НА ДОМ НЕ ВЫДАЕТСЯ

На правах рукописи

АЮШЕЕВА Наталья Николаевна

# ИССЛЕДОВАНИЕ И РАЗРАБОТКА МОДЕЛЕЙ И МЕТОДОВ ПОИСКА ИНФОРМАЦИОННЫХ ОБРАЗОВАТЕЛЬНЫХ РЕСУРСОВ В ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Н. Аюш –

Москва – 2004

Работа выполнена в Восточно-Сибирском государственном технологическом университете.

| | |
|---|---|
| Научный руководитель: | кандидат технических наук, доцент Найханова Л.В. |
| Официальные оппоненты: | доктор технических наук, профессор Шеремет И.А. |
| | кандидат технических наук, доцент Троицкий И.И. |
| Ведущая организация: | Государственный научно-исследовательский институт информационных технологий и телекоммуникаций «Информика» |

Защита диссертации состоится «*20*» *января* 2005 г. в *14.30* часов на заседании диссертационного совета Д 212.141.10 в зале Ученого Совета Московского государственного технического университета им. Н.Э.Баумана по адресу: 105005, г.Москва, ул. 2-ая Бауманская, д.5.

С диссертацией можно ознакомиться в научной библиотеке МГТУ им. Н.Э.Баумана.

Ваши отзывы в 2-х экземплярах и заверенные печатью, просим высылать по указанному адресу.

Автореферат разослан «9» декабря 2004 г.

Ученый секретарь
диссертационного совета,
к.т.н, доцент Иванов С.Р.

НТБ МГТУ им. Н.Э. Баумана
1698500
Аюшеева Н.Н. Исследование

-501

НА ДОМ НЕ ВЫДАВАТЬ

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**мы.** Исследования в области автоматизации информационного поиска полнотекстовых документов ведутся более тридцати лет. За это время данное направление работ превратилось из узкоспециализированной тематики в одну из ключевых областей информатики. Основоположником теории информационного поиска по праву считается Дж. Солтон. Основные положения этой теории, изложенные им в 70-х гг. XX века, считаются канонами информационного поиска и нашли применение в большинстве существующих поисковых систем.

С появлением и активным использованием глобальной сети Интернет задачи информационного поиска несколько видоизменились, стало необходимым учитывать природу сети Интернет, которой свойственны огромный объем доступной информации, её разнородность, высокий процент временной информации, отсутствие контроля за качеством информации. Все это явилось предпосылками того, что процессы перевода традиционных источников информации в форму ресурсов Сети получили новое «звучание» в плане придания им организации учета (хотя бы частичной), близкой по своей сути к традиционному учету в библиотеках.

В настоящее время в Сети существует множество локальных информационных ресурсов и их организация в единое общедоступное информационное пространство, является составной частью формирования цивилизованного информационного общества. Доступность к информационным ресурсам при наличии систем, обеспечивающих полноту и корректность их поиска, может в существенной степени повлиять на интенсивность и качество образования, а также на качество и объемы научных исследований. Отмеченное выше может быть достигнуто за счет решения трех задач: обеспечения возможности удаленного доступа к информационным ресурсам, оптимизации размещения в сети часто используемой информации и глобализации ресурсов. Анализируя данные задачи можно отметить, что первая задача на сегодняшний день практически решена, при решении второй задачи возникает больше организационных вопросов, чем научных. Рассматривая же третью задачу, можно сказать, что для ее решения необходимо наработать механизм, который бы позволил ресурсу, выставленному в любом месте земного шара, автоматически регистрироваться в глобальной системе учета информационных ресурсов и при этом он должен стать доступным для поисковых систем.

Очевидно, что глобальный учет информационных ресурсов должен вестись по областям знаний, а возможно и по отраслевому принципу. В своей работе мы не ставим перед собой нахождение всеобъемлющего решения данной задачи и ограничимся только решением некоторых вопросов, связанных с образовательной деятельностью. Актуальность этой задачи можно обосновать тем, что в настоящее время в области образования наблюдается «всплеск» появления разнообразных информационных образовательных ресурсов на многочисленных сайтах. Этот факт отчасти объясняется функционированием с 2001 года Федеральной Целевой Программы «Развитие единой образовательной информационной среды». При этом доступность к образовательным ресурсам полностью определяется возможностями

существующих поисковых систем. Хорошо известно, что поисковые системы на запрос пользователя выдают очень большой список электронных документов, имеющихся в Сети. Причем в этом списке, как правило, наблюдается огромная избыточность, что, в свою очередь, не позволяет такой поиск охарактеризовать как эффективный.

Следует отметить, что данная работа появилась в результате выполнения темы «Разработка республиканской электронной библиотеки публикаций научных и образовательных учреждений на базе портала Регионального ресурсного центра информатизации образования Республики Бурятия», выполняемой в рамках федеральной целевой программы «Развитие информационных ресурсов и технологий». В данной работе рассматриваются вопросы разработки единого распределенного центра доступа и хранения информационных ресурсов. Такой центр (в дальнейшем депозитарий) предназначен для хранения качественных информационных образовательных ресурсов. Причем в нем организация поиска должна быть иной, чем в обычных поисковых системах. Она должна обеспечивать минимизацию избыточности найденного множества документов и обладать высокой точностью поиска.

Рассматривая далее вопрос по депозитарию можно отметить, что одну часть информационных образовательных ресурсов составляют полнотекстовые документы, являющиеся электронным представлением бумажных изданий, а вторую часть - программные продукты, ресурсы Интернет и т.п. Исходя из этого, можно сделать вывод о том, что депозитарий должен иметь два вида хранилищ: первое – для полнотекстовых документов и второе – для всех остальных видов ресурсов. Тогда метакаталог должен обеспечивать поиск ИОР, не являющихся полнотекстовыми документами, а поисковая система - поиск полнотекстовых документов. Поэтому на основе накопленных теоретических и практических знаний в области информационного поиска для создания единого распределенного центра доступа и хранения информационных ресурсов необходимо разработать модели метакаталога и поисковой системы, которые должны обеспечивать поиск ИОР по метакаталогу депозитария.

Очевидно, что хорошая поисковая система должна находить все нужные документы и ни одного лишнего, т.е. обладать, в первую очередь, хорошими качественными характеристиками. Для сравнения эффективности различных методов информационного поиска обычно используются два параметра: точность (precision) – доля релевантного материала в ответе поисковой системы и полнота (recall) – доля найденных релевантных документов в общем числе релевантных документов коллекции. Их значения должны стремиться к 100%. Однако стопроцентное качество поиска невозможно, поэтому необходимо разработать методы, позволяющие повысить данные характеристики по сравнению с характеристиками существующих поисковых систем.

Основываясь на вышесказанном, можно определить **цель работы**, которая состоит в исследовании и разработке методов поиска информационных ресурсов, учитывающих их образовательную направленность, а также позволяющих разработать эффективную поисковую систему депозитария.

Для достижения поставленной цели исследования проводились по следующим основным ***направлениям***:

1. Выбор базовой схемы метаописания ИОР и разработка модели данных метакаталога информационных образовательных ресурсов.

2. Исследование и разработка метода индексирования полнотекстового документа, содержащего научные, учебные и учебно-методические материалы.

3. Исследование и разработка метода информационного поиска на основе структурно-лингвистического подхода семантического анализа полнотекстового документа.

**Методы исследования.** В работе использованы методы теории множеств, теории графов, теории искусственных нейронных сетей, искусственного интеллекта.

**Научная новизна** диссертационной работы заключается в том, что:

1. Разработан метод индексирования полнотекстового документа, основанный на оригинальном способе построения семантической сети, позволяющей учитывать семантику документа при формировании его поискового образа.

2. Разработан метод информационного поиска, основанный на сопоставлении графов запроса и поискового образа документа для выявления степени релевантности документа, и позволяющий уменьшить мощность множества релевантных документов, образующих отклик на затребываемую в запросе информацию, за счет применения кластеризации этого множества.

3. Получена модель метакаталога, спецификация которого учитывает образовательную направленность информационных ресурсов, также создана модель поисковой системы, позволяющая повысить точность результатов поиска.

**Практическая ценность** исследования состоит в том, что полученные результаты могут быть применены при разработке двух компонентов депозитария информационных образовательных ресурсов: метакаталога и поисковой системы, удовлетворяющие требованиям, предъявляемым к их разработке, среди которых важнейшим является повышение точности отклика, включающего документы действительно релевантные запросу пользователя.

**Внедрение результатов работы.** Результаты работы в виде разработанного программного и лингвистического обеспечения используются в Межотраслевом НИИ «Интеграл». Кроме того, результаты исследования вошли в материалы отчетов по госбюджетной научно-исследовательской работы «Теоретические и прикладные вопросы разработки интегрированных интеллектуальных информационных систем. Этап: Основные аспекты методологии построения интеллектуальных информационно-поисковых систем» (ГР № 01.200.205060; Инв.№ 02.200305099), по проекту «Разработка республиканской электронной библиотеки публикаций научных и образовательных учреждений на базе портала Регионального ресурсного центра информатизации образования Республики Бурятия» (Научная программа «Развитие информационных ресурсов и технологий», подпрограмма «Оптимизация ресурсного обеспечения системы образования. Индустрия образования»), а также по НИР «Исследование и разработка методов и алгоритмов полнотекстового поиска информации в системе образовательных порталов», выполненной в 2002 году по гранту Правительства Республики Бурятия для молодых ученых.

**Публикации.** Основные результаты диссертационной работы опубликованы в 10 печатных работах общим объемом 5,75 п.л., из которых 1 отчет о НИР, 5 статей, 2 тезиса докладов, 1 учебное пособие и 1 свидетельство об официальной регистрации программы для ЭВМ.

**Апробация результатов исследования**. Основные положения диссертационной работы докладывались и обсуждались на международной научной конференции «Информация-Коммуникация-Общество» (Санкт-Петербург, 11-12 ноября 2003 г.), международной научной конференции «VI Энгельмейеровские чтения» (Москва, 2003 г.), Всероссийской научно-практической конференции «Российская школа и Интернет» (Санкт-Петербург, 2002 г.), Третьей, Четвертой, Пятой Всероссийских научно-технических конференциях «Теоретические и прикладные вопросы современных информационных технологий» (Улан-Удэ, 2002-2004 гг.), Третьей Всероссийской научно-практической конференции-выставке «Единая образовательная информационная среда: проблемы и пути развития» (Омск, 2004 г.), Всероссийской научно-практической конференции «Проблемы качества, безопасности и диагностики в условиях информационного общества» (Сочи, 2004 г.). Материалы диссертации были использованы при подготовке учебного курса «Основы интернет-технологий» и нашли применение в учебном процессе ВСГТУ.

**Структура и объем работы.** Диссертация состоит из введения, четырех глав, заключения, списка литературы и пяти приложений. Работа содержит 158 страниц машинописного текста, 15 рисунков и 26 таблиц. Список литературы содержит 164 наименования. Объем приложений составляет 57 страниц.

## СОДЕРЖАНИЕ РАБОТЫ

Во **введении** приводится обоснование актуальности темы, формулируются основные задачи исследования, кратко излагается содержание работы и перечисляются основные ее результаты, выносимые на защиту.

В **первой главе** работы выполнено описание проблемной ситуации, связанной с созданием единого центра хранения и доступа к ИОР (депозитария), приведены анализ существующих систем сетевого поиска, обзор методов индексирования и информационного поиска, рассмотрены подходы к организации метакаталога и определены основные направления исследования диссертационной работы.

В описании проблемной ситуации дается определение депозитария, уточняется его роль в создании общедоступного информационного пространства, описывается его назначение, а также приводятся функциональные и системные требования к разработке таких его компонент, как метакаталог ИОР и поисковая система. На наш взгляд, в рамках решения задачи глобализации информационных ресурсов для сферы образования должна быть разработана распределенная система депозитариев ИОР. Все ресурсы, разрабатываемые на региональном уровне, должны регистрироваться в метакаталоге регионального депозитария и храниться в регионе. При этом должна быть предусмотрена автоматическая регистрация нового ИОР в «едином» метакаталоге федерального депозитария. В свою очередь региональные депозитарии с заданной периодичностью должны обращаться к «единому» метакаталогу и описания новых ИОР копировать в собственный.

В данной работе рассматривается задача создания метакаталога регионального депозитария, модель которого разрабатывается в рамках первого направления исследования. Для этого рассмотрены подходы к организации метакаталога, средства и технологии его создания. Идея метакаталога, как организующего и управляющего звена в информационной системе, легла в основу предложенной

нами интерпретации метакаталога. В работе под метакаталогом понимается каталог метаописаний объектов информационной деятельности. На наш взгляд метаописание должно заполняться при регистрации и помещении ресурса в хранилище. При этом оно должно учитывать требования всех категорий пользователей депозитария, которые различаются по уровням образования и видам профессиональной деятельности, и отражать область знаний, виды учебных занятий и т.п. Поэтому при создании метакаталога информационных ресурсов образовательного характера должны быть учтены многие аспекты публикуемых научных и учебно-методических материалов сферы образования. Кроме того, спецификация метаописания информационных ресурсов должна поддерживать существующие стандарты на описание метаданных. К основным требованиям, которые нужно выполнять при разработке модели данных спецификации метаописания, относятся: однозначное определение каждого ИОР; адекватность описываемому ресурсу; возможность осуществления мониторинга использования ИОР; открытость спецификации. Это обусловило применение методологии Information Engineering при разработке модели данных.

Для обеспечения доступа к затребованному ИОР, зарегистрированному в метакаталоге, необходимо разработать систему поиска. В работе в результате анализа состояния современных информационно-поисковых систем выполнена их классификация. При этом рассмотрены поисковые каталоги, поисковые и метапоисковые машины, которые делятся в зависимости от назначения на многоцелевые и специализированные, в зависимости от объекта и цели поиска – на фактографические и документальные, от архитектуры – на системы поиска с централизованной архитектурой и распределенные системы. Аналитический обзор систем поиска показал, что традиционно сетевые поисковые системы состоят из двух компонент, которые обеспечивают соответственно автономный и распределенный поиск. Предлагаемая организация метакаталога обуславливает наличие только одного компонента в поисковой системе депозитария, который осуществляет автономный поиск. При этом метакаталог, по которому выполняется поиск, является централизованным, хотя по своей архитектуре депозитарий относится к распределенным системам с децентрализованной архитектурой. Создаваемая модель поисковой системы имеет документальный тип и специализированный характер, так как поисковая система осуществляет поиск информационных образовательных ресурсов, имеющих вид полнотекстовых документов.

На основе вышеизложенного поисковая система должна обеспечивать поиск полнотекстовых документов в коллекции метакаталога, их ранжирование по степени релевантности запросу пользователя и выборку документов. При этом система поиска должна осуществлять выборку документов, имеющих наиболее высокую степень релевантности, а мощность множества отобранных документов должна быть минимальной за счет уменьшения числа «шумовых» документов, попадающих в отклик.

В рамках первой главы проведен обзор моделей поиска: булева модель из теоретико-множественных моделей, векторная модель из алгебраических и вероятностная модель. Каждая модель поиска, как правило, состоит из двух частей. Первая часть отвечает за формальное представление документа и запроса в виде их

поисковых образов на основе использования методов индексирования. В связи с этим в работе рассмотрены существующие методы индексирования. Наибольшее распространение получили методы, использующие для описания содержания документа ключевые слова, а также статистические и синтаксические закономерности естественно-языковых текстов. В этой группе методов внимание уделено лингвистическим методам, включающим интуитивно-прагматические, синтаксические, семантические методы. Сделан вывод, что наиболее эффективным является применение комбинации статистических методов с другими методами, из которых нами выбран семантический метод, позволяющий отразить в поисковом образе семантику индексируемого документа. Вторая часть модели направлена на поиск релевантных документов, их ранжирование и формирование результата поиска в виде отклика. В этой связи рассмотрены методы индексного (двоичного) поиска, статистические методы и методы поиска, основанные на базах знаний, которые соответствуют той или иной модели информационного поиска.

Методы поиска и индексирования документов должны быть согласованы. Для того, чтобы отклик имел хорошие показатели точности и полноты необходимо построить поисковый образ документа, отражающий его содержание. Наиболее адекватно передает смысл документа семантическая сеть. Поэтому для построения формальной модели поиска использована векторная модель, как модель, подходящая для описания графовых структур.

Таким образом, для построения поисковой системы необходимо сформировать поисковый образ документа $I^D$ и поисковый образ запроса $I^Q$. $I^D$ играет роль прообраза $I^Q$. Необходимо найти такой способ интерпретации $I^Q$ на $I^D$, который позволяет найти наименьшее покрытие множества $D^R$, включающее в себя документы $d_i$, обладающие наилучшими характеристиками степени релевантности $\alpha^D$. При этом областью поиска является метакаталог депозитария ИОР, в котором хранится множество $\{I^D\}$. Для представления поискового образа использована модель представления знаний в виде семантической сети, которая позволяет достаточно точно отобразить смысловое содержание полнотекстового документа на основе выделения термов текста и анализа отношений между ними.

В соответствии с вышеизложенным, второе направление исследований в работе связано с разработкой метода построения поискового образа документа, а третье – с формированием отклика, включающего минимальное количество «шумовых» документов.

Во **второй главе** рассмотрены вопросы разработки модели метакаталога информационных образовательных ресурсов. Для решения этих вопросов предложена спецификация метаописания информационных образовательных ресурсов, описана классификация ИОР, разработана модель данных спецификации. Описываются основные функции метакаталога ИОР, в том числе традиционные методы поиска по нему: атрибутный, контекстный и их комбинация.

Спецификация метаданных представляет собой набор элементов метаданных, содержащих описание электронных ресурсов, которое может распознать и проинтерпретировать вычислительная система. Значимость метаданных обуславливается рядом задач, решаемых ими для широкого использования информационных ресурсов в различных видах деятельности: нахождение нужной информации; предоставление информации пользователю в удобной для него форме;

обеспечение прав собственности создателям информационных ресурсов; возможность сопровождения электронной информации.

Для решения каждой из указанных вопросов необходимы разные множества метаданных, которые могут образовывать как непустые, так и пустые пересечения. В целом, различают два множества метаданных, которые сформированы в зависимости от уровня детализации описываемого объекта:

1) метаданные описания контента, которые охватывают описание всех аспектов данного информационного объекта, как отдельной сущности. Иногда их дополнительно подразделяют на структурные и описательные;

2) административные метаданные, объединяющие различные группы метаданных, и отличающиеся между собой целью, которая достигается с их помощью. Например, некоторая группа метаданных позволяет владельцу ресурса проводить четкую и гибкую политику в отношении информационного объекта, включая авторизацию, аутентификацию, управление авторскими правами, доступом, другая группа служит для идентификации и категоризации объектов в рамках специальной коллекции или организации. Существует группа административных метаданных, которая может использоваться для позиционирования данного информационного ресурса в контексте множества подобных документов, информационно-поисковой системы, предметной области и т.д.

В рамках данной работы интерес представляют метаданные, предназначенные для описания информационных объектов, интегрируемых в образовательный процесс.

В последние годы было разработано множество разнообразных спецификаций метаданных, которые позволяют различным организациям и группам пользователей использовать метаданные для решения специфических задач. В работе рассмотрены следующие стандарты на спецификации метаданных:

- международная модель Дублинское ядро (Dublin Core, DC), которая является попыткой определить основной набор элементов описания информационных ресурсов;
- концептуальная схема Learning Object Metadata (LOM), разработанная Комитетом по стандартизации образовательных технологий Института инженеров по электротехнике и электронике (IEEE);
- модель The Gateway to Educational Materials (GEM), разработанная Министерством образования США, Национальной библиотекой образования и группой исследователей Сиракузского университета;

Для российского сегмента системы образования учеными Российского государственного университета инноваций и технологий предпринимательства А.И. Башмаковым и В.А. Старых предложено расширение схемы LOM с целью адаптации данного стандарта к особенностям российской системы образования. Данное расширение является полным, содержательным метаописанием информационных образовательных ресурсов. Вместе с тем, существующая избыточность и наличие параметров, оценка которых субъективна, не позволяют широко использовать данную схему для практической реализации.

На основе существующих решений по созданию спецификаций описания информационных ресурсов и классификации информационных образовательных ресурсов, в работе предложена спецификация метаданных. В спецификации

используются следующие типы данных, совместимые с типами данных, определенными в расширении схемы LOM:

– контейнер; составной элемент, включающий подчиненные элементы; допустима вложенная структура контейнера; предназначен для группировки других элементов;

– строка; представляет собой последовательность символов;

– значение из словаря; служит для задания значений, выбираемых из словарей; образуется двумя данными типа строка, в первом из них содержится идентификатор словаря, а во втором указывается значение из этого словаря.

Предлагаемая спецификация метаданных об ИОР содержит все элементы метаданных набора Dublin Core. Соответствие элементов, которое доказывает семантическую интероперабельность предлагаемой схемы, показано в таблице 1.

Таблица 1

Соответствие элементов предложенной схемы набору элементов метаданных Dublin Core (DC)

| Элемент DC | Элемент спецификации метаданных |
|---|---|
| Идентификатор | 1.2 Уникальное обозначение ИОР |
| Название | 1.1 Заглавие (наименование) ИОР |
| Язык | 1.3 Язык представления |
| Описание | 1.4 Аннотация ИОР |
| Предмет | 1.5 Ключевые слова или 2.4 Дисциплина |
| Зона действия | 4.4 Охватываемый объем |
| Тип ресурса | 3.6 Тип ИОР |
| Дата | 1.12 Дата |
| Создатель | 1.6 Автор ИОР |
| Издатель | 1.11 Каталогизатор |
| Формат | 4.5 Представление ИОР |
| Правовые аспекты | 1.10 Права доступа к ИОР |
| Отношение | 1.13 Связанный ИОР |
| Источник | 1.7 Поставщик |

В предложенной схеме метаданных большинство элементов имеют соответствующие элементы в расширении схемы LOM. Однако строгого согласования данных схем не наблюдается, поскольку в предлагаемую схему метаданных введены новые элементы, не имеющие аналогов ни в расширении схемы LOM, ни в других схемах, предназначенных для описания информационных ресурсов образовательного характера.

В состав нововведенных элементов входят следующие элементы:

– Категория ИОР (2.1);
– Учебное назначение ИОР (2.2);
– Вид обучения по ИОР (2.5);
– Способ программной реализации (3.2);
– Структурно - логическое построение ИОР (4).

Введение этих элементов позволяет максимально учесть образовательную направленность ИОР при решении всех задач, возложенных на метаданные. Кроме того, в данной схеме более четко определяется тип ИОР путем указания класса,

подкласса и вида ИОР.

По элементам, входящим в состав контейнеров «Общие сведения об ИОР» и «Назначение ИОР», можно находить различные ИОР без акцентирования на технические, технологические и структурно-логические аспекты построения ИОР. Это позволяет выделить множество атрибутов ИОР, по которым целесообразно осуществлять поиск.

Таким образом, приведенное описание (представление) ИОР для сферы образования обладает спецификой, которое обеспечит возможность отражения:

– логической и физической структур ИОР;

– информации для взаимодействия ИОР с системой управления учебным процессом, в рамках которого выполняются настройка ИОР на текущие условия применения и конкретного обучаемого, а также фиксация хода и результатов его работы;

– правил, определяющих методику работы с ИОР (порядок навигации по его компонентам, оценивание действий обучаемого);

– педагогических характеристик (уровень образования, целевая аудитория, сложность, контактное время и т.д.), необходимых для принятия решения о включении ИОР в состав контента, покрывающего учебный план или программу;

– информации для пополнения регионального депозитария информационных образовательных ресурсов.

По предложенной спецификации метаописания ИОР разработана по методологии Information Engeniring (стандарт IDEF1X) модель данных, которая позволяет однозначно идентифицировать ИОР и учитывать требования различных категорий пользователей депозитария.

**Третья глава** диссертации посвящена вопросам разработки методов индексирования и поиска полнотекстовых документов.

При решении этих вопросов считаются заданными множество документов $D$, лингвистическое обеспечение $S$ и запрос $Q$. Необходимо построить поисковые образы запроса $I^Q$ и документов $I^D$ и на основе их сопоставления сформировать отклик $\widetilde{D}$ на запрос $Q$, характеризующийся показателем точности поиска $p > 0{,}75$.

Для решения этой задачи построена формальная модель информационного поиска, описываемая тройкой:

$$M_{IP} = < I^D, I^Q, \Psi >,$$

где $I^D$ – тип представления документов, $I^Q$ – тип представления поисковых запросов, $\Psi$ – модель интерпретации запроса $I^Q$ на документ $I^D$.

Поисковый образ документа (ИОР) $I^D$ – двухкомпонентная модель:

$$I^D = < M, G >,$$

где $M$ – метаописание документа; $G$ – семантическая сеть документа.

Поисковый образ запроса $I^Q$, или семантическая сеть $G^Q$ запроса $Q$, определяется парой:

$$I^Q = G^Q = < U^Q, V^Q >,$$

где $U^Q$ – множество вершин $G^Q$, которые содержат термы запроса $t^Q$;

$V^Q$ – матрица смежности, задающая дуги между вершинами.

Модель интерпретации запроса $I^Q$ на документ $I^D$:

$$\Psi = < \Psi_{POD}, \Psi_{OS} >,$$

где $\Psi_{POD}$ – компонент построения поискового образа документа в части формирования его семантической сети, а $\Psi_{OS}$ – компонент построения отклика на запрос пользователя.

***Построение поискового образа документа***. При решении данной задачи исходными данными считаются:

1) множество документов $D = \{d_i | i=1..n, n$ – количество документов$\}$;

2) лингвистическое обеспечение $S = < S_1, S_2, ..., S_{13} >$. В $S$ включены словари готовых (неизменяемых) словоформ, основ частей речи, окончаний частей речи, флективных классов, морфологической информации и схемы субстантивных именных словосочетаний русского языка.

Необходимо:

1) выделить в документе множество $T_i^D$ термов $t_j$ и определить для $\forall\ t_j \in T_i^D$ частоту $f_j^t$ его встречаемости в тексте документа $d_i$ $(j=1..|T_i^D|)$;

2) построить взвешенную семантическую сеть $G = (U, V, W)$, где $U$ – множество вершин сети, $V$ – множество дуг сети, $W=\{ w_{xy}^G \mid w_{xy}^G$ – вес дуги, отражающий семантическую близость вершин $u_x$ и $u_y\}$;

3) найти множество $T_i'$ термов $t_j$ документа $d_i$ таких, что они являются ключевыми словами, характеризующими данный документ.

Ограничения:

1) множество $T_i' \subset T_i^D$ и $t_j \in T_i'$ при $w_j^t < \mu$, где $\mu$ - пороговое значение весов $w_j^t$ термов $t_j$;

2) $f_j^t > 1$ для $\forall\ t_j \in T_i'$, где $f_j^t$ - частота встречаемости терма $t_j$.

В работе понятие семантической близости двух вершин сети определяется на основе структурной близости в тексте документа термов, принадлежащих этим вершинам, при этом атомарным элементом структуры текста считается одно предложение $e \in E_i$, где $E_i$ – текст документа $d_i$. Вес дуги, или сила связи, $w_{xy}^G$ определяется на основе частоты совместной встречаемости термов вершины $u_x$ и термов вершины $u_y$ в структурных элементах текста $E_i$.

***Формирование множества лексем L***. Для этого применяется лексический и морфологический анализы документа $d$ с применением словарей $S$. При реализации морфологического анализа использован декларативный подход, предложенный Белоноговым Г.Г. В результате выполнения вышеназванных анализов формируются поток лексем $L = \{l_i \mid i=1..k$, $k$ – общее количество лексем в потоке$\}$, множество $L^S=\{ \rho_i^{ls} \mid i=1..k'$, $k'$ – количество разновидностей лексем в потоке, $k' \le k\}$ векторов лексем, вектор $\rho_i^{ls}$ описывает статические характеристики лексемы $l_i$, и множество $L^V=\{ \rho_i^{lv} \mid i=1..k\}$ векторов, представляющих динамические характеристики лексемы $l_i$, зависящие от контекста.

***Формирование множества $T^D$ термов документа d***. На данной фазе из потока лексем $L$ с их характеристиками $L^S$ и $L^V$ формируется множество $T^D$. Согласно Белоногову Г.Г. и Беловольской Л.А. термами полнотекстового документа

являются именные субстантивные словосочетания, поэтому решение основывается на поиске таких словосочетаний. В общем случае именные словосочетания включают в свой состав следующие классы слов: существительные, прилагательные, предлоги и сочинительные союзы. Количество слов в именных словосочетаниях колеблется от двух до пятнадцати. В работе используются 18 шаблонов словосочетаний. Все словосочетания выделяются посредством регулярных выражений. В результате формируется множество термов $T^D=\{t_j \mid j=1..N^t,\ N^t$ – количество всех выделенных в документе термов$\}$. Каждый терм описывается вектором $\tau_j =< n^t_j, t_j, a_j >$, где $n^t_j$ – уникальный номер вектора терма; $t_j$ – терм-словосочетание; $a_j$ – адрес терма в тексте $E$. Множество векторов $\tau_j$ образует характеристики термов $T^H$.

***Построение взвешенной семантической сети G документа d.*** Подход построения семантической сети $G$ заключается в следующем. При изучении текстов научного, учебного или учебно-методического характера можно заметить, что, как правило, некоторый затрагиваемый в тексте предмет (явление, свойство и т.п.) рассматривается с различных сторон, что выражается его описанием различными именными субстантивными словосочетаниями. Например, если описывается *сетевое сообщество*, то оно может составлять следующие словосочетания: сетевое образовательное сообщество, образовательное сообщество, сетевое научно-образовательное сообщество и т.д. В этих словосочетаниях общим словом (в дальнейшем будем называть его «несущим») является слово «*сообщество*». На наш взгляд, все они образуют некоторую окрестность в области семантики документа. Таким образом, если сформировать по каждой окрестности фрагмент (подграф) $G^F$ графа семантической сети $G$, то семантическая сеть документа $G$ может быть построена как объединение фрагментов $G^F$.

Данный подход реализуется следующим образом. Несущее слово $b^t$ словосочетания $t$ определяется как слово, выраженное именем существительным, частота встречаемости которого будет наибольшей среди существительных рассматриваемого словосочетания. Тогда на множестве $T^D$ зададим отношение $R = \{(t_i, t_j) \mid t_i$ и $t_j$ имеют одно несущее слово$\}$, которое позволяет получить множество классов эквивалентности $[T^D]_R$. Обозначим $T^R_k$ - $k$-ый класс эквивалентности. Для $\forall t_{ik} \in T^R_k$ формируется вектор $\theta_{ik}$:

$$\theta_{ik} =< b^t_i, t_{ik}, f^t_i, p^t_i >,$$

где $t_{ik}$ – $i$-тый терм-словосочетание $k$-того класса эквивалентности; $f^t_i$ – частота встречаемости $t_{ik}$ в документе; $p^t_i$ – значимость $t_{ik}$, которая рассчитывается по формуле:

$$p^t_i = f^t_i / |T^R_k|.$$

Физический смысл класса эквивалентности $T^R_k$ заключается в том, что он описывает $k$-ую окрестность в области семантики документа, т.е. является основой формирования фрагмента $G^F_k = (U^F_k, V^F_k, W^F_k)$, где $U^F_k$ – множество вершин фрагмента $k$, которыми являются термы-словосочетания, $V^F_k$ – множество дуг

фрагмента, $W_k^F = \{ w_{ij}^{F_k} \} \mid w_{ij}^{F_k}$ – вес дуги, который отражает силу связи между $t_i$ и $t_j$ и интерпретируется как семантическая близость термов}. Вес дуги $w_{ij}^{F_k}$ рассчитывается по формуле:

$$w_{ij}^{F_k} = \log_{\varphi_k} ( \sum_{m=1}^{4} \lambda_{ijm} \cdot f_both_{ijm} ),$$

где $\lambda_{ijm}$ – коэффициент вида связи $m$ между $t_i$ и $t_j$ ($\lambda_{ijm}$=1, если $t_i$ и $t_j$ встречаются в одной главе; $\lambda_{ijm}$=2, если $t_i$ и $t_j$ встречаются в одном разделе; $\lambda_{ijm}$=3, если $t_i$ и $t_j$ встречаются в одном параграфе; $\lambda_{ijm}$=4, если $t_i$ и $t_j$ встречаются в одном предложении); $f_both_{ijm}$ – частота совместной встречаемости термов-словосочетаний $t_i$ и $t_j$ в связи $m$; $\varphi_k$ - основание логарифма, равное максимуму подлогарифмической функции: $\varphi_k = 10 \cdot \max_{t_i \in U_k^t}(f_i^t)$.

Для каждого фрагмента $G_k^F$ формируется характеризующий его вектор:

$$\xi_k = < w_k^b, f_k^b, f_k, p_k^F >,$$

где $w_k^b, f_k^b$ – вес и частота встречаемости несущего слова $b$ фрагмента, $f_k$ – частота встречаемости фрагмента, $p_k^F$ – значимость фрагмента $G_k^F$ в документе.

Таким образом, семантическая сеть состоит из фрагментов $G_k^F$ и описывается матрицей смежности $W$, которая содержит веса дуг $w_{xy}^G$ между $x$–ым и $y$–ым фрагментами, вычисляющиеся аналогично расчету весов $w_{ij}^{F_k}$ дуг фрагментов.

***Выделение множества ключевых документов T′.*** Выделение списка ключевых слов документа $T'$ из множества $T$ осуществляется следующим образом.

1. В каждом фрагменте $G_k^F$ выбирается терм $t_{ik}$ такой, что его вес $w_i = \max_{t_i \in U_k^t}(w_i)$, $w_i \in [0; 1]$ и рассчитывается по формуле:

$$w_i = \log_{f_\Sigma} f_i,$$

где $f_\Sigma$ – сумма $f_i$; $f_i$ – частота встречаемости терма $t_i$.

2. Для уменьшения мощности множества $T'$ определяется пороговое значение $\mu$:

$$\mu = \sqrt{\frac{1}{N} \sum_{i=1}^{N} w_i} \qquad (3)$$

где $N$ – мощность множества $T'$.

Формула (3) получена на основе экспериментальных данных. При проведении экспериментов в качестве порога были взяты среднее значение весов $w_i$, его удвоенное произведение и корень. Анализ отклонений доли выделенных ключевых слов от их средних значений при разных $\mu$ показал, что доля выделенных ключевых слов в потоке словоформ в документах разного объема при использовании формулы (3) достаточно постоянна, и среднее значение отклонений в этом случае является наименьшим.

Таким образом, структура модуля построения поискового образа документа имеет два блока: первый блок осуществляет препроцессорную обработку индексируемого документа, второй – построение его семантической сети $G$.

***Построение отклика на запрос пользователя***. При решении данной задачи исходными данными являются:

1) множество поисковых образов $I^D$ документов, включающие семантические сети $G$;

2) запрос $Q$, заданный одним из следующих способов:

– простое предложение $e$ на естественном (русском) языке;

– множество $U^Q = \{t^Q | t^Q$ – терм запроса $Q\}$;

3) лингвистическое обеспечение $S = < S_1, S_2, ..., S_{13} >$.

Необходимо:

1) построить поисковый образ $I^Q$ запроса $Q$;

2) найти такое множество $D^R = \{d_i | \alpha_i^G \to \max_{d_i}(\alpha_i^G)$, $\alpha_i^G$ – степень релевантности документа $d_i$ запросу $Q$, рассчитанная по семантической сети $G_i$ };

3) сформировать отклик $\tilde{D}$ из множества $D^R$, такой что $p > 0{,}75$, где $p$ – показатель точности отклика.

Построение поискового образа $I^Q$ осуществляется аналогично построению поискового образа $I^D$ и описывается множеством $U^Q = \{ t_i^Q | \ i = 1..m$, $m$ – количество термов в запросе $Q\}$.

***Поиск множества релевантных запросу Q документов***. В множество релевантных документов $D^R = \{ d_i^R \}$ включаются $d_i^R$, имеющие $\alpha_i^M > 0$ или $\alpha_i^G > 0$, где $\alpha_i^M$ и $\alpha_i^G$ – показатели степени релевантности документа поисковому запросу, которые рассчитываются по каждому документу $d_i$. Первый показатель рассчитывается на основе анализа метаописания информационного образовательного ресурса, а второй – интегральный показатель степени релевантности, основанный на вычислении смысловой близости документа $d_i$ и запроса $Q$ по семантической сети $G_i$ и рассчитывается по формуле (4):

$$\alpha_i^G = \frac{\sum\limits_{(t_i,t_j),t_i \in G_x^F, t_j \in G_y^F} w_i^t \cdot w_j^t \cdot w_{xy}^F + \sum\limits_{(t_i,t_j),t_i,t_j \in G_k^F} w_i^t \cdot w_j^t \cdot w_{ab}^{F_k}}{K} \qquad (4)$$

где $w_i^t$ - вес терма $t_i$; $w_j^t$ - вес терма $t_j$; $w_{xy}^F$ – вес дуги (сила связи) между фрагментами $G_x^F$ и $G_y^F$; $w_{ab}^{F_k}$ – вес дуги (сила связи) между термами-словосочетаниями $t_i$ и $t_j$, которые принадлежат одному и тому же фрагменту $G_k^F$, при этом $a$ и $b$ – их порядковые номера во фрагменте; $K$ - количество фрагментов семантической сети, в которых встретились термины поискового запроса $\{t^Q\}$.

Веса $w_i^t$ термов поискового запроса вычисляются по правилам:

1) если терм $t^Q \in U^Q$ является ключевым словом документа, т.е. $t^Q = t_j$, а $t_j \in T'$, то его вес $w_i^t$ в документе $d_i$ принимается равным весу $w_j$ ключевого слова $t_j$;

2) если терм $t^Q \in U^Q$ не является ключевым словом документа, т.е. $t^Q \notin T'$, то вес терма $t^Q$, найденного в некотором фрагменте $G_k^F$, рассчитывается по формуле (5):

$$w_i^t = p_k^F \cdot f^t / f_k , \qquad (5)$$

где $p_k^F$ – значимость фрагмента $G_k^F$ сети, которому принадлежит терм $t^Q$; $f$ –

частота встречаемости терма; $f_k$ – частота встречаемости фрагмента $G_k^F$.

Таким образом, будут рассчитаны степени соответствия поисковому запросу множества релевантных документов.

***Ранжирование документов на основе кластерного анализа и формирование отклика*** $\widetilde{D}$. Для этого используется подход, предложенный в работе Некрестьянова И. и Пантелеевой Н., основанный на кластеризации коллекции документов для повышения качества поиска. В данной работе мы выполняем кластеризацию множества релевантных документов $D^R$, которая разбивает его на группы близкородственных документов (кластеры) на основе вычисления степени соответствия между подвергающимися кластеризации объектами посредством метода К-средних. В качестве меры расстояния между двумя документами используется Евклидово расстояние:

$$z(d_i, d_j) = \sqrt{(\alpha_i^M - \alpha_j^M)^2 + (\alpha_i^G - \alpha_j^G)^2}$$

где $d_i$ и $d_j$ – документы коллекции, $\alpha_i^M$, $\alpha_i^G$ - показатели степени релевантности документа $d_i$, $\alpha_j^M$, $\alpha_j^G$ - показатели степени релевантности документа $d_j$.

Очевидно, что кластер, имеющий наибольшие значения интегральных показателей $\alpha_k^M, \alpha_k^G$, содержит наиболее релевантные запросу документы, кластера образуют отклик документов $\widetilde{D}$.

**Четвертая глава** диссертации посвящена описанию программного обеспечения, необходимого для апробации разработанных методов и которое состоит из трех программ. Эти программы позволяют осуществить экспериментальную проверку решения трех поставленных в работе задач. При этом приняты следующие ограничения:

1) на поиск:

- поиск ИОР, представленных полнотекстовыми документами, с помощью поисковой системы:
- поиск других видов ИОР посредством атрибутного и/или контекстного поиска, поддерживаемых метакаталогом ИОР;

2) на представление документов:

- индексируемый документ представляется только своей текстовой составляющей в формате txt;
- язык представления документов – русский;
- вид документов (ИОР) – монография, учебно-методическое пособие, отчет о НИР, диссертация и т.п.

Программа Metacatalog позволяет создавать метаописания различных информационных образовательных ресурсов. Реализованная схема метаописания включает основные элементы предложенной в работе схемы метаданных. Апробированы все основные функции метакаталога, к которым относятся: обновление метаописания и поиск ресурсов по метаданным. Для выполнения вычислительных экспериментов разработано программное обеспечение (прототип) метакаталога информационных образовательных ресурсов с использованием технологии XML, при этом реализованы основные функции по работе с метакаталогом и традиционные методы поиска по метакаталогу ИОР: атрибутный,

контекстный, атрибутно-контекстный. Описание программного обеспечения приведено в четвертой главе данной работы.

Программа IndexingPro позволяет сформировать второй компонент поискового образа полнотекстового документа. Результаты экспериментальной проверки работы программы показывают, что семантические сети документов, построенные в результате работы программы, адекватно отражают содержание документа.

Программа KoxonenNet осуществляет поиск документов в коллекции документов метакаталога, вычисляет степени релевантности документов информации, затребованной в запросе, выполняет их кластеризацию, и позволяет выбрать кластер, содержащий документы с высшей степенью релевантности.

Анализ результатов работы программ IndexingPro и KoxonenNet выполнен на основе метода экспертной оценки, который показал, что сформированные поисковые образы отражают семантику документа на 70%, средние показатели точности и полноты отклика равны 0,86 и 0,76 соответственно.

Таким образом, предложенные в работе методы индексирования, поиска и ранжирования являются жизнеспособными, полученные результаты адекватно отражают семантику индексируемых документов, что доказывается проведенными экспериментами.

В **приложениях** приведены проанализированные спецификации метаданных, лингвистическое обеспечение морфологического анализа, основные результаты экспериментальной проверки разработанных методов.

## ЗАКЛЮЧЕНИЕ

В работе получены следующие научные и практические результаты:

1. Предложена спецификация метаописания информационных образовательных ресурсов, базирующаяся на международном стандарте Learning Object Metadata и на выполненной в работе классификации информационных образовательных ресурсов. Элементы данной спецификации достаточно полно отражают образовательный характер описываемого ресурса.

2. Предложена двухкомпонентная структура поискового образа документа, основу которого составляет взвешенная семантическая сеть полнотекстового документа, адекватно отражающая семантику этого документа.

3. Разработана двухкомпонентная модель поиска. Первый компонент осуществляет построение семантической сети документа, второй – построение отклика на запрос пользователя. Это позволяет отделить поиск полнотекстовых документов от поиска других видов ИОР.

4. Разработан подход к индексированию документа, основанный на оригинальном способе формирования семантической сети документа.

5. Предложены методы построения и ранжирования отклика поисковой системы на запрос пользователя, которые основаны на анализе семантических сетей запроса и документов коллекции депозитария. Применение кластеризации релевантных документов позволяет получить семантически близкие документы в одном кластере. Содержимое кластера с наибольшим средним значением интегрального показателя степени релевантности образует отклик поисковой системы, обладающий высоким показателем точности.

6. Проведена экспериментальная проверка разработанных моделей и методов, результаты которой подтверждают основные положения работы.

7. Полученные результаты могут быть применены при разработке поисковой системы регионального депозитария информационных образовательных ресурсов.

**Результаты диссертации отражены в следующих работах:**


1. Найханова Л.В., Аюшеева Н.Н., Евдокимова И.С. Концепция построения поисковой системы информационно-образовательных ресурсов // Теоретические и прикладные вопросы современных информационных технологий: Материалы третьей всероссийской научно-технической конференции. – Улан-Удэ, 2002. – С. 170-173.

2. Основы интернет-технологий: Учебное пособие / Аюшеева Н.Н., Бильгаева Н.Ц., Найханов В.В. и др. – Улан-Удэ: Изд-во ВСГТУ, 2002 г. – 106 с.

3. Найханова Л.В., Аюшеева Н.Н., Евдокимова И.С. Методологические основы отраслевой поисковой системы // Сборник трудов Второй Всероссийской научно-практической конференции «Российская школа и Интернет». – СПб, 2002. – С.23.

4. Аюшеева Н.Н. Метод индексирования полнотекстовых документов // Теоретические и прикладные вопросы современных информационных технологий: Материалы Четвертой Всероссийской научно-технической конференции. - Улан-Удэ, 2003. – С.174-176.

5. Аюшеева Н.Н. О результатах исследования методов индексирования // Информация – Коммуникация – Общество (ИКО-2003): Тезисы докладов и выступлений Международной научной конференции. – СПб., 2003. – С.34-36.

6. Найханова Л.В., Аюшеева Н.Н., Шаманаев А.В. Исследование методов определения пороговой частоты // Теоретические и прикладные вопросы современных информационных технологий: Материалы Пятой Всероссийской научно-технической конференции. – Улан-Удэ, 2004. – С.36-39.

7. Аюшеева Н.Н. Схема метаданных метакаталога информационных образовательных ресурсов // Проблемы качества, безопасности и диагностики в условиях информационного общества: Тезисы докладов Всероссийской научно-технической конференции. – Сочи, 2004. – С. 187.

8. Аюшеева Н.Н. Выделение словосочетаний для индексирования полнотекстовых документов // Единая образовательная информационная среда: проблемы и пути развития: Материалы третьей Всероссийской научно-практической конференции. – Омск, 2004. – С. 283-285.

9. Теоретические и прикладные вопросы разработки интегрированных интеллектуальных информационных систем. Этап: Основные аспекты методологии построения интеллектуальных информационно-поисковых систем: Отчет о НИР (промежуточный) / ВСГТУ; Рук. Бильтриков В.Н. – ГР № 01.200.205060; Инв.№ 02.200305099. – Улан-Удэ, 2002. – 40 с. – Отв. исп. Аюшеева Н.Н.

10. Свидетельство об официальной регистрации программы для ЭВМ № 2004612385. Комплекс программ «Индексирование полнотекстовых документов и кластеризация релевантных поисковому запросу документов» / Найханова Л.В., Аюшеева Н.Н., Шаманаев А.В. – М.: Всероссийское агентство по патентам и товарным знакам, 2004.