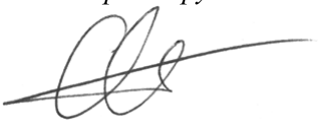
На правах рукописи



ОКРОПИШИН АНТОН ЕВГЕНЬЕВИЧ

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МОДЕЛЕЙ И СРЕДСТВ ОБЕСПЕЧЕНИЯ ДОКУМЕНТАЛЬНОГО ПОИСКА В РАСПРЕДЕЛЕННЫХ ГЕТЕРОГЕННЫХ ИНФОРМАЦИОННЫХ РЕСУРСАХ

05.13.01 – «Системный анализ, управление и обработка информации
(в информационных системах)»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва - 2013

Работа выполнена в «Национальном исследовательском ядерном университете «МИФИ»

Научный руководитель: доктор технических наук, профессор **Максимов Николай Вениаминович**

Официальные оппоненты: доктор физико-математических наук, профессор, зав. кафедрой Автоматизированных систем управления НИТУ "МИСиС" **Кривоножко Владимир Егорович**

кандидат технических наук, доцент, заместитель генерального директора по информационной политике ЗАО «Региональный Сетевой Информационный Центр» **Храмцов Павел Брониславович**

Ведущая организация: Институт проблем информатики Российской академии наук

Защита состоится «25» декабря 2013 г. в 15 часов 00 минут на заседании диссертационного совета Д 212.130.03 при Национальном исследовательском ядерном университете «МИФИ», расположенном по адресу: 115409, г. Москва, Каширское шоссе, 31.
С диссертацией можно ознакомиться в библиотеке НИЯУ МИФИ
Автореферат разослан «25» ноября 2013г.
Отзывы и замечания по автореферату в двух экземплярах, заверенные печатью, просьба высылать по вышеуказанному адресу на имя учёного секретаря диссертационного совета.

Ученый секретарь диссертационного совета, доктор технических наук, доцент  **Леонова Н.М.**

Общая характеристика работы

Актуальность исследования. Неотъемлемым атрибутом современного общества в последние десятилетия стало непрерывное увеличение объемов информации, представленной на электронных носителях и организованной в виде множества разнообразных распределенных документальных ресурсов. Становится очевидным, что развитие средств поиска информации не может компенсировать возрастающую как количественно, так и качественно сложность ее обработки. При этом, несмотря на создание все более совершенных систем управления информационными ресурсами (ИР) в рамках отдельно взятых электронных библиотек (ЭБ), на уровне информационного пространства в целом остается не решенной одна из основных задач любой информационной системы – предоставление пользователю нужной ему информации в удобной и доступной для него форме, обеспечивающей максимальное соответствие его личным потребностям, в том числе по требованиям к полноте и точности поиска.

Поэтому организация современных специализированных средств доступа к опубликованным отечественным и зарубежным результатам научной деятельности, исследований и экспериментов является залогом высоких темпов развития науки и техники. Актуальность этого отражена и указами президента РФ, предписывающими, в частности, создание единой базы данных о научно-исследовательских и опытно-конструкторских работах[1,2].

Целью диссертационной работы является разработка моделей и средств унифицированного доступа к гетерогенным распределенным информационным ресурсам, обеспечивающим оптимизацию процесса поискового взаимодействия пользователя с ресурсами с учетом особенностей задач информационного обеспечения научных исследований.

Основными задачами являются**:**

- исследование процессов поискового взаимодействия в среде распределенных гетерогенных информационных ресурсов;

---

[1] Поручение Президента Российской Федерации от 4 января 2010 г. № Пр-22

[2] Поручение Президента Российской Федерации от 1 августа 2008 г. № Пр-1572

- разработка моделей информационной совместимости ресурсов;
- разработка объектной модели информационного ресурса для задач распределенного документального поиска;
- разработка механизма обеспечения интероперабельности ИР, использующего унифицированные описания ресурсов, включающего трансляцию поискового запроса с языка поисковых запросов (ЯПЗ) исходного ресурса на язык целевого ресурса и сопоставление схем данных взаимодействующих ресурсов;
- проектирование и разработка средств унифицированного доступа к распределенным гетерогенным информационным ресурсам, включая формирование прототипа репозитория описаний ИР.

Объектом исследования являются распределенные гетерогенные информационные ресурсы, доступные для поискового взаимодействия по сети, такие как документальные базы данных, электронные библиотеки, каталоги издательств, поисковые машины, а также характеристики этих ресурсов с точки зрения организации автоматизированного доступа к ним.

Предметом исследования являются:

- совокупность способов взаимодействия с информационными ресурсами;
- механизмы установления соответствий между элементами данных при работе с ИР.

Экспериментальной базой являются промышленные информационные ресурсы, а также полученные автором результаты экспериментальных исследований поисковых процессов в среде гетерогенных ИР, проводимых в рамках НИР[3,4] и ОКР[5], а также лабора-

---

[3] Федеральная целевая программа «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2012 годы» в Центре информационных технологий и систем по проекту «Создание информационно-аналитической системы регистрации, учёта, обработки и хранения отчётных документов по НИОКР, выполняемым ФГУП и ОАО, с целью проведения мониторинга состояния и основных тенденций и направлений развития научных исследований и разработок, выполняемых компаниями государственного сектора, в том числе направленных на реализацию приоритетных направлений развития науки, технологий и техники в Российской Федерации, а также критических технологий Российской Федерации»

торных практикумов и учебно-исследовательских работ студентов в НИЯУ МИФИ и РГГУ.

Методы исследования. Основные результаты работы получены с использованием методов теории множеств, теории вероятностей, математической статистики и системного анализа.

Научная новизна результатов работы.

- модель метаинформационной совместимости, позволяющая ввести расстояние на основе меры различия между любой парой схем данных, отражающее точность отображения схем данных при переходе от одного ресурса к другому;
- модель лингвистической совместимости, позволяющая ввести расстояние для пар языков поисковых запросов (ЯПЗ) на основе меры их различия, отражающее потерю смысла поискового запроса при переходе к иному синтаксису и структуре данных;
- модель лексической совместимости, дающая вероятностную оценку близости ресурсов по используемой лексике, отражающую зависимость результатов поиска от попарного пересечения словарей ресурсов.

Практическая значимость результатов работы:

- модель метаинформационной совместимости позволяет рассчитать близость между схемами данных взаимодействующих ресурсов, обеспечивая оценку целесообразности использования ассоциированного ресурса и, тем самым, позволяя снизить избыточность выдачи;
- модель лингвистической совместимости позволяет количественно оценить адекватность преобразования поискового запроса, выполняемого в соответствии с синтаксисом и набором метаданных целевого ИР, что обеспечивает возможность взаимного ранжирования поисковых результатов, получаемых из нескольких ИР;

---

[4] РФФИ, грант 11-09-13128 офи-м-2011-РЖД. «Моделирование и разработка распределенных гетерогенных информационных ресурсов онлайн-информирования пассажиров»

[5] Опытно-конструкторская работа по теме: «Создание единой государственной информационной системы мониторинга процессов аттестации научных и научно-педагогических кадров высшей квалификации» для разработки подсистемы «Шлюз с ЕФБД НИОКР» (мероприятие 5.1 ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы»)

- модель лексической совместимости ресурсов дает вероятностную оценку лексической близости ИР, которая при переадресации запроса используется для обоснования выбора ресурса;
- объектная модель информационного ресурса, обеспечивающая построение унифицированного объектно-ориентрованного описания ресурса, используемого для ранжирования потенциально полезных ресурсов и преобразования запросов в соответствии с требованиями целевого ИР;
- совокупность программных инструментов позволяет обеспечить пользователей средствами поддержки управления поиском в ИР, обеспечивая возможность обращения к внешним ресурсам не только с использованием запросов на ЕЯ, но и запросов, содержащих булевы и контекстные операторы ЯПЗ, что в значительной степени повышает точность выдачи и, в отдельных случаях, например для Internet-поисковых машин, на 2-3 порядка снижает количество документов в выдаче.

На защиту выносятся:

- модель метаинформационной совместимости ресурсов и мера, позволяющая определить совместимость схем данных для пар ресурсов;
- модель лингвистической совместимости ресурсов и мера, позволяющая определить совместимость ИПЯ различных ИР;
- модель лексической совместимости ресурсов и мера, отражающая близость лексики ИР, обусловленной тематикой;
- объектная модель, алгоритм и объектно-ориентированное описание ресурса, обеспечивающие управление процессом переадресации поисковых запросов с учетом различий в схемах данных, а также в формах и синтаксисе ЯПЗ.

Достоверность полученных результатов и адекватность моделей подтверждаются корректностью математического аппарата, а именно элементов теории множеств, теории вероятностей и системного анализа, использованных в диссертационной работе; а также сопоставлением с экспериментальными данными, полученными путем компьютерного моделирования и путем внедрения в составе конкретных информационных систем.

Апробация работы. Основные результаты работы докладывались и обсуждались на конференциях:

1. Научная сессия МИФИ-2009. XIII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании»;
2. 7-я Курчатовская молодёжная научная школа 2009;
3. Международная научно-практическая конференция 2009 «Математика, информатика, естествознание в экономике и в обществе»;
4. XIX международная конференция-выставка «Информационные технологии в образовании» 2009;
5. Научная сессия НИЯУ МИФИ-2010. XIV выставка-конференция «Телекоммуникации и новые информационные технологии в образовании»;
6. IX Международная научно-практическая конференция-выставка «Единая образовательная информационная среда: направления и перспективы развития электронного и дистанционного обучения 2010»;
7. XX международная конференция-выставка «Информационные технологии в образовании» 2010;
8. Научная сессия НИЯУ МИФИ-2011. XV выставка-конференция «Телекоммуникации и новые информационные технологии в образовании»;
9. Научная сессия НИЯУ МИФИ-2012;
10. «Технические и программные средства систем управления, контроля и измерения» (УКИ'12): Конференция с международным участием, 2012;
11. Научная сессия НИЯУ МИФИ-2013.

Реализация результатов работы:

- модель информационной совместимости разнородных информационных ресурсов, в частности модель лексической совместимости, а также объектная модель ресурса использованы в Федеральном государственном автономном научном учреждении «Центр информационных технологий и систем органов исполнительной власти» (ФГАНУ ЦИТиС) в рамках опытно-конструкторской работы по теме: «Создание единой государственной информационной системы мониторинга процессов аттестации научных и научно-педагогических кадров высшей квалификации» для разработки подсистемы «Шлюз с ЕФБД НИОКР» (мероприятие 5.1 ФЦП «Иссле-

дования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы»);

- модель лингвистической совместимости, объектная модель ресурса и объектно-ориентированное описание ИР использованы в Федеральном государственном автономном образовательном учреждении высшего профессионального образования «Национальный исследовательский ядерный университет «МИФИ» в рамках проекта автоматизации Центра информационно-библиотечного обеспечения учебно-научной деятельности;
- модели лингвистической и метаинформационной совместимости информационных ресурсов, объектная модель и объектно-ориентированное описание ресурсов, а также программно-информационные средства поддержки поиска в распределенных гетерогенных информационных ресурсах использованы в ИНИОН РАН в составе информационного портала для организации поиска в локальных реферативных БД с возможностью трансляции и передачи запроса в ассоциированные внешние ИР.

Публикации. По теме диссертации опубликовано 16 статей, а также получено два свидетельства о государственной регистрации программ для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы (85 наименований), а также приложений (содержит 148 страниц текста, 39 рисунков, 11 таблиц).

Содержание работы

Во **введении** обосновывается актуальность проведения исследования, а также сформулированы цели и задачи работы.

В **первой главе** приводится краткое изложение проблемы информационного поиска в среде распределенных гетерогенных документальных ресурсов, сводящейся к необходимости подбора адекватных ресурсов по тематическому и видовому признаку, обеспечению доступа к гетерогенным ресурсам за счет унификации представления данных. Выделяются следующие свойства, определяющие эффективность взаимодействия с ресурсами:

- особенности прикладного поискового интерфейса и языка поисковых запросов;

- способы представления документов и используемые группы метаданных;
- тематические и видовые спектры ресурса;
- характер лексики, используемой в документах ресурса;
- сетевые протоколы и программы-клиенты, используемые для взаимодействия с ресурсом.

Проблема поддержки распределенного поиска имеет давнюю историю. Существующие решения классифицируются по признаку взаиморасположения индексируемых данных, поисковых механизмов и способа взаимодействия унифицированного ресурса с первоисточниками следующим образом: ресурсы, агрегирующие данные, имеющие собственный поисковый механизм (их формирование и наполнение административно и технически ограничено), ресурсы-каталоги, взаимодействующие с первоисточниками и возвращающие ссылки на результаты поиска (наиболее широко распространены, но не обеспечивают достаточного для рассматриваемой области качества поискового взаимодействия) и унифицированные поисковые оболочки с возможностью объединения и ранжирования результатов (в чистом виде в настоящее время фактически не функционируют). Таким образом, необходима разработка комплексного решения на основании системного подхода, учитывающего определенные выше свойства ресурсов, и обеспечивающего, с одной стороны, возможность динамического ассоциирования любых технически доступных ресурсов, а с другой, адекватное для рассматриваемой области качество поиска в них. Проведенный анализ публикаций и проектов показал отсутствие готовых комплексных решений, что обусловлено в первую очередь значительно более сложными, нежели в других областях, схемами элементов данных ресурсов научной информации, и более высокими требованиями к полноте и точности поиска.

Основной объект, участвующий в поисковом взаимодействии, принадлежит классу «информационный ресурс» $\widehat{R}:\{K,L\}$, который задается множеством хранимых знаний K, характеризующихся формой представления, и множеством средств манипулирования знаниями L, обеспечивающих доступ к ресурсу. Множе-

ство взаимодействующих ресурсов $I = \left\{ I_{ij} \in \widehat{I} \right\}$, где $\widehat{I}$ – класс «поисковое взаимодействие ИР», определяется как:

$$\widehat{I} : \left\langle R, S, R \right\rangle \quad (1)$$

$$S = \left\{ S_{ij} \right\}, S_{ij} = R_i \times R_j \quad (2)$$

Здесь каждый ресурс $R_i, R_j \in \widehat{R}$ ввиду симметричности взаимодействия может выступать и как источник, и как потребитель информации в процессе их взаимодействия S_{ij}.

Исходя из уровневой схемы сетевого взаимодействия, совместимость ресурсов будем рассматривать в следующих аспектах:

- техническая совместимость – по способам реализации функций обращения к ресурсу;
- информационная совместимость, которая подразделяется на:
  - лингвистическую совместимость – по языковым средствам работы с информацией, хранящейся в ресурсе;
  - метаинформационную совместимость – по интерфейсам доступа на уровне элементов данных (точек входа);
  - лексическую – по близости используемой в документах лексики, определяемой тематикой ресурса.

Для сравнительной оценки информационного содержания ресурсов рассмотрим характеристики: спектра видов документов $T_i = \left\{ \lambda_{ij} \right\}$ и тематического спектра $H_i = \left\{ \kappa_{ij} \right\}$ для *i*-го ресурса; где λ_{ij} – характеризует вероятность нахождения документов *j*-го вида в составе ресурса, κ_{ij} – характеризует вероятность нахождения документов *j*-ой тематической рубрики в составе ресурса.

Для спецификации спектра видов документов T_i вводится показатель α_i^w, отражающий наличие тех или иных видов документов, и характеристика равномерности β_i^w.

$$\alpha_i^w = \frac{\sqrt{\sum_{j=1}^{t}\left(\lambda_{ij}^{\alpha}\right)^2} - 1}{\sqrt{t} - 1} \in [0,1] \qquad (3)$$

$\lambda_{ij}^{\alpha} \in \{0,1\}$ – дискретная величина, отражающая факт встречаемости документов *j*-го вида в *i*-м ресурсе, t – мощность множества всех рассматриваемых видов документов. Вырожденный случай, соответствующий единственному виду документов ($t = 1$), не отвечает практике и поэтому исключается из рассмотрения.

$$\beta_i^w = \frac{\left(1 - \sqrt{\sum_{j=1}^{t}\left(\lambda_{ij}^{\beta}\right)^2}\right)}{\sqrt{\sum_{j=1}^{t}\left(\lambda_{ij}^{\beta}\right)^2}\left(\sqrt{t} - 1\right)} \in [0,1] \qquad (4)$$

$\lambda_{ij}^{\beta} \in [0;1]$ – вероятность нахождения документов *j*-го вида в *i*-м ресурсе.

Определим интегральную характеристику полноты представления и равномерности видового распределения документов в ресурсе как $\Lambda_i^w = \frac{\left(\alpha_i^w + \beta_i^w\right)}{2}$.

Аналогично, для спектра тем *i*-го ресурса $\Lambda_i^h = \frac{\left(\alpha_i^h + \beta_i^h\right)}{2}$.

В работе проведен эксперимент, в рамках которого группой экспертов осуществлялся информационный поиск в восьми ИР по некоторому набору тем с учетом видов документов. Полученные в результате эксперимента значения интегральных характеристик $0 < \Lambda_i^w < 0,54$ и $0,33 < \Lambda_i^h < 0,95$ подтвердили, что ни один ресурс не обладает исчерпывающей полнотой ни по видам, ни по темам документов. Для обеспечения унифицированного доступа к разнородным ресурсам научной информации за счет их объедине-

ния в рамках единой поисковой среды в работе сформулирована следующая совокупность задач:

- построение описания ИР на основании объектной модели информационного ресурса, учитывающей специфические характеристики ресурса, отвечающие за взаимодействие с ним;
- создание механизма трансляции исходного поискового запроса, записанного в синтаксисе $L_1^Q \in L^Q$ (L^Q – множество синтаксисов ЯПЗ), к синтаксису $L_2^Q \in L^Q$ ЯПЗ целевого ИР, на основе модели лингвистической совместимости ресурсов. Т.е. необходимо построить отображения f_T исходного ЯПЗ в целевой: $f_T : L^Q \mapsto L^Q, L_2^Q = f_T(L_1^Q)$;
- определение в рамках модели метаинформационной совместимости функции, позволяющей определить степень соответствия элементов данных, относящихся к двум классам;
- определение в рамках модели лексической совместимости степени близости пар ресурсов по лексике.

Во **второй главе** рассматриваются вопросы метаинформационной совместимости ресурсов. Результаты эксперимента по использованию при поиске основных элементов данных, принадлежащих набору Dublin Core (DC) показали, что две трети ресурсов допускают использование при поиске менее половины от всех элементов набора, два элемента из десяти не используются ни в одном из рассмотренных ресурсов, при этом динамика прироста совокупного словаря элементов данных показала сравнительно быстрое его наполнение.

При решении задачи метаинформационной совместимости ресурсов выделяются две стратегии формирования глобальной схемы данных (схемы-медиатора), обеспечивающей установление соответствий элементов данных из разных схем. Стратегия Global-As-View (GAV) предполагает формирование глобальной схемы на основании схем локальных источников, а стратегия Local-as-View (LAV) – введение глобальной схемы независимо от локальных. Последний подход допускает работу в условиях, когда заранее не известен набор ИР.

В работе используется комбинированная стратегия. Для построения схемы-медиатора используется стратегия GAV. Затем на

основе принятого в библиографической науке принципа разделения документа (и, соответственно, метаданных его описывающих) на ряд областей[6] строится классификация с нечетким определением и неоднозначным основанием деления (что объясняется различиями в существующих стандартах), представляемая древовидной структурой, содержащей элементы метаданных. Использование такой классификации далее происходит в соответствии со стратегией LAV, т.е. опираясь на свойства полученной классификации и принципы ее построения, по мере необходимости, осуществляется ее развитие и уточнение.

Очевидно, что каждый элемент в разных ИР может встречаться неоднократно в пределах всего дерева и именоваться поразному. Для формализации и определения меры введем понятие класса θ_A элементов данных, как абстрактного элемента, не относящегося к какой-либо конкретной схеме. Тогда в соответствии с полученной структурой, абстрактные элементы, детализирующие данный, будем называть нижестоящими классами, для которых он будет вышестоящим классом. В качестве оценки метаинформационной эквивалентности (с точки зрения замены одних элементов данных другими) введем понятие расстояния ρ как меры различия между классами элементов данных θ_A и θ_B, которое может быть определено по их координатам в структуре, т.е.:

$$\exists\Theta : \theta_A = \Theta(a_1, a_2, \ldots, a_k, 0, \ldots, 0) = \Theta(A), k = \overline{0,n} \quad (5)$$

Для расстояния между двумя соседними в иерархии классами будем иметь:

$$\rho(A,B) \sim \frac{1}{2^{k-1}} \quad (6)$$

Функция расстояния между классами с координатами $A = (a_1, \ldots, a_k, 0, \ldots, 0)$ и $B = (b_1, \ldots, b_l, 0, \ldots, 0)$ можно определить следующим образом:

---

[6] Способы такого разделения определяются в различных стандартах (например, ГОСТ 7.1-2003, МЕКОФ, USMARC, и т.д.), но не всегда совпадают.

$$\rho(A,B)=\begin{cases}0 & ,\text{если } A=B^{*(|l-k|)}\\ \dfrac{1}{2^{k-1}}\left(1-\dfrac{1}{d_A}\right) & ,\text{если } A^*=B\\ \dfrac{1}{2^{k-1}} & ,\text{если } (A^*=B^{*(|l-k|+1)})\wedge(A\neq B^{*(|l-k|)})\\ \rho(A,A^*)+\rho(A^*,B) & ,\text{иначе}\end{cases} \quad (7)$$

Здесь d_A – количество нижестоящих классов для A^*.

Практическое обеспечение метаинформационной совместимости реализуется основанной на этой модели глобальной таблицей классов элементов данных (ТКЭД). На основе установленных соответствий между унифицированными элементами из таблицы и реальными элементами данных, принадлежащими конкретным ресурсам, можно производить замены имен элементов данных при трансляции запроса. Исходя из обобщенной модели машинного поиска, элементарный запрос представляется как:

$$q=\left\{q^F,q^C,q^T\right\} \quad (8)$$

где q^F – область поиска, q^C – критерий сравнения, q^T – термин запроса (простой или составной) и его маркеры (кавычки и т.п.), и квалификаторы. Тогда язык запросов $L^Q=\left\{o,q,z\right\}$ в целом описывается следующими атрибутами: o – множество допустимых операторов-связок, q – множество элементарных запросов, z – правила совместного использования терминов и операторов в запросе (синтаксис ЯПЗ).

Определение расстояния между парами ЯПЗ L_1^Q и L_2^Q сводится к определению расстояний между их соответствующими компонентами.

Расстояние между двумя классами $\beta_1=\mathrm{B}(c_1^u,c_1^d,c_1^s)$ и $\beta_2=\mathrm{B}(c_2^u,c_2^d,c_2^s)$ определим следующим образом:

$$\eta(\beta_1,\beta_2)=\frac{c^{u0}\eta^u(c_1^u,c_2^u)+c^{d0}\eta^d(c_1^d,c_2^d)+c^{s0}\eta^s(c_1^s,c_2^s)}{c^{u0}+c^{d0}(1-p^d)+\frac{1}{2}c^{s0}}\in[0,1] \quad (9)$$

Где:

- c^u – признак оператора-связки булевого типа;
- c^d – признак оператора указания расстояния;
- c^s – признак оператора учета порядка следования.

$$\eta^u(c_1^u,c_2^u)=\begin{cases}0, если\ c_1^u=c_2^u\\ 1, иначе\end{cases} \quad (10)$$

$$\eta^d(c_1^d,c_2^d)=\begin{cases}0, если\ c_1^d\le c_2^d\\ 1-p^d, иначе\end{cases}, 0<p^d<1 \quad (11)$$

$$\eta^s(c_1^s,c_2^s)=\begin{cases}0, если\ c_1^s\le c_2^s\\ 1/2, иначе\end{cases} \quad (12)$$

Где c^{d0}, c^{u0} и c^{s0} – весовые коэффициенты расстояний между отдельными признаками. p^d – вероятность сохранения смысла запроса при переходе к оператору, не учитывающему расстояние между терминами.

Определим функцию расстояния для двух множеств доступных операторов критериев:

$$\varepsilon_C(\varsigma_1,\varsigma_2)=\frac{|\varsigma_1\setminus\varsigma_2|}{|\varsigma_1\cup\varsigma_2|}\in[0,1], \quad (13)$$

где $\varsigma_i=\{\varsigma_{ij}\}$ – множество всех доступных операторов критерия ς_{ij}.

Функцию расстояния для двух множеств доступных квалификаторов и маркеров определим, как:

$$\varepsilon_T(\sigma_1,\sigma_2)=\frac{|\sigma_1\setminus\sigma_2|}{|\sigma_1\cup\sigma_2|}\in[0,1], \quad (14)$$

где $\sigma_i = \{\sigma_{ij}\}$ – множество всех доступных квалификаторов и маркеров термина.

Расстояние как меру различия для пар языков, имеющих синтаксисы L_1^Q и L_2^Q, определим следующим образом:

$$\mu(L_1^Q, L_2^Q) = \mu^o \eta^\Sigma(\beta_1, \beta_2) + \mu^F \rho_1^\Sigma(A_1, A_2) + \\ + \mu^C \varepsilon_C(\varsigma_1, \varsigma_2) + \mu^T \varepsilon_T(\sigma_1, \sigma_2) \quad (15)$$

$$\eta^\Sigma(\beta_1, \beta_2) = \frac{\sum_{\beta_{1i} \in \beta_1} \min_{\beta_{2i} \in \beta_2} \left(\eta(\beta_{1i}, \beta_{2j})\right)}{|\beta_1|} \quad (16)$$

$$\rho_1^\Sigma(A_1, A_2) = \frac{\sum_{A_{1i} \in A_1} \min_{A_{2i} \in A_2} \left(\rho_1(A_{1i}, A_{2j})\right)}{|A_1|} \quad (17)$$

Где μ^o, μ^F, μ^C, μ^T – весовые коэффициенты (константы), определяющие влияние расстояния между теми или иными компонентами языков на расстояние между L_1^Q и L_2^Q в целом; $\mu^o + \mu^F + \mu^C + \mu^T = 1$;

ρ_1 – нормированная функция расстояния между двумя классами элементов данных.

Для оценки лексической совместимости (как одной из сторон информационной), характеризующей близость по содержанию, тематике и, следовательно, используемой лексике, при переходе от одного ресурса к другому в процессе распределенного поиска предложено использовать меру l_X, определяемую на основе вероятности $P(Q,b)$. Где $P(Q,b)$ – вероятность того, что документ D, формально релевантный произвольно заданному запросу Q по полю A ресурса R_1, будет релевантен запросу Q по полю B ресурса R_2.

Используя теоретико-множественную модель, было получено, что зависимость искомой оценки l_X от длины запроса и количества терминов в нем из словаря B имеет следующий вид:

$$l_X = \frac{1}{\alpha\left(b/Q\right)} P\left(Q,b\right) \tag{18}$$

$$P\left(Q,b\right) = \frac{\sum_{D=h}^{A} \sum_{g=h}^{\min(f,D)} \sum_{a=h}^{\min(Q,D)} \sum_{x=h}^{\min(a,b,g)} C_g^x C_{D-g}^{a-x} C_{f-g}^{b-x} C_{A-D-f+g}^{Q-a-b+x}}{\sum_{D=h}^{A} \sum_{g=0}^{\min(f,D)} \sum_{a=h}^{\min(Q,D)} \sum_{x=0}^{\min(a,b,g)} C_g^x C_{D-g}^{a-x} C_{f-g}^{b-x} C_{A-D-f+g}^{Q-a-b+x}} \tag{19}$$

где b – количество терминов в запросе, принадлежащих словарям полей A и B; Q – количество терминов в запросе.

$\alpha\left(b/Q\right) = \dfrac{c}{\left(b/Q\right)^{\gamma}}$ – коэффициент, учитывающий неравномерность распределения частот терминов в базе ресурса. График зависимости $P(Q,b)$ в координатах PQ (значение вероятности P и длина запроса Q) приведен на рисунке 1.

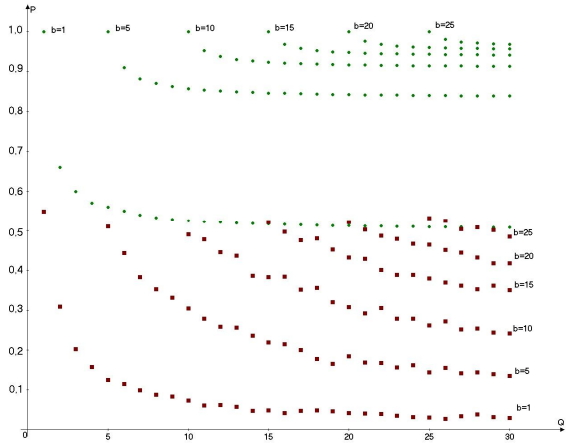


Рис. 1

Для рассмотренных в рамках данного эксперимента ИР методом аппроксимации значений отношений теоретических и экспериментальных значений вероятности для различных b/Q (рисунок 2) получены следующие значения для констант: $c = 1,82$ и $\gamma = 0,67$.

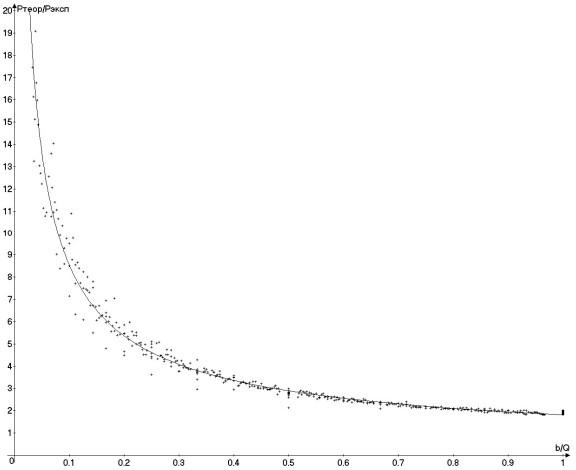


Рис. 2

Данная оценка (18) позволяет судить о лексической близости (в смысле индексирования) двух различных ресурсов и, следовательно, обеспечивает инструмент для ранжирования ИР.

В **третьей главе** рассматривается объектная модель ресурса, отражающая поисковое взаимодействие в рамках сетевой эталонной модели взаимодействия открытых систем (OSI). Показывается возможность отображения объявленных параметров ресурса на параметры протоколов трех верхних уровней OSI.

Сеансовый уровень обеспечивает функции идентификации ресурса и включает описание характеристик поисковой сессии, унификация формы и состава которых позволит обеспечить техническую совместимость ресурсов при поиске.

На представительском уровне задаются и используются характеристики, описывающие содержание ресурса, его форму и представление (основывающиеся на модели данных ресурса).

На прикладном уровне механизмы взаимодействия с ресурсом представляются объектами «запрос» (различающиеся по формам и типам) и «ответ», где, исходя из практики АИПС, можно выделить три возможные формы: документальная выдача, справочные и аналитические данные и файлы.

Такое представление ресурса может применяться к различным типам ИР вне зависимости от их внутренней структуры и организации хранящихся данных.

Формализуя свойства путем описания отдельных типов запросов, их синтаксиса и используемых переменных, а также описывая схемы данных на уровне отдельных элементов и, соотнося их с классами элементов данных, можно обеспечить формирование описания ресурса достаточного для задач обеспечения автоматизированного поискового взаимодействия в среде распределенных гетерогенных ресурсов.

Для обеспечения интероперабельности в соответствии с объектной моделью ресурса предложено объектно-ориентированное описание, средствами которого представляются такими свойствами ресурса, как: имя ресурса, параметры протокола взаимодействия, включая адрес поискового интерфейса, характеристики базы ресурса, включая используемые схемы данных и элементы данных в рамках них, синтаксис используемых запросов, включая виды поисковых запросов (описания которых составляются с использованием шаблонов регулярных выражений), а также параметры ответов, получаемых от ресурса, включая описание областей возвращаемых документов и переменных. В структуру описания также заложены параметры, отвечающие за особенности передачи запроса, выявленные в ходе проведенного анализа поисковых интерфейсов различных ИР.

В **четвертой главе** рассматривается технология проведения информационного поиска в распределенных гетерогенных информационных ресурсах. Проводится ее поэтапный анализ и выделяются основные аспекты, касающиеся вопросов автоматизации процесса поиска в нескольких ресурсах, такие как формулирование запроса, выбор подходящего ресурса для поиска, преобразование запроса с учетом установленных соответствий элементов данных, а также отправка запроса ассоциированному ресурсу и обработка ответа от него.

Приводится описание разработанных программных средств, обеспечивающих автоматизированный распределенный поиск в гетерогенных ИР, реализующих как техническую (средствами программных клиентов для отдельных протоколов), так и информационную совместимость (на основе моделей метаинформационной, лингвистической и лексической совместимости) ассоциированных ресурсов, свойства которых специфицированы в соответствии со структурной моделью ресурса и хранятся в разработанном репозитории ресурсов.

Описываются результаты экспериментального поиска, проведенного по множеству распределенных ресурсов, включающего электронные библиотеки, ресурсы университетов, издательств, а также поисковые машины, в рамках которого для обращения к ассоциированным ресурсам использовались разработанные программные средства. Результаты показали повышение качества документальной выдачи более чем в два раза.

В **заключении** приведены основные результаты исследования:

1. Проведен анализ публикаций и проектов по проблемам поддержки пользователя при работе в среде гетерогенных информационных ресурсов, показавший отсутствие готовых комплексных решений.
2. Выведены основные характеристики поискового взаимодействия в среде гетерогенных распределенных ИР и определены интегральные характеристики для оценки информационного содержания ресурса.
3. Проведены экспериментальные исследования, позволившие оценить использование элементов данных при поиске в различных ИР.
4. Построена модель метаинформационной совместимости, в рамках которой определена функция расстояния как мера различия для классов элементов данных.
5. Построена модель лингвистической совместимости, в рамках которой определена функция расстояния как мера различия для пар ЯПЗ, используемая в процедурах преобразования запроса из синтаксиса одного ЯПЗ в синтаксис другого.
6. Построена модель лексической совместимости ресурсов и получена оценка этой совместимости, характеризующая близость лексики ИР, обусловленной тематикой.

7. Разработана объектная модель ресурса, учитывающая основные характеристики ресурса в задачах обеспечения распределенного поиска. На ее основе разработано объектно-ориентированное представление ИР, обеспечивающие их интероперабельности.

8. Разработана унифицированная поисковая оболочка, основанная на предложенных моделях и реализующая технологию распределенного поиска, более двух лет функционирующая в составе АИПС xIRBIS в режиме промышленной эксплуатации. Разработан инструмент, обеспечивающий взаимное отображение схем элементов данных для различных ресурсов. С использованием разработанных программных средств проведена серия экспериментальных поисков в ассоциированных информационных ресурсах, результаты которых показали повышение качества документальной выдачи за счет обеспечения адекватного требованиям целевого ресурса преобразования поискового запроса, что подтверждает эффективность разработанных методов и средств.

В **приложениях** представлены следующие дополнительные материалы.

В приложении 1 приводятся экспериментальные данные, использованные для оценки рассеяния документов по ресурсам.

В приложении 2 приводятся экспериментальные данные, характеризующие возможность использования элементов данных при поиске в различных ИР.

В приложении 3 приведены данные, использованные для экспериментального расчета меры различия для двух ЯПЗ.

В приложении 4 приводятся преобразования, проведенные в рамках вывода функции меры лексической совместимости.

В приложении 5 приводится вывод функции зависимости для верхней оценки вероятности в рамках модели лексической совместимости.

В приложении 6 приводится диаграмма классов объектной модели ресурса.

В приложении 7 приведена структура разработанного объектно-ориентрованного описания в нотации XML-схем.

В приложении 8 приведены акты о внедрении полученных автором результатов.

Публикации:

1. Максимов Н.В., Голицына О.Л., Васина Е.Н., Резниченко П.И. Окропишин А.Е. Документальная информационно-аналитическая система xIRBIS 4.0 // Свидетельство о государственной регистрации программы для ЭВМ №2008611511 от 25.03.2008 г.

2. Максимов Н.В., Васина Е.Н., Голицына О.Л. Окропишин А.Е. Документальная информационно-аналитическая система // Научная сессия МИФИ-2009. XIII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании». Сборник научных трудов. М.: МИФИ, 2009. – С.140-141

3. Максимов Н.В., Васина Е.Н., Голицына О.Л., Резниченко П.И., Окропишина О.В., Окропишин А.Е. Интегральная информационная система поддержки научных исследований и процессов управления научными кадрами // Научная сессия МИФИ-2009. XIII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании». Сборник научных трудов. М.: МИФИ, 2009. – С.25-26

4. Максимов Н.В., Окропишина О.В., Окропишин А.Е. Дескриптивное представление предметных областей научных исследований // Инновационные технологии когнитивного управления в экономике, менеджменте и образовании. Межвузовский сборник научных трудов. – М.: «РЭА», 2009. –Вып. 2. с.190-197.

5. Максимов Н.В., Бебчук М.Б. Окропишин А.Е. Об одном подходе к обеспечению совместимости ИПЯ в задачах документального поиска в распределенных гетерогенных информационных ресурсах // Инновационные технологии когнитивного управления в экономике, менеджменте и образовании. Межвузовский сборник научных трудов. – М.: «РЭА», 2009. –Вып. 2. с.21-27.

6. Максимов Н.В., Окропишина О.В., Окропишин А.Е. Об одном подходе к созданию информационной среды, обеспечивающей возможность формирования и управления индивидуальной образовательной траекторией // Математика, информатика, естествознание в экономике и в обществе / Труды международной научно-практической конференции. Том 1- М.: МФЮА, 2009. - С. 55-58, ISBN 978-5-94811-139-1

7. Окропишина О.В. Окропишин А.Е. Дескриптивное представление знаний // Информационные технологии в образовании.

XIX международная конференция-выставка: Сборник трудов. Ч. II. – М.:МИФИ, 2009. – С. 25-28.

8. Окропишин А.Е. Методы и средства организации документального поиска в распределенных гетерогенных информационных ресурсах // Научная сессия МИФИ-2010. Сборник научных трудов. XIV выставка-конференция «Телекоммуникации и новые информационные технологии в образовании». М.: НИЯУ МИФИ, 2010. – С.156-157

9. Максимов Н.В., Голицына О.Л., Амосов П.А., Окропишина О.В. Окропишин А.Е. О введении и использовании информационно-лингвистических средств в единой образовательной среде научно-исследовательского университета // Единая образовательная информационная среда: направления и перспективы развития электронного и дистанционного обучения : материалы IX Международной научно-практической конференции-выставки (Новосибирск, 22-24 сентября 2010 г). – Новосибирск : Изд-во НГТУ, 2010. – С.95-97.

10. Окропишин А.Е. Средства и технологии документального поиска в образовании и научных исследованиях // Информационные технологии в образовании. XX Международная конференция-выставка: Сборник трудов. Ч. VI. – М.: МИФИ, 2010. – С.34-36.

11. Степанова Е.Б., Окропишина О.В., Окропишин А.Е. Болотин Е.И., Амосов П.А. Разработка стандартизованных модулей учебно-методических комплексов по дисциплинам // Научная сессия НИЯУ МИФИ-2011. Сборник научных трудов. XV выставка-конференция «Телекоммуникации и новые информационные технологии в образовании».М.: НИЯУ МИФИ, 2011. - Т.1 – С.131-132

12. Окропишин А.Е. Средства и технологии распределенного документального поиска в информационно-образовательных средах // Научная сессия НИЯУ МИФИ-2011. Сборник научных трудов. XV выставка-конференция «Телекоммуникации и новые информационные технологии в образовании». М.: НИЯУ МИФИ, 2011. - Т.1 – С.148-149

13. Окропишин А.Е. Применение модели вариативности индексирования для задач документального поиска в распределенных гетерогенных информационных ресурсах // Научная сессия НИЯУ МИФИ-2011. Сборник научных трудов, Т.3. – М.: МИФИ, 2011. – с.138-139.

14. Максимов Н.В., Голицына О.Л., Окропишина О.В. Окропишин А.Е. Подсистема аналитической обработки документальной

информации // Свидетельство о государственной регистрации программы для ЭВМ №2011611694 от 22.02.2011 г.

15. Максимов Н.В., Окропишина О.В., Передеряев И.И. Окропишин А.Е. Использование технологии автоматизированного формирования понятийной структуры предметной области научного исследования в задачах управления научными кадрами // Вестник РГГУ. Научный журнал. Серия «Управление» – М.: «Российский государственный гуманитарный университет», 2011. –№. 4. – С.175-185.

16. Строгонов В.И., Максимов Н.В., Голицына О.Л., Болотин Е.И., Окропишин А.Е. Модели и эффективность распределенного поиска в документальных информационных ресурсах // Системы управления и информационные технологии, №1(47), 2012. – С. 78-83.

17. Строгонов В.И., Максимов Н.В., Голицына О.Л., Окропишин А.Е. Интегральный подход к формированию и использованию распределенных гетерогенных информационных ресурсов для онлайн-информирования пассажиров // «Технические и программные средства систем управления, контроля и измерения» (УКИ'12): Конференция с международным участием (16-19 апреля 2012 г., Москва, Россия). – Москва: Изд-во ИПУ РАН, 2012. – С.77.

18. Строгонов В.И., Максимов Н.В., Окропишин А.Е. Модель информационного ресурса как объекта поискового взаимодействия // Системы управления и информационные технологии, 2012. 1(50) С. 183-186.