержание документов. Краткое содержание содержит 3 наиболее значимых в документе предложения. Сравнение слов выполняется с помощью встроенного изменяемого пользователем в ручном режиме тезауруса. Существует возможность классификации документов на основе правил. Пакет программ Discovery Server отслеживает и указывает на присутствие других пользователей

в сети. Отслеживая предпочтения пользователя, пакет имен рекомендательную функцию по предложению интересных тем пользователю. Система хранения профилей и мониторинг совместно обеспечивают возможность выявления неявных знаний. [49]

Inxight SmartDiscovery 3

Пакет программ SmartDiscovery разработан компанией Inxight и является многоуровневым сервером, который классифицирует текст документов различных форматов и БД. Система классификации должна быть задана в формате XML.

SmartDiscovery включает модули синтаксической обработки текста для анализа более 25 языков с учетом синтаксиса, среди которых арабский, фарси, иврит, русский и корейский. Используется технология обработки запросов на естественном языке исследовательского центра PARC Xerox. SmartDiscovery Server позволяет извлекать ключевые слова, также выявляет в документах фразы определенных типов, такие как даты и географические названия. Анализатор синтаксиса содержит индексы совместного употребления существительных. Например, поиск по ключевому слову «масло» выдает названия разных видов масел, а также расширить результаты совместно употребимыми словами, такими как виды и стоимость масла. Правила синтаксического анализатора при анализе текста, не объедены какой-либо системой классификации, а также не могут быть изменены. Построение индексов поиска и ассоциаций на основе набора правил осуществляется без участия человека. Для расширения базы знаний анализатора и спектр выявляемых им фраз необходимо обновление версии пакета. SmartDiscovery использует визуализацию данных Star Tree для отображения/редактирования системы классификации. Форма поиска и форма визуализации расположены на одной странице[49].

Open Text Livelink 9.1

Пакет программ Open Text Livelink содержит средства управления знаниями и платформу для коллективной работы. Главной функцией Livelink является управление контентом и документами. Livelink также реализует функции автоматизации бизнес–процессов и документооборота организации, проводя интеграцию программ и приложений сторонних разработчиков. Система классификации может быть импортирована из XML-файла либо создана. Классификации и категории при вводе создают профиль, который содержит все правила, использованные для извлечения ключевых фраз. Имеется возможность внесения правил пользователями. Результаты поиска можно оценивать с целью более точного выбора документов. Система оценки применяется и к тем пользователям, которые осуществляют навигацию по документам. Таким образом, система имеет возможность поиска экспертов и людей с общими интересами. Это может рассматриваться как функция доставки неявно сформулированных знаний. В поисковом запросе используются ключевые слова, запросы на естественном языке, метаданные документов, логические выражения. Livelink имеет функцию создания краткого содержания документов, которое рассматривается как отдельный элемент при поиске. Реализована возможность выбора репозиториев для поиска, среди которых репозитории сторонних программ управления документооборотом (например, Documentum). Используется множество атрибутов поиска, поэтому не существует строгих ограничений или правил построения запроса. Например, для осуществления поиска возможен ввод ключевых слов и имени автора вместе. Livelink имеет возможность поиска по естественно-языковому запросу. Особенностью этой функции отмечается возможность определения части репозитория знаний для поиска. Если пользователь ищет документы определенного автора, то поиск осуществляется по документам, которые имеют соответствующее описание. Управление данными осуществляется моделью полномочий, через которую транслируются результаты поиска и

классификации документов. Livelink не имеет встроенных функций изменения системы классификации и средств визуального просмотра данных.

Таким образом, в перечисленных решениях обеспечивается интеграция на основе таких подходов как консолидация и распределение данных, недостатки которых были описаны в предыдущих разделах. Следует также отметить, что для данных в этом случае определяется лишь унифицированных формат представления, а их семантика в большинстве случаев остается недоступной для машинной обработки. Это затрудняет построение систем осуществляющих проведение процедур анализа информации или поддержку принятия решений. Нереализуемым также является вопрос расширения информационных баз систем, которое может приводить к появлению новых информационных элементов и их атрибутов. Такая ситуация может потребовать одновременного обновления как структур интегрируемых данных, так и приложений, осуществляющих их обработку, что является трудновыполнимым в рамках изменяемых предметных областей. Также следует отметить, что отражение любого изменения предметной области и/или бизнес-процессов организации влечет необходимость привлечения экспертов.

На основе вышеизложенного материала был проведен сравнительный анализ интересующих нас возможностей существующих методов и средств поиска нормативно-справочной информации организаций. Результаты сравнительного анализа сведены в таблицу 2.

Таблица 2 – Возможности существующих решений в области эффективного доступа к документам организаций

	Архивариус 3000	Fulcrum PC DOCS	Documentum i4	Knowled-ge Station	DocsFusion	Требования к разработке
Накопление знаний	-	+	+	+	-	+
Автоматический анализ данных	-	-	+	-	+	-
Автоматическое формирование СМПО	-	-	-	-	-	+
Семантический поиск	+	+	-	-	-	+
Автоматизированное профилирование пользователей	-	-	-	-	-	+
Отсутствие эксперта	+	-	-	-	-	+

На сегодняшний день не существует систем, а также методов их формирования и функционирования, позволяющих на высоком уровне автоматизации организовать эффективный поиск и навигацию в рамках документооборота организаций. В свою очередь совершенствование процессов поиска и навигации по нормативно-справочной информации организации необходимо проводить с учетом следующих требований:

1. Поддержка мультипредметности. Предметная область обладает динамикой по структуре, содержанию и объему данных. Разрабатываемая

система должна обладать возможностью обработки, в т.ч. интеграции, разнородных данных предметной области. Данные для организации и функционирования должны быть представлены в унифицированном виде, тем не менее, данная унификация не должна препятствовать возможности адаптированного представлении информации.

2. Отсутствие необходимости в привлечении экспертов для формирования и функционирования ИС. Большой объем и динамика данных в современных организациях накладывают ограничения для привлечения человека для формирования и интеграции формализованных знаний. Разрабатываемая система должна обладать возможностью обработки разнородных данных предметной области без привлечения экспертов по знаниям.

3. Персонализация и адаптация к пользователю. Персонализация определяет то, каким образом информационные службы должны адаптироваться к предпочтениям пользователя. В системах обеспечения поиска и навигации такой подход должен быть ориентирован на предпочтения пользователей, явно связанные с предметной областью деятельности пользователя. Взаимодействие с пользователем должно быть организовано с учетом предпочтений пользователя, а также с возможностью автоматизации получения предпочтений пользователя.

Таким образом, вследствие перечисленных выше особенностей, для совершенствовании процессов поиска нормативно-справочной информации, с учетом особенностей предметной области, целесообразно использовать методы системного анализа, математического моделирования, методов и моделей интеллектуальных систем.

Выводы по главе 1

Повышение эффективности доступа к информационным ресурсам организаций является сложной многоаспектной проблемой, требующей учета специфики предметной области;

Существующие подходы к интеграции информационных ресурсов обеспечивают лишь структурную интеграцию или используют простые шаблонные модели, не позволяющие в полной мере представлять и оперировать формальной семантикой;

Семантическая интеграция данных предполагает разработку семантического представления информационных ресурсов в виде их онтологий, при этом необходимым условием является привлечение экспертов по знаниям, а также заранее известный перечень ресурсов для интеграции, что труднодостижимо в реальных условиях.

Практически все существующие системы, повышающие эффективность информационного доступа, оперирующие формальными знаниями, ориентированы на жестко заданную модель предметной области, а также подразумевают привлечение экспертов по знаниям, как на этапе создания, так и функционирования, что увеличивает стоимость владения подобными системами. Кроме того, используемые подходы лишены адаптивности к различным категориям пользователей и не справляются со своими задачами при существующем темпе роста количества документов, а также динамикой самой предметной области современных ораганизаций.

Требуется разработка новых эффективных методов обработки информации, интеграции и обработки знаний, методов человеко-машинного взаимодействия, комплексное применение которых позволит повысить уровень автоматизации процессов интеграции разнородных ресурсов с целью повышения эффективности навигации и поиска информации.

ГЛАВА 2. Подход и модели построения корпоративных информационных систем, основанных на формализованных знаниях

В данной главе представлен поход к построению и концептуальная модель мультипредметной информационной системы организаций, основанной на формализованных знаниях, полученной на основе результатов анализа существующих решений в области разработки, внедрения и эксплуатации существующих информационных систем организаций. Представлено формальное определение динамической мультипредметной информационной системы с обратной связью. Рассмотрена роль пользовательского опыта в контексте получения обработки и передачи информации. Представлена модель предпочтений пользователя мультипредметной информационной системы.

2.1. Подход к построению мультипредметных информационных систем, основанных на знаниях

Большинство функционирующих организаций обладают системами сбора и хранения отдельной документации, данных об объемах производства, однако не представляется возможным наглядно и обозримо провести систематизацию имеющейся информации. Этому также способствует жесткая организация каналов распределения информации. Однако, как отмечается в [5], обмен информацией позволяет провести унификацию данных, а также стимулирует поиски новых подходов к решению управленческих задач. «Организации, в которых лучше поставлено дело по сбору внешней информации и ее внутреннему распределению, могут лучше спрогнозировать динамику рыночных тенденций и действовать без промедления, более обоснованно принимать решения. Первоочередной является информация о новой продукции, технологических процессах и применяемых стратегиях, ее использование уменьшает у предприятий степень риска столкнуться с непредвиденными ситуациями.»

Опираясь на анализ существующих информационных систем и требования к разработке, приведенные в главе 1, перечислим основные принципы подхода к построению мультипредметных информационных систем, основанных на формализованных знаниях.

Мультипредметность. Разрабатываемая информационная система должна обладать возможностью поиска, сбора, хранения, обработки и интеграции разнородных данных предметной области без привлечения экспертов по знаниям. Данные для организации и функционирования мультипредметной информационной системы должны быть представлены в унифицированном виде, тем не менее, данная унификация не должна препятствовать возможности адаптированного к различным категориям пользователей представлении информации.

Использование формализованного представления знаний. Исходные данные могут быть представлены в различных форматах. Для использования в мультипредметной информационной системе все данные должны быть приведены к структурированному, машиночитаемому формату, также понятному человеку.

Автоматизация получения и обновления знаний. Ввиду большого количества документов и скорости изменения нормативно-справочной информации организации, состав и количество метаданных в виде формализованного представления НСИ также могут меняться с течением времени. Также влияние на НСИ оказывают пользователи, которые с ней работают. В связи с этим сервисы, входящие в мультипредметную информационную систему должны быть способны к автоматизированному поддержанию актуальности формализованного представления НСИ.

Ориентированность на пользователя. Для персонализации результатов работы системы, система должна использовать имеющуюся и накопленную информацию о пользователе и его предпочтениях. Такой информацией

является: категория пользователя (например, принадлежность определенному отделу организации), статистика и семантическая обработка его запросов, взаимодействие с навигацией информационной системы.

Использование знаний пользователей. Автоматизированное формирование сложной семантической структуры, описывающей понятия и процессы организации становится возможным благодаря использованию знаний пользователей–экспертов в различных предметных областях. В силу специализации подразделений организаций каждый сотрудник способен выступать в роли эксперта, обладающими необходимыми знаниями соответствующего подразделения. Данное обстоятельство позволяет не привлекать экспертов по знаниям для формирования формализованных знаний, как это предлагается в существующих решениях.

Адаптация пользовательского интерфейса. Возможность адаптации пользовательского интерфейса информационной системы промышленного предприятия к различным категориям пользователей, являющимися представителями различных подразделений организации способствует повышения эффективности навигации в ИС.

Данные принципы легли в основу модели мультипредметной динамической информационной системы организаций с обратной связью.

Далее представлена концептуальная модель мультипредметной информационной системы промышленного предприятия, применение которой позволяет повысить эффективность работы с информацией на основе использования технологий получения, интеграции и сопровождения семантических метаданных, семантического поиска, семантического профилирования пользователей. Концептуальная модель мультипредметной информационной системы представлена на рисунке 5.

Рисунок 5 – Концептуальная модель мультипредметной информационной системы

Целью основанной на знаниях мультипредметной информационной системы является обеспечения процесса поиска пользователем нормативно-справочной информации, представленной в виде текстовых документов. При этом повышение эффективности данного процесса достигается путем повышения полноты и точности поиска документов, за счет автоматизированного формирования интегрированной семантической модели предметной области и реализации в рамках этой модели интерфейсной навигации и поиска документов.

Компонентами мультипредметной информационной системы, основанной на знаниях, являются: семантическая модель предметной области (СМПО), модель предпочтений пользователей (МПП), база индексов текстовых документов организации, навигационный и поисковый интерфейсы.

Входными данными, обеспечивающими функционирование мультипредметной информационной системы организации, являются текстовые документы организации и статистика работы с системой пользователя. Выходными данными являются результаты поиска по запросу и навигационная структура интерфейса.

Семантическая модель предметной области формируется в результате интеграции семантических образов документов организации. Семантический

образ документа – семантическая сеть, множество вершин которой составляют понятия СМПО, присутствующие в документе, множество ребер – множество двухместных отношений над понятиями. Семантическая модель предметной области содержит проинтегрированные семантические образы всех документов информационной системы в унифицированном виде, без привязки к формату и физическому размещению.

Модель предпочтений пользователя - семантическая сеть, множество вершин которой составляют понятия СМПО, которыми оперирует пользователь, множество ребер – множество взвешенных двухместных отношений над понятиями, вес которых характеризует значимость семантического отношения между понятиями для пользователя, определенную на основе статистики его взаимодействия с системой. МПП служит «эталоном» в процессе формирования пользовательского интерфейса и его оценки как степени соответствия структуры интерфейса модели предпочтений пользователя. МПП формируется на основе обработки запросов пользователя и статистики его работы. Запросы пользователей представлены множеством понятий предметной области, представленных множеством ключевых слов. Взаимодействие пользователя и мультипредметной информационной системы может быть представлено следующим алгоритмом:

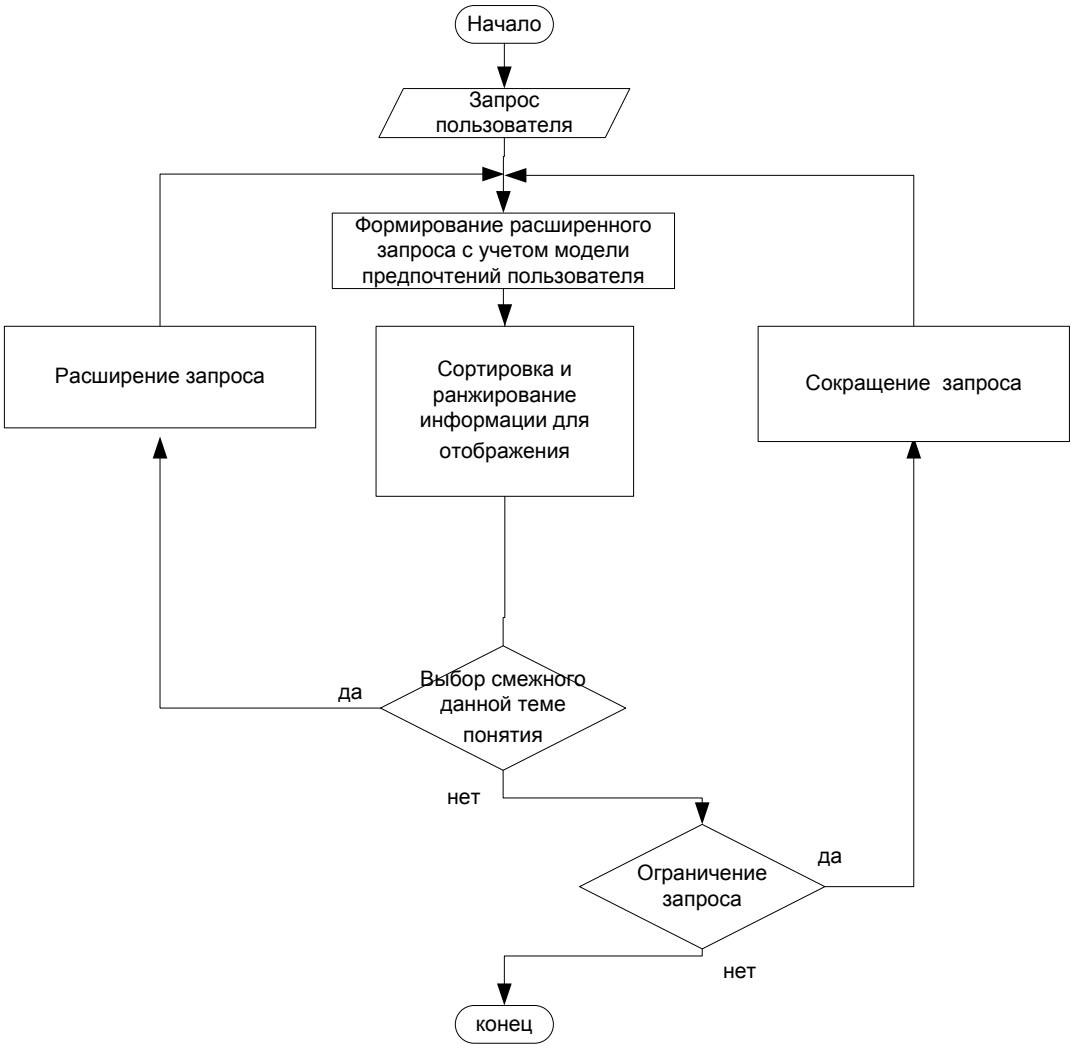


Рисунок 6 – Алгоритм взаимодействия пользователя и ИС

Взаимодействие заключается в итеративном расширении запросов пользователя на основе МПП, обеспечении возможности коррекции запроса средствами адаптивного интерфейса, а также учета пользовательских предпочтений путем коррекции весовых коэффициентов между понятиями МПП. При совместном употреблении в запросе пар понятий, входящих в состав МПП, изменяются весовые коэффициенты отношений между данными понятиями.

Семантическая модель предметной области, модель предпочтений пользователей и база индексов документов составляют системные информационные базы, используемые в дальнейшем для формирования

навигационно-поискового интерфейса, обеспечивающего эффективный (в смысле скорости и релевантности) доступ пользователя к требуемым данным.

Поисковый интерфейс содержит поле для ввода запроса пользователя, поле для визуализации расширенного запроса с возможностью участия пользователя (изменения запроса) в процессе поиска, и поле для вывода результатов поиска.

2.2. Сценарная модель мультипредметной информационной системы

Для разработки и описания функционирования мультипредметной информационной системы масштаба предприятия целесообразным является использования унифицированного языка моделирования UML (англ. «Unified Modeling Language»). Язык графического описания объектного моделирования UML применяется в области разработки программного обеспечения, моделирования бизнес-процессов, системного проектирования и отображения организационных структур. UML также является открытым стандартом, который использует графические обозначения с целью построения абстрактных моделей проектируемой системы, называемой UML-моделью. Язык UML создан для описания, визуализации, проектирования и документирования крупных программных систем.

Язык UML подразумевает использования структурных диаграмм, диаграмм поведения и диаграмм взаимодействия компонентов разрабатываемой программной системы.

Наибольший интерес в аспекте определения и реализации функциональных требований на основе имеющихся принципов представляют диаграммы вариантов использования (сценариев) (англ. Use case). Данная диаграмма отражает отношения между актёрами (в роли которого может выступать человек, внешняя сущность, класс, другая система) и прецедентами (выполняемыми системой действиями) и является составной частью модели прецедентов, позволяющей описать систему на концептуальном уровне.

Актёры не могут быть связаны друг с другом (за исключением отношений обобщения/наследования),

Диаграмма сценариев использования мультипредметной информационной системы представлена на рисунке 7:

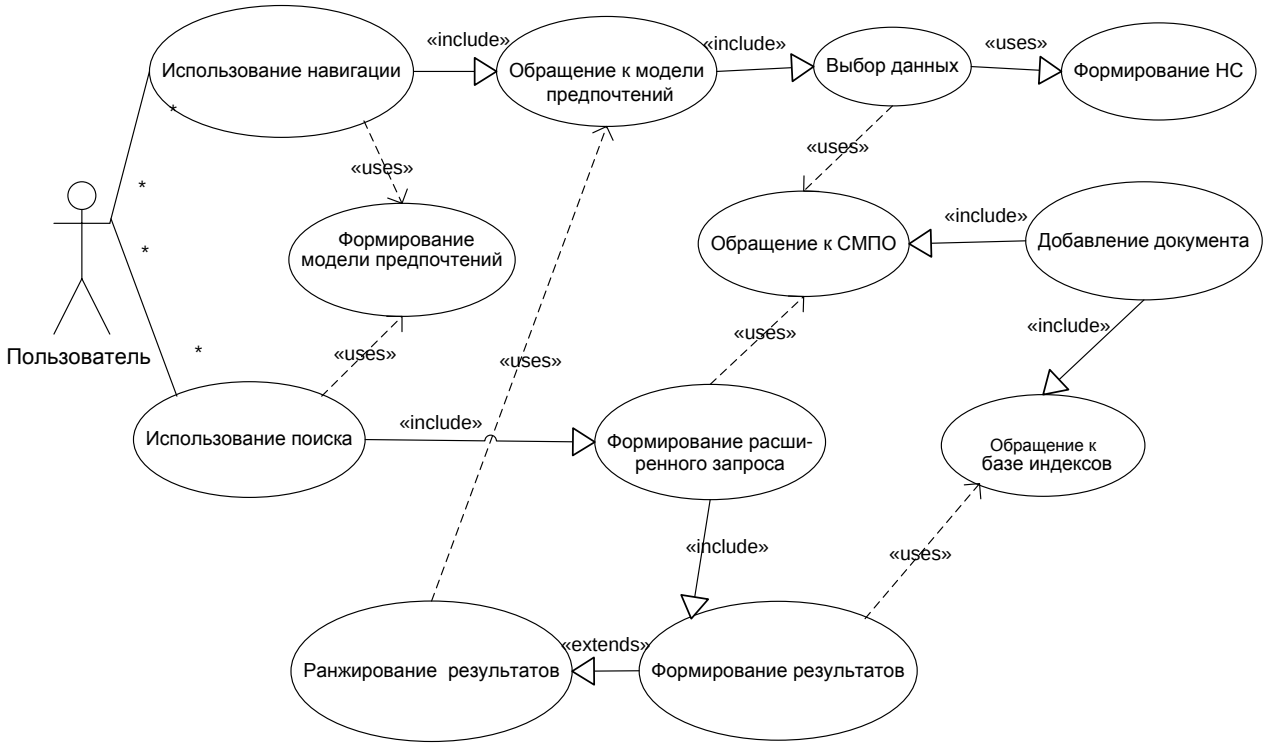


Рисунок 7 – Диаграмма сценариев использования мультипредметной информационной системы

Сценарий «Использование навигации» позволяет пользователям осуществлять навигацию в информационной системе, при этом осуществляется в автоматическом режиме адаптация структуры навигации в соответствии с выбранной моделью предпочтений пользователя. Также при использовании навигации формируется и сама модель предпочтений пользователя.

Сценарий «Использование поиска» позволяет осуществить поиск документов по запросу, включает формирование расширенного запроса на основе семантической модели предметной области, а также ранжирование результатов на основе модели предпочтений пользователей.

Рассмотрим более подробно реализуемые мультипредметной информационной системой сценарии – сценарий навигации и сценарий поиска.

Сценарий навигации

Навигация в ИС это процесс доступа пользователя к необходимой информации путем осуществления навигационного поиска по заданной структуре классификации информации.

Диаграмма сценариев использования навигации представлена на рисунке 8:

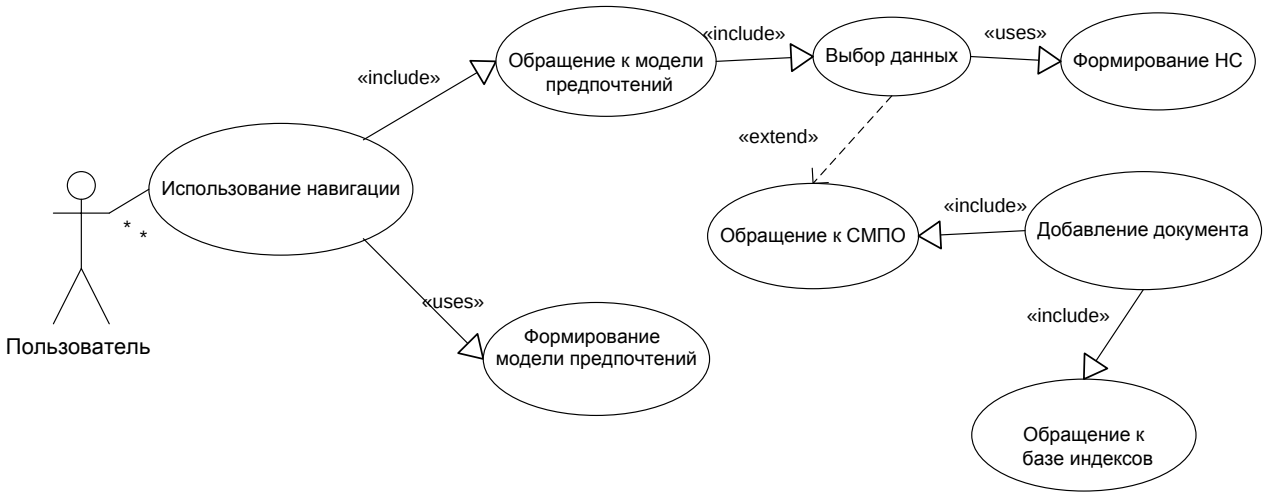


Рисунок 8 - Диаграмма сценария использования навигации

Для реализации навигации с учетом мультипредметной специфики необходимо учитывать предпочтения пользователя с целью формирования модели предпочтений. Модель предпочтений пользователя используется для адаптации информационного содержания интерфейса и ранжирования результатов поиска информации. Более подробно модель предпочтений пользователей будет представлена далее.

Сервис навигации должен учитывать модель предпочтений пользователя при формировании навигационной структуры. Однако навигация должна быть осуществима и для неперсонифицированного пользователя, например, в случае еще не полностью сформированной модели предпочтений пользователя.

Данный сценарий подразумевает использование заранее заданной общей структуры навигации в рамках информационной системы с последующей ее адаптации к различным категориям пользователей. Структурой навигации будем называть древовидную структуру, элементы которой связывают множество информационных элементов информационной системы в группы в соответствии с некоторыми классификационными признаками. В соответствии с принципами адаптации к пользователю и использования знаний пользователя целесообразным является использование в качестве классификационных признаков, разделяющих все множество информационных элементов, понятий, которыми оперирует тот или иной пользователь. Для этого сценарий осуществления навигации расширяется выбором модели пользователя мультипредметной информационной системы.

Далее выполняется выбор данных, соответствующих модели предпочтений пользователя и формировании на основе полученных данных и модели предпочтений пользователя навигационной структуры интерфейса для последующего предъявления последней пользователю.

Сценарий поиск

Сценарий «Использование поиска» позволяет осуществить поиск документов по запросу, включает формирование расширенного запроса на основе семантической модели предметной области, а также ранжирование результатов на основе модели предпочтений пользователей. На рисунке 9 представлена диаграмма сценария использования поиска.

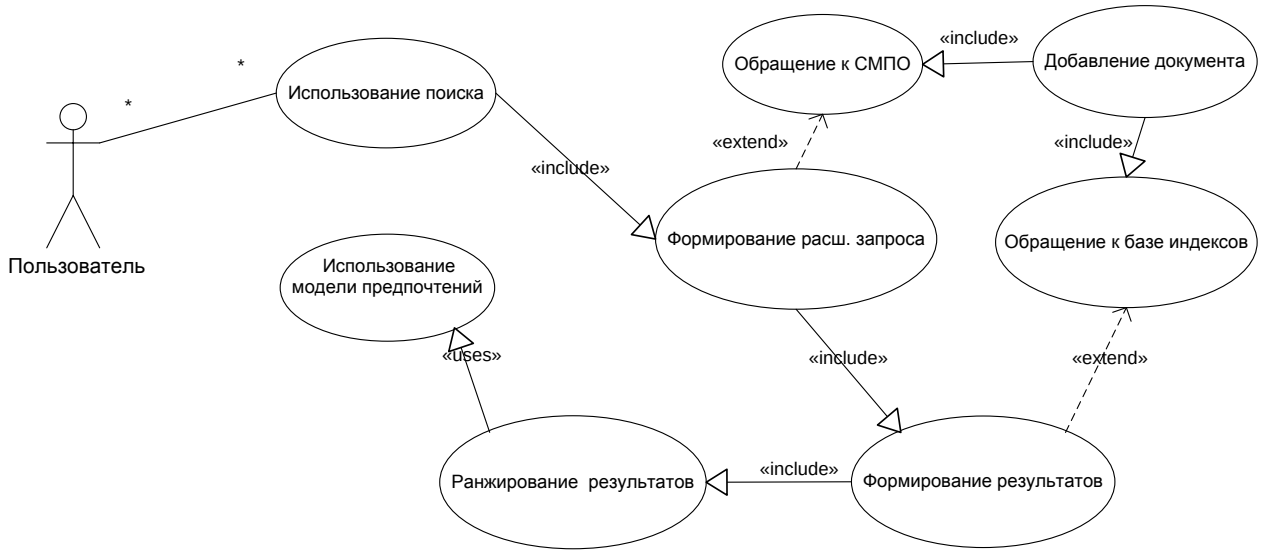


Рисунок 9 - Диаграмма сценария использования поиска

Аспектами повышения эффективности поиска документов являются время, точность результатов поиска, полнота результатов поиска.

Время осуществления поиска включает в себя время, затрачиваемое пользователем на ввод запроса, время вывода результатов, время, затачиваемое пользователем на ознакомление с результатами, коррекцию запроса и выполнения следующих итераций. В плане сокращения времени, необходимого для формирования запроса, используется автоматическое расширение запроса пользователя на основе семантической модели предметной области. В плане сокращения времени, необходимого на ознакомление с результатами используется автоматизированное ранжирование результатов на основе модели предпочтений пользователей. Для сокращения времени, необходимого для изменения расширенного запроса, необходимо разработать соответствующий интерфейс, визуализирующий расширенный запрос с возможностью его итеративного изменения.

Точность результатов поиска, предъявляемых пользователю, понимается, как строгое соответствие результатов выполнения запроса (поиска) запросу, введенному пользователем. В случае, когда запрос не может быть четко сформулирован в терминах предметной области, с указанием всех возможных

ограничений и контекстов, целесообразно говорить о соответствии результатов информационной потребности, или ожиданиям пользователя как о пертинентности результатов поиска. Пертинентность в данном работе рассматривается как соответствие результатов поиска модели предпочтений пользователя. Следует отметить, что запрос в большинстве случаев может быть сформулирован различными способами, поэтому для учета персональных особенностей пользователя результаты выполнения поиска ранжируются в соответствии с моделью предпочтений пользователей. Кроме того, на повышение точности поиска влияет ограничение области поиска в плане удаления заведомо несоответствующих модели предпочтений пользователя результатов поиска. Ограничение области поиска осуществляется на основании включения в расширенный запрос субтрактивных отношений модели предпочтений пользователя – отношений между концептами, имеющими отрицательный весовой коэффициент.

Полнота результатов поиска обеспечивается построением расширенного запроса на основе семантической модели предметной области. Семантическая модель предметной области используется для получения списка концептов, смежных в рамках модели концептам, представленным ключевыми словами запроса. Так же в запрос должны включатся концепты, являющиеся транзитивными для остальных концептов запроса.

Таким образом, использование субтрактивных отношений позволяет автоматизировать процесс добавления ограничений в расширенный запрос. Учет семантической модели предметной области и модели предпочтений пользователей позволяет учесть контекст используемых ограничений, а автоматизация данного процесса снимает необходимость в выработке и вводе ограничений непосредственно пользователем. Общая схема метода поиска может быть представлена следующей последовательностью шагов:

1. Ввод пользователем запроса в виде множества ключевых слов.

2. Формирование расширенного запроса, содержащего отношения и соответствующие запросу концепты семантической модели предметной области.

3. Вывод множества документов, соответствующих расширенному запросу.

4. Ранжирование множества документов с учетом модели предпочтений пользователя.

Таким образом, в основе эффективности формирования и функционирования динамической мультипредметной информационной системы организации лежат следующие аспекты:

1. Формирование в автоматизированном режиме семантической модели предметной области (СМПО).
2. Реализация семантического поиска с адаптивным ранжированием результатов и автоматизированным ограничением области поиска.
3. Реализация интерфейсной навигации в виде формирования адаптивных пользовательских интерфейсов на основе семантического профилирования пользователей.

2.3. Формальное определение мультипредметной информационной системы

На основе вышесказанного, представляется возможным формально определить мультипредметную информационную систему, как информационную систему, которая субъективно полезна, в приблизительно равной степени, для более чем одной категории пользователей, при этом его полезность для всех прочих категорий существенно ниже [64].

Пусть $U = \{u_i\}$ - множество категорий пользователей, $R = \{r_i\}$ - множество информационных ресурсов. Определим на их декартовом произведении некоторую функцию, характеризующую полезность ресурса для соответствующей категории пользователей:

$$f : R \times U \to \Re, \tag{2}$$

где $\Re$ - множество вещественных чисел.

Определим понятие «проблемно-ориентированный информационный ресурс» от обратного. Проблемно-ориентированным будем называть ресурс r, НЕ удовлетворяющий условию «равнополезности» для всех категорий пользователей:

$$\forall u_i, u_j \in U', f(r,u_i) \approx f(r,u_j) \tag{3}$$

Монопредметным информационным ресурсом будем называть ресурс r, удовлетворяющий условию:

$$\exists u_i \in U : f(r,u_i) >> f(f,u_j), \forall u_j \in U, j \neq i \tag{4}$$

Мультипредметным ресурсом будем называть информационный ресурс, удовлетворяющий условию:

$$\exists U' \subseteq U : \forall u_i, u_j \in U', f(r,u_i) \approx$$

$$\approx f(r,u_j) \wedge f(r,u_i) >> f(r,u_k), \forall u_k \in U \setminus U' \tag{5}$$

Знак приближенного равенства функций может быть определен следующим образом:

$$f(r,u_i) \approx f(r,u_j) \Leftrightarrow \left| f(r,u_i) - f(r,u_j) \right| \leq d, \tag{6}$$

где d – константа, задающая порог идентичности субъективной полезности ресурса.

Очевидно, что при данных определениях монопредметные и мультипредметные ресурсы относятся к категории проблемно-ориентированных.

2.4. Модель предпочтений пользователей мультипредметной информационной системы

Для реализации возможности адаптации информационного содержания, а так же реализации адаптивного поиска необходима модель предпочтений пользователя информационной системы.

Модель предпочтений пользователя используется для адаптации информационного содержания интерфейса и ранжирования результатов поиска информации. С целью снижения объемов хранимых системных данных модель строится не для каждого пользователя в отдельности, а для групп пользователей. Далее приведено формальное определение категорий пользователей.

Будем понимать под информационным ресурсом коллекцию документов, содержимое которых (контент) оперирует в смысле синтаксиса языка документа некоторым множеством понятий, складывающихся тем или иным образом в логическую систему. Логическая система образуется заданием на данном множестве понятий C различных семантических связей L, определяющих допустимые с точки зрения создателя документа способы взаимной интерпретации понятий из C: [23]

$$KB = \{C, L\}, \tag{7}$$

где C - множество понятий (концептов), L - множество отношений над понятиями.

Подобную систему в современной ИТ-науке принято называть онтологией [97]. Назовем систему, описывающую контент информационного ресурса *онтологией ресурса*.

Обратим внимание, что отношения на множестве понятий онтологии могут быть как симметричными, так и асимметричными. При этом один из концептов, участвующих в двухместном асимметричном отношении, может рассматриваться как *атрибут* другого.

Модель пользователя, использующего информационный ресурс, также может быть представлена в виде логической системы – *онтологии пользователя*. Данная онтология характеризует взаимосвязь понятий с точки зрения пользователя. Отметим, что (вследствие, вероятно, социальной природы человека) представления различных людей об одной предметной области в целом мало отличаются, что выражается в схожести структур различных пользовательских онтологий. Однако люди, различающиеся по различным признакам, таким, например, как принадлежность к социальным группам, возраст, пол, область профессиональной деятельности и другие, в процессе жизнедеятельности, как правило, оперируют различными фрагментами своих онтологий с разной интенсивностью. Для практического использования в рамках современных информационных систем данные зависимости должны быть формализованы.

Для определенного профессионального сообщества людей имеют место общие представления о некоторых объектах или задачах. Эта общность выражается в схожем ранжировании атрибутов понятий по значимости. При этом наиболее важные для пользователя атрибуты играют роль свойств, идентифицирующих объект. Например, в ментальной модели человека категории «управленец» экземпляр понятия «Карьерный экскаватор» будет идентифицироваться значениями атрибутов, характеризующих производительность и стоимость владения. В то же время для пользователя

отдела технического планирования экземпляр этого же понятия идентифицируется атрибутами, описывающими аспекты технического обслуживания – объем и время работ, требования к квалификации персонала, и т.п.

В этой связи можно определить на множестве атрибутов понятий отношение порядка, определяющее значимость атрибута для данного пользователя. Тогда некоторое количество наиболее значимых атрибутов (в представлении конкретного человека) будет идентифицировать объект окружающего мира как принадлежащий к тому или иному классу.

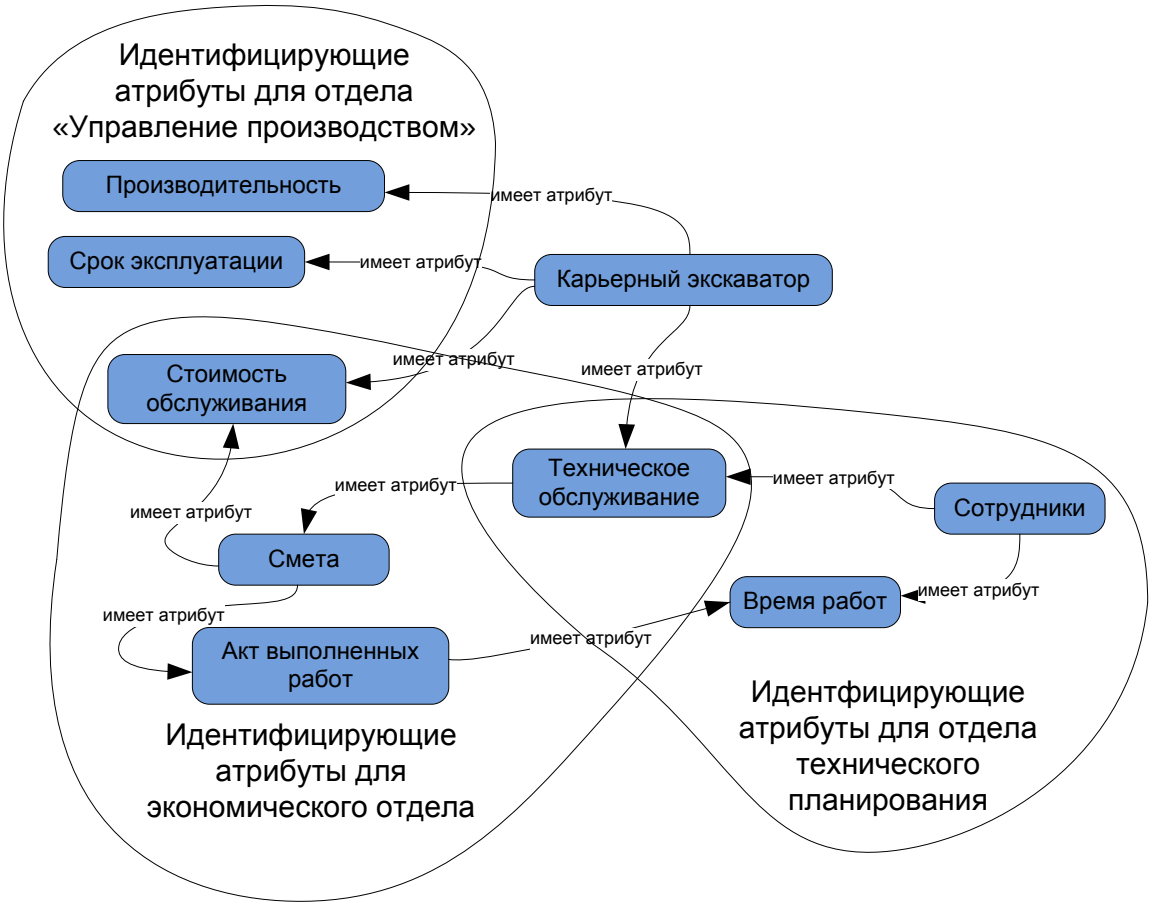


Рисунок 10 - Идентифицирующие атрибуты понятия «Карьерный экскаватор» для различных категорий пользователей

Пусть *C*- некоторое множество понятий (концептов), *U* – множество пользователей. Каждый концепт *c* имеет множество атрибутов:

$$A(c) = \{a(c)_i\}, a(c)_i \in C, i = \overline{1, N_c} \quad (8)$$

Упорядочив множество атрибутов по убыванию степени их значимости для пользователя *u*, получим последовательность, характеризующую его представление о данном концепте:

$$A^u(c) = \{a^u(c)_i\}, i = \overline{1, N_c} : a^u(c)_i \varphi^u a^u(c)_j, \forall i \le j, \quad (9)$$

где φ^u - отношение, задающее значимость атрибутов для пользователя *u*; $a\varphi^u b$ означает, что «для пользователя *u a* не менее значим, чем *b*».

Определим группу пользователей, имеющих схожие представления о понятиях из некоторого множества *C*. Назовем подобную группу пользовательской *категорией k*-го порядка на множестве концептов *C*, и определим ее следующим образом:

$$U_C^k = \{u \mid \{a^u(c)_i\} = \{a^{u'}(c)_i\}, i = \overline{1, k}, \forall c \in C, \forall u' \in U_C^k\} \quad (10)$$

Модель предпочтений некоторой *k*-й группы пользователей представлена взвешенным мультиграфом:

$$UM_k = \{C, L_k\}, \quad (11)$$

где *C* – множество вершин графа, представляющих понятия СМПО, $L_k = \{l_k{}^{ijm}\}$ – множество взвешенных дуг, вес которых характеризует значимость семантического отношения *m*-го типа между *i*-м и *j*-м понятиями для *k*-й категории пользователей.

В совокупности, МПП всех пользователей образуют фрагмент семантической модели предметной области мультипредметной ИС, представляющий собой мультиграф с векторными весами дуг. Матрица инцидентности мультиграфа имеет размерность 3:

$$M_I : C \times L \times U_k \rightarrow w_k^{ij}, \qquad (12)$$

где C – множество вершин графа, представляющих понятия СМПО, L - множество дуг, задающих отношения над C, U_k – множество категорий пользователей. Элементами матрицы являются весовые коэффициенты w_k^{ij}, задающие вес связи между концептами c_i и c_j для k-й категории пользователей.

2.5. Архитектура мультипредметной информационной системы

Архитектура мультипредметной информационной системы представлена на рисунке 11.

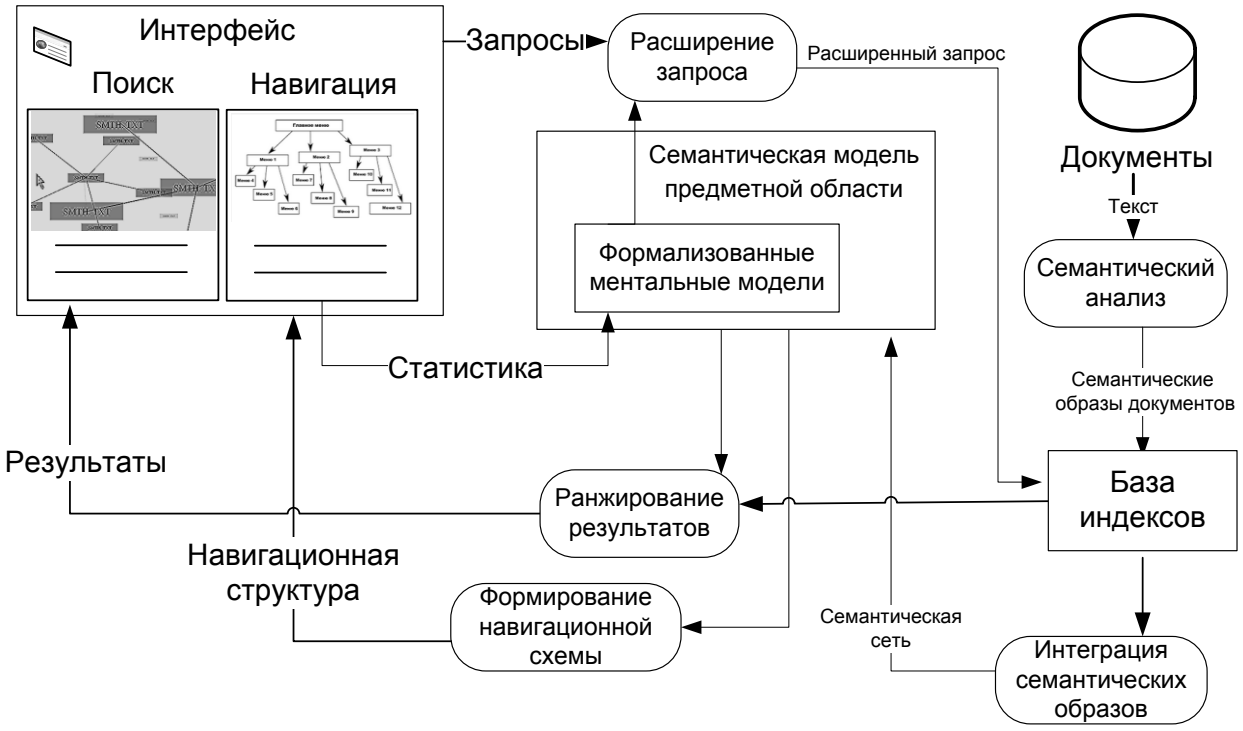


Рисунок 11. Архитектура мультипредметной информационной системы

Основу МИС составляют семантическая модель предметной области (СМПО), формализованная ментальная модель пользователей (ФММ), база индексов документов НСИ. Семантическая модель предметной области - это особого рода база данных, разработанная для оперирования знаниями (метаданными). Формирование СМПО – процесс ее изменения и последующего

уточнения пользователями. Изменение СМПО подразумевает получение новых знаний, например, в результате интеграции семантических образов документов ИС, или экспертных знаний пользователей, полученных в результате взаимодействия с СМПО. Достоинства и недостатки моделей представления знаний представлены в главе 1. В данной работе используется модель представления знаний в виде семантической сети. Семантические сети – нагруженные ориентированные мультиграфы, которые удобны для представления концепций и их взаимоотношений и являются наиболее общей формой представления знаний, что обуславливает их использование для описания мультипредметной информации организаций.

Входными данными, обеспечивающими функционирование МИСП, являются документы предприятия и статистика работы пользователя. Выходными данными являются результаты обработки запросов и навигационная структура интерфейса. СМПО формируется в результате интеграции семантических образов, вновь появляющихся в информационных хранилищах документов, и последующего уточнения модели на основе формализованных ментальных моделей пользователей. Особенностью мультипредметной информационной системы является адаптация информации для различных категорий пользователей. Мультипредметные информационные системы направлены на комплексное удовлетворение разнородных ограниченных сообществ пользователей, определяемых возрастом, профессиональными или досуговыми интересами, и т.п. В работе [18] такие сообщества определены как категории пользователей.

Развитие современных концепций и методов человеко-машинного взаимодействия, например концепции «пользователь как эксперт»[64], позволяет ставить не только задачи передачи знаний о предметной области от пользователей, но и проводить уточнение автоматически сформированных формализованных знаний, что является весьма актуальной задачей в связи с ростом объема информации хранимой современными информационными

системами организаций, а также необходимостью автоматизации процессов получения и обработки данной информации.

В следующей главе будут представлены методы формирования и функционирования мультипредметных информационных систем организаций, применение которых позволяет повысить эффективность механизмов информационного поиска нормативно-справочной информации и эффективность человеко-машинного взаимодействия:

Метод автоматизированного динамического формирования семантической модели предметной области мультипредметных информационных систем, использующий опыт пользователей для уточнения автоматически сформированных знаний. Метод основан на интеграции существующих формализованных знаний, результатов семантического анализа новых документов и моделей предпочтений пользователей.

Метод поиска, обеспечивающий автоматизированное расширение запроса и оценку релевантности результатов поиска на основе совместного анализа модели предпочтений пользователя и семантической модели предметной области с учетом субтрактивных отношений.

Метод интерфейсной навигации для формирования пользовательских интерфейсов мультипредметной информационной системы, адаптированных для различных категорий пользователей. Повышение эффективности человеко-машинного взаимодействия обеспечивается за счет отображения модели предпочтений пользователей на автоматически формируемую навигационную структуру интерфейса.

Выводы по главе 2

Повышение эффективности информационных систем промышленных предприятий представляется возможным за счет оперирования формальными знаниями и реализацией на основе формализованных знаний компонентов информационной системы организаций, использование которых позволит сократить время доступа пользователей к требуемой информации.

Представлена модель динамической мультипредметной информационной системы организации, центральным звеном которой является сформированная в автоматизированном режиме СМПО организации. СМПО является хранилищем мета-информации и позволяет организовать сервисы эффективного доступа (поиска и навигации) пользователей информационной системы к требуемой информации.

Автоматизированное формирование семантической структуры, описывающей понятия и процессы промышленного предприятия становится возможным благодаря использованию инструментов семантического анализа и использованию опыта пользователей, выступающих в роли экспертов, что позволяет не привлекать экспертов по знаниям для формирования формализованных знаний.

Мультипредметная информационная система организации должна обладать возможностью автоматической адаптации содержания информационной системы к различным категориям пользователей, являющимися представителями различных структурных подразделений организации. С целью повышения эффективности механизмов информационного поиска нормативно-справочной информации и эффективность человеко-машинного взаимодействия необходимо разработать методы формирования и функционирования, основанных на знаниях мультипредметных информационных систем организаций.

ГЛАВА 3. Методы формирования и функционирования мультипредметных информационных систем

3.1. Метод автоматизированного формирования семантической модели предметной области информационной системы на основе принципа «пользователь как эксперт»

В данном разделе представлен метод автоматизированного формирования СМПО организаций, отличительными особенностями которого являются:

1. Уровень автоматизации, выраженный в отсутствии необходимости в привлечении экспертов в процессе формировании СМПО, что позволяет применять метод для динамических предметных областей.
2. Использование знаний пользователя для уточнения автоматически сформированных формализованных знаний, интегрируемых в СМПО. В классическом понимании интеграция формализованных знаний подразумевает объединение нескольких экспертно созданных онтологий [45], при этом не ставиться вопрос достоверности исходных онтологий. Данный метод позволяет рассматривать пользовательские знания как инструмент для проведения уточнения, что делает возможным автоматизированную интеграцию семантических образов документов информационной системы организации.
3. Использование СМПО как основы для организации доступа к данным информационной системы, а именно осуществления поиска и навигации.

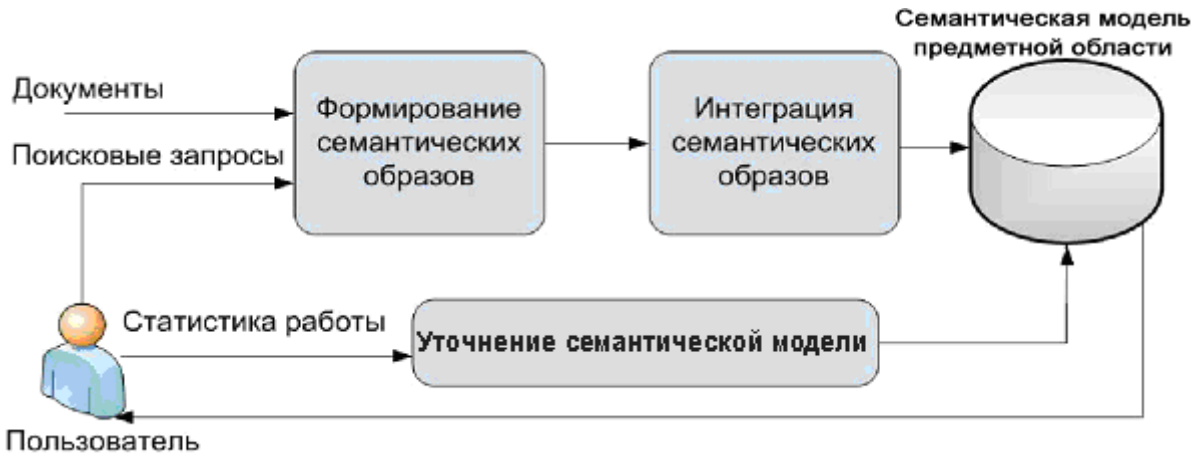


Рисунок 12- Метод динамического формирования семантической модели предметной области на основе принципа «пользователь как эксперт»

Источниками знаний для формирования СМПО являются тезаурус русского языка, семантические образы документов организации. Полученные в результате семантического анализа образы документов подлежат интеграции с существующими знаниями СМПО. Интеграция формализованных знаний основывается на модифицированном методе интеграции на основе составной семантической метрики, предложенной в работе [45]. Задачи поддержки актуальности и уточнения знаний решаются путем расширенного взаимодействия с пользователем в рамках его работы с пользовательским интерфейсом. Пользователи различных категорий выступают в роли источника знаний, а процесс уточнения в данном случае представляется как соотнесение автоматически сформированной модели предметной области с представлением о предметной области пользователя.

3.2.1. Формальное описание семантической модели предметной области

Формально СМПО представлена неоднородной n-арной семантической сетью:

$$KB = \{C, L, Tp\}, \tag{13}$$

$$Tp = \{synonymOf, HyponymOf, associateWith, subStract\},$$

где C - множество концептов, L - множество отношений над концептами. Tp –

множество типов отношений. Основой СМПО выступает тезаурус русского языка, расширяемый семантическими образами документов, составляющих контент информационной системы.

3.2.2. Интеграция семантических образов документов организации

Семантический образ документа – семантическая сеть, множество вершин которой составляют концепты СМПО, присутствующие в документе, множество ребер – множество двухместных отношений над концепами. Информационный элемент – документ информационной системы предприятия. Получение семантических образов документов происходит путем анализа документов ИС.

Процесс получения семантического образа включает насколько этапов обработки текста документа, являющихся уровнями лингвистического анализа: графематический, морфологический, синтаксический, семантический. Результаты работы каждого уровня используются следующим уровнем анализа в качестве входных данных. Целью графематического анализа является выделения элементов структуры текста: параграфов, абзацев, предложений, отдельных слов и т. д. Целью морфологического анализа является определение морфологических характеристик слова и его основной словоформы. Особенности анализа сильно зависят от выбранного естественного языка. Целью синтаксического анализа является определение синтаксической зависимости слов в предложении. Следует отметить, что в связи с присутствием в русском языке большого количества синтаксически омонимичных конструкций, наличием тесной связи между семантикой и синтаксисом, процедура автоматизированного синтаксического анализа текста является трудоемкой. Сложность алгоритма увеличивается экспоненциально при увеличении количества слов в предложении и числа используемых правил.

Для определения типа отношений используется тезаурус предметной области. В работе был использовании русскоязычный тезаурус WordNet 3.0.

[80] Для определения нормальной формы слов был использован грамматический словарь русского языка А.А. Зализняка, содержащий приблизительно 100 тыс. базовых словоформ русского языка с их полным морфологическим описанием.

Далее представлен алгоритм анализа текста документа.

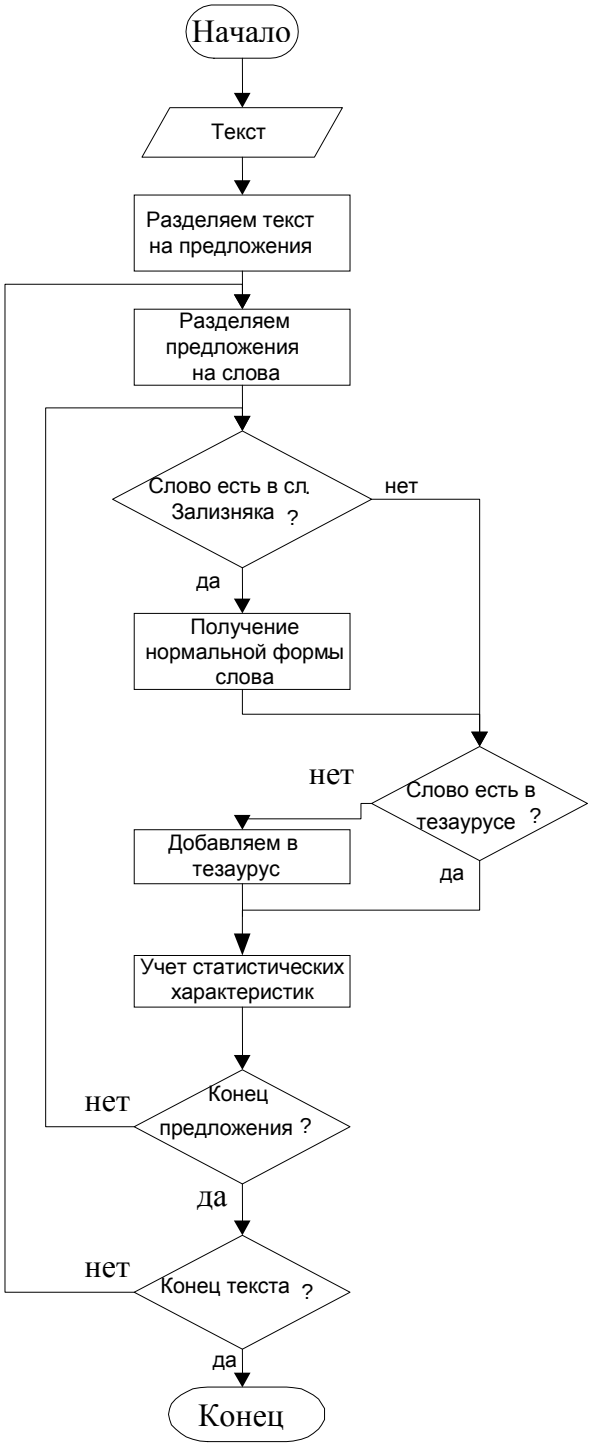


Рисунок 13 - Алгоритм анализа текста документа

Результатом анализа документа является семантическая сеть, представляющая множество слов документа и отношений между ними.

Интеграция семантических образов осуществляется путем пополнения СМПО недостающими фрагментами. И состоит из нескольких этапов - поиск, оценка сходства и добавление новых знаний (рисунок 14).

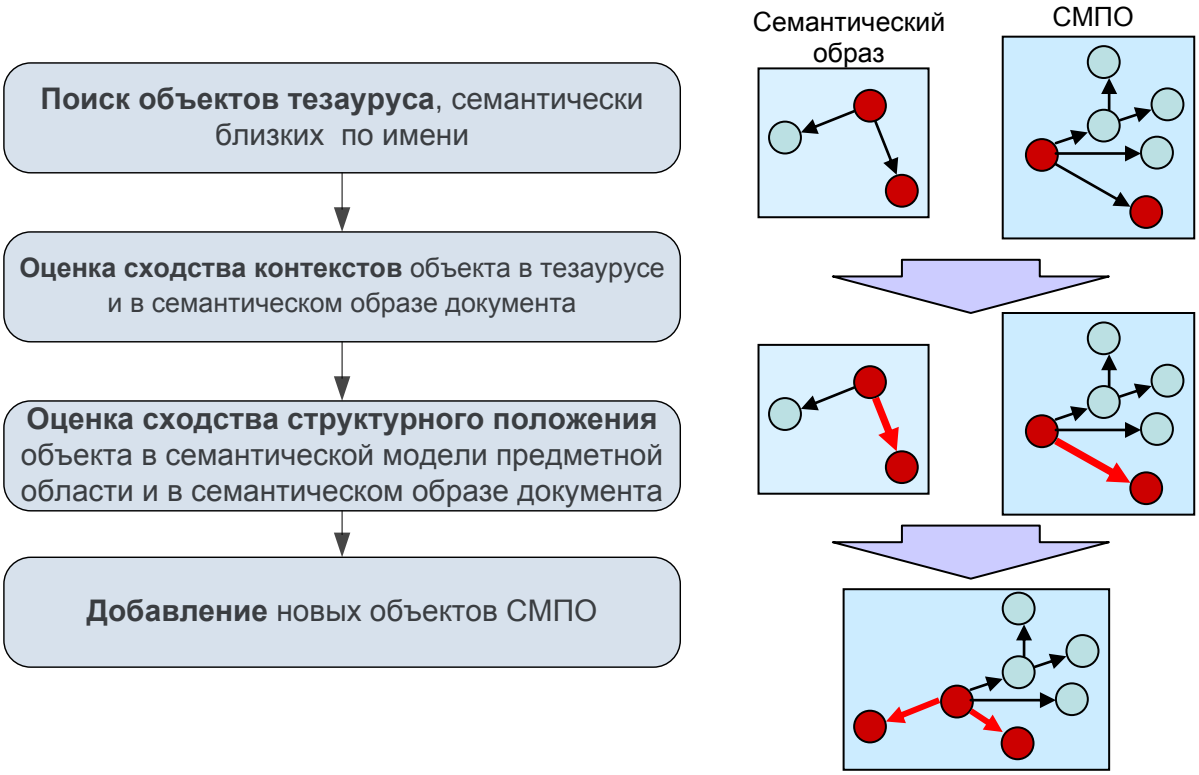


Рисунок 14 - Этапы интеграции результатов семантического анализа документов в СМПО

Если отсутствуют концепты, совпадающие по имени, то объектом оценки со стороны документа выступает контекст понятия. Контекстом понятия являются все связанные с ним понятия. Таким образом, механизм добавления основан на выделении схожего контекста, а интеграция происходит путем расширения существующей СМПО – ее дополнения новыми понятиями и отношениями.

Процесс формирования семантической модели предметной области предприятия на основе коллекции документов информационной системы и расширяемого тезауруса состоит из следующих этапов:

1. Формирование семантического образа документа. Семантический образ

задан семантической сетью, полученной статистическими и лексико-грамматическими методами обработки текста:

$$D = \{C^D, L^D\}, C^D \subset C, L^D \subset L, \quad (14)$$

где C^D - множество концептов, выделенных в документе, L^D - множество отношений вида (2), выделенных в документе.

2. Интеграция семантических образов в СМПО на основе модифицированной составной семантической метрики [45] следующим образом:

   I. Вычисление оценки сходства имен концептов документа и СМПО:

$$\forall c_i : Eq(c_i, c_j) = \frac{x / len}{\max(len_i, len_j)}, \; c_i \in C^D, c_j \in C^{KB}, \quad (15)$$

$$i = \overline{1, N_D}, \; j = \overline{1, N_{KB}}$$

   где *Eq()* – функция оценки сходства имен двух концептов, *x* – длина эквивалентной цепочки символов в именах понятий, *len* – длина имен концептов.

   II. Вычисление оценки сходства контекста концептов документа с контекстом СМПО:

$$\forall c_i : Syn(c_i, c_j) = \frac{\left| C_{syn}^D \cap C_{syn}^{KB} \right|}{\left| C_{syn}^{KB} \right|}, c_i \in C^D, c_j \in C^{KB}, \quad (16)$$

$$i = \overline{1, N_D}, \; j = \overline{1, N_{KB}}$$

   где *Syn()* – функция оценки сходства контекста двух концептов, C_{syn}^D, C_{syn}^{KB} - множество синонимов концепта c_i и c_j соответственно.

   III. Вычисление оценки сходства структурного положения концептов документа с контекстом СМПО:

$$\forall c_i : Poseq(c_i, c_j) = \frac{\left|C_{Hyp}^{D} \cap C_{Hyp}^{KB}\right|}{\left|C_{Hyp}^{KB}\right|}, c_i \in C^D, c_j \in C^{KB}, \quad (17)$$

$$i = \overline{1, N_D}, j = \overline{1, N_{KB}}$$

где *Poseq()* – функция оценки сходства структурного положения двух концептов, $C_{Hyp}^{D}, C_{Hyp}^{KB}$ - множество гиперонимов концепта c_i и c_j соответственно.

IV. Добавление концептов на основании результатов вычисления пороговой функции от среднего трех оценок:

$$f(c_i, c_j) = \frac{a \cdot Eq(c_i, c_j) + b \cdot Poseq(c_i, c_j) + c \cdot Syn(c_i, c_j)}{3} > z, \quad (18)$$

$$c_i \in C^D, c_j \in C^{KB}, i = \overline{1, N_D}, j = \overline{1, N_{KB}},$$

где *z* – значение пороговой функции, *a,b,c* – некоторые коэффициенты, которые определяются экспертно, исходя из объема и разнородности предметной области.

3. Уточнение СМПО пользователями осуществляется путем изменения весовых коэффициентов существующих отношений между понятиями. Данный процесс инициируется при совместном использовании двух понятий в одном пользовательском запросе. Величина изменения весового коэффициента определяется следующим образом:

$$dw(c_i, c_j) = \frac{\sum_{<c_i, c_j> \in (ACW)} (w(c_i, c_j))}{|ACW|} \quad (19)$$

- для весового коэффициента отношения между совместно использованными понятиями, и

$$dw(c_i, c_j) = -\frac{\sum_{c_i \in ACW, c_j \in CW} (w(c_i, c_j))}{|CW / ACW|} \quad (20)$$

- для весового коэффициента отношения между остальными (неиспользованными) понятиями, где $dw(c_i, c_j)$ - величина изменения весового коэффициента, $w(c_i, c_j)$ - значение весового коэффициента отношения между понятиями c_i и c_j, ACW и CW – множества отношений с совместно использованными и неиспользованными понятиями.

Предложенный метод предлагается использовать для автоматизированного формирования и обеспечения актуального состояния семантической модели предметной области производственного предприятия в условиях динамических изменений предметной области без привлечения эксперта по знаниям. На основе полученной модели решаются задачи обеспечения эффективного доступа пользователей к НСИ предприятия, а именно реализуются методы поиска и интерфейсной навигации.

3.3. Метод формирования модели предпочтений пользователя мультипредметной информационной системы

Одним из возможных путей адаптации ИС к пользователю является построение модели пользовательских предпочтений. В работе [88] предложено использовать модель пользователя, полученную на основе опроса пользователя для оптимизации информационного поиска мультимедиа-файлов. В работе [106] предложено использовать ассоциативную лексическая сеть отношений между словами для моделирования когнитивных процессов пользователя поисковой системы и оптимизации запроса. В [23] предложен способ автоматического получения предпочтений пользователя в виде ментальной модели на основе учета статистики взаимодействия пользователя с информационной системой.

Применительно к задачам навигации и поиска модель предпочтений пользователя позволит автоматизировано уточнить контекст запроса и ограничить область поиска за счет использования субтрактивных отношений. модель предпочтений пользователя (МПП) представляет собой ассоциативную семантическую сеть, множество вершин которой составляют понятия предметной области, которыми оперирует пользователь, множество ребер – множество взвешенных двухместных отношений над понятиями. МПП формируется автоматическом режиме на основе обработки запросов пользователя и статистики его работы с информационной системой. Взаимодействие пользователя с информационной системой может быть представлено следующим алгоритмом:

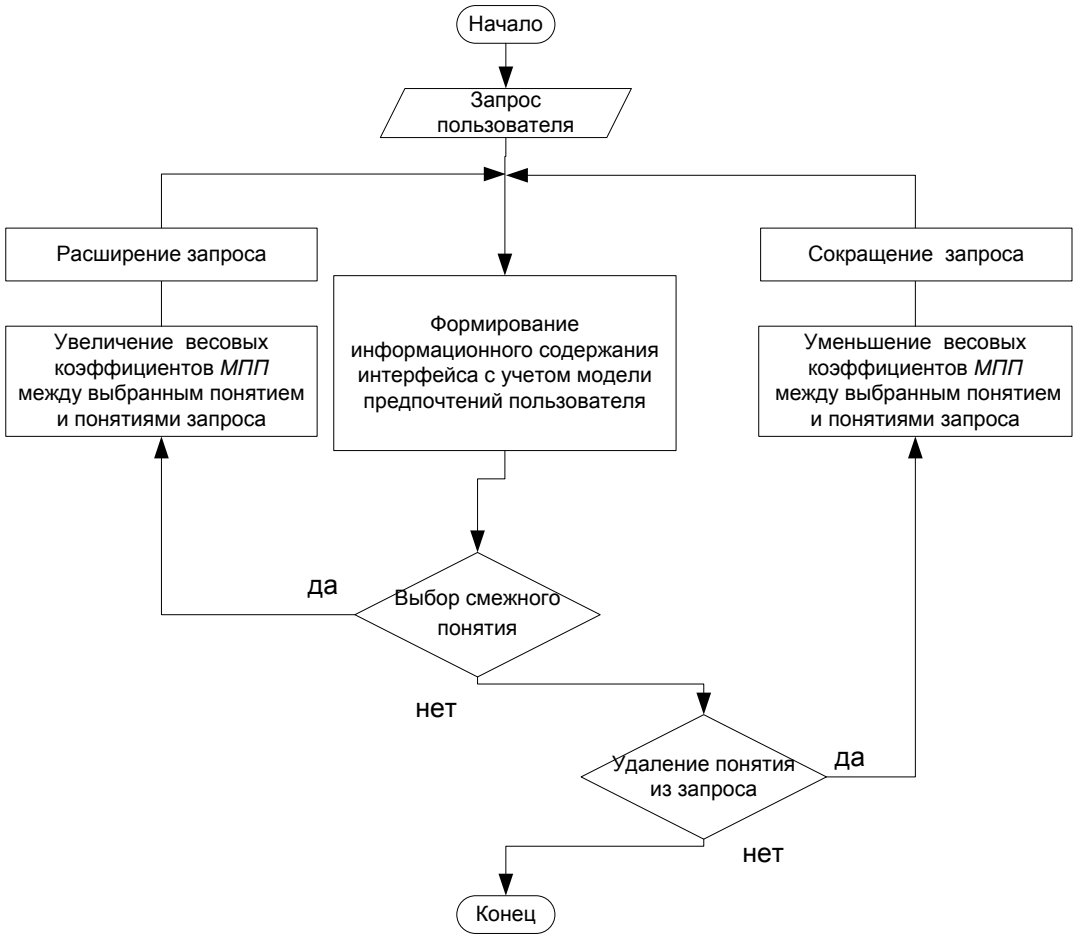


Рисунок 15 - Алгоритм взаимодействия пользователя и ИС

Взаимодействие заключается в итеративном расширении запросов пользователя на основе МПП, обеспечении возможности коррекции запроса, а также учета пользовательских предпочтений путем коррекции весовых коэффициентов между понятиями МПП. При совместном употреблении в запросе пар понятий, входящих в состав МПП, увеличиваются весовые коэффициенты отношений между данными понятиями. Превалирование одного понятия из модели над другим задается весовыми коэффициентами дуг формализованной ментальной модели. Особенностью такого взаимодействия является возможность задания отрицательных весовых коэффициентов (субтрактивных отношений), обозначающих отсутствие значимости данного контекста понятия для пользователя.

Модель предпочтений некоторой *k*-й группы пользователей представлена в разделе 2.4 взвешенным мультиграфом:

$$UM_k = \{C, L_k\}, \tag{21}$$

где C – множество вершин графа, представляющих понятия СМПО, $L_k = \{l_k{}^{ijm}\}$ – множество взвешенных дуг, вес которых характеризует значимость семантического отношения *m*-го типа между *i*-м и *j*-м понятиями для *k*-й категории пользователей.

В совокупности, МПП всех пользователей образуют фрагмент семантической модели предметной области мультипредметной ИС, представляющий собой мультиграф с векторными весами дуг. Матрица инцидентности мультиграфа имеет размерность 3:

$$M_I : C \times L \times U_k \rightarrow w_k^{ij}, \tag{22}$$

где C – множество вершин графа, представляющих понятия СМПО, L - множество дуг, задающих отношения над C, U_k – множество категорий

пользователей. Элементами матрицы являются весовые коэффициенты w_k^{ij}, задающие вес связи между концептами c_i и c_j для k-й категории пользователей. Более подробно взаимодействие пользователя с информационной системой будет рассмотрено далее.

3.4. Метод интерфейсной навигации в мультипредметных информационных системах

С ростом масштабов информационных систем, в смысле объемов хранимой информации, все более актуальной становится задача оптимальной структуризации информационных элементов, имеющихся в системе. Любая информационная система (ИС) содержит множество информационных элементов (документов, html-страниц), адресованных конечному пользователю. Современные информационные системы отличаются большим количеством информационных элементов. Вместе с тем данное множество не статично – практически любая ИС в настоящее время обладает динамикой своего содержания, что приводит к затруднениям пользователя в плане доступа к требуемой информации. Целью навигационного интерфейса является структуризация и обесепение доступа пользователю к информационным элементам. От вида навигационной структуры (логики организации, упорядоченности меню, количества разделов и их иерархии) зависит количество вариантов перебора, необходимых для доступа к искомому информационному элементу.

В данном разделе представлен метод интерфейсной навигации, обеспечивающий автоматизированное формирование адекватных различным категориям пользователей интерфейсов и генерацию сложных поисковых запросов в мультипредметных информационных системах. Общая схема метода представлена на рисунке 16.

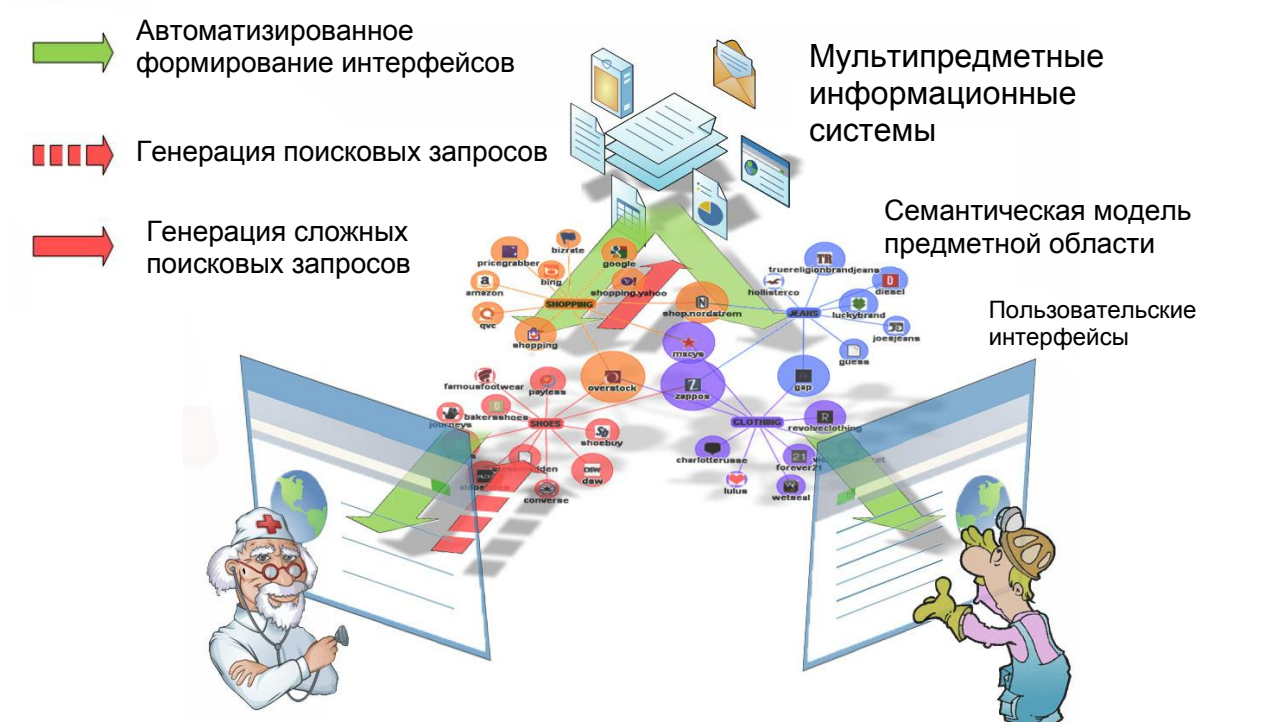


Рисунок 16 - Метод интерфейсной навигации

Навигационная структура информационного ресурса представляет собой иерархическую систему, которая должна решать следующие задачи:

1. Логически объединять отдельные информационные блоки,
2. служить оглавлением разделов информационного ресурса и давать представление об информации, которая находится на нем,
3. отображать элементы управления, с помощью которых пользователь переходит в раздел для получения заинтересовавшей его информации.

Информационность навигационной структуры заключается в том, чтобы кратко показать посетителю информационный потенциал ресурса.

Далее представлены модель навигационного интерфейса, ограничения на его структуру и процесс формирования навигационной структуры, адекватной модели предпочтений пользователей.

3.4.1. Модель навигационного интерфейса

Назначением пользовательского интерфейса ресурса является обеспечение доступа пользователя к информационным элементам. Такой доступ в ИС реализуется двумя путями:

С помощью поискового механизма, позволяющего выбрать информационные элементы, удовлетворяющие заданным пользователем критериям. Будем называть интерфейс такого типа «*поисковым*».

С помощью некоторой навигационной структуры, реализующей функцию каталога. Назовем такой тип интерфейса «*навигационным*».

Интерфейс навигационного типа имеет две основных составляющих – это внешнее оформление (дизайн) и навигационная структура. Если качество первого компонента является исключительно субъективной категорией и вряд ли может быть оценено формально, то для оценки качества навигационной структуры ресурса можно использовать степень ее соответствия модели предпочтений пользователя. При этом должна рассматриваться семантика навигационной структуры (семантическая структура интерфейса). Чем более схожими являются семантическая структура интерфейса и модель предпочтений пользователя, тем более удобным и понятным для конечного пользователя будет интерфейс. Последний в этом случае будет способен «предугадывать» образ мыслей пользователя и визуализировать фрагмент понятийной системы ожидаемым для него способом. Будем далее именовать меру соответствия семантической структуры интерфейса ментальной модели пользователя *когнитивностью* интерфейса [23]. Далее рассматривается формализованное описание навигационной структуры и основанная на нем, требующая оптимизации, количественная оценка когнитивности пользовательского интерфейса.

Итак, пользовательский интерфейс представляет собой пару:

$$UI = \langle I, s \rangle, \tag{23}$$

где I – множество информационных элементов; s – навигационная структура.

Навигационная структура определяет иерархию групп информационных элементов (ИЭ) или доступных для пользователя действий. При этом на каждом

уровне иерархии исходное множество информационных элементов (будем полагать, что доступное пользователю действие является частным случаем ИЭ) делится на подмножества в соответствии с одним или несколькими классификационными признаками. В качестве классификационных признаков используются атрибуты понятий предметной области. Очевидно, что при использовании на одном уровне навигационной структуры нескольких признаков, полученные множества ИЭ могут пересекаться. Введем следующие обозначения:

$\Gamma^l = \{G_i^l\}$ - множество разделов *l*-го уровня навигационной структуры;

G_i^l - *i*-я группа информационных элементов *l*-го уровня навигационной структуры;

$P^l = \{p_i^l\}$ - множество классификационных признаков, используемых для формирования групп ИЭ на *l*-м уровне навигационной структуры.

Использование информационной системы представляет собой, по сути, поиск некоторых информационных элементов по имеющемуся у человека образу. При этом образ, чаще всего, неточный: в нем специфицируются лишь некоторая часть идентифицирующих атрибутов. Вследствие этого, пользователь с разной степенью уверенности может предполагать в какой из групп ИЭ на некотором уровне навигационной структуры находится искомый элемент. Эта уверенность тем выше, чем более точно представляет пользователь потенциальное содержимое группы. Введем следующую функцию, задающую числовую оценку степени уверенности пользователя *u* (чем выше значение, тем выше степень уверенности):

$$p^u : \Gamma^l \to [0,1] \tag{24}$$

Оценка времени, требуемого для доступа к искомому информационному элементу в рамках навигационной структуры на *l*-м уровне, будет равна

$$O(\frac{\max_i \left|G_i^l\right|}{p^u(G_i^l)}). \quad (25)$$

Таким образом, при прочих равных, степень уверенности пользователя в принадлежности информационного элемента к той или иной группе определяет качество интерфейса в смысле скорости доступа к требуемой информации.

Сделаем следующее предположение: если для формирования навигационной структуры на некотором уровне иерархии используются идентифицирующие атрибуты, то пользователь с высокой долей уверенности сможет определить, в какой группе находится искомый информационный элемент. Обозначим через $w^u(a) \in [0,1]$ нормированный вес атрибута *a* в модели предпочтений пользователя *u*. Тогда, с учетом указанного предположения:

$$p^u(G_i^l) = \max_{a \in P^l} w^u(a) \quad (26)$$

То есть мы предполагаем, что если на *l*-м уровне используется несколько классификационных признаков для группирования информационных элементов, то пользователь оперирует той частью навигационной структуры, которая определяется наиболее значимым с его точки зрения атрибутом понятия верхнего уровня.

Пусть навигационная структура интерфейса имеет глубину $\hat{l}$ уровней. Тогда в качестве количественной оценки когнитивности интерфейса для пользователя *u* может использоваться сумма:

$$\sum_{l=1}^{\hat{l}} p^u(G_i^l) \quad (27)$$

Данная мера может использоваться для оценки уже существующих интерфейсов, когда известно значение $\hat{l}$. Для решения же прямой задачи, то есть структуризации исходного множества информационных элементов в рамках навигационной структуры, требуется учитывать дополнительные ограничения. Эти ограничения обусловлены психологией восприятия человека, ограничивающей максимальное количество одновременно эффективно воспринимаемых объектов. Вследствие этого необходимо ограничивать размер группы информационных элементов, а также глубину навигационной структуры.

С учетом сказанного, оптимальная для пользователя *u* структура интерфейса есть решение следующей задачи с ограничениями:

$$\max_{s} \sum_{l=1}^{\hat{l}(s)} p^u(G_i^l), g(s) \le K, \hat{l}(s) \le K' \qquad (28)$$

Здесь $\hat{l}(s)$ - количество уровней в навигационной структуре *s*; $g(s)$ - максимальный размер группы информационных элементов $\hat{l}(s)$-го в навигационной структуре *s*; *K* – когнитивная константа, определяющая максимальное число одновременно предъявляемых пользователю информационных элементов для их эффективного восприятия; *K'* – когнитивная константа, определяющая максимальное число уровней навигационной структуры, в рамках которых поиск информации для пользователя остается комфортным.

3.4.2. Ограничения на структуру пользовательского интерфейса

Тенденции развития современных ИТ включают исследования, направленные на изучение сложности и удобства использования (англ. «usability») пользовательских интерфейсов, выдвигаются формальные критерии оценки сложности интерфейсов. В данном разделе выдвигается предположение

о наличии оптимальных значений количества уровней и количества групп информационных элементов навигационной структуры.

Существующие оценки пользовательских интерфейсов можно разделить на две группы – экспериментальные и формальные оценки. Среди методик оценки сложности интерфейсов первой группы следует отметить методы экспертных оценок [111], анкетирования пользователей [130, 112, 87], оценки, базирующиеся на экспериментальных данных [84, 120], например, времени выполнения операции, количество совершенных ошибок. Однако, наибольший интерес представляют формальные методы оценки сложности интерфейса, так как позволяют проводить исследование еще на этапе проектирования, а так же позволяют проводить оценку адаптивных динамических пользовательских интерфейсов[67]. Среди формальных оценок, можно выделить оценку информационного поиска, информационную производительность модели KLM-GOMS, оценку сложности системы по количеству объектов и классов [85], оценку сложности визуального представления интерфейса XAOS — Actions[123], оценку на основе метаданных об интерфейсе (LOC-CC модель измерения сложности)[74].

Следует отметить, что все перечисленные методики, за исключением [74] оценивают фактически представленный интерфейс, и не подразумевают возможности синтеза на их основе пользовательских интерфейсов с заданными качествами. Вместе с тем, целенаправленное формирование навигационной структуры интерфейса, адаптированной к модели предпочтений пользователя, позволяет значительно повысить эффективность работы с информационной системой. В работе[66] проводилось исследование навигационной структуры интерфейсов нескольких популярных веб-ресурсов на предмет выявления превалирующих семантических связей навигационной структуры. Исследование показало, что даже в весьма популярных информационных системах вопросу соответствия структуры навигации ожиданиям пользователя не уделяется должного внимания.

В предыдущем разделе мы показали, что оптимальная навигационная структура должна быть такова, чтобы пользователь мог с наивысшей уверенностью делать предположения о содержимом того или иного раздела (группы) информационных элементов. Это позволяет сократить предельное количество вариантов перебора в процессе поиска интересующего пользователя информационного элемента. Вместе с тем, это количество существенным образом зависит и от независимых от семантики характеристик навигационной структуры – количества разделов, их размеров, количества уровней в структуре.

Далее рассмотрим сложность интерфейса с точки зрения количества вариантов перебора для поиска пользователем требуемого информационного элемента, а именно зависимость количества вариантов перебора от количества групп информационных элементов на одном уровне и глубины навигационной структуры. Обозначим N – количество информационных элементов, k – количество равных по мощности групп n_i^l на одном уровне навигационной структуры, l – количество уровней навигационной структуры,

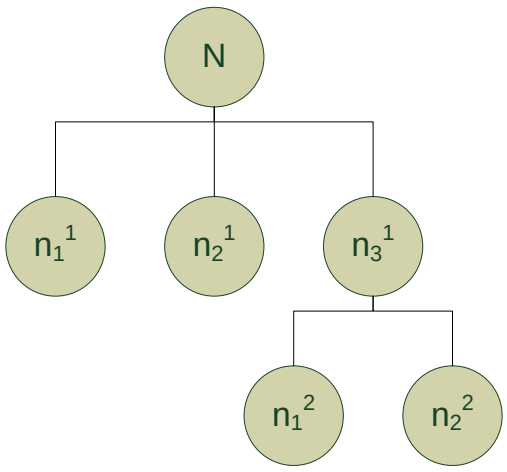


Рисунок 17 - Навигационная структура

В случае, когда классифицирующие признаки однозначны и позволяют произвести разбиение информационных элементов на равные по мощности

множества (группы), зависимость количества вариантов перебора от количества информационных элементов и можно выразить формулой:

$$f(N,l,k) = \frac{N}{k^l} + (k-1)l \quad , \tag{29}$$

где N - количество информационных элементов, k - количество групп информационных элементов на одном уровне, l - количество уровней иерархии навигационной структуры.

Логично предположить, что разбиение на равные группы информационных элементов по некому классификационному признаку, позволяет сократить количество вариантов перебора, однако при большом значении k данная зависимость малозначительна (рисунок 18).

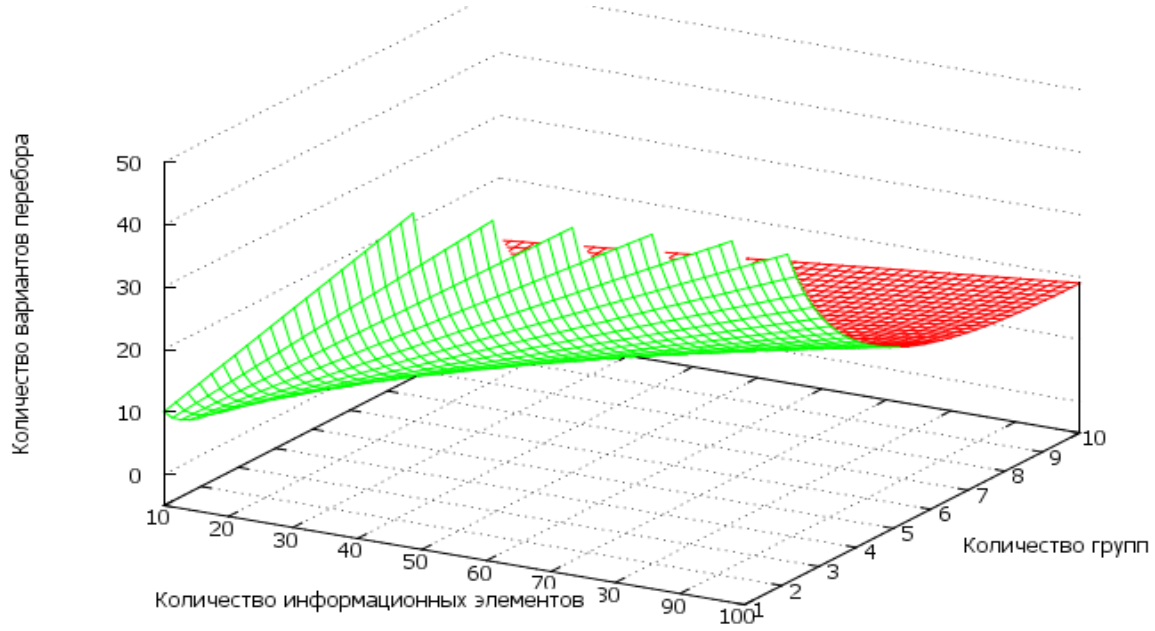
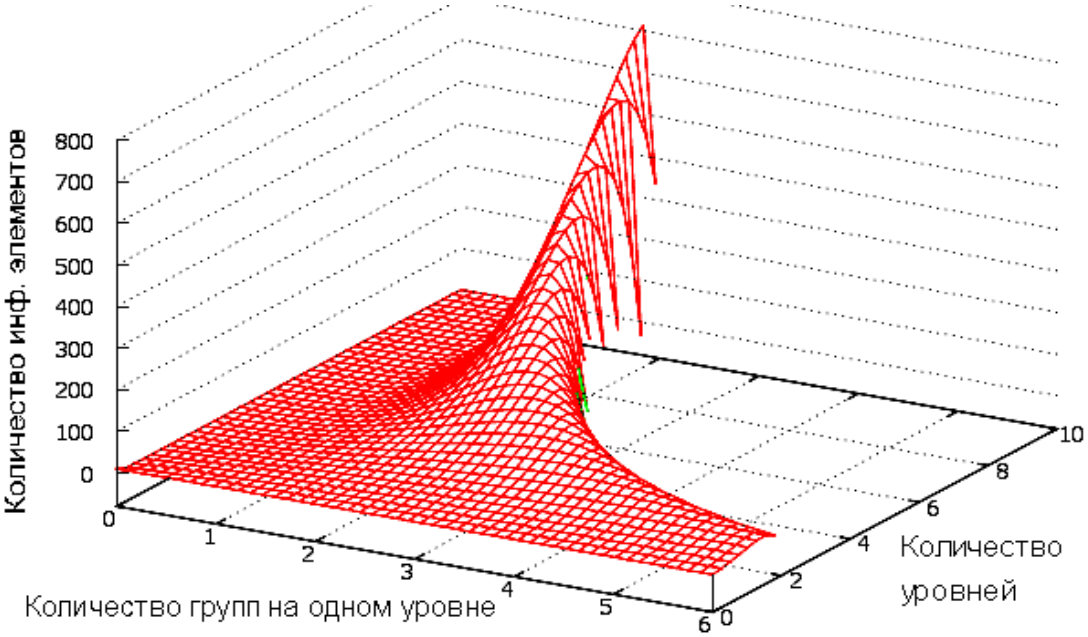


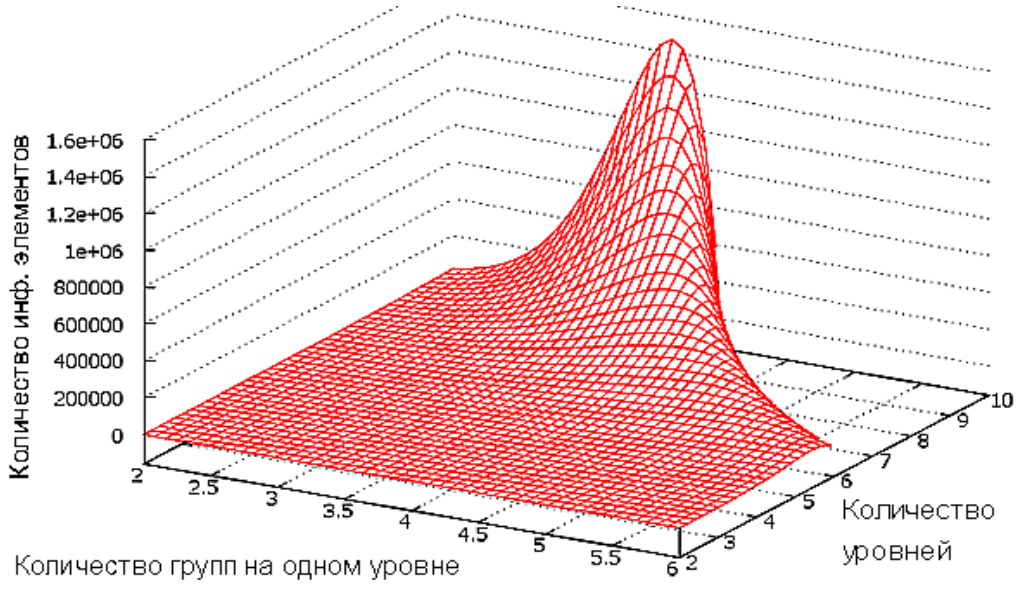
Рисунок 18 - Зависимость количества вариантов перебора от количества информационных элементов

Ограничим количество вариантов перебора, тогда $N = (F - kl + l)k^l$, где F – количество вариантов перебора. Тогда при различных трудозатратах

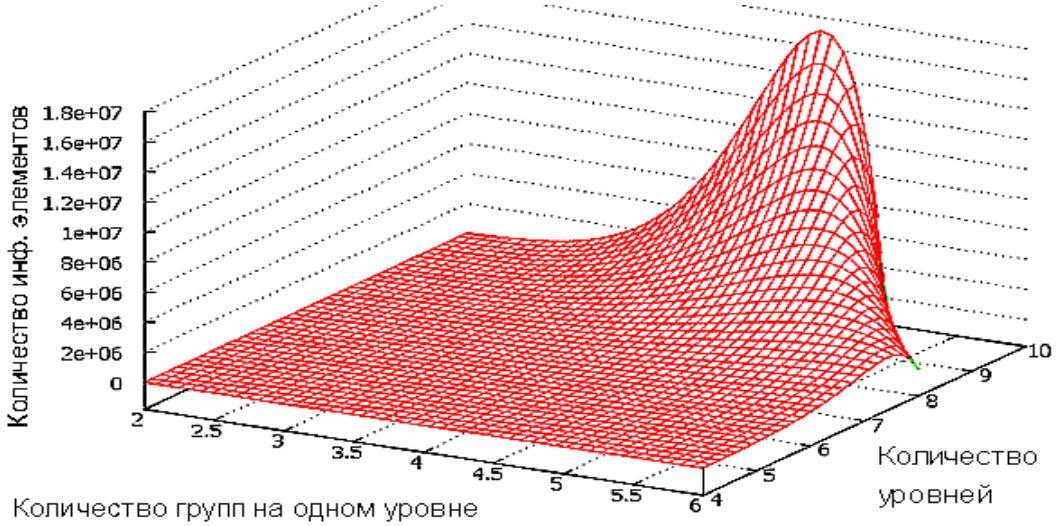
пользователя, выраженных *F*, в навигационную структуру можно «уместить» N информационных элементов:



а)



б)



в)

Рисунок 19 - Зависимость количества информационных элементов от количества уровней иерархии навигационной структуры при фиксированном значении количества вариантов перебора F=10(а), F=30 (б), F=40(в)

Таким образом, можно предположить, что при наличии навигационной структуры с параметрами *k,l*, пользователь, осуществив F действий, способен осуществить навигацию среди N информационных элементах, связанных навигационной структурой.

Если известно количество информационных элементов, то зависимость вариантов перебора от формы навигационной структуры примет следующий вид:

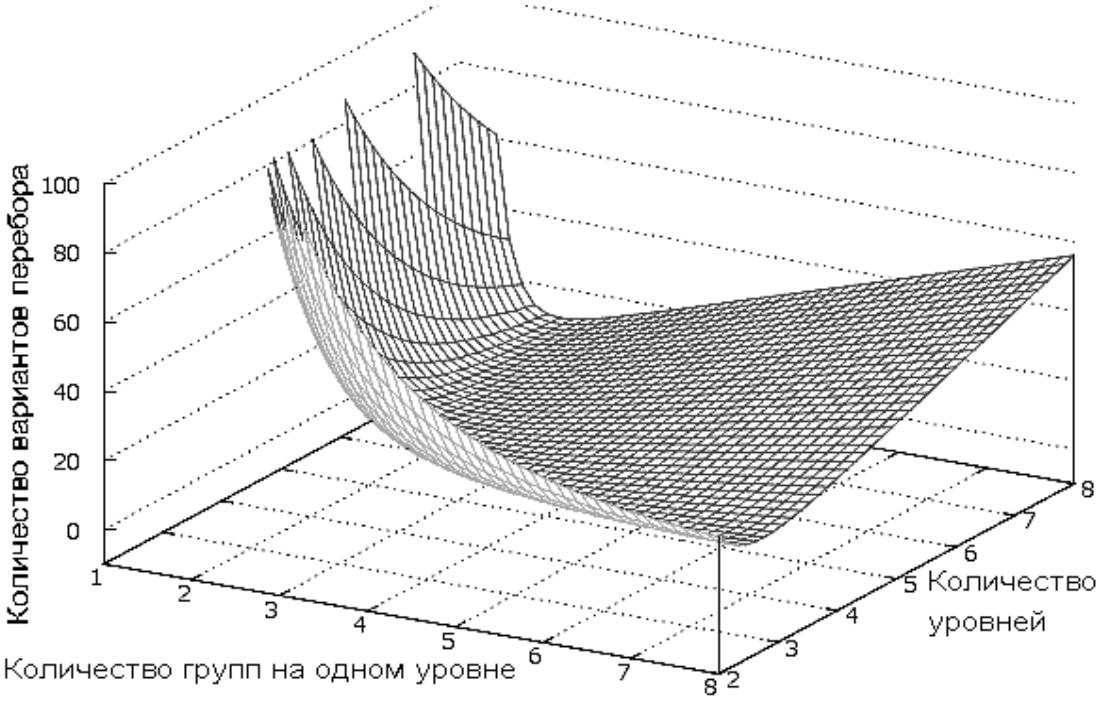


Рисунок 20 - Зависимость количества вариантов перебора от количества уровней иерархии и количества групп на одном уровне навигационной структуры при фиксированном N=1000

Таким образом, в навигационной структуре, содержащей N информационных элементов, существуют близкие к оптимальным значения k и l, позволяющие получить доступ к требуемому информационному элементу за минимальное количество действий пользователя.

В данном разделе рассмотрена зависимость количества действия пользователя от формы навигационной структуры в задаче доступа к требуемому информационному элементу. Результаты позволяют сделать вывод о существовании оптимальных значений количества групп информационных элементов на одном уровне, и количества уровней иерархии навигационной структуры.

3.4.3. Метод формирования навигационной структуры, адекватной модели предпочтений пользователей

Для реализации метода интерфейсной навигации, удовлетворяющего приведенной ранее формулировке задачи синтеза пользовательского интерфейса, предлагается соответствующая процедура, основанная на модели предпочтений пользователя. Процедура содержит несколько этапов:

1. Определение текущей информационной потребности пользователя на основе модели пользовательских интересов и текущего запроса:

$$UQ = f_m(Q, UM_k),\ \forall c_i : \exists l : (c_i \in Q) \wedge (c_j \in UM_k) \wedge (\overline{w_k} > 0),$$
$$Q = \{c_i\}, l = < c_i, c_j, tp, \overline{w} >, i = \overline{1, N_Q},\ j = \overline{1, N_L}, \quad (30)$$

где UM_k – модель предпочтений *k-ой* категории пользователей; Q - запрос, f_m - функция, ставящая соответствие запросу фрагмент модели предпочтений, $\overline{w_k}$ - *к-ый* компонент вектора весовых коэффициентов, l – отношение между концептами c_i и c_j в модели UM_k.

2. Определение множества информационных элементов интерфейса, соответствующих текущей информационной потребности:

$$G = \{g \mid \exists c_i = g\}, c_i \in UQ, \quad (31)$$

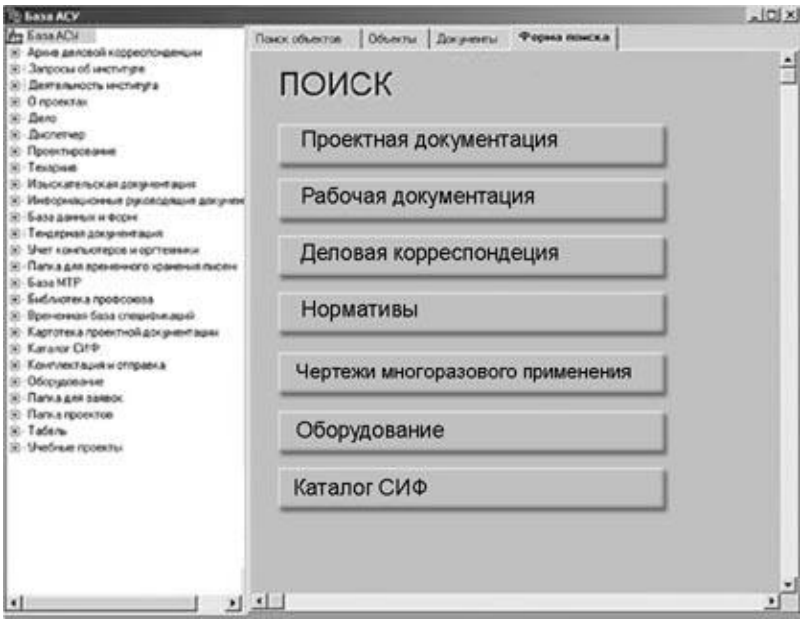
где G - множество информационных элементов навигационной структуры; c_i - концепты текущей информационной потребности пользователя.

3. Разбиение множества информационных элементов интерфейса на подмножества и их ранжирование в соответствии с весовыми коэффициентами модели пользовательских интересов.

$$G_i^d = \{g_k \mid (\forall g_k, g_m \exists l \in L : w(c_k, c_m) > x) \wedge (\exists g_z : \exists l_{Hyp}(c_k, c_z), l_{Hyp}(c_m, c_z) \in L)\},$$
$$k, m, z = \overline{1, N_G}, z = \overline{1, N_L}, G_i^d \subset G, L \in UM, \quad (32)$$

где G_i^d - i-я группа информационных элементов d-го уровня навигационной структуры; $l_{Hyp}(c_k, c_z)$ - отношение гипонимии в модели интересов пользователя; $w(c_k, c_m)$ - весовой коэффициент отношения l над концептами c_k, c_m; x – порог вхождения информационного элемента в навигационную структуру; d – количество уровней навигационной структуры, задается на основе ограничений на максимальное число информационных элементов.

На основе групп множеств G формируется интерфейс мультипредметной информационной системы, отражающих информационное содержание ресурсов в терминах СМПО с учетом модели предпочтений пользователей. На рисунке 5 представлен адаптивный пользовательский интерфейс, динамически изменяющий навигационную структуру в зависимости от модели предпочтений пользователя.

а) 

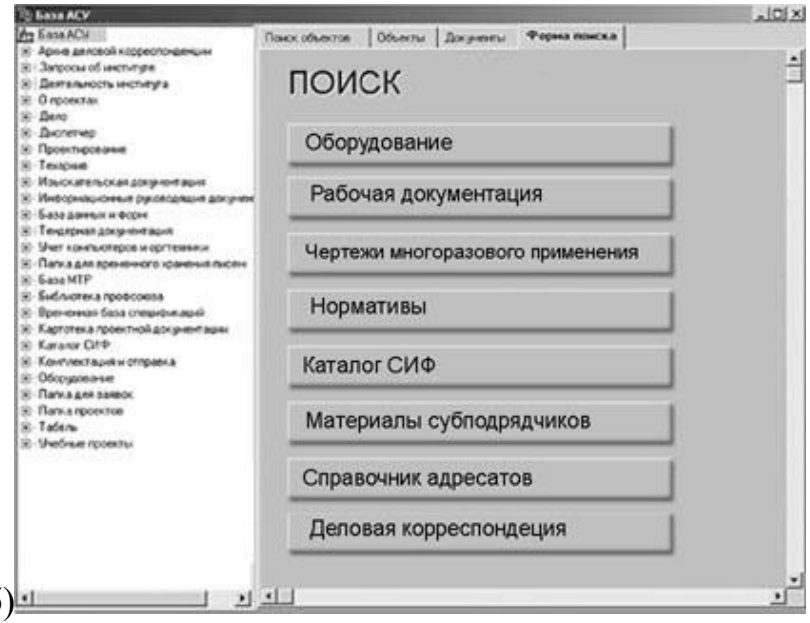
б) 

Рисунок 21 - Вид пользовательского интерфейса для категорий пользователей отделов «Управление производством» (а) и «Техническое планирование» (б)

Алгоритм функционирования адаптивного пользовательского интерфейса подразумевает итеративную коррекцию информационной потребности пользователя и визуализацию соответствующих данной информационной потребности информационных элементов. В свою очередь каждое действие пользователя инициирует изменение весовых коэффициентов отношений модели предпочтений пользователей.

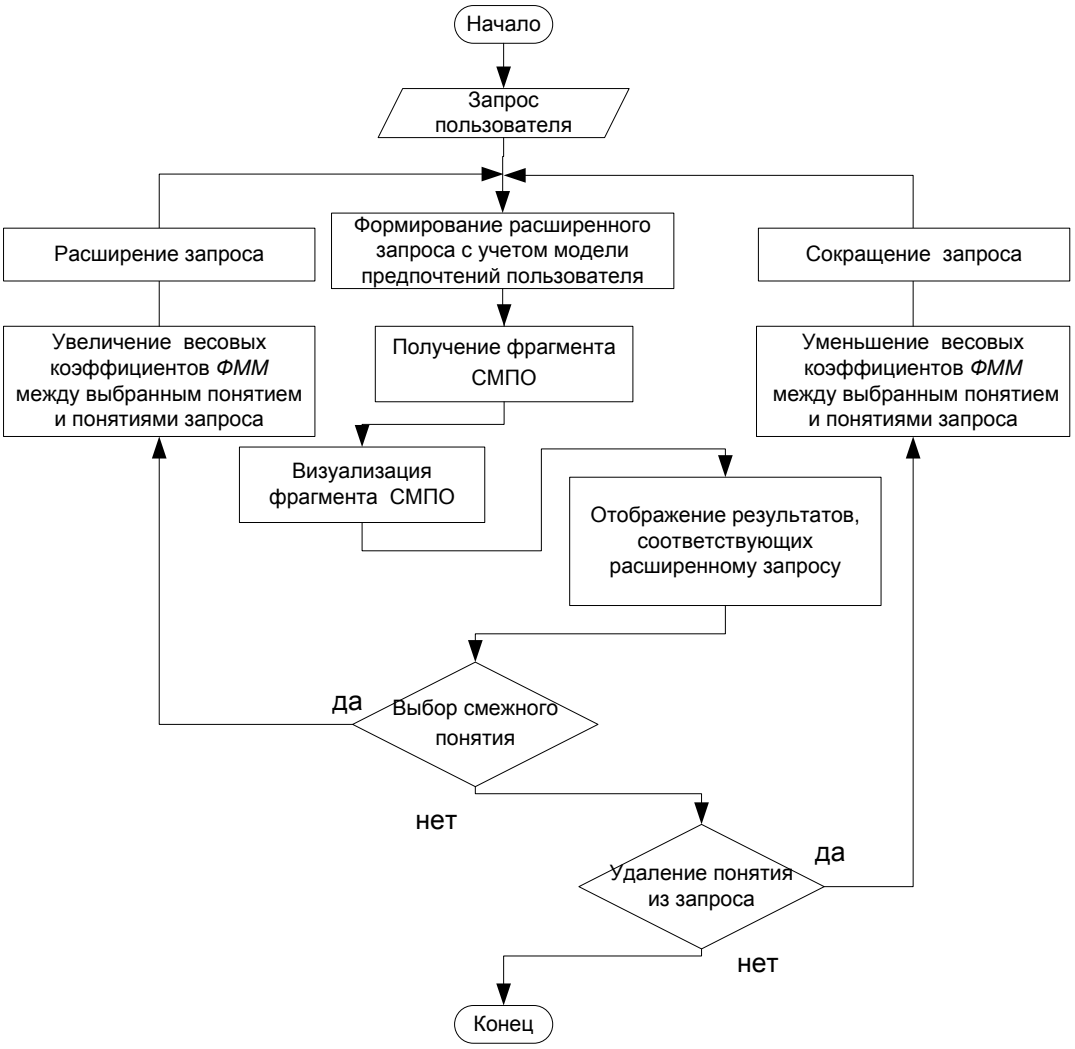


Рисунок 22 - Алгоритм функционирования адаптивного пользовательского интерфейса

Взаимодействие заключается в итеративном расширении запросов пользователя на основе модели предпочтений пользователя, обеспечении возможности коррекции запроса средствами адаптивного интерфейса, а также учета пользовательских предпочтений путем коррекции весовых коэффициентов между концептами модели предпочтений пользователя. При совместном употреблении в запросе пар концептов, входящих в состав модели предпочтений пользователя, увеличиваются весовые коэффициенты отношений между данными понятиями. Системные информационные базы используются в дальнейшем для формирования навигационно-поискового интерфейса, обеспечивающего эффективный, в смысле скорости и релевантности, доступ

пользователя к требуемым данным. Последовательность действий, описывающих взаимодействие компонентов мультипредметной ИС и пользователя для навигационного (а), и поискового (б) способа взаимодействия представлена следующими диаграммами:

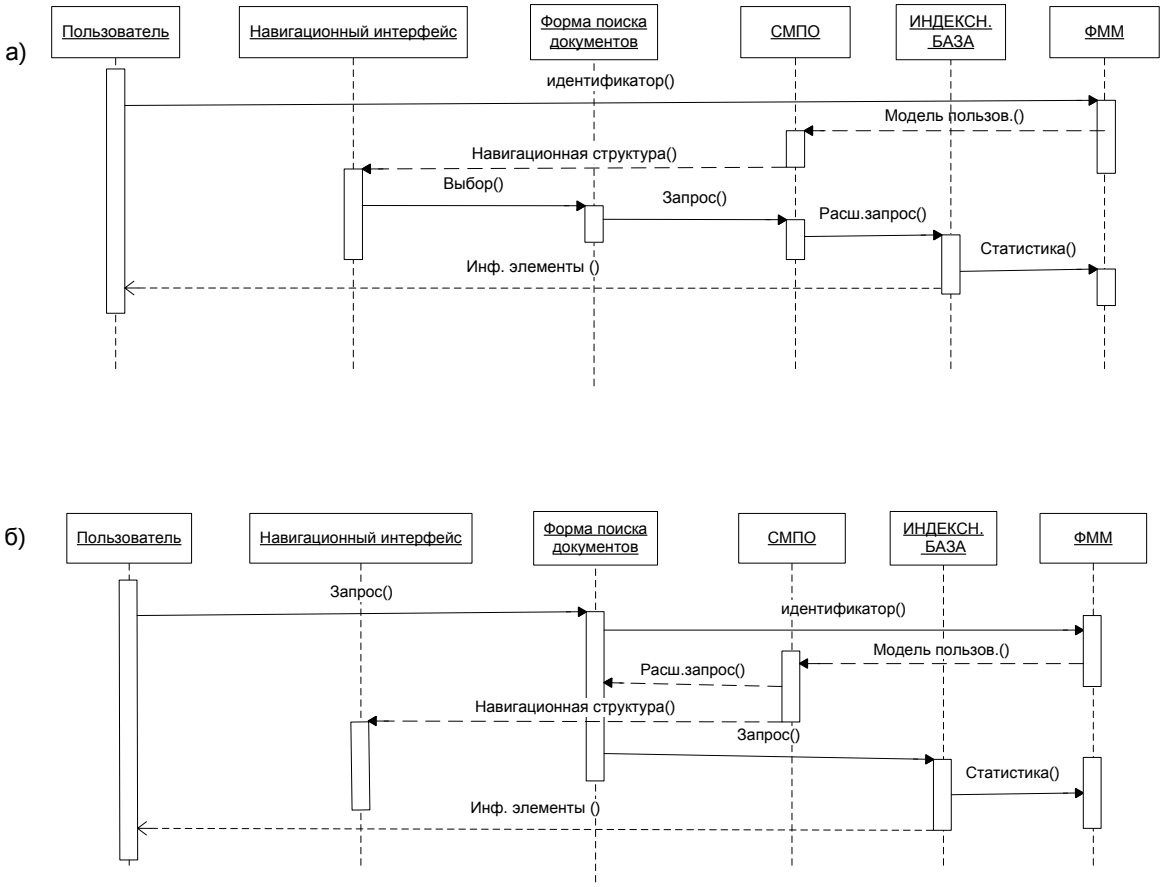


Рисунок 23 - Диаграмма последовательности действий, а) навигация б) поиск

Таким образом, пользователь может в процессе работы влиять путем коррекции модели предпочтений на представление СМПО, адаптируя знания в соответствии со своими представлениями об отображаемых концептах.

3.4.4. Визуализация формализованных знаний на основе методов визуального анализа информации

Одной из существенных проблем, возникающих при попытке создания информационного ресурса, объединяющего большой объем информации,

является сложность интерпретации и обработки данных. Проблема усугубляется также и тем, что данные могут иметь «многомерный» характер, то есть могут быть структурированы множеством способов: например по временной шкале, по принадлежности тому или иному источнику, по ассоциированной с ними области знаний, и т.д. Возможности современных программных средств и вычислительной техники в ряде случаев безнадежно уступают человеческому мозгу, в частности в образном мышлении, а также при обработке многомерных данных, требующей рассмотрения имеющейся информации в различных, часто неявных, контекстах. С другой стороны, человеческий мозг, обладающий мощным потенциалом творческого, ассоциативного мышления, сильно уступает возможностям вычислительных машин в смысле обработки больших объемов информации. Обеспечить при обработке подобных данных эффективное совместное использование, как человеческого мозга, так и вычислительных машин призваны активно развиваемые в настоящее время методы визуального анализа информации [127].

Для визуализации сетевых структур, например, при отображении расширенного запроса, используются методы визуального анализа данных иерархических образов с возможностью динамического проецирования, интерактивной фильтрации и масштабирования. Представление сетевых структур на основе методов визуального анализа информации позволияет реализовать интерфейс человеко-машинного взаимодействия, обладающий следующими возможностями:

Одномоментное (симультанное) восприятие области поиска. При использовании простых интерфейсов, например, таких как строка ввода запроса, пользователю необходимо знать терминологию для конкретизации объекта поиска, выполнять множество запросов, итеративно корректируя свой запрос. При отображении, например, фрагмента семантической сети, пользователь имеет возможность увидеть смежные искомым понятия.

Динамическое отслеживание запроса во время работы с интерфейсом. Навигация в сети позволяет точнее выразить информационную потребность пользователя, определяемую как множество концептов семантической сети, выбранных/отмеченных пользователем.

Улучшенное зрительное восприятие информации за счет пространственной модели визуализации. В силу возможной сложности визуализации граф отображается в псевдотрехмерном пространстве с возможностью поворота и перемещения виртуальной сцены.

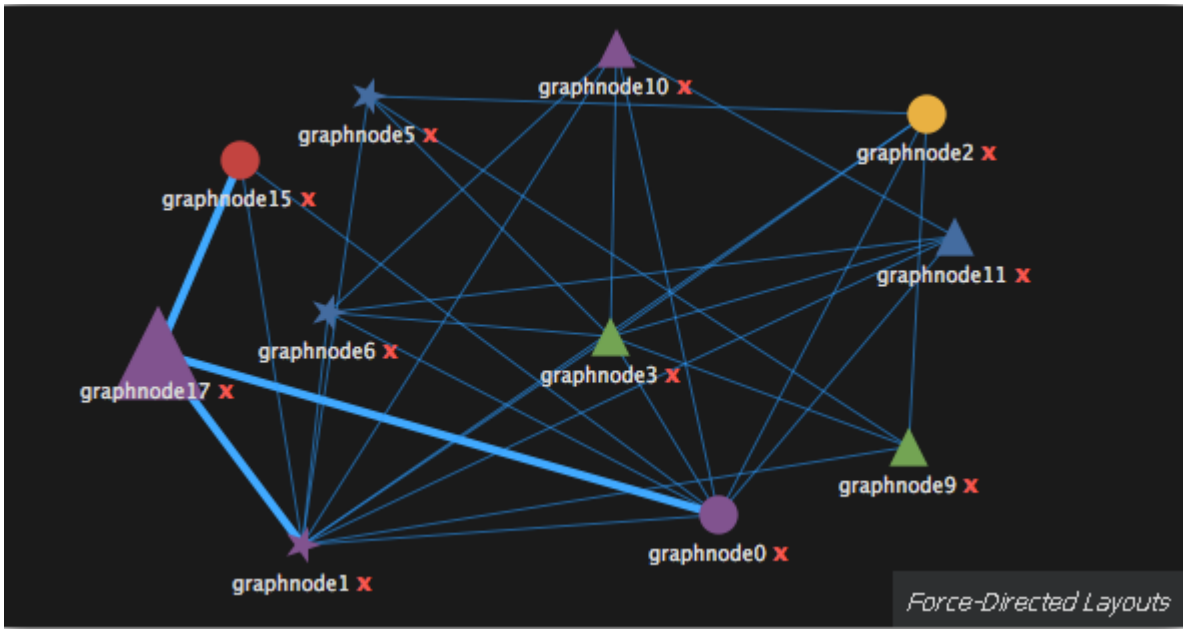


Рисунок 24 - . Пример визуализации графа библиотекой JavaScript InfoVis Toolkit

Максимальное исключение пересекающихся связей между вершинами (концептами семантической сети) осуществляется путем разделения отображаемого множества концептов на подмножества и последующее их размещение на отдельных плоскостях многомерного интерфейса пользователя

$$SN = \langle N_1 \cup N_2 \cup ... \cup N_n, W \rangle, N_i \cap N_j = \varnothing, \forall N_i \exists W(N_i, N_j) \neq \varnothing, \quad (33)$$

где SN – отображаемый фрагмент семантической сети, N_i – непересекающиеся подмножества концептов исходной семантической сети, W – множество связей семантической сети.

Деление концептов семантической сети на множества N производиться по типам отношений семантической сети, которые условно делятся на горизонтальные (связывающие концепты одной плоскости), и вертикальные (связывающие концепты различных плоскостей).

Отнесения типа отношения к одному из видов является экспертной задачей, целью которой является обеспечения интуитивно понятной навигации пользователя. Обычно к горизонтальным отнесятся отношения типа ассоциаций и коннотаций, к вертикальным – отношения часть-целое, принадлежности, классификации и т.п. На рисунке 25 представлен пример визуализации фрагмента сети.

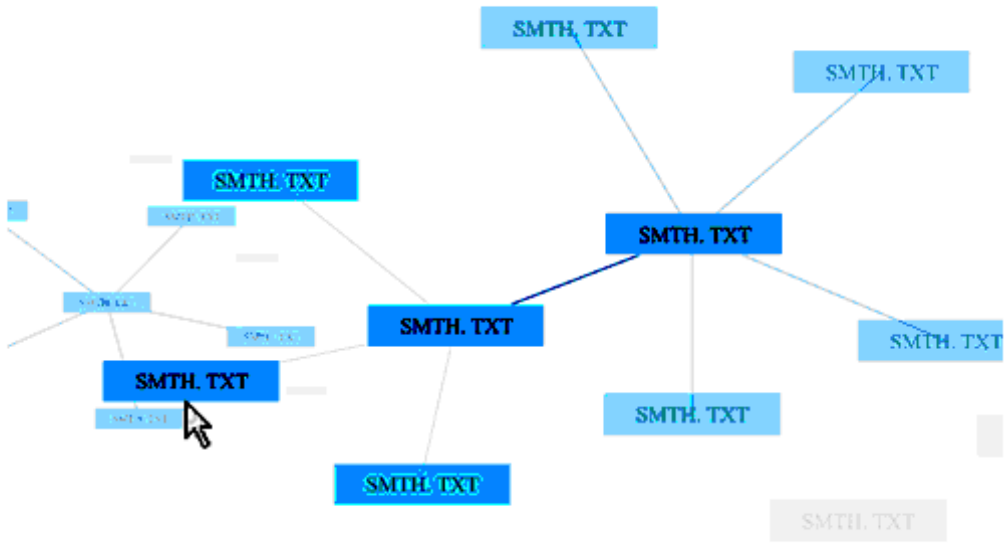


Рисунок 25 - Пример представления фрагмента СМПО

Горизонтальные отношения отражают отношения набора элементов семантической сети, находящихся на одной плоскости интерфейса, вертикальные отношения реализуют связь плоскостей интерфейса друг с

другом. Горизонтальные связи – связи типа ассоциаций и коннотаций, вертикальные – связи часть-целое, отношения принадлежности и классификации. Таким образом, семантически близкие концепты размещены в одной плоскости, а переход на другую плоскость осуществляется преимущественно в процессе уточнения/обобщения запроса пользователем.

Применение технологий визуального анализа данных является перспективным в условиях роста количества данных для отображения.

3.5. Метод поиска на основе семантической сети с субтрактивными связями

Несмотря на интенсивное развитие методов информационного поиска, участие пользователя в процессе поиска остается малоизученным направлением. Роль индивидуальности пользователя относительно как оценки результатов, так и механизма поиска информации отмечается в работах [104, 109, 73], в [104] отмечается предпочтение пользователей в доступе к информации путем информационно-поисковых систем (ИПС), нежели прямой навигации. В [78] отмечается зависимость удовлетворения информационной потребности от эффективности ИПС, опыта и характеристик пользователя. В работах [99,109] рассмотрено вовлечение пользователя в процесс поиска, предложена концепция «human–computer information retrieval» (HCIR), включающая различные аспекты информационного поиска и человеко-машинного взаимодействия. Исследование [73] показывает, что учет неявной обратной связи в виде поведение пользователя при ранжировании результатов позволяет увеличить эффективность поиска на 21%.

В данном разделе представлен метод информационного поиска, позволяющий повысить точность поиска[81] за счет исключения непертинентных[81] результатов. Пертинентность в данной работе рассматривается как соответствие результатов поиска модели предпочтений позьзователя. Ограничение области поиска осуществляется на основании включения в расширенный запрос субтрактивных отношений модели предпочтений пользователя. Субтрактивные отношения – это отношения между концептами, имеющие отрицательный весовой коэффициент. Использование субтрактивных отношений позволяет автоматизировать процесс добавления ограничений в расширенный запрос. Использование семантической модели предметной области и модели предпочтений пользователей позволяет учесть

контекст используемых ограничений, а автоматизация данного процесса снимает необходимость в выработке и вводе ограничений непосредственно пользователем. Общая схема метода поиска представлена на рисунке:

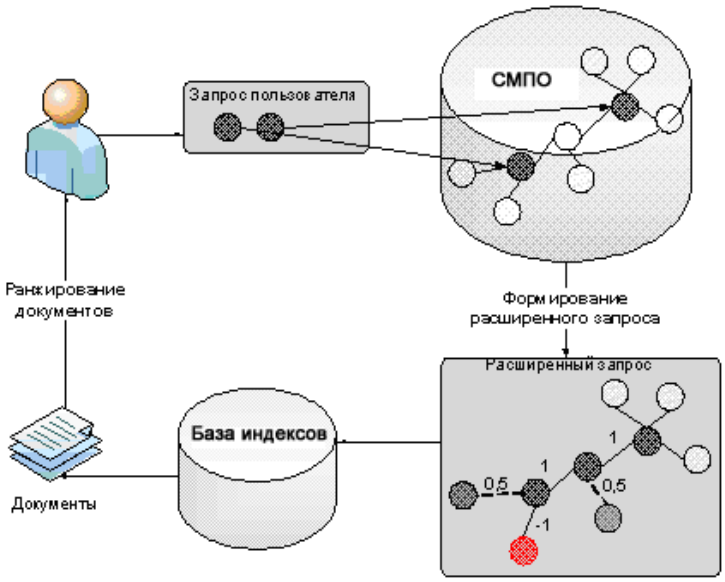


Рисунок 26 - Схема метода поиска

Механизм ограничения области поиска существует со времен реализации первых поисковых машин [86]. Ограничение результатов поиска реализуется существующими поисковыми системами с помощью языка запросов. Однако использование языка запросов вызывает трудности для пользователя в силу необходимости выработки и ручного ввода ограничений при формировании запроса. Так же от мощности и семантики множества ограничений зависит точность поиска – пользователь, не обладающий знаниями всех контекстов используемых ограничений, исключает заведомо перспективные результаты. И наоборот, отсутствие ограничений вынуждает пользователя просматривать множество нерелевантных результатов поиска.

Ввиду вышесказанного, актуальным является разработка метода поиска, исключающего непертинентные пользователю результаты в автоматизированном режиме, с учетом контекста используемых ограничений.

Во время формулировки запроса существуют несколько трудностей со стороны пользователя:

Первая из них – сформировать информационную потребность в виде множества ключевых слов, при этом со стороны пользователя запрос должен быть кратким. Со стороны информационной системы запрос должен содержать максимально полное описание объекта поиска, например, максимальное количество ключевых слов их искомого документа, в случае поиска по ключевым словам. В случае избытка результатов поиска, а также в случаях, когда информационная потребность не может быть четко сформулирована, например, в силу наличия множества синонимов, омонимов в лексическом описании объектов поиска, целесообразно область результатов поиска сокращать не путем уточнения запроса, а путем исключения заведомо ненужных результатов. Данный подход используется в существующих методах поиска, однако механизм его работы весьма несовершенен – отсекаются из рассмотрения страницы, содержащие любую словоформу слова запроса, перед которым есть специальный символ. При этом могут исключаться из рассмотрения результаты, содержащие слово вовсе не в контексте запроса. Т.е. эффективность поиска зависит от объема и семантики множества исключающих слов в контексте текущего запроса, очевидно, что чем данное множество больше, тем точнее можно определить информационную потребность пользователя.

Второй аспект - возможность пользователя в каждом конкретном запросе влиять на множество ограничений. Для пользователя это весьма трудоемкая задача, требующая знания целевой и смежной предметных областей. В данном методе предлагается использовать модель предпочтений пользователей для хранения ограничений в виде субтрактивных отношений – отношение между понятиями модели предпочтений пользователя, имеющими отрицательный весовой коэффициент. Это позволит учитывать ограничения при формировании расширенного запроса в автоматизированном режиме. Также необходимость отображения смежных ключевым словам понятий (отображение расширенного запроса) накладывает функциональное требование к интерфейсу в виде

возможности отображения и изменения расширенного запроса в процессе поиска.

Метод поиска включает 3 составляющих: Модель документа, модель запроса и функцию соответствия между ними. Документ представлен фрагментом СМПО и множеством ключевых слов в базе индексов, выделенных семантическим анализатором на этапе индексации (рис. 27).

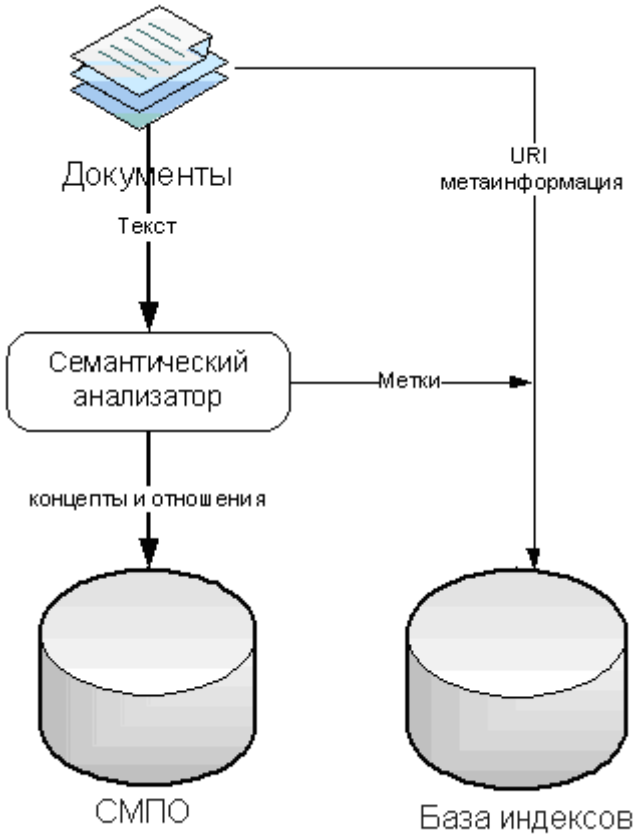


Рисунок 27 - Представление документов

Запрос представлен в виде множества ключевых слов *Q*.

Процесс поиска документов, соответствующих запросу состоит из следующих этапов:

1. Формирование запроса в терминах СМПО:

1.1. Формирование расширенного запроса, содержащего отношения и соответствующие запросу концепты СМПО:

$$EQ = f_q(Q, KB) = \{C^Q, L^Q \mid (Eq(c_i^Q, c_j^{KB}) > 1 - \varepsilon)\}, \quad (34)$$

$$C^Q \subset C, L^Q \subset L \;\; i = \overline{1, N_Q}, j = \overline{1, N_{KB}},$$

где *KB* – СМПО, C^q - множество концептов СМПО, содержащихся в запросе, L^Q - множество отношений над концептами C^q, f_q - функция, ставящая соответствие запросу фрагмент СМПО, *Eq* – функция оценки сходства имен двух концептов, ε – погрешность оценки сходства концептов.

1.2. Расширение запроса с учетом весовых коэффициентов отношений и субтрактивных отношений, ограничивающих контекст запроса:

$$EQ = \{C^Q, L^Q\} \bigcup \{C', L' \mid l : c_i \in C^Q, c_j \in C', \left|\overline{w_k}\right| > x\}, \quad (35)$$

$$C' \subset C, L' \subset L, l \in L',$$

где C' - множество концептов СМПО, связанных с концептами множества C^Q отношениями из множества L', $\overline{w_k}$ - *k-ая* компонента вектора весовых коэффициентов отношения *l*, *x* –коэффициент включения отношения в расширенный запрос.

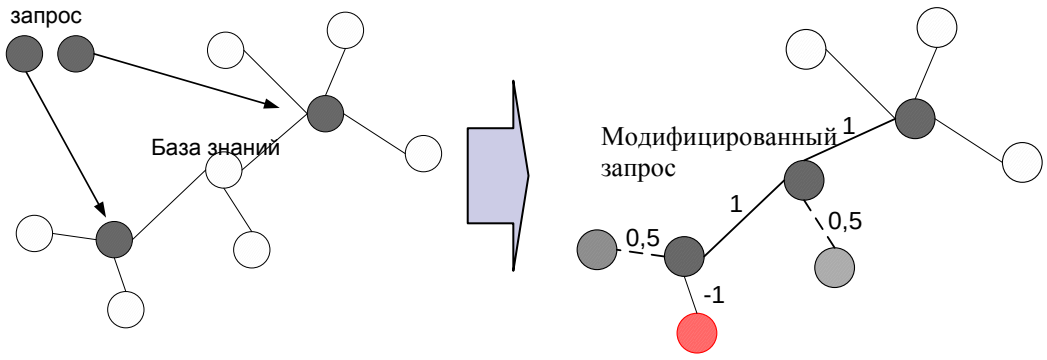


Рисунок 28 - Модифицированный запрос

2. Получение множества документов *D*, соответствующих расширенному запросу:

$$D = \{d_i \mid C^{d_i} \bigcap C^Q \neq \varnothing\},\ i = \overline{1,n}, \tag{36}$$

где C^{d_i} - множество концептов СМПО, присутствующих в документе d_i, C^Q множество концептов СМПО, присутствующих в запросе *EQ*.

3. Ранжирование множества документов с учетом весовых коэффициентов отношений:

$$R(d_k) = \sum_{L_{d_k}} (f_u(\overline{w}_k, r) - \sum_{L'_{d_k}} (f_u(\overline{w}_k, r),, \tag{37}$$

$$L_{d_k} = \{l^d \mid (c_i, c_j \in d_k) \wedge (tp \in \{synonymOf, HyponymOf, associateWith\})\}$$
$$L'_{d_k} = \{l^d \mid (c_i, c_j \in d_k) \wedge (tp \in \{subStract\})\},\ i, j = \overline{1,n}, k = \overline{1,m}, tp \in Tp$$

где $f_u(\overline{w}_k, r)$ - функция получения компоненты вектора весовых коэффициентов отношений из множества L_{d_k} между концептами c_i и c_j, присутствующими в документе d_k, для категории пользователей *r*. *Tp* – множество типов отношений. Таким образом, документы, в которых присутствуют субтрактивные отношения, будут иметь меньший приоритет после ранжирования. Результатом ранжирования является упорядоченное по убыванию оценки *R* множество документов, представляющих результаты поиска.

Алгоритм процесса поиска представлен на рисунке 29.

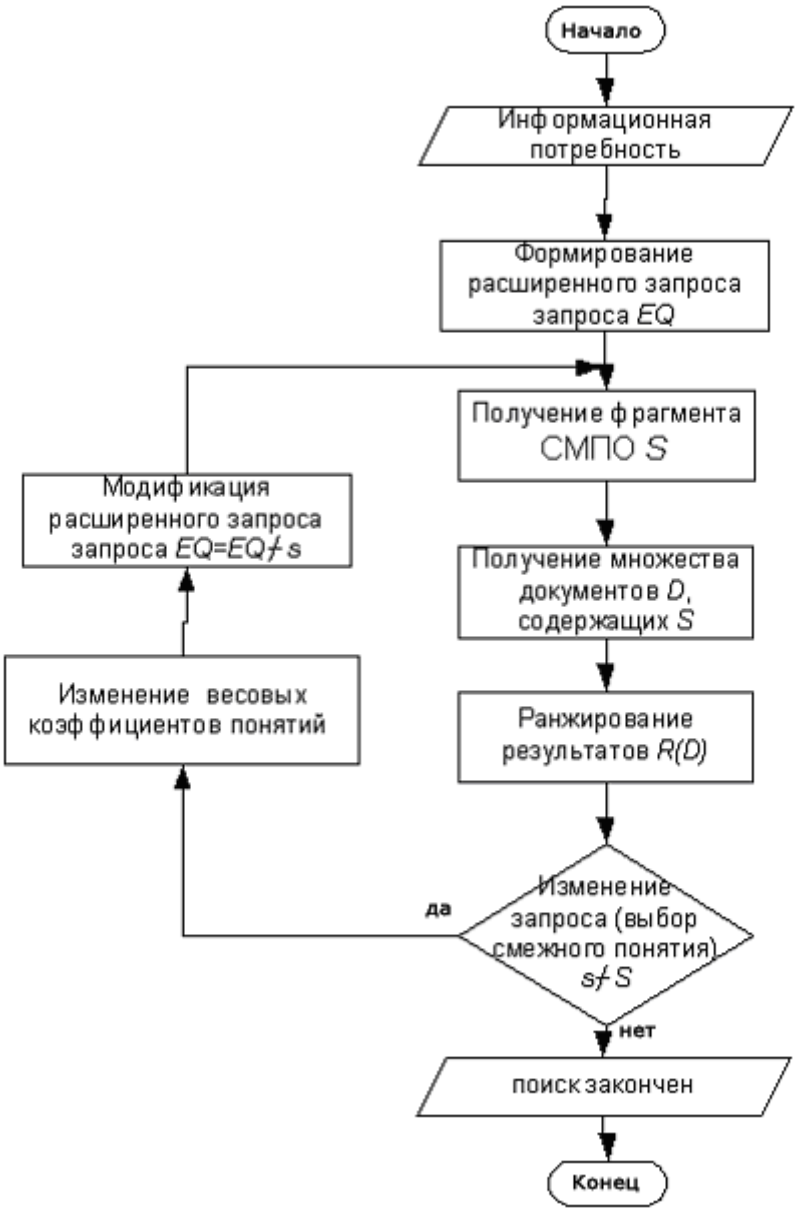


Рисунок 29 – Алгоритм поиска

Особенность данного метода состоит в автоматизации процессов изменения запросов, а также ограничения области поиска, что позволяет учитывать дополнительные ограничения в запросе, которые, в свою очередь, могут добавляться в запрос в автоматизированном режиме.

Выводы по главе 3

Рост объемов информации, обрабатываемой современными информационными системами, обуславливает необходимость развития технологий оперативного доступа в ней. Одним из путей решения данной проблемы является построения интерфейсов, способных предоставить пользователю необходимый функционал для оперирования большими массивами данных.

Применение описанных метода и технологий в мультипредметных информационных систем промышленных предприятий, позволяет повысить эффективность доступа сотрудников организаций к требуемым документам. Метод динамического формирования семантической модели предметной области позволяет сформировать в автоматизированном режиме динамическую семантическую модель предметной области и обеспечить процессы поддержания ее актуальности. В процессе взаимодействия пользователя в рамках метода интерфейсной навигации он может влиять на СМПО, проводя процессы уточнения и адаптации знаний в соответствии со своими представлениями о предметной области. Метод поиска позволяет сформировать в автоматизированном режиме расширенный запрос, ограничить область поиска и ранжировать результаты с учетом модели предпочтений пользователя.

ГЛАВА 4. Применение методов мультипредметных информационных систем в рамках документооборота организаций

На основе предложенных в предыдущих главах методов создан программный комплекс, реализующий следующие функции:

1. Информационный поиск с автоматическим расширением запроса, и ограничением области поиска на основе модели предпочтений пользователей;
2. семантическое индексирование и интеграцию текстовых документов;
3. интерфейсной навигации на основе модели предпочтений пользователей;
4. формирование модели предпочтений пользователей информационной системы.

Программный комплекс выполнен в виде набора веб-сервисов, архитектура программного комплекса представлена на рисунке 30.

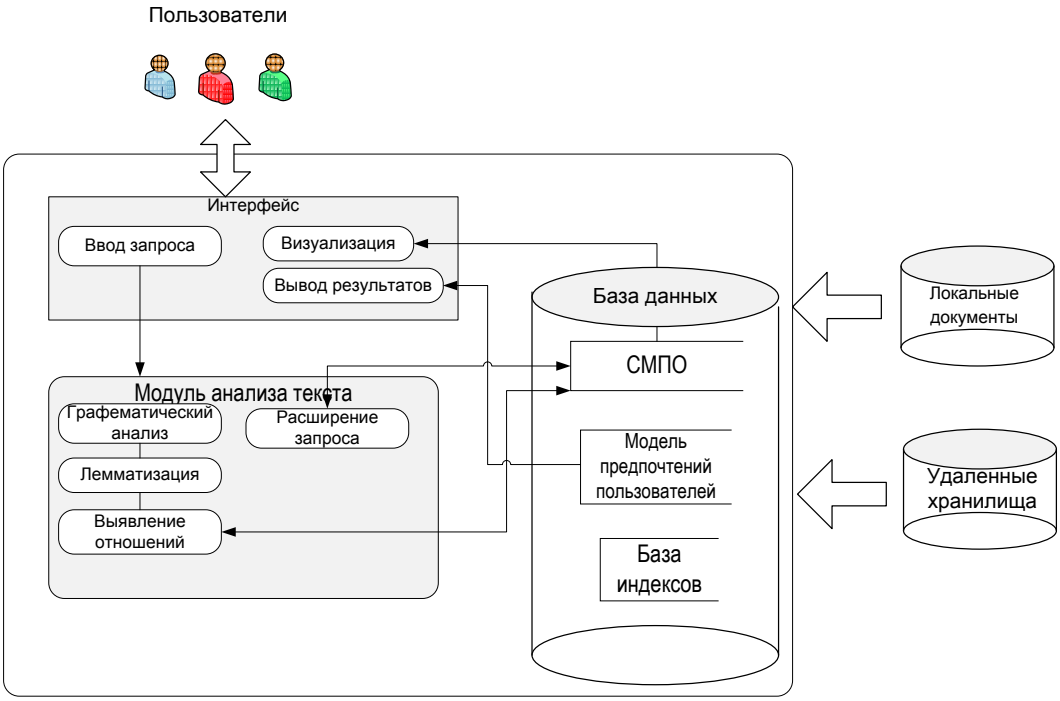


Рисунок 30 - Архитектура программного комплекса

Базу данных мультипредметной информационной системы составляют семантическая модель предметной области, модель предпочтений пользователей, база индексов текстовых документов организации. Входными данными, являются

документы, размещенные на локальных и удаленных хранилищах, и статистика работы пользователя. Выходными данными являются результаты обработки запросов и навигационная структура интерфейса. МПП формируется на основе обработки запросов пользователя и статистики его работы. Модуль анализа текста используется для формирования семантической модели предметной области, а также при формировании расширенного запроса.

4.1. Реализация сервиса поиска информации

Сервис поиска информации по запросу пользователя, выполняет следующие функции:

1. Формирование расширенного запроса пользователя на основе введенных ключевых слов;
2. вывод документов, соответствующих расширенному запросу пользователя;
3. ранжирование списка результатов поиска в соответствии с моделью предпочтений пользователей.

Сервис поиска информации реализует взаимодействие с пользователем через форму поиска, представленную на рисунке 31.

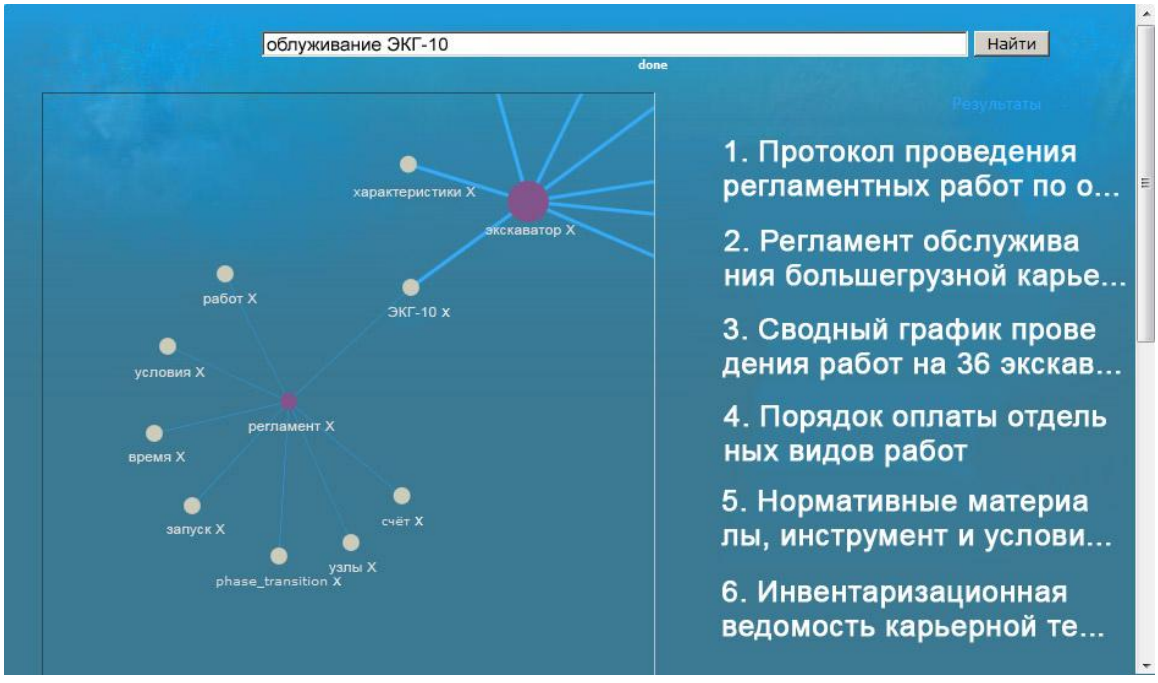


Рисунок 31 – Форма поиска

Форма поиска содержит поле для ввода ключевых слов, поле отображения расширенного запроса в виде графа, вершины которого обозначают концепты СМПО, соответствующие расширенному запросу, и поле отображения результатов. Поле результатов отображает список найденных документов.

На рисунке 32 представлен алгоритм функционирования сервиса поиска.

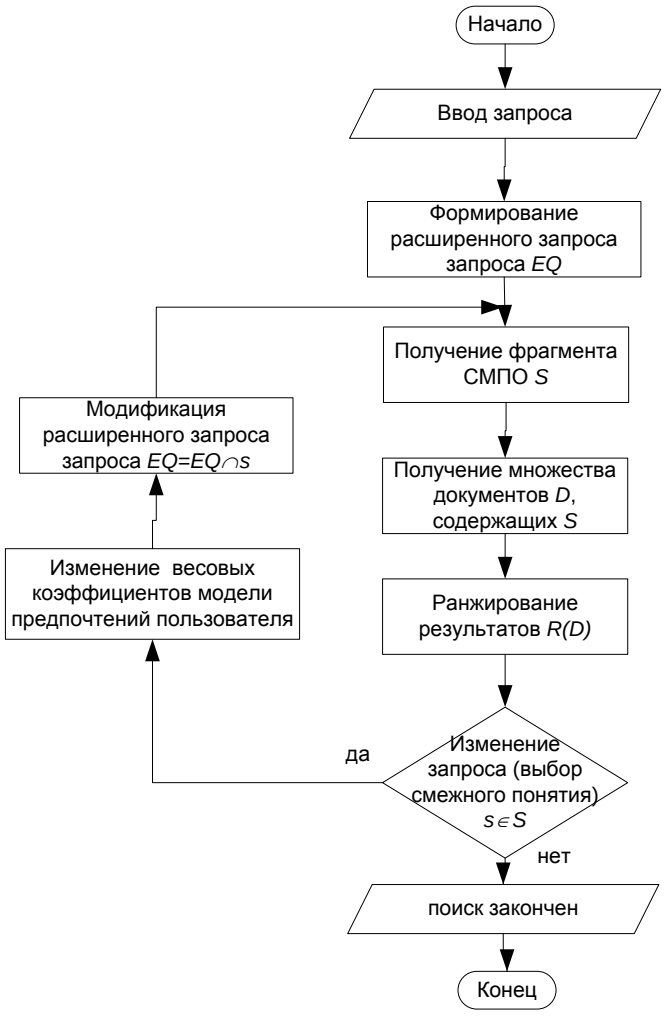


Рисунок 32 - Алгоритм функционирования сервиса поиска

Сервис поиска информации реализован средствами интерпретируемого скриптового языка программирования PHP. Основной особенностью языка являются возможности для работы с текстом, генерации HTML-страниц на веб-сервере, поддержка форматов XML и JSON. В области Веб-программирования PHP отличается скоростью выполнения, функциональностью, распространению исходных кодов на основе лицензии PHP. PHP отличается наличием ядра и

подключаемых модулей для работы с базами данных, сокетами, динамической графикой, криптографическими библиотеками, документами формата PDF. В PHP переданные сценарию параметры автоматически станут переменными сценария, с которыми можно работать, как с обыкновенными переменными. То же самое происходит с переменными окружения сервера. PHP поддерживает взаимодействие с различными базами данных (MySQL, PostgresSQL, Sybase, Informix, др.). База данных реализована средствами СУБД MySQL.

Функция формирования запроса в терминах СМПО представлена в листинге 1:

```
$query=$_REQUEST['query'];                    //запрос
lems = explode(" ", $query);                  //разбиение запроса
на слова
foreach ($lems as $lem_num => $lem) {
    $fl_last=-1;
    $last_id=$pr_id;
    $lem =ereg_replace("\(","",$lem);
    $lem =ereg_replace("\)","",$lem);
    $lem =ereg_replace(",","",$lem);
    $lem =ereg_replace(";","",$lem);
    $lem =ereg_replace(" ","",$lem);
$query1 = "SELECT * FROM zalizniak WHERE MATCH
(init,forms) AGAINST ('".$lem."');"; // запрос на поиск
слова в грамматическом словаре русского языка А. А.
Зализняка
$result = mysql_query($query1)or die("Query failed: " .
mysql_error());
if (mysql_num_rows($result)== 0){
$query_tez = "SELECT * FROM tezaurus WHERE  (name) LIKE
('".$lem."');";}    // запрос на поиск начальной формы
слова в тезаурусе
else{
$query_tez = "SELECT * FROM tezaurus WHERE  (name) LIKE
('".substr($lem,0,-2)."%');";}
$result_tez  =  mysql_query($query_tez)  or  die("Query
failed: " . mysql_error());
```

Листинг 1 – Формирование запроса в терминах СМПО

Строка запроса разбивается на ключевые слова, которые нормализуются с помощью словаря русского языка Зализняка А.А. В качестве основы СМПО на начальном этапе, а также для определения синонимов используется русскоязычный тезаурус WordNet версии 3.0.

Далее происходит расширение запроса путем поиска отношений в СМПО, частью которых являются слова запроса, и формирование графа расширенного запроса в формате JSON для последующей визуализации. Листинг приведен ниже.

```
for ($j=0; $j < $i ;$j++){ //поиск слов запроса в тезаурусе
  $result4 = mysql_query("SELECT `name` FROM `tezaurus` WHERE id=".$ids[$j][0]."");
  while ($row4 = mysql_fetch_object($result4))
    { $name=$row4->name." "; }   // формирование графа запроса
    if ($p2fl==1) {$json=$json.",";}
    $json=$json.'{ "adjacencies" :    [ ';
    $p2fl=1;
    $res_A = mysql_query("SELECT * FROM `links` WHERE id1=".$ids[$j][0])" OR id2=".$ids[$j][0])" AND w0>".$w);
        $links[$j][0]=mysql_num_rows($res_A);
    //количество связей
        $links[$j][1]=$ids[$j][0];
        $jj=2;
        $pfl=0;
        $branch_count=0;
        while ($r_A = mysql_fetch_object($res_A)) {
            $links[$j][$jj]=$r_A->id2;
            $jj=$jj+1;
            $branch_count=$branch_count+1;
                        if ($branch_count < $branch_n)
    {
            $nameto="error";
            $res_b = mysql_query("SELECT * FROM `tezaurus` WHERE id=".$r_A->id2."");      while ($rowb = mysql_fetch_object($res_b))   {
            $nameto=$rowb->name." ";
            $tf=$rowb->TFIDF;}
            if ($pfl==1) {$json=$json.",";}
        $json=$json.'{                                "nodeTo": "'.$nameto.'","nodeFrom": "'.$ids[$j][0].'", "data ": {} }';
        $snow[]=$r_A->id2;
        $pfl=1;  }
        $json=$json.'], "data":  { "$color": "#83548B", "$type":  "circle"  }, "id":  "'.$ids[$j][0].'", "name": "'.$name.'"  } ';
```

Листинг 2 - Расширение запроса и формирование графа расширенного запроса

Для визуализации графа расширенного запроса использована библиотека интерактивной визуализации JavaScript InfoVis Toolkit.

Силовые (Force-directed) методы рисования направленных графов представляют собой класс алгоритмов для вывода графа в удобном виде. Их цель заключается в позиционировании узлов графа в двумерном или трехмерном пространстве таким образом, чтобы все ребра имеют одинаковую длину и наименьшее количество пересечений. Для достижения этого имитируется сила натяжения ребер, а движение ребер и узлов заключается к сведению к минимуму энергию натяжения ребер. Граф, визуализирующий расширенный запрос, представленный на рисунке 31 является интерактивным, при выборе вершины соответствующее понятие СМПО включается в запрос, при удалении вершины ограничивается область поиска – исключаются результаты, содержащее данное понятие.

Далее выводится список результатов, соответствующий расширенному запросу и его ранжирование в зависимости от степени значимости понятия (весового коэффициента, определяемого функцией mmod()) в модели предпочтений пользователя.

```
docs=array();                    // массив документов
for ($j=0; $j < $i ;$j++)  {
  $result4 = mysql_query("SELECT * FROM `keywords` WHERE
kwid=".$ids[$j][0]."");
  while ($row4 = mysql_fetch_object($result4)){
  $doc=$row4->docid;
  for ($x=0; $x < $c ;$x++){
     if($doc == $docs[$x][0]) {
     $docs[$x][1]=$docs[$x][1]+1*mmod($ids[$j][0]); }    }
     if ($fl==-1)     {
     $docs[$c][0]=$doc;
     $docs[$c][1]=0;
     $c++;    }    }    }
  sort($docs[1]);
  for ($j=0; $j < $c ;$j++){      //          ранжирование
результатов
$result4  =  mysql_query("SELECT  *  FROM  `docs`  WHERE
docid=".$docs[$j][0]."");
     while ($row4 = mysql_fetch_object($result4)){
     $title=$row4->title;
     $url=$row4->url;
     if ($docs[$j][1]>10) $rang="2";
     if ($docs[$j][1]>100) $rang="1";
     $res_count=$res_count+1; // вывод результатов
```

```
echo '<div align="left"> <a href="'.$url.'"><h3>'.$res_count.'. '.$title.'</h3> '.$url.'</a> </div>' ; } }
```

Листинг 3 – Вывод и ранжирование результатов поиска

Форма поиска отображает фрагмент СМПО в виде расширенного запроса, и реализует итеративную коррекцию запроса за счет выбора пользователем вершин отображаемого графа. Форма поиска позволяет реализовать одномоментное восприятие, интерактивность (корректировка запроса в процессе работы с пользователем), а так же возможность ограничения области поиска путем удаления отображаемых концептов. Отображаемые вершины графа символизируют понятия СМПО, соответствующие расширенному запросу пользователя. Возле каждого понятия присутствует символ удаления понятия из расширенного запроса (Рисунок 33).

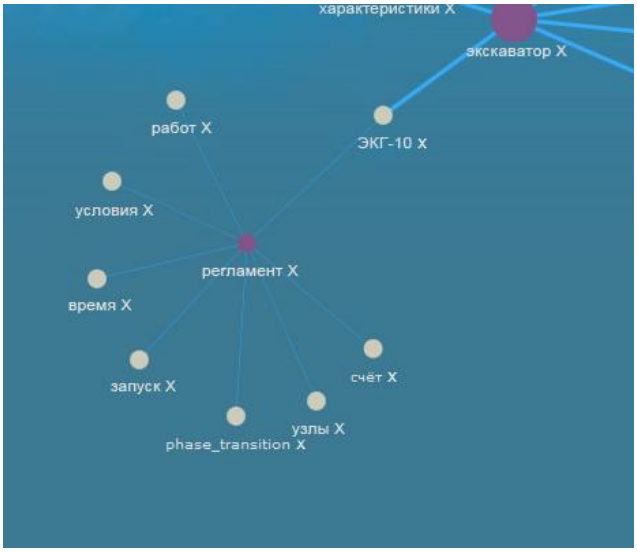


Рисунок 33 - Граф, отражающий фрагмент СМПО

Пользователь может включить в запрос отображаемые в вершинах графа понятия путем их выбора, либо удалить данные понятия кликом по значку удаления. Каждое действие пользователя инициирует изменение запроса пользователя путем включения или исключения понятий, соответствующих вершинам отображаемого графа.

4.2. Реализация сервиса семантического индексирования

Сервис семантического индексирования формирует семантические образы документов в виде множества понятий и отношений между ними. Семантическую модель предметной области на основе нормативно-справочной информации организации составляют семантические образы основных документов, среди которых:

1. Нормативно-правовые акты;
2. требования безопасности;
3. регламенты работ;
4. правила организации работ;
5. технологические карты.

Исходными данными для построения семантических образов документов является текст индексируемого документа. Алгоритм построения семантических образов документов содержит несколько этапов:

1. Разбиение исходного текста на предложения;
2. разбиение предложений на слова;
3. удаление стоп слов, спецсимволов на основе статистических характеристик текста;
4. определение начальных форм с использованием грамматического словаря русского языка (А. А. Зализняк);
5. определение отношений между словами;
6. запись в «таблицу ключевых слов документа» и «таблицу связей ключевых слов».

Для определения нормальной формы слов был использован грамматический словарь русского языка Андрея Анатольевича Зализняка. Словарь содержит приблизительно 100 тыс. базовых словоформ русского языка с их полным морфологическим описанием. Для определения отношений между словами был использован тезаурус WordNet версии 3.0. Тезаурус WordNet

содержит около 155 000 слов, 207 000 отношений. Структура тезауруса представлена на рисунке 34.

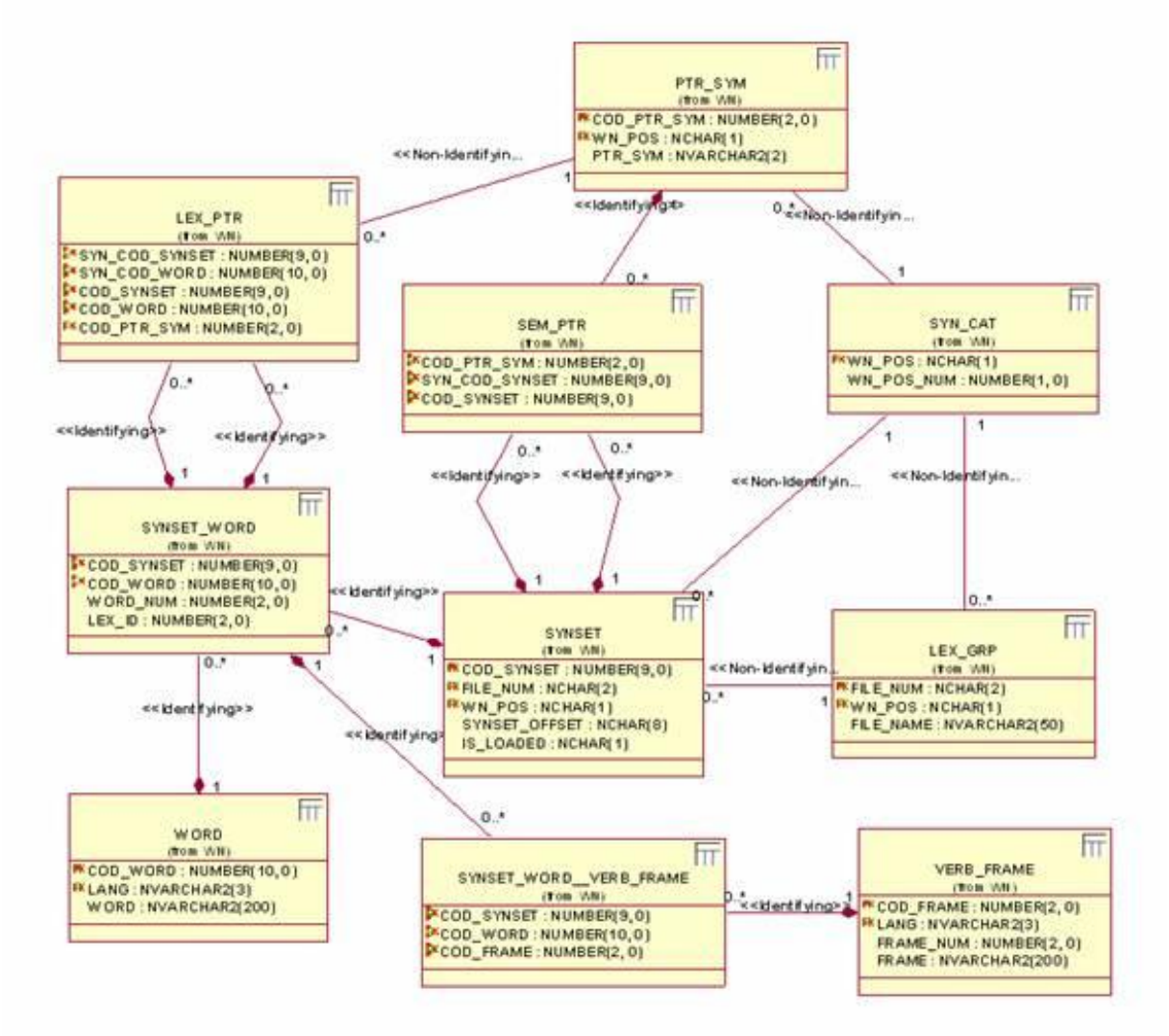


Рисунок 34 - Структура тезауруса WordNet

WordNet – это тезаурус, реализованный моделью семантической сети. WordNet содержит 4 сети для следующих частей речи: существительных, глаголов, прилагательных и наречий. Базовой словарной единицей в WordNet является синонимический ряд (англ. «sinset»). В синонимические ряды объединены слова близкого значения. Синонимические ряды являются узлами сети. Каждый синонимический ряд содержит определение. Такая организация позволяет учитывать различные значения и часть речи одного слова или словосочетания может появляться более чем в одном синонимическом ряду и иметь более одной категории части речи.

Помимо синонимов и синонимичных словосочетаний каждый синонимичный ряд содержит указатели, описывающие отношения между синонимичными рядами. Синонимичные ряды в тезаурусе WordNet связаны между собой отношениями гиперонимии, гипонимии, часть-целое, меронимии, антонимии, а также контекстную связь, обозначающую наличие неявной связи между синонимичными рядами.

Далее представлены листинги получения семантического образа документа.

```
    $data = strip_tags($data); //лемматизация текста
    $data = preg_replace('/\s\s+/', ' ', $data);
    $data = preg_replace('/\s-\s/', ' ', $data);
    $data = preg_replace('/ё+/ui', 'е', $data);
    $data = preg_replace('/,/', ' ', $data);
    $data = preg_replace('/[.]+/', '.', $data);
    $data = preg_replace('|\?+|', '.', $data);
    $data = preg_replace('|\!+|', '.', $data);
    $data = trim($data);
    $data = mb_convert_case($data, MB_CASE_LOWER, "UTF-
8");
    $con = db_connect();
    $query_stop = "SELECT * FROM stop_words"; //
удаление стоп-слов
    $del = mysql_query($query_stop);
    $del = db_result_to_array($del);
    foreach ($del as $re) {$gt .= $re[stop_word].'|';}
    $gt = substr($gt, 0, strlen($gt)-1);
    $pattern_stop = '/\s('.$gt.')\s/ui';
    for ($i = 1; $i <= 4; $i++) {
    $data = preg_replace($pattern_stop, ' ', $data);   }
    $pattern_znaki = '/[^a-zA-Za-яА-Я0-9-.]+/ui';
    $data = preg_replace($pattern_znaki, ' ', $data);
    $sentence = explode(".", $data, -1);
    $data = preg_replace('/[.]/', NULL, $data);
    $data = explode(" ", $data); //разбиение текста на
слова
    foreach ($data as $word_num => $word) {
    if (strlen($word) < 4) {unset($data[$word_num]);}  }
    foreach ($data as $word_num => $word) {
        $rt = get_normal_form($word);    //приведение к
нормальной форме
        if (count($rt) != 0) {
            foreach ($rt as $qwe) {
            $text_nf[] = $qwe[init];}     }
        else $text_nf[] = $word;      }
```

Листинг 3 - Лемматизация текста

После лемматизации производится поиск существующих и добавление новых отношений между словами документа.

```
$query_tez = "SELECT * FROM tezaurus WHERE  (name) LIKE
('".$lem."');";}
$result_tez = mysql_query($query_tez); // поиск слов в
СМПО
if (mysql_num_rows($result_tez)== 0) {
  $lem = preg_replace("/(\n \s{2,})/"," ",$lem);
if((strlen($lem)>4))//&&(preg_match("/\s/",$lem)==0))  {
  $q="SELECT id FROM tezaurus ORDER BY id DESC LIMIT 1";
  $qf = mysql_query($q);
  $rid;
  while ($r = mysql_fetch_object($qf))    {
    $rid=$r->id; }
    $rid=$rid+1;
    $q= "(".$rid.",'".$lem."','0','2','1','1','0')";
    $q="INSERT                INTO                tezaurus
(id,name,part,source,TF,DF,TFIDF) VALUES ".$q.";";
    // добавление в СМПО новых слов
        $qf = mysql_query($q);
while ($rowtez = mysql_fetch_object($result_tez))   {
    for ($j=0; $j < $i ;$j++){
    if($rowtez->id == $ids[$j][0])    {
    $fl=$j;$ids[$j][1]=$ids[$j][1]+1;}}
    if ($fl==-1)  {
    $ids[$i][0]=$rowtez->id;
    $ids[$i][1]=1;
    $ids[$i][2]=$rowtez->part;
    $i++; }} }   }
```

Листинг 4 - Поиск и добавление понятий и отношений в СМПО

4.3. Реализация сервиса интерфейсной навигации на основе СМПО и формировании модели предпочтений пользователей

Сервис интерфейсной навигации обеспечивает формирование формализованной ментальной модели категории пользователей, а так же обеспечивает выработку навигационной структуры интерфейса, адаптировано для различных категорий пользователей.

Формирование ментальной модели происходит путем учета статистики использования концептов СМПО в запросах пользователя, а так же корректирования расширенного запроса. На рисунке представлен алгоритм функционирования адаптивного пользовательского интерфейса.

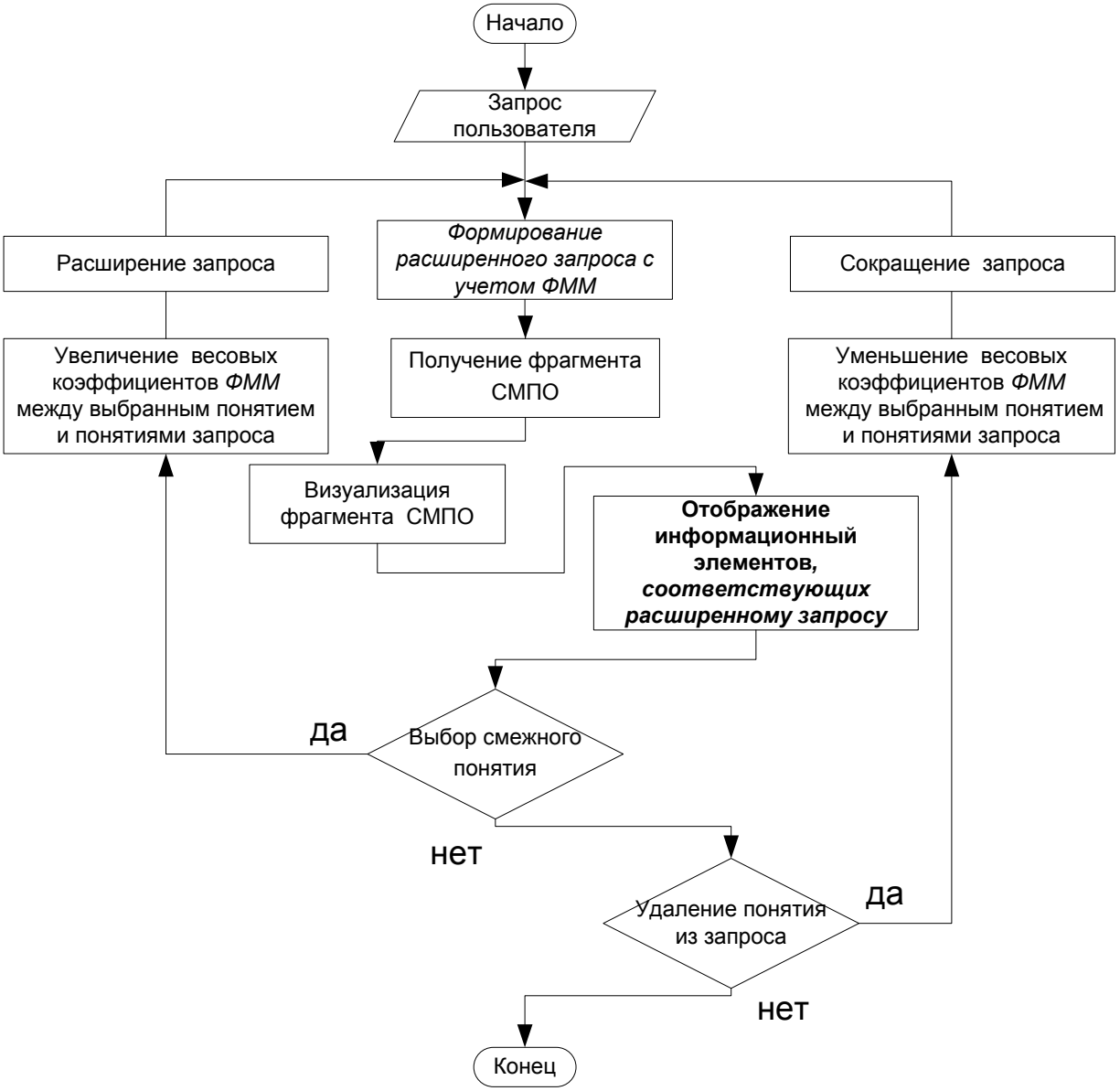


Рисунок 35 - Алгоритм функционирования адаптивного пользовательского интерфейса

Далее представлен листинг формирования модели предпочтений пользователя.

```
    $kolword = array_count_values($text_nf);
    array_multisort($kolword, SORT_DESC);
    foreach ($kolword as $word => $kol) {
    $ty = search_word_in_db($word, $user_id, "words");
    if (count($ty) == 0) {
$query_add = "INSERT INTO words (id_user, word, tf,df)
VALUES ('".$user_id."', '".$word."', '".$kol."','1')";
    mysql_query($query_add);
    if ($check_url) {
    $nnn = mysql_insert_id();
    $words_of_doc1[] = $nnn;
```

```
    $polar .= $nnn.',';}      }
    else {
    foreach ($ty as $qwer) {
    $oi = $kol + $qwer[tf];
    if ($check_url) {
    $words_of_doc1[] = $qwer[id];
    $polar .= $qwer[id].',';}     }
    $query_freq = "UPDATE words SET tf='".$oi."' WHERE
id_user='".$user_id."' AND word='".$word."'";
        mysql_query($query_freq);}  }
```

Листинг 5 - Формирование модели предпочтений пользователя

Изменение весовых коэффициентов осуществляется при совместном использовании нескольких концептов, например, при совместном употреблении понятий пользователем в одном запросе.

```
foreach ($lems as $lem_num => $lem) {
$fw   =   search_word_in_db($lems[$lem_num],   $user_id,
"words");
$sw   =   search_word_in_db($lems[$lem_num+1],   $user_id,
"words");
foreach   ($fw   as   $kv1)   {$k   =   $kv1[id];   $mmmm   =
$kv1[word];}
foreach   ($sw   as   $kv2)   {$j   =   $kv2[id];   $nnnn   =
$kv2[word];}
if ($k > $j) {$teemp = $k; $k = $j; $j = $teemp;};
 if ($k != $j) {
$search_id   =   "SELECT   *   FROM   neighbor   WHERE
id_user='".$user_id."'    AND    word_1='".$k."'    AND
word_2='".$j."'";
    $sel = mysql_query($search_id);
    $sel = db_result_to_array($sel);
    if (count($sel) == 0) {
$add_neighbor = "INSERT INTO neighbor (id_user, word_1,
word_2,   count)   VALUES   ('".$user_id."',   '".$k."',
'".$j."', '1')";
    mysql_query($add_neighbor);   }
    else {
    $up_neighbor = "UPDATE neighbor SET count = count + 1
WHERE  id_user='".$user_id."'  AND  word_1='".$k."'  AND
word_2='".$j."'";
        mysql_query($up_neighbor);   }   }   }
```

Листинг 6 - Изменение весового коэффициента в ментальной модели пользователя

Навигационная структура формируется на основе модели предпочтений пользователя. Ниже представлен фрагмент листинга вывода навигационной структуры в виде графа на основе на основе СМПО и модели предпочтений пользователя.

```
  $query4 = "SELECT * FROM `words` WHERE id='".$row-
>id."' LIMIT 1 ;";
  $result4 = mysql_query($query4) or die("Query failed: "
. mysql_error());
     while ($row4 = mysql_fetch_object($result4))  {
     $from=$row4->word;
     $from_id=$row4->id;  }
  $query2 ="SELECT * FROM `weight` WHERE id_user =
'".$query."' AND id_w1= '".$row->id."' AND weight >
'".$weight_min."' ORDER BY 'weigt' DESC LIMIT 10;";
if($query=="all") $query2 ="SELECT * FROM `weight` WHERE
id_w1= '".$row->id."' AND weight > '".$weight_min."'
ORDER BY 'weigt' DESC LIMIT 10;";
  $result2 = mysql_query($query2);
  while ($row2 = mysql_fetch_object($result2))     {
     $tfdf2 = 1;
     if($row2->df!=0)     {
  $tfdf2 = ($row2->tf)/($row2->df);
  $update_tfdf = "UPDATE words SET tfdf = tf/df WHERE
id_user='".$user_id."' AND id='".$row->id."'";
  mysql_query($update_tfdf);
if(($tfdf2>$tfdf_min2)&($tfdf2<$tfdf_max2))    {
  $query3 = "SELECT * FROM `words` WHERE id='".$row2-
>id_w2."' LIMIT 1 ;";
  $result3 = mysql_query($query3) or die("Query failed: "
. mysql_error());
while ($row3 = mysql_fetch_object($result3)) {
  $to=$row3->word;
  $to_id=$row3->id;
  $linkFlag=1;
  $data=$data.'  {    "adjacencies":  [],       "data":  {
"$color": "#70A35E",   "$type": "square"  },        "id":
"'.$to_id.'",   "name": "'.$to.'"  }, ';  }
  $asj=$asj.'{    "nodeTo":    "'.$to_id.'",    "nodeFrom":
"'.$from_id.'", "data": {} },'; }}
  $asj=substr($asj,0,-1).'],
  "data": { "$color": "#83548B", "$type": "circle" },
  "id": "'.$from_id.'", "name": "'.$from.'"    },';
  if ($linkFlag == 1) $json=$json.$asj;
  $linkFlag=0;   }   }
  $data = substr($data,0,-1);
  $json = $json.$data;
```

Листинг 7 - Формирование структуры навигации в формате JSON

На рисунке 36 представлена диаграмма последовательности действий, характеризующая действия пользователя и реакцию информационной системы.

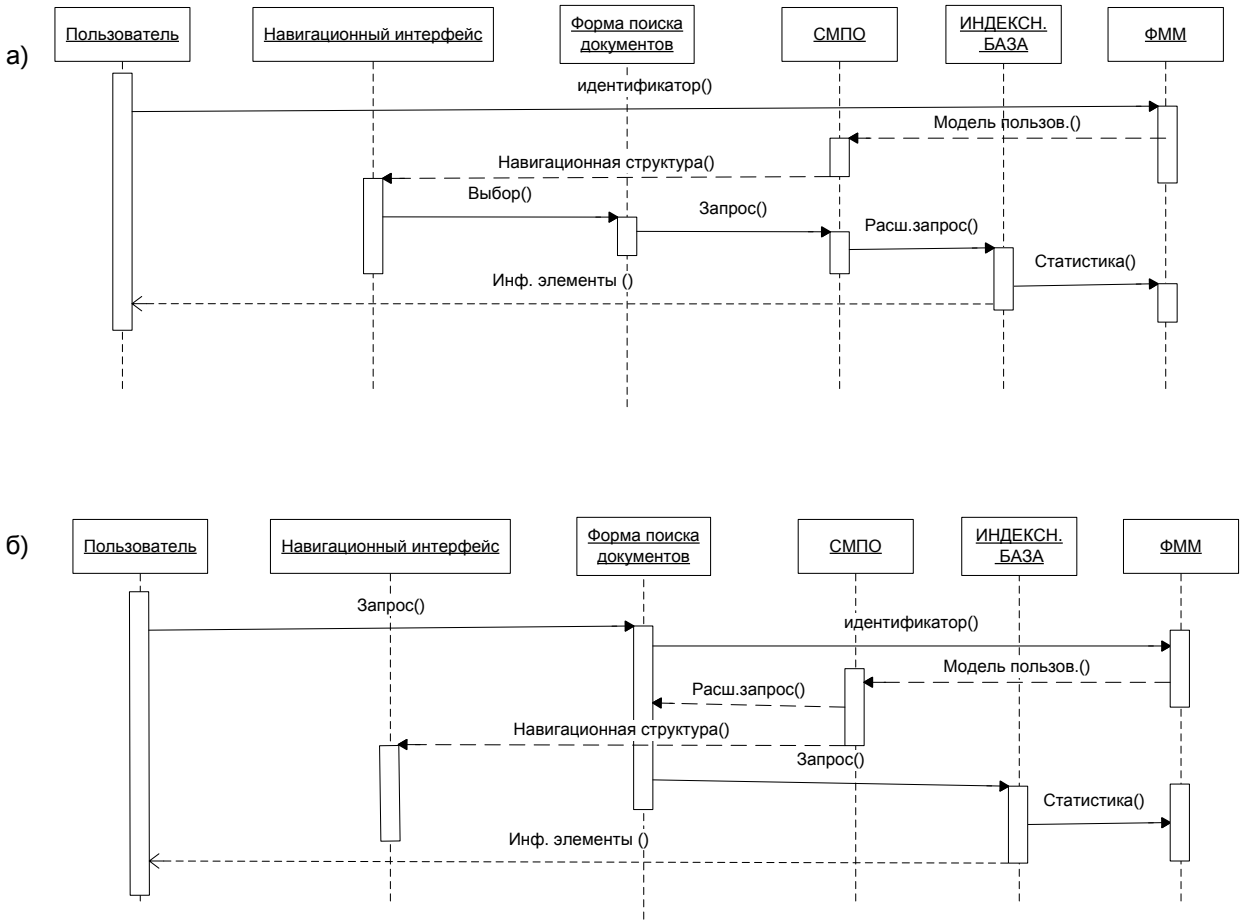


Рисунок 36 - Диаграмма последовательности действий а) навигация б) поиск

Каждое действие пользователя инициирует изменение весовых коэффициентов модели предпочтений пользователя и влечет изменение навигационной структуры.

4.4. Апробация мультипредметной информационной системы в рамках документооборота организаций.

Предложенные методы были опробованы в рамках документооборота организаций Мурманской области. СМПО мультипредметной информационной системы на начальном этапе составляет русскоязычный тезаурус WordNet версии 3.0, расширяемый результатами работы семантического анализатора над коллекцией документов организаций. Апробация результатов работы проводилась на предприятии АО «Апатит», а также в рамках документооборота

ФГБОУ ВПО «Петрозаводский государственный университет». В рамках пробации была сформирована СМПО нормативно-справочной информации организации. На рисунке 37 представлены графики, отражающие динамику формирования СМПО.

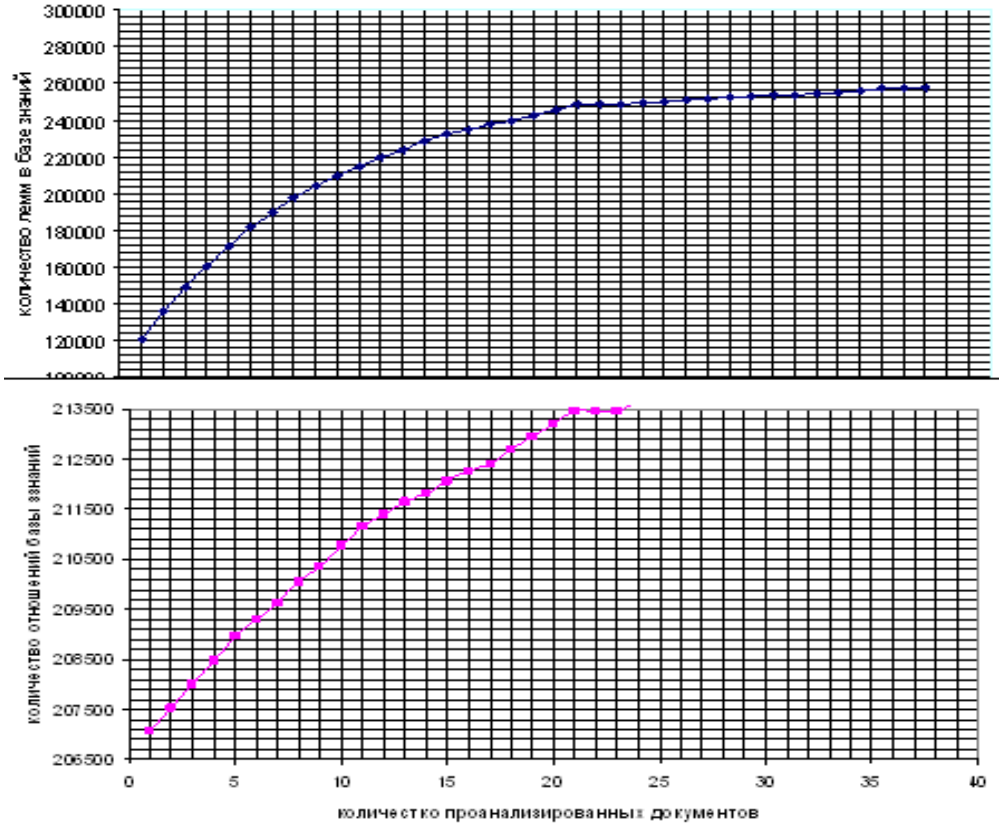


Рисунок 37 - Формирование семантической модели предметной области

С ростом количества обработанных документов сокращается темп роста количества новых (отсутствующих в СМПО) понятий при относительном сохранении динамики увеличения количества отношений. Данное наблюдение позволяет судить о формирования терминологической базы коллекции документов. Всего было проиндексировано около 800 документов, на графиках представлена статистика пополнения СМПО первыми 40 документами.

На рисунке 38 представлен график зависимости встречаемости новых слов в поступающих документах.

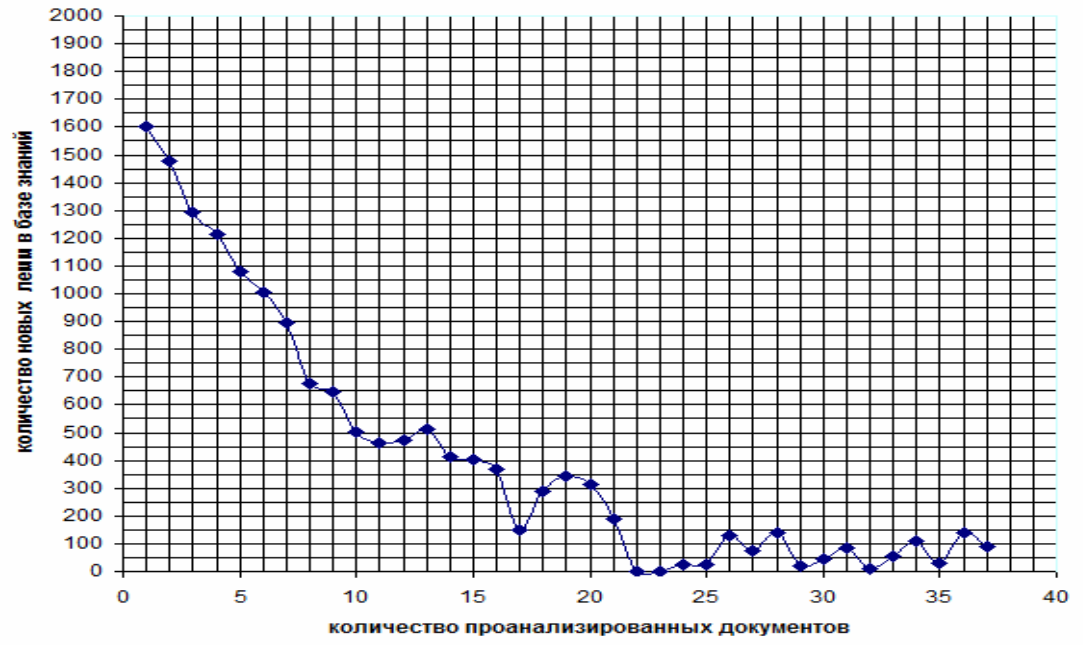


Рисунок 38 - Формирование СМПО

Эффективность предложенных методов была проверена путем натурного эксперимента с привлечением тестовой выборки пользователей, которым предлагалось решить схожие информационно-поисковые задачи с помощью существующих систем поиска с использованием интерфейса строки ввода и с помощью адаптивного интерфейса мультипредметной информационной системы.

Для оценки характеристик адаптивного пользовательского интерфейса был использован метод GOMS[85], заключающиеся в сравнении временных характеристик выполнения пользователями производственных задач в рамках традиционного и разрабатываемого интерфейса. Целью пользователей являлось получение доступа к требуемому документу. Среднее время выполнения операции представлены в таблице 3.

Таблица 3 – Оценка показателей методов поиска

Оцениваемые характеристики	Навигация с использованием традиционного интерфейса	Навигация с использованием адаптивного интерфейса
Среднее время, затраченное на навигационный поиск, с.	31,8	20,85

Для оценки эффективности предложенного метода поиска осуществлялся поиск по заранее проиндексированной коллекции объемом 14 тыс. документов. Экспертами оценивались результаты выполнения 10 запросов. В качестве критериев оценки выступали скорость поиска – время, затраченное на удовлетворение информационной потребности, выраженной одним запросом; точность – соответствие результатов запросу; и полнота результатов – полнота охвата документов с упоминанием об объекте поиска.

$$Precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}, \quad \text{Recal} = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}, \tag{38}$$

где D_{rel} - множество релевантных документов в базе индексов, D_{retr} - множество документов, найденных системой.

Для оценки альтернатив экспертам была предложена лингвистическая шкала измерений.

Оценка *i-й* альтернативы производилась *j-м* экспертом по формуле:

$$v_{ij} = 1 - \frac{(l-1)}{k}, \tag{39}$$

где l – индекс значения лингвистической шкалы; k – количество значений этой шкалы.

Для оценки *i-й* альтернативы n экспертами используется формула:

$$s_i = \sum_{j=1}^{n} v_{ij} \tag{40}$$

Результаты оценки характеристик приведены в таблице 4.

Таблица 4 - Оценка характеристик методов поиска

Оцениваемые характеристики	Используемый метод поиска документов	
	Поиск по ключевым словам с использованием интерфейса строки ввода, нормированная оценка	Поиск с использованием метода поиска информации с учетом субтрактивных отношений, нормированная оценка
Оценка скорости выполнения поиска	0,5	0,8
Точность результатов	0,9	0,9
Полнота результатов	0,7	0,9
Среднее значение оценок	0,7	0,9

Высокая точность поиска с использованием интерфейса строки ввода обусловлена знакомством пользователей с предметной областью и, как следствие, малой неопределенностью при формировании запроса, а также относительно малым объемом коллекции документов. Тем не менее, результаты экспериментов позволяют сделать вывод о корректности и обоснованности использования динамической автоматически формируемой СМПО для

реализации адаптивного интерфейса и метода поиска информации с учетом субтрактивных отношений в корпоративных информационных системах.

Выводы по главе 4

В главе представлены особенности реализации методов формирования и функционирования мультипредметных информационных систем, полученных в предыдущих главах.

Реализован информационный поиск с автоматическим расширением запроса, и ограничением области поиска на основе модели предпочтений пользователей. Разработанная форма поиска позволяет реализовать одномоментное восприятие, интерактивность (корректировка запроса в процессе работы с пользователем), а также возможность ограничения области поиска путем удаления отображаемых концептов. Отображаемые вершины графа символизируют понятия СМПО, соответствующие расширенному запросу пользователя.

Сервис семантического индексирования формирует семантические образы документов в виде множества понятий и отношений между ними. Для формирования семантической модели предметной области были использованы грамматический словарь русского языка Андрея Анатольевича Зализняка, содержащий приблизительно 100 тыс. базовых словоформ русского языка с их полным морфологическим описанием, и русскоязычный тезаурус WordNet. Проведенный эксперимент по формированию СМПО организаций показал достаточность наличия базовых единиц русского языка (при применении авторского метода формирования семантической модели предметной области) в задачах информационного и навигационного поиска.

Разработанный сервис интерфейсной навигации обеспечивает выработку навигационной структуры интерфейса, адаптировано для различных категорий пользователей, а также формирование формализованной ментальной модели категории пользователей. Формирование ментальной модели происходит путем

учета статистики использования концептов СМПО в запросах пользователя, а также при корректирования расширенного запроса в форме поиска. Навигационная структура в свою очередь формируется на основе модели предпочтений пользователя и СМПО.

Результаты опытного внедрения показали, что разработанная мультипредметная информационная система позволяет эффективнее и качественнее находить требуемую информацию, что подтверждает достоверность теоретических исследований.

ЗАКЛЮЧЕНИЕ

В диссертационной работе содержится решение научной задачи разработки методов интерфейсной навигации и поиска нормативно-справочных документов в корпоративных информационных системах. В ходе исследования получены следующие результаты:

1. Разработан метод автоматизированного формирования семантической модели предметной области информационной системы организаций на основе принципа «пользователь как эксперт», заключающийся в интеграции на основе модифицированной составной семантической метрики разнородных источников знаний и последующего уточнения знаний пользователями;
2. Разработан метод поиска информации на основе формализованных знаний, учитывающий весовые коэффициенты отношений семантической модели предметной области для различных категорий пользователей и субтрактивные отношения, ограничивающие область поиска;
3. Разработаны метод интерфейсной навигации, реализующий динамическое формирование адаптивного интерфейса, реализующего обратную связь с пользователем;
4. Создан комплекс программных средств для повышения эффективности доступа к документам организаций, отличающийся использованием методов, способных к автоматическому уточнению и адаптированному представлению информации организаций. Результаты проведённого анализа эффективности разработанных методов мультипредметных корпоративных информационных систем с использованием разработанного программного обеспечения показали, что при осуществлении информационного поиска обеспечивается сокращение времени, необходимое для доступа к необходимой информации примерно в 1,5 раза, и увеличение полноты результатов примерно в 1,3 раза при сохранении точности результатов информационного поиска.

Полученные результаты соответствуют п. 3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем» и п. 7 «Человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения.» паспорта специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Определения

В настоящей диссертационной работе применяют следующие термины с соответствующими определениями.

Информационный ресурс — источник информации, пригодный для удовлетворения информационных потребностей какого-либо лица.

Информационный элемент — часть информационной системы.

Контекст понятия — семантическое окружение понятия в семантической модели предметной области.

Онтология — формально представленные на базе концептуализации знания, предполагающие описание множества объектов и понятий, знаний о них и связей между ними.

Сервис — механизм для предоставления доступа к одной или многим возможностям информационной системы, в котором доступ предоставляется с помощью описанного заранее интерфейса.

Семантическая сеть — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними.

Тезаурус – разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т. п.) между лексическими единицами.

Веб-сервис — программная система, идентифицируемая строкой URI, чьи общедоступные интерфейсы определены на языке XML.

Список обозначений и сокращений

НСИ – Нормативно-справочная информация

БД – База данных

СУБД – Система управления базами данных

АСУТП – Автоматизированные системы управления технологическими процессами

УЗ – Управление знаниями

ЭС – Экспертная система

ЕЯ – Естественный язык

КМ – Knowledge Management

NLP – Natural Language Processing

HTML – HyperText Markup Language

URL – Uniform Resource Locator

URI – Universal Resource Identifier

XML – Extensible Markup Language

RDF – Resource Definition Framework

OWL – Web Ontology Language

SPARQL – Protocol and RDF Query Language

SQL – Structured query language

MES – Manufacturing Execution System

ERP – Enterprise Resource Planning

OLAP – On-Line Analytic Processing

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Александров, В.В. Инфологическая система формирования семантических понятий инвариантных по отношению к естественно-языковому окружению в Интернет среде / Александров В.В. Кулешов С.В., Цветков О.В, Левашкин С.П. // Программируемые инфокоммуникационные технологии. Сборник статей/ Под ред. В.В.Александрова, В.А.Сарычева. – М.:Радиотехника, 2009. – С. 5-10.

2. Александров, В.В. Цифровая технология инфокоммуникации. Передача, хранение и семантический анализ текста, звука, видео/ Александров В.В., Кулешов С.В., Цветков О.В. – СПб.: Наука, 2008. – 244 с.

3. Архивариус 3000 – поиск документов [Электронный ресурс]. – Режим доступа: http://www.likasoft.com/ru/document-search/ (дата обращения: 01.03.2016).

4. Берков, В. Ф. Вопрос как форма мысли / В. Ф. Берков. – Минск.: БГУ, 1972.

5. Бушуева, Л.И. Маркетинговые информационные системы в управлении предприятием [Электронный рессурс] // Вест. научно-исследовательского центра корпоративного права, управления и венчурного инвестирования Сыктывкарского государственного университета. – Режим доступа: http:// www.syktsu.ru (дата обращения: 01.03.2016).

6. Виттих, В. А. Разработка первой очереди системы управления регионом с применением мультиагентных технологий/ Виттих В. А., Светкина Г. Д., Скобелев П. О., Волхонцев Д. В., Гриценко Е. А., Никитин А. Н., Сурнин О. Л., Шамашов М. А.// Труды VI Международной конференции по проблемам управления и моделирования сложных систем, Самара: Самарский научный центр (СНЦ) РАН, 2004, С. 346-351.

7. Виттих, В.А. Применение мультиагентных технологий при создании распределенной системы взаимодействия в сфере государственного

управления/ Виттих В. А., Волхонцев Д. В., Горбенко А. В., Караваев М. А., Ревин П. М., Скобелев П. О., Сурнин О. Л., Шамашов М. А. // Труды VII Международной конференции по проблемам управления и моделирования сложных систем. — Самара: СНЦ РАН, 2005, С. 366-373.

8. Вовченко, А.Е. Анализ и сравнение систем интеграции неоднородных информационных ресурсов/ Вовченко А.Е., Калиниченко Л.А. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 10 Всероссийской научной конференции "RCDL-2008". – Дубна: ОИЯИ, 2008. – С. 246-251.

9. Воройский, Ф.С. Информатика. Новый систематизированный толковый словарь-справочник. – 3-е изд., перераб. и доп. – М.: ФИЗМАТЛИТ, 2003. – 760 с. ISBN 5-9221-0426-8.

10. Гаврилова, Т.А. Визуальные методы работы со знаниями: попытка обзора / Гаврилова Т.А., Гулякина Н.В. // ИИ и принятие решений. 2008. №1. С. 19–33.

11. Гаврилова, Т.А. Технология проектирования интеллектуальных систем/ Гаврилова Т.А., Гулякина Н.В., Голенков В.В. // Информационные системы и технологии (IST’2009): материалы V Междунар. конф.-форума в 2-х ч. Ч.2 – Минск: А.Н. Вараксин, 2009. – С.93-96.

12. Гаврилова, Т.А. Базы знаний интеллектуальных систем /Т.А. Гаврилова, В.Ф. Хорошевский. – СПб. : Изд-во «Питер», 2001. – 382 с.

13. Гаврилова, Т.А. Использование онтологии в системах управления знаниями [Электронный ресурс] – Режим доступа: http://big.spb.ru/publications/bigspb/kni/use_ontology_m_suz.shtml (дата обращения: 01.12.2015).

14. Гаврилова, Т.А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем / Т.А. Гаврилова // Новости искусственного интеллекта, 2003. – №2. – С. 24-30.

15. Гаврилова, Т.А. Извлечение и структурирование знаний для экспертных систем / Гаврилова Т.А., Червинская К.Р. //. – М.: Радио и связь, 1992.– 200 с.

16. Голенков, В.В. Графодинамические модели параллельной обработки знаний: принципы построения, реализации и проектирования / Голенков В. В., Гулякина Н.А. // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS–2012): материалы II Междунар. научн.-техн. конф. редкол. : В. В. Голенков (отв. ред.) [и др.]. –Минск : БГУИР, 2012. – 548 с.

17. Диковицкий, В.В. Извлечение знаний пользователя и верификация знаний самоорганизующихся информационных систем с обратной связью //Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS–2012): материалы II Междунар. научн.-техн. конф. редкол. : В. В. Голенков (отв. ред.) [и др.]. –Минск : БГУИР, 2012. – С. 205-206.

18. Диковицкий, В.В. Современные методы создания мультипредметных веб-ресурсов на базе визуализации и обработки формализованной семантики/ Диковицкий В.В., Ломов П. А., Сепеда-Эррера Р. Р., Шишаев М. Г. // Вестник Кольского научного центра РАН. 2011. №3. С.63-73.

19. Диковицкий, В.В. Концепция двунаправленного семантического поиска / Диковицкий В.В., Ломов П.А. , Шишаев М.Г. //Прикладные проблемы управления макросистемами: мат. VIII Всерос. Школы - семинара, 29 марта – 2 апреля 2010 г.

20. Диковицкий, В.В. Метод визуального семантического поиска на базе семантической сети с субтрактивными связями: Современные проблемы прикладной информатики: сб. науч. трудов междунар. науч.-практ. конф. 25-27 мая 2011 г. / Диковицкий В.В., Шишаев М.Г. // отв. ред. И.А. Брусакова, И.Л. Андреевский. – СПб.: Изд-во Политехн. Ун-та, 2011 г. сс.158-161

21. Диковицкий, В.В. Обработка текстов естественного языка в моделях поисковых систем / Диковицкий В.В., Шишаев М.Г. // Труды Кольского научного центра РАН. Информационные технологии. – 2010. – С. 29-34.

22. Диковицкий, В.В. Применение метода семантического поиска на основе семантической сети с субтрактивными связями для реализации сервисов интернет-портала/ Диковицкий В.В., Шишаев М.Г.// Интеллектуальные системы и технологии: современное состояние и перспективы. Сборник научных трудов Международной летней школы-семинара по искусственному интеллекту для студентов, аспирантов и молодых ученых (Тверь – Протасово, 1-6 июля 2011 г.) – Тверь: Изд-во Тверского государственного технического университета, 2011. С. 196-200

23. Диковицкий, В.В. Технология формирования адаптивных пользовательских интерфейсов для мультипредметных информационных систем промышленных предприятий/ Диковицкий В.В., Шишаев М.Г. //

Информационные ресурсы России. 2014. № 1(137). С. 23–26.– ISSN 0204–3653.

24. Диковицкий, В.В. Система интеграции ВЕБ-ресурсов: Модуль динамического формирования поисковых запросов: тез. докл. XII Межрег. научно-практ. конф.,15-17 апр. 2009 г. – Апатиты: КФ ПетрГУ, 2009. –Ч.1. – С.8.

25. Диковицкий, В.В. Метод информационного поиска на основе динамической расширяемой базы знаний // Труды Кольского научного центра РАН. 4/2012(11). Информационные технологии. Выпуск.3, С. 85-88.

26. Диковицкий, В.В. Методы интеллектуальной обработки и представления информации в мультипредметных информационных системах промышленных предприятий // Труды СПИИРАН. 2015. Вып. 42. С. 56-76.

27. Диковицкий, В.В. Семантическое профилирование пользователей в задаче информационного поиска // Труды Кольского научного центра РАН.

Информационные технологии. –Вып.6. –3/2015(29). –Апатиты: Издательство КНЦ РАН, 2015. – С.54-58- ISBN 978-5-91137-317-7

28. Добров, Б.В. Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие / Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. // – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2008. – 172 с.

29. Епифанов, М.Е. Индуктивное обобщение в ассоциативных сетях / М. Е. Епифанов // Известия АН СССР. Техническая кибернетика.– 1984.– №5.– С.132-146.

30. Жихарев, А.П. Методология интеграции и государственного регулирования информационных ресурсов (Региональный аспект): Автореферат диссертации на соискание ученой степени доктора экономических наук – Москва, 2008.

31. Журавлев, С.В. УИС «РОССИЯ».Автоматическое тематическое индексирование полнотекстовых документов / С.В. Журавлев, Б.В. Добров //Материалы научно-практической конференции «Проблемы обработки больших массивов неструктурированных текстовых документов», 2001.

32. Золотова, Г.А. Коммуникативная грамматика русского языка / Г.А. Золотова, Н. К. Онипенко, М. Ю. Сидорова //Институт русского языка РАН им. В. В. Виноградова. – М., 2004 – 544 с.

33. Иберла, К. Факторный анализ. Пер. с нем. В. М. Ивановой. М.: Статистика, 1980 – 398 с.

34. Избачков, Ю.С. Информационные системы: Учебник для вузов. / Избачков Ю.С., Петров В.Н. // – СПб.: Питер, 2005.–656 с.

35. Калиниченко, А.В. Сущность проблемы анализа текста в полнотекстовых поисковых системах. Подходы и пути решения [Электронный ресурс] – Режим доступа: http://www.jurnal.org/articles/2010/inf12.html (дата обращения: 01.12.2015).

36. Клещев, А.С. Математические модели онтологий предметных областей. Часть 3. Сравнение разных классов моделей онтологий. / Клещев А.С., Артемьева И.Л. // Научно–техническая информация, Сер. 2. Информационные процессы и системы, № 4, 2001, С. 10–15.

37. Когаловский, М.Р. Интеграция данных в информационных системах. // Стандарты в проектах современных информационных систем. Сб. трудов III-й Всероссийской практической конф., М., 2003, С. 83-85.

38. Когаловский, М.Р. Перспективные технологии информационных систем / М.Р. Когаловский. –М.: Компания АйТи, 2003. – 288 с.

39. Коробейников, П.А. Исследование семантической структуры навигационных интерфейсов типовых веб - ресурсов / П.А. Коробейников, М.Г. Шишаев // Труды Кольского научного центра РАН. Информационные технологии. –Вып. 4. – 5/2013. –С.98-102.

40. Кудинов, А. Хранилище данных как основа корпоративной интеграции // PC Week/RE, 2006, №20.

41. Кулешов, С.В. Ассоциативно-онтологический подход к обработке текстов на естественном языке/ Кулешов С.В., Зайцева А.А., Марков В.С.// Интеллектуальные технологии на транспорте. 2015. №4 С.40-45.

42. Кучуганов, В. Н. Элементы теории ассоциативной семантики // Управление большими системами. Сборник трудов. Выпуск 40: М.: ИПУ РАН, 2012. – 328 с. С. 30-48.

43. Лифшиц, Ю. Модели информационного поиска [Электронный ресурс] – Режим доступа: http://yury.name/internet/03ianote.pdf (дата обращения: 05.05.2016).

44. Ломов, П.А. Технология упрощенного представления OWL онтологий для их использования в графических пользовательских интерфейсах / Ломов П. А., Шишаев М. Г., Диковицкий В. В. // Сборник трудов конференции «Инженерия знаний и технологии семантического веба – 2012». – СПб: НИУ ИТМО, 2012. – С. 54 - 66.

45. Ломов, П.А. Онтологическая модель государственного и муниципального управления для проведения семантической интеграции информационных ресурсов/ Ломов П.А. , Диковицкий В.В., Шишаев М.Г. // Прикладные проблемы управления макросистемами: сборник научных трудов . Российская Академия Наук (М.), Институт системного анализа; ред.: Ю. С. Попков, В. А. Путилов. – М. : КРАСАНД : URSS. – 2010. С. 118-132

46. Ломов, П.А., Преобразование OWL–онтологии для визуализации и использования в качестве основы пользовательского интерфейса / П.А.Ломов, М.Г. Шишаев, В.В. Диковицкий // Научный журнал «Онтология проектирования» – №3–2012. – Самара: Новая техника, 2012, С. 49-61 ISSN 2223-9537

47. Ломов, П.А. Метод и технологии семантической обработки информации для государственного и муниципального управления: дис. кандидата технических наук: 05.13.10 / Ломов Павел Андреевич; [Место защиты: ИСА РАН].– Москва, 2011.– 178 с.

48. Некрестьянов, И.С. Латентно–семантический анализ: Введение в латентно–семантический анализ [Электронный ресурс] – Режим доступа: http://meta.math.spbu.ru/~igor/papers/lsa-prg/node2.html (дата обращения 02.02.2016).

49. Системы управления знаниями: обзор [Электронный ресурс] // CRN/RE («ИТ-бизнес»), выпуск №17 (190). – Режим доступа: http://meta.math.spbu.ru/~igor/papers/lsa-prg/node2.html (дата обращения: 01.12.2015).

50. Осипов, Г.С. Семантический поиск в сети Интернет средствами поисковой машины Exactus/ Осипов Г.С., Тихомиров И.А., Смирнов И.В. // Труды 11-ой национальной конф. по искусственному интеллекту КИИ-2008. – 2008. – С. 323-328.

51. Осипов, Г.С. Методика оценки эффективности систем информационного поиска / Осипов Г.С., Выборнова О. Е., Завьялова О.С.,

Смирнов И.В., Тихомиров И.А. // Сборник трудов VI международной конференции Интеллектуальный Анализ Информации ИАИ'2006, г. Киев, С. 214-227.

52. Осипов, Г.С., Реляционно-ситуационный метод поиска и анализа текстов и его приложения/ Г.С. Осипов, И.В. Смирнов, И.А. Тихомиров// ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРИНЯТИЕ РЕШЕНИЙ. № 2.–2008. С.3-10.

53. Осипов, Г.С. Семантический поиск в сети интернет средствами поисковой машины Exactus [Электронный ресурс] /Г.С. Осипов, И.А. Тихомиров, И.В. Смирнов// – Режим доступа: http://www.raai.org/cai-08/files/cai-08_exhibition_31.doc (дата обращения: 01.12.2015).

54. Поспелов, Д.А. Данные и знания. Представление знаний // Искусственный интеллект. Кн.2: Модели и методы: Справочник – М.: Радио и связь. – С.7-13.

55. Поликарпов, А.А. Опыт построения контекстуального словаря и анализ его устройства/ Поликарпов А.А., Бушуева О.В. //Теоретические и практические проблемы прикладной лингвистики.– М., 1988.– 76 с.

56. Скворцов, Н. А. Вопросы согласования неоднородных онтологических моделей и онтологических контекстов. Труды Симпозиума «Онтологическое моделирование», г. Звенигород, 19–20 мая 2008 г. Ред. Калиниченко Л.А. – М: ИПИ РАН, 2008. – С. 149-166. – ISBN 978-5-902030-54-6

57. Смирнов, А.В. Онтологии в системах искусственного интеллекта: способы построения и организации / А. В. Смирнов, М. П. Пашкин, Н. Г. Шилов, Т. В. Левашова // Новости искусственного интеллекта.—2002.—№1. - С.3–13

58. Солтон, Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979.

59. Серебряков, В.А. Основы конструирования компиляторов / В.А. Серебряков, М.П. Галочкин М.: Едиториал УРСС, 2001. – 224 с.

60. Статистические и динамические экспертные системы: Учебное пособие/ Э.В. Попов, И.Б. Фоминых, Е.Б. Кисель, М.Д. Шапот. – М.: Финансы и статистика, 1996. –320 с.

61. Тихонов, В. Архитектура метапоисковых систем [Электронный ресурс] – Режим доступа: http://www.cmsmagazine.ru/library/items/internet_info/metasearch/ (дата обращения: 05.05.2016).

62. Тузовский, А. Ф. Разработка систем управления знаниями на основе единой онтологической базы знаний // Известия ТПУ. №2. – 2007

63. Черняк, Л. Интеграция данных: синтаксис и семантика. // Открытые системы 10/2009– С.24-30.

64. Шишаев, М.Г. Использование концепции «User as an expert» в разработке мультипредметных веб-ресурсов, основанных на онтологиях / М.Г. Шишаев, П.А. Ломов, В.В. Диковицкий // Труды ИСА РАН: Информационные технологии. Системное моделирование. Численные методы решения. Компьютерный анализ текстов. Том 62. Выпуск 3. – М. КРАСАНД, 2012, С. 40-47.

65. Шишаев, М.Г. Использование онтологий для независимого от реализации представления бизнес-логики прикладной ИС. / Шишаев М.Г. Диковицкий В.В., Попова Л.П., //Информационные технологии в региональном развитии .Под редакцией Путилова В.А. – Апатиты: издательство КНЦ РАН, 2009. – Вып. IX. – С. 51-55

66. Шишаев, М.Г. Исследование семантической структуры навигационных интерфейсов типовых веб-ресурсов/ П.А. Коробейников, М.Г. Шишаев// Труды КНЦ РАН 5/2013. С.98-102

67. Шишаев, М.Г. Формализация задачи построения когнитивных пользовательских интерфейсов мультипредметных информационных ресурсов/

Диковицкий В.В., Ломов П.А, Шишаев М.Г. // Вестник Кольского Научного Центра 3/2011 - С. 62-72.

68. Электронная версия словаря А.А. Зализняка [Электронный ресурс]. – Режим доступа: http://www.morfologija.ru (дата обращения: 20.04.2016).

69. Federated Search. Интерфейс полнотекстового поиска в нескольких репозиториях [Электронный ресурс] – Режим доступа: www.coredge.com (дата обращения: 05.11.2015).

70. IBM Collaboration Solution. Система коммуникации предприятия [Электронный ресурс] – Режим доступа: www.lotus.com/km (дата обращения: 05.11.2015).

71. MESA. Международное сообщество компаний-производителей [Электронный ресурс] – Режим доступа: www.mesa.org (дата обращения: 05.11.2015).

72. Acker, L. Extracting viewpoints from knowledge bases / L. Acker, B. Porter // The 12th National Conference on Artificial Intelligence, – 1994. Pp. 547–552.

73. Agichtein, E. Improving web search ranking by incorporating user behavior information/ E. Agichtein, E. Brill, S. Dumais// SIGIR ’06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. – USA,2006. – Pp. 19–26.

74. Alsmadi, I. GUI Structural Metrics / I. Alsmadi, M. Al-Kabi // The International Arab Journal of Information Technology, vol. 8, No. 2. – 2011. – Pp. 124-129.

75. Arens, Y. «Query Processing in the SIMS Information Mediator» / Y. Arens, C. Knoblock, N. Hsu N // Advanced Planning Technology: AAAI Press. – 1996. – Pp. 61-69.

76. Atkinson, K. E. An Introduction to Numerical Analysis, Second Edition. John Wiley, New York – 1989.

77. Averbukh, V.L. Toward formal definition of conception adequacy in visualization. // IEEE Symposium on Visual Languages. Italy – 1997. Pp. 46–47

78. Al-Maskari, A. A review of factors influencing user satisfaction in information retrieval / Azzah Al-Maskari, Mark Sanderson // Journal of the American Society for Information Science and Technology Journal of the American Society for Information Science and Technology Volume 61, Issue 5. – 2010. Pp. 859–868.

79. Baader, F. The description logic handbook: theory, implementation and applications / F. Baader, D. McGuinness, D. Nardi, Patel Schneider. – Cambridge University Press, 2003.

80. Balkova,V. Russian WordNet. From UML-notation to Internet/Intranet Database Implementation/ V. Balkova, A. Sukhonogov, S. Yablonsky // Proceedings GWC 2004. –Masaryk University – 2004. – Pp. 31–38.

81. Baeza-Yates, R. Modern Information Retrieval / R. Baeza-Yates , B. Ribeiro-Neto // ACM Press Series/Addison Wesley, New York, 1999. – 513 p.

82. Berners-Lee, T. The Semantic Web / T. Berners-Lee, J. Hendler, O. Lassila // Scientific American. – 2001. Pp. 29-37.

83. Bevan, N. International Standards for HCI and Usability // International Journal of Human-Computer Studies.– 2001.– 55 (4). – P. 533-552.

84. Bevan, N., Measuring usability as quality of use // Journal of Software Quality Issue.– 1995.– P.115-140.

85. Bonsiepe, G.A., A Method of Quantifying Order in Typographic Design, Journal of Typographic Research, Vol. 2. – 1968. –Pp.203-220.

86. Brin, S. The Anatomy of a Large-Scale Hypertextual Web Search Engine [Электронный ресурс]/ Sergey Brin, Lawrence Page// Stanford InfoLab – Режим доступа: http://infolab.stanford.edu/pub/papers/google.pdf (дата обращения: 01.12.2015).

87. Brooke, J. SUS: A ‘quick and dirty’ usability scale / J. Brooke, P. W. Jordan, B. Thomas, B.A. Weerdmeester, I.L. McClelland // Usability evaluation in industry.– London – 1996. –Pp.189-194.

88. Chai, W. Using User Models in Music Information Retrieval Systems / W. Chai, B. Vercoe // Proceedings of ISMIR. 2000.

89. Davenport, T. H. Saving IT's Soul: Human Centered Information Management / *Harvard Business Review* **72** (2). – 1994. Pp. 119–131.

90. Dou, D. Ontology translation by ontology merging and automated reasoning/ Dou D. McDermott D. Qi P. // EKAW’02 workshop on Ontologies for Multi-Agent Systems. Spain, – 2002. – pp. 3 - 18.

91. Etzold, T. SRS: Information retrieval system for molecular biology data banks/ Etzold T., Ulyanov A., Argos P. // Methods in Enzymology, Volume 266. – 1996. Pp. 114-128.

92. Extensible Markup Language (XML) 1.0, W3C Recommendation 10.02.1998. [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/1998/REC-xml-19980210 (дата обращения: 07.10.2015).

93. Gangemi, A. An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies / Gangemi A, Pisanelli DM, Steve G.// Data & Knowledge Engineering 31(2). – Pp.183–220.

94. Gennari, JH. The evolution of Protégé: an environment for knowledge-based systems development / Gennari JH, Musen MA, Fergerson, RW // International Journal of Human-Computer Studies, vol. 58. – 2003. – pp. 89-123.

95. Gruber, T. A translation approach to portable ontology specifications // Knowledge Acquisition, Vol. 5. – 1993. – Pp. 199 - 220.

96. Gruber, T.R. A Translation Approach to Portable Ontology Specifications [Электронный ресурс]. – Режим доступа: http://tomgruber.org/writing/ontolingua-kaj-1993.pdf. (дата обращения: 07.10.2015).

97. Guarino, N. Formal Ontology and Information Systems // Proc. 1st Int'l Conference on Formal Ontology in Information Systems. – 1998.

98. Guha, R. V. Semantic search / Guha, R. V., McCool, R., and Miller, E. // Proc. of the 12th International World Wide Web Conference (WWW 2003). Hungary – 2003. – Pp. 700-709.

99. Kelly, D. Methods for Evaluating Interactive Information Retrieval Systems with Users. [Электронный ресурс] / Foundations and Trends in Information Retrieval: Vol. 3: No. 1–2, Pp. 1-224. – Режим доступа: http://dx.doi.org/10.1561/1500000012 (дата обращения: 07.12.2015).

100. Kuropka, D. Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und -Retrieval mit relationalen Datenbanken, 2004, 242 p. –ISBN 3-8325-0514-8.

101. Lenzerini M. Data Integration: A Theoretical Perspective" // In PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems – 2002 – pp. 233-246.

102. Levy, A. Logic-based techniques in data integration // Logic-Based Articial Intelligence, Kluwer Academic Publishers, Dordrecht – 2000 – Pp. 575 - 595.

103. Lewis, J. Evaluation of Procedures for Adjusting Problm-Discovery Rates Estimated from Small Samples // The International Journal of Human-Computer Interaction 13(4). –2001.–Pp.445-479.

104. Liawa, Information retrieval from the World Wide Web: a user-focused approach based on individual experience with search engines / Computers in Human Behavior 22. – 2006. Pp. 501–517.

105. Lenzerini, M. «Data Integration: A Theoretical Perspective» // In Proc. PODS'02. pp. 233-246.

106. Wettler, M. Cognitive processes in information retrieval: production rules and lexical nets/ Wettler, M., Glockner-Rist,A. // Mental Models and Human-Computer Interaction. – 1991. Pp. 243-255.

107. Manning, C. Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schütze // Cambridge University Press. – 2008. — ISBN 0-521-86571-9

108. Manolescu, I.XML Queries over Heterogeneous Data Sources / Manolescu I., Florescu D., Kossman D. Answering // Proceedings Of the 27th VLDB Conference. – Italy. – 2001. – Pp. 241 - 250.

109. Marchionini, G. . Toward Human-Computer Information Retrieval Bulletin [Электронный ресурс]// Bulletin of the American Society for Information Science. Режим доступа: – http://www.asis.org/Bulletin/Jun-06/marchionini.html (дата обращения: 07.12.2015).

110. McGuinness, DL. An environment for merging and testing large ontologies/ McGuinness DL, Fikes R, Rice J, Wilder S // Proc. 7th International Conference on Principles of Knowledge Representation and Reasoning. – 2000. Pp 483-493.

111. Molich, R. Improving a human-computer dialogue / Molich, R., Nielsen, J. // Communications of the ACM 33 , 3 (March). –1990.–Pp.338-348.

112. Nielsen J. A mathematical model of the finding of usability problems / Nielsen J., Landauer T.K. // CHI ‘93 Proceedings of the INTERACT ‘93 and CHI ‘93 Conference on Human Factors in Computing Systems. – Pp.206-213.

113. Noy N.Prompt: Algorithm and tool for automated ontology merging and alignment / Noy N., Musen A // Proceedings of the Seventeenth National Conference on Artificial Intelligence. – 2000. pp. 450–455.

114. OWL – Web Ontology Language. Overview, 2004. [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/owl-features (дата обращения: 01.11.2015).

115. Paton, N. W. Query processing in the TAMBIS Bioinformatics Source Integration System/ Paton, N. W., Stevens, R. D., Baker, P. G // Proceedings of 11th International Conference on Scientific and Statistical Database Management (SSDBM), IEEE Press, pp. 138–147.

116. Fitts, Paul M. The information capacity of the human motor system in controlling the amplitude of movement // Journal of Experimental Psychology, volume 47, number 6. – 1954. –Pp.381–391.

117. Pinto, HS. Some Issues on Ontology Integration/ Pinto HS, Gómez-Pérez A, Martins JP // Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends. Vol. 18, Stockholm, Sweden. – 1999, – pp. 1 - 12.

118. Pinto, HS. A methodology for ontology integration/ Pinto HS, Martins JP.// Proceedings of the International Conference on Knowledge Capture K–CAP2001. – 2001. – Pp. 131-138.

119. Resource Description Framework (RDF) [Электронный ресурс] // RDF Working Group. — 2014. — Режим доступа: http://www.w3.org/RDF/ (дата обращения: 20.04.2016).

120. Sauro J. Benchmarks For User Experience Metrics [Электронный ресурс] – Режим доступа: www.measuringusability.com/blog/ux-benchmarks (дата обращения: 01.11.2015).

121. Shishaev,M.G. Architecture and Technologies of Knowledge-Based Multi-Domain Information Systems for Industrial Purposes / V.V. Dikovitsky, M. G. Shishaev , N. V. Nikulina // Automation Control Theory Perspectives in Intelligent Systems. Proceedings of the 5th Computer Science On-line Conference 2016 (CSOC2016), Vol 3. – 2016. Pp. 359-369.

122. SPARQL Query Language for RDF W3C Candidate Recommendation 14 June 2007. [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/rdf-sparql-query (дата обращения: 05.11.2015).

123. Stickel, C. The XAOS Metric – Understanding Visual Complexityas measure of usability/ Stickel C., Ebner M., Holzinger A // Work &Learning, Life & Leisure, Springer. – 2010. – Pp.278-290.

124. Studer, R. Knowledge Engineering: Principles and Methods/ Studer R., Benjamins V.R., Fensel D. // In Data & Knowledge Engineering, 25, 1998. – Pp. 161 –197.

125. Schwartz, L. Thematic relations and case linking in Russian. Thematic relations (syntax and semantics, 21), ed. by W. Wilkins, New York: Academic Press. 1988. – Pp. 167-189.

126. The Gene Ontology Consortium. 2000. Gene ontology: Tool for the unification of biology. Nat. Genet. 25, pp. 25–29.

127. Thomas, J.J. Illuminating the Path: The Research and Development Agenda for Visual Analytics / J.J. Thomas K.A. Cook (Eds.) / / IEEE Press, – 2005.

128. Turner, C. W. Determining usability test sample size / Turner, C. W., Lewis, J. R., and Nielsen, J.// International Encyclopedia of Ergonomics and Human Factors Boca Raton, FL: CRC Press. – 2006. – Pp.3084-3088.

129. Vakkary, P. eCognition and changes of search terms and tactics during task performance // RIAO'2000.

130. Virzi R. Refining the test phase of usability evaluation: how many subjects is enough?– Human Factors – Special issue: measurement in human factors archive, Volume 34, Issue 4. – 1992. –Pp.457-468.

131. Wache H., An integration method for the specification of rule-oriented mediators / Wache H., Scholz Th., Stieghahn H., Konig-Ries B. // Proceedings of the 1999 International Symposium on Database Applications in Non-Traditional Environments, Kyoto, 1999 – pp. 109 - 112.

132. Wache H. Ontology-Based Integration of Information – A Survey of Existing Approaches / Wache H., Vogele T., Visser U., Stuckenschmidt H.// Proceedings of the IJCAI–2001 Workshop: Ontologies and Information Sharing. – Seattle, WA. – 2001. – Pp. 108–117.

133. Wielinga, B. Framework and Formalism for Expressing Ontologies / B. Wielinga etc.// ESPRIT Project 8145 KACTUS, Free University of Amsterdam Deliverable, DO1b.1. – 1994.

134. Zhang, Y. Undergraduate students' mental models of the Web as an information retrieval system //Journal of the American Society for Information Science and Technology, 59(13). – 2008. Pp. 2087-2098

135. The NIST Definition of Cloud Computing (SP 800-145) [Электронный ресурс]. – Режим доступа: http://csrc.nist.gov/publications/PubsSPs.html#800-145 (дата обращения: 07.04.2016).

ПРИЛОЖЕНИЕ 1. Акты внедрения

Министерство образования и науки РФ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Петрозаводский государственный университет»
Кольский филиал

Факультет информатики и прикладной математики

Справка
об использовании результатов диссертационной работы
на соискание ученой степени кандидата технических наук
Диковицкого Владимира Витальевича

Настоящая справка свидетельствует о том, что результаты диссертационного исследования Диковицкого Владимира Витальевича на тему «Методы интерфейсной навигации и поиска нормативно-справочных документов в корпоративных информационных системах» были апробированы в системе электронного документооборота Кольского филиала федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Петрозаводский государственный университет».

На основе сформированной семантической модели предметной области были апробированы процедуры навигации и выполнения семантического поиска документов. Применение данных результатов позволяет снизить трудозатраты, связанные с подготовкой справочно-информационных документов, а также повысить оперативность выполнения данных процессов.

Зам. декана факультета ИПМ
к.т.н. / Быстров В.В./

«__4__» _декабря_ 2015 г.

Подпись Быстрова Виталия Викторовича
по месту работы удостоверяю.
Ведущий специалист
КФ ПетрГУ Н.С. Петренко
04.12.2015

ФОСАГРО

Акционерное общество «Апатит»
(АО «Апатит»)

184250, Российская Федерация, Мурманская область, город Кировск, ул, Ленинградская, дом 1
Тел.: +7(81531) 3 22 50, Факс: +7(81531) 3 17 02, телетайп 126735 «Лава», e-mail: apatit@phosagro.ru, www.phosagro.ru
ОКПО 00203938, ОКТМО 47712000, ОГРН 1025100561012, ИНН/КПП 5103070023/997350001

_____ № _____

На № _____ от _____

Справка

об использовании результатов исследований, полученных в диссертационной работе Диковицкого В.В. «Методы интерфейсной навигации и поиска нормативно-справочных документов в корпоративных информационных системах»

Результаты диссертационного исследования Диковицкого Владимира Витальевича на тему «Методы интерфейсной навигации и поиска нормативно-справочных документов в корпоративных информационных системах» были апробированы в АО «Апатит».

В рамках апробации результатов исследования на основе межведомственного электронного документооборота была сформирована семантическая модель предприятия. Использование полученной модели позволяет предоставить сотрудникам отдела кадров актуальную информацию в полном объеме для решения задач отдела.

На основе созданной модели были апробированы процедуры формирования запросов на получение данных, в частности выполнения семантического поиска требуемых документов.

Учет мультипредметного характера информации отдела кадров позволяет сократить время доступа к нормативно-справочным документам, что позволяет повысить эффективность работы организации.

Заместитель директора по персоналу
и социальной политике
АО «ФосАгро – Череповец»
в г.Кировске

Дирекция по персоналу и социальной политике

Д.Н. Мартынов