На правах рукописи

АМАМРА РУШДИ АХМАД

РАЗРАБОТКА МЕТОДОВ И АЛГОРИТМОВ ТЕМАТИЧЕСКИ ОРИЕНТИРОВАННОГО РАСПРЕДЕЛЕННОГО ПОИСКА ИНФОРМАЦИИ В ГЛОБАЛЬНЫХ СЕТЯХ ТИПА ИНТЕРНЕТ

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

САНКТ-ПЕТЕРБУРГ 2002

Работа выполнена в Санкт-Петербургском Государственном Техническом Университете.

Научный руководитель:

доктор технических наук, профессор Шкодырев В. П.

Официальные оппоненты:

Доктор технических наук, профессор Александров А. М.

Кандидат технических наук, Лазарев А. Г.

Ведущая организация:

Библиотека Российской Академии Наук

г. Санкт-Петербург

Защита состоится "___" мая 2002 г. в 16 час. 00 мин. на заседании диссертационного совета Д 212.229.18 при Санкт-Петербургском Государственном Техническом Университете по адресу: 195251, Санкт-Петербург, ул. Политехническая, 29, корпус № 9, ауд.325.

С диссертацией можно ознакомиться в библиотеке университета

Автореферат разослан "___" _____ 2002 г.

Ученый секретарь
диссертационного совета Шашихин В.Н.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Переход к информационному обществу XXI в. породил концентрацию информации в глобальных компьютерных сетях, распределенной по хранилищам – базам данных. Для полного использования потенциальных возможностей работы с информацией в сетях необходима организация эффективного ее поиска, что на сегодняшний день остается нерешенной проблемой. Беспрецедентный рост информации в глобальных информационных компьютерных сетях резко обостряет проблему создания высокоэффективных информационно-поисковых систем (ИПС). Существующие системы – AltaVista, Google и т.д., как правило, основаны на централизованной архитектуре, имеющей жесткие ограничения на проведение поиска.

Хорошо известно, что поисковые системы с централизованной архитектурой не в состоянии индексировать всю информацию, опубликованную в Интернет: скорость роста информации значительно превосходит возможности индексирования любой централизованной поисковой системы, чьи ресурсы всегда ограничены. Эти системы не обладают свойствами высокой точности и оперативности, в результате пользователь получает много лишней информации.

Необходимо найти масштабируемое решение данной проблемы. Возможное решение – использование распределенных поисковых систем с децентрализованной архитектурой. Распределенная поисковая система состоит из множества компонент нескольких типов: тематических коллекций (индексов); агентов, формирующих тематические индексы; брокеров, выбирающих тематические индексы и документы в соответствующем индексе. Различные компоненты в рамках единой системы могут принадлежать различным независимым владельцам, конкурирующим друг с другом на рынке информационных услуг. Именно эта конкурентная борьба и экономическая заинтересованность позволяют привлечь ресурсы, необходимые и достаточные для проведения индексирования всей опубликованной в Интернет информации и организации ее эффективного поиска.

Как показано в диссертационной работе, основные характеристики перечисленных компонент отвечают свойствам и требованиям глобальных информационных сред, что открывает возможность построения систем, в частности поисковых, способных эффективно работать в указанных средах.

В рамках системы с централизованной архитектурой вся информация хранится в

одном индексе, что накладывает определенные ограничения на возможность проведения поиска, учитывающего тематическую специфику искомой информации. В распределенной системе информация хранится в большом числе различных тематически ориентированных коллекций. Тематическая ориентация коллекции позволяет повысить качество поиска для тех пользователей, тематика информационных потребностей которых соответствует тематической направленности данной коллекции. В этом случае можно использовать более эффективные "интеллектуальные" методы поиска, сохранять информацию о пользователях, часто пользующихся услугами данной коллекции. Кроме того, тематическая ориентация коллекции позволяет использовать для сканирования Интернет тематических агентов, ориентированных на поиск документов по заданной относительно узкой тематике. Это позволяет значительно эффективнее использовать ресурсы, выделяемые на индексирование опубликованной в Интернет информации.

Диссертационная работа посвящена решению проблемы построения эффективной распределенной системы поиска релевантной информации в глобальных компьютерных сетях на основе развития тематических агентов, брокеров и выявления информационных потребностей конкретного пользователя с использованием метода вероятностного латентного семантического индексирования.

Целью настоящей диссертационной работы является разработка принципов построения, архитектуры, методов и алгоритмов поиска для функционирования в распределенных поисковых системах, *повышающих точность и оперативность результатов поиска* информации в сложных глобальных средах гипертекстовой информации типа Интернет.

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

- Разработка алгоритма и архитектуры тематического сетевого робота;
- Разработка алгоритма и архитектуры брокера, осуществляющего маршрутизацию запросов пользователя;
- Разработка настраиваемого пользовательского интерфейса;

Предметом исследования являются

- Построение описания тематики индекса на основе использования метода вероятностного латентного семантического индексирования.

- Маршрутизация запросов в рамках системы распределенного поиска, производимого брокером.
- Выявление информационных интересов конкретного пользователя на основе анализа тематической принадлежности документов, возвращаемых системой поиска в ответ на запросы пользователя.
- Оценка эффективности применения метода вероятностного латентного семантического индексирования.

***Методы исследования**:*

В основу проводимых исследований положены методы теории вероятности, теории графов, линейной алгебры, семантического анализа, теории принятия решения и тестирование на реальных данных в Internet.

Научная новизна и основные результаты

В процессе исследований и теоретических обобщений получены следующие результаты:

1. Разработан и реализован информационный агент и его архитектура для формирования тематических коллекций (индексов). При этом решены некоторые новые задачи и даны новые решения для ранее известных задач. В том числе:
   - Предложено новое построение тематической фильтрации потока документов с использованием двух типов фильтров, рекомендующих документ для включения в коллекцию (индекс); фильтр ядра индекса, построенный на основе анализа содержимого коллекции, и фильтр запросов, построенный на основе анализа архива пользовательских запросов к данному индексу, который позволяет тематике коллекции не устаревать с течением времени
   - Предложен новый итерационный метод оценки значимости (вычисления весов) термов из запросов пользователей и доказана его сходимость.
   - Предложен новый метод формирования и управления очередью ссылок на документы (для обхода Internet), подлежащих загрузке из сети агентом при формировании тематического индекса, на основе динамического оценивания вероятности того, что документ, на который указывает данная ссылка, релевантен заданной тематике.

2. Модифицирован к особенностям распределенной поисковой системы и реализованы архитектура и алгоритм маршрутизации запросов пользователей. В том числе:
   - Предложен новый способ оценивания весов термов в запросах и документах. При решении задачи маршрутизации запросов, запрос сопоставляется с описаниями индексов, в которые этот запрос может быть направлен для поиска. Обычно при формировании описания индекса используется частотная модель для вычисления весов термов. В данной работе предложен способ взвешивания термов, основанный на использовании метода вероятностного латентного семантического индексирования.
   - Выведено новое, более простое для проверки достаточное условие оптимальности найденного решения задачи маршрутизации запросов. Известное условие оптимальности требует построения ряда числовых последовательностей и проверки того, что все они являются монотонно возрастающими. Найдено более простое эквивалентное условие, показано, что оно выполняется во всех практически важных случаях.
3. Предложен и реализован новый подход к построению настраиваемого на пользователя интерфейса и его сценарий в поисковой системе. В том числе:
   - Предложен новый метод выявления информационных потребностей пользователя;
   - Разработан метод оценивания релевантности документа, возвращаемого поисковой системой в ответ на запрос пользователя, информационным потребностям данного пользователя.

Применение разработанных методов и алгоритмов в компонентах распределенных поисковых систем повысило адаптацию и обеспечило высокую релевантность результатов поиска.

Положения, выносимые на защиту:

- Архитектура и алгоритм тематически ориентированного поиска (в средах гипертекстовой информации) тематическим агентом на основе вероятностного латентного семантического индексирования, формирующего тематический индекс.
- Архитектура и алгоритм брокера, осуществляющего маршрутизацию запросов пользователя для выбора тематических коллекций на основе вероятностного латентного семантического индексирования и оптимального распределения ресурсов.

- Сценарий и алгоритм работы интерфейса пользователя, выявляющий информационные потребности конкретного пользователя.
- Архитектура распределенной децентрализованной поисковой системы для поиска в гипертекстовой информационной среды.
- Метод индексирования тематического индекса, позволяющий повысить качество поиска информации.

Практическую ценность диссертационной работы представляют следующие результаты:

- Разработанный метод интеллектуального поиска релевантной информации в гипертекстовой информационной среде (Интернет), позволяющий существенно сократить объем просматриваемой при поиске информации за счет ориентации поиска в перспективных для тематического индекса направлениях.
- Комплекс программ, реализующих алгоритм тематического агента, служащий основой при формировании тематического индекса (коллекции) поисковых систем нового поколения с более высокой точностью поиска, меньшей нагрузкой на сетевые ресурсы, способностью настройки на интересы пользователя-владельца.
- Комплекс программ, реализующих алгоритм маршрутизации запросов, пользователей позволяющий существенно повысить точность релевантности поиска за счет максимизации доходов пользователя и оптимального распределения ресурсов выделенных пользователем для поиска.
- Сценарий интерфейса пользователя, который облегчает получением пользователем общего впечатления о результатах поиска за счет хранения, анализа и кластеризации данных о тематической принадлежности документов, возвращаемых системой поиска в ответ на запрос пользователя.
- Программный модуль, реализующий разработанный сценарий интерфейса пользователя, который служит для выявления информационных потребностей пользователя, сокращает время навигации пользователя по результатам поиска в ответ на его запрос за счет сопровождения каждого документа тематической меткой.
- Практические рекомендации для проектирования действующих поисковых систем с децентрализованной архитектурой на основе разработанных вариантов функционирования распределенных децентрализованных поисковых систем на базе тематического

агента, брокера, тематического индекса, интеллектуального интерфейса пользователя.

Апробация работы.

Основные результаты работы докладывались и обсуждались на III научно-методической конференции "Internet-технологии и современное общество" (Санкт-Петербург, 2000), международной научно-методической конференции "Телематика" (Санкт-Петербург, 2000 и 2001), международной конференция по мягким вычислениям и измерениям (Санкт-Петербург, 2000 и 2001).

Внедрение результатов работы.

Результаты работы реализованы в виде программного обеспечения и использованы в Российской Национальной Библиотеке и в информационно-справочной системе АТЛАНТ и на СПб. Фьючерской бирже.

Публикации. Результаты, полученные в работе, нашли отражение в 12 печатных работах, из них две – в журнале "Приборостроение", семь работ были опубликованы в сборниках научных трудов международных конференций.

Структура работы. Диссертационная работа состоит из введения, трех глав и заключения, изложенных на 132 страницах, содержит 16 рисунков, 4 таблицы и 7 приложений на 78 страницах; всего 210 страниц.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы, сформулированы цель и задачи исследования.

В **первой главе** проведен краткий анализ основных свойств глобальной гипертекстовой информационной среды WWW, где подчеркиваются такие ее качества, как сверхвысокая распределенность, децентрализованность, структурированность и гигантский объем хранимой в ней информации на основе гипертекстовой модели данных, что позволяет сделать вывод о необходимости построения автоматических ПС, облегчающих пользователю задачу поиска релевантной информации.

Дается общее описание функций и требований к поисковой системе WWW, обзор существующих в настоящее время поисковых систем (ПС), выведена обобщенная

архитектура классических ПС, определены вытекающие из этой архитектуры принципиальные недостатки: в рамках системы с централизованной архитектурой вся информация хранится в одном индексе, что накладывает определенные ограничения на возможность проведения поиска, учитывающего тематическую специфику искомой информации. Определены также основные проблемы, препятствующие полному удовлетворению информационных потребностей пользователя. Выявленные недостатки позволяют сделать вывод о необходимости разработки ПС новой архитектуры, адаптированной к основным свойства глобальных информационных систем и информационных потребностей пользователя.

Рассмотрены известные модели представления данных, среди которых булева модель, модель векторного пространства, разнообразные вероятностные пространства. Определено, что метод вероятностного латентного семантического индексирования представляет собой относительно новую и неисследованную модель, которая может найти широкое применение в области информационного поиска и удовлетворять требованиям к решению поставленных задачам.

На основании материала первой главы делается вывод, что масштабируемым решением проблем поиска информации может являться система распределенного поиска с децентрализованной архитектурой. Предложенная модель распределенной ПС состоит из компонентов нескольких типов: агентов, тематических индексов, брокеров. Кроме того, различные компоненты в рамках единой системы могут принадлежать различным владельцам, что позволит привлечь ресурсы, необходимые для проведения индексирования опубликованной в Интернет информации.

Вторая глава посвящена разработке алгоритма и архитектуры тематического информационного агента в информационной среде гипертекстовой организации. Описывается общая методология компонентов: анализатора ядра индекса, анализатора архива запросов, генератора тематических фильтров; компоненты, обеспечивающая фильтрацию документов и компоненты, управляющую очередью ссылок.

Для реализации процедуры тематически-ориентированного поиска разработана модель агента, важнейшая задача которого, состоит в пополнении коллекции (индекса) новыми релевантными ее тематике документами. Как правило, работа такого агента начинается с некоторого множества HTML документов (ядра коллекции), заданных администратором коллекции. Далее агент загружает документы, на которые ссылаются уже

включенные в коллекцию документы, и рекомендует к включению в коллекцию те из них, которые проходят через фильтр, сформированный на основе анализа ядра коллекции. Такой подход имеет свои недостатки. Например, ядро коллекции может не отражать некоторые важные темы (особенно недавно появившиеся) и, следовательно, агент будет пропускать документы, которые следовало бы включить в коллекцию. В данной работе для устранения этого недостатка предлагается при фильтрации документов использовать два фильтра и рекомендовать в коллекцию документ, прошедший хотя бы через один из них. Первый фильтр отражает содержимого ядра коллекции и представляет тем самым интересы ее администратора. Второй основан на анализе архива запросов пользователей, полученных данной коллекцией за определенный период времени. Запросы пользователей отражают информационные потребности сообщества пользователей, которые должны учитываться при пополнении коллекции новыми документами. Это позволит поддерживать в течение всего времени жизни коллекции ее адекватность интересам пользователей. Построение обоих фильтров основано на использовании относительно нового метода вероятностного латентного семантического индексирования (PLSI). Использование данного метода позволяет выявить в коллекции заданное число латентных факторов, представляющих более узкие подтемы, затронутые в документах коллекции. Для каждого фактора можно оценить силу связи данного фактора с каждым словом из словаря коллекции и с каждым ее документом. Все это дает возможность выявить для данной коллекции наиболее важные темы и наиболее важные для каждой темы слова, которые и формируют фильтр, основанный на анализе ядра коллекции.

Фильтрация новых документов: Новый документ, загруженный из Интернет, рекомендуется коллекции, если он проходит хотя бы через один из фильтров: фильтр коллекции и фильтр запросов. При построении фильтра коллекции выполняется анализ ядра коллекции методом PLSI. Обозначим через $\boldsymbol{Z = \{z_i | i = 1, \ldots, k\}}$ так называемые скрытые (латентные) факторы - идентификаторы относительно узких тем, представленных в ядре коллекции. Метод PLSI позволяет вычислить оценки для следующих вероятностей: $\boldsymbol{P(z_i)}$ - вероятность того, что случайно выбранный из ядра коллекции документ относится к тематике $\boldsymbol{z_i}$ (вес фактора z); $\boldsymbol{P(d_j | z_i)}$ - вероятность то го, что документ $\boldsymbol{d_j}$ относится к тематике $\boldsymbol{z_i}$ (сила связи документа $\boldsymbol{d}$ с фактором z); $\boldsymbol{P(t_j | z_i)}$ - вероятность того, что для заданной тематики $\boldsymbol{z_i}$, терм $\boldsymbol{t_j}$ лучше всего соотносится с тематикой $\boldsymbol{z_i}$ (сила связи терми-

на t с фактором z). Для всех термов из словаря ядра индекса $T(K(D))$ предлагается вычислять их веса по следующей формуле

$$W(t) = \sum_{z \in Z} P(z)P(t \mid z),\ t \in T(K(D)).$$

В качестве фильтра коллекции выбирается заданное количество слов из словаря коллекции с наибольшими весами.

Фильтр запросов отражает информационные потребности всего сообщества пользователей. В этот фильтр входят только те слова, которые встречаются в архиве запросов. Среди них имеются как слова из словаря коллекции, так и новые слова, представляющие новые темы, не представленные в ядре коллекции. Веса слов из словаря коллекции уже вычислены при построении фильтра коллекции. Для вычисления весов новых слов используется следующий подход. Построим граф G, вершинами которого являются все слова, встречающиеся в архиве запросов. Две вершины t_i и t_j соединены ребром ($(t_i\ t_j) \in E$ - множество ребер), если эти два терма встречается вместе хотя бы в одном запросе. Вес нового слова равен среднему арифметическому весов всех слов, являющихся соседями данного слова в графе G. Для вычисления этих весов используется метод простой итерации. Сходимость метода гарантируется, если каждое новое слово встречается с каким-либо из слов из словаря коллекции хотя бы в одном запросе из архива запросов. В противном случае итерационный процесс прерывается по достижении заданного числа итераций. При этом веса некоторых новых слов могут быть оценены с большой погрешностью, но эта погрешность будет уменьшаться при появлении новых запросов, включающих данные новые слова. В фильтр запросов включается заданное число слов (встречающихся в архиве запросов) с наибольшими весами. При сопоставлении нового документа с фильтром вычисляется скалярное произведение tf профайла документа с фильтром. Документ проходит через фильтр, если это скалярное произведение превышает заданный порог.

Алгоритм функционирования агента содержит следующие основные этапы. **1)** Генерация фильтра коллекции. **2)** Генерация фильтра запросов. **3)** Инициализация дерева URL. На этапе инициализации формируется дерево URL с двумя уровнями. На первом уровне - корень дерева, на втором - узлы, содержащие стартовые URL заданные администратором коллекции. Каждому узлу υ приписывается оценка $p(\upsilon)$ вероятности того, что ссылка из соответствующего документа указывает на документ релевантный те-

матике коллекции. Для стартовых URL на этапе инициализации эти оценки принимаются равными 1. **4)** Выбор URL по алгоритму: выбор сайтов, для которых $t_{текущее} \geq t_{посещение} + \delta$ (в экспериментах $\delta = 5$ сек); выбор такого υ (из выбранных сайтов), для которого $p(\upsilon)$ максимально**.** Если URL из υ уже загружен, то: выбирается случайным образом нерассмотренной ссылки из документа с данным URL; формируется новый узел υ', родитель которого – узел υ, принимается оценка $p(\upsilon') = 1$ и вычисляется переход к выбору сайтов, для которых $t_{текущее} \geq t_{посещение} + \delta$. Далее - загрузка документа с заданным URL. **5)** Загрузка и фильтрация документа. С помощью программы wget загружается документ с заданным URL. Выполняется разбор текста документа - выделяются ссылки на другие html документы и вычисляется tf -профайл загруженного документа. **6)** Модификация дерева URL: проверяется документ в узле υ' на прохождение фильтров (релевантность индексу). Если документ не релевантен, то: вероятность $p(\upsilon') = 0$ или $p(\upsilon')$ уменьшается и полагается равной $p(\upsilon) = \dfrac{1 + links_+(\upsilon)}{1 + links_+(\upsilon) + links\_(\upsilon)}$, где – υ родитель υ'; величина $links_+(\upsilon)$ равна числу проверенных релевантных ссылок из документа в узле υ, а $links\_(\upsilon)$ - числу проверенных нерелевантных ссылок из того же документа. Если документ релевантен, то: вероятность $p(\upsilon') = 1$ или $p(\upsilon')$ увеличивается и будет равна $p(\upsilon) = \dfrac{1 + links_+(\upsilon)}{1 + links_+(\upsilon) + links\_(\upsilon)}$.

Для выбора оптимальных значений порогов используемых в фильтре коллекции и в фильтре запросов был проведен эксперимент. В качестве ядра коллекции была выбрана небольшая коллекция по тематике *информационного поиска*. В качестве архива запросов использовались 100 относительно коротких фраз, выбранных из других документов, связанных с тематикой информационного поиска. Были выбраны 25 пар значений пороговых величин. Для каждой пары агент стартовал с одного и того же множества стартовых URL и останавливался после загрузки $n_+ = 200$ прошедших фильтрацию документов. Средняя итоговая точность работы агента оценивается величиной 0.793. В проведенных экспериментах максимальная точность была достигнута при пороге фильтре ядра индекса $T_{index} = 0.1$ и пороге для фильтра запросов $T_{queries} = 0.5$.

Третья глава посвящена разработке алгоритма и архитектуры брокера, осу-

ществляющего маршрутизацию запросов пользователя (выбор тематического индекса).

В системе распределенного поиска брокер является ключевым звеном, решающим задачу оптимального распределения ресурсов, выделенных пользователем на выполнение поиска. Каждый индекс имеет свои расценки на поиск, зависящие от объема собранной в данном индексе информации и от качества, обеспечиваемого данной коллекцией поиска. В связи с независимостью коллекций, возможна ситуация, когда имеется несколько коллекций, тематика которых близка тематике запроса. В этом случае на брокер возлагается основная функция оценки числа релевантных документов, которые пользователь может получить от каждой коллекции в ответ на его запрос.

Архитектура брокера. Пользователь предоставляет брокеру (через интерфейс пользователя) следующую информацию: запрос, состоящий из нескольких ключевых слов (терминов); общее число документов, которые он желает получить в ответ на запрос (***nd***); потери пользователя от получения и просмотра нерелевантных документов (штраф за получение нерелевантного документа - C^{-}); доход от получения релевантного документа (C^{+}).

Репозитарий хранит информацию о зарегистрированных тематических индексах. Брокер использует его информацию для решения задач маршрутизации запроса. Репозитатий хранит следующую информацию о каждой коллекции: описание коллекции (некоторое множество термов с весами и некоторая статистика); стоимость доставки одного документа пользователю (сюда входят и затраты на индексирование документов, поиск и т.п. - C_i^d);

Основные компоненты брокера: **1)** Анализатор - формирует оценку числа документов, релевантных запросу пользователя, для каждой из коллекций, представленных в репозитарии. **2)** Маршрутизатор - решает основную задачу: оценивает оптимальное число документов, которые следует запросить у каждой из коллекций. **3)** Подсистема рассылки запросов и получения результатов - обеспечивает связь с выбранными коллекциями, пересылает им запросы пользователя с сопроводительной информацией, получает результаты и передает их пользователю после некоторой предварительной обработки (объединение, ранжирование, кластеризация и т.п.).

Основной алгоритм.

1. Оценка числа документов, релевантных запросу пользователя: Обозначим через $D = \{D_1, \ldots, D_n\}$ множество коллекций, в которых выполняется поиск. Для оценки

числа документов в коллекции D_i релевантных запросу q, предлагается следующая формула $R_i = C\sum_{t\in q} W_{t,i}V_{t,i}$, где C – некоторая константа; $W_{t,i}$ - вес терма t в запросе q при направлении его в коллекцию D_i; $V_{t,i}$ - вес терма t в коллекции D_i. Для вычисления указанных весов используется следующая формула: $W_{t,i} = \frac{(k'+1)f_t'}{k'L_i' + f_t'}$ Здесь k'- некоторая константа; f'_t - число вхождений терма t в запрос q; L'_i - число термов запроса q, деленное на среднее число термов в документе в коллекции D_i. Параметр k' позволяет регулировать влияние величин f'_t и L'_i. Для вычисления $V_{t,i}$ используется следующая формула $V_{t,i} = \sum_{d\in D_i} \frac{(k+1)f_{t,d}}{kL_{d,i} + f_{t,d}} \log\left(\frac{N_i - n_{t,i} + h}{n_{t,i} + h}\right)$. Здесь k - некоторая константа; $f_{t,d}$ - число вхождений терма t в документе d; $L_{d,t}$ - число термов документа d, деленное на среднее число термов в документе коллекции D_i; N_i - число документов в коллекции D_i; l_d - длина документа d; $n_{t,i}$ - число документов, в коллекции D_i, содержащих терм t, $h = ½$.

При реальной работе брокера не вся указанная выше информация будет ему доступна. Часть данных будет представлена в интегрированном виде как описание коллекции. В частности, из описания коллекции D_i брокер получит следующую информацию: $L_{avr}(i)$ - среднее число термов в документах коллекции D_i, $V_{t,i}$. По заданному запросу q, используя данное описание коллекции, можно вычислить оценку числа релевантных документов R_i с точностью до выбора констант C и k'. Выбор этих констант принадлежит брокеру.

Выбор константы k и величины $f_{t,d}$ производится на стороне коллекции. Величина $f_{t,d}$ характеризует силу связи между документом d и термом t. Частота вхождений терма в документ не является идеальной характеристикой, описывающей эту связь. Модель вероятностного латентного семантического индексирования дает более реалистичные оценки силы связи документа d и терма t. На втором этапе наших экспериментов, вместо выражения $f_{t,d}$ использовалась величина $f_{plsi\,t,d} = \frac{P(t,d)}{\sum_{t'\in q} P(t',d)} lenght\,(d)$, где $P(t,d) = \sum_z P(z)P(t\mid z)P(d\mid z)$. В соответствии с принципом максимального прав-

доподобия, $P(z)$, $P(d|z)$ и $P(t|z)$ определяются путем максимизации функции правдоподобия $L = \sum_{d \in D_i} \sum_{t \in T_i} f_{t,d} \log P(t,d)$. На этапе максимизации функции L использовался стандартный метод *оценивания-максимизации*. На каждой итерации выполняется шаг оценивания $P(z \mid d,t) = \frac{P(z)P(d \mid z)P(t \mid z)}{\sum_{z'} P(z')P(d \mid z')P(t \mid z')}$, а затем шаг *максимизации*

$$P(t \mid z) = \frac{\sum_d f_{t,d} P(z \mid d,t)}{\sum_{d,t'} f_{t',d} P(z \mid d,t')}; \quad P(d \mid z) = \frac{\sum_t f_{t,d} P(z \mid d,t)}{\sum_{d',w} f_{t,d'} P(z \mid d',t)}; \quad P(z) = \frac{\sum_{d,t} tf(d,t) P(z \mid d,t)}{\sum_{d,t} tf(d,t)}.$$

В экспериментах начальные значения для функций $P(z)$, $P(d|z)$ и $P(t|z)$ выбирались случайным образом, и сходимость наблюдалась после 50-70 итераций.

2. Алгоритм маршрутизации: решает задачу оптимального выбора коллекций, в которые будет направлен запрос пользователя. На первом этапе определяется качество поиска в D_i функцией $EP_i(s)$ - математическим ожиданием доли релевантных документов среди первых s документов (точность), полученных из D_i в ответ на запрос: $EP_i(s) = \frac{P_i^0 \cdot R_i}{R_i + sP_i^0}$. Здесь R_i - оценка числа релевантных документов в коллекции D_i, вычисленная в предыдущем пункте, а P_i^0 - вероятность того, что первый полученный из D_i документ релевантен запросу. Мы полагали $P_i^0 = 1$. На втором этапе оценивается стоимость получения заданного числа s документов из D_i.: $EC_i(s) = s(C_i^d + C^-) - sEP_i(s)(C_i^d - C^-)$. где ($EC_i(s)$ задает математическое ожидание итоговой стоимости получения s документов из коллекции D_i.

Брокер должен выбрать некоторое множество коллекций и направить запрос в выбранные коллекции с указанием числа документов, которые должна вернуть каждая коллекция. Решение задачи ищется в виде вектора $\bar{s} = (s_1, \ldots, s_n)$, где каждое s_i – неотрицательное число и $\sum_{i=1,\ldots,n} s_i = nd$ На оптимальном векторе $\bar{s}$ должен достигаться минимум стоимости поиска $\sum_{i=1,\ldots,n} EC_i(s_i)$. Вычислим оценку стоимости получения k-го по порядку документа ,возвращенного из коллекции D_i в ответ на запрос пользователя: $\Delta_{i,k} = EC_i(k) - EC_i(k-1)$, если $k>0$, $\Delta_{i,k} = 0$.

Предположим, что $\Delta_{i,k} \le \Delta_{i,k+1}$ для всех k *и* i. В этом случае оптимальное решение может быть найдено с помощью следующего алгоритма. На шаге m, $m = 0, \ldots, nd$ алго-

ритма вычисляется оптимальное решение для задачи получения m документов. На шаге $m = 0$ решение $s = (0, \ldots, 0)$. На шаге $m > 0$ рассматриваются векторы $s^{(j)} = s + e_j$, где e_j есть вектор размерности n, в котором все компоненты кроме j равны 0, а $s_j = 1$. В качестве оптимального решения на шаге m выбирается вектор $s^{(j)}$, который является униформным. То есть для всех $i = 1, \ldots, n$ выполняется неравенство $\Delta_{i, s_i^{(j)}+1} \geq \max_l \Delta_{i, s_l^{(j)}}$. После выбора такого $s^{(j)}$ далее в качестве s рассматривается $s^{(j)}$.

Цель экспериментов состояла в сравнении двух моделей, используемых при построении описания коллекций. Первая модель – частотная. Вторая модель – вероятностное латентное семантическое индексирование. Кроме задачи выбора оптимальной модели, рассматривалась также задача настройки брокера – задача выбора ряда параметров, управляющих его работой. При выборе достаточно большого значения параметра $k = 100$ как для частотной модели, так и для модели вероятностного латентного индексирования оптимальное значение параметра $k' \approx 25$. При этом усредненная точность поиска, обеспечиваемая брокером, равна соответственно $P = 0.478$ для частотной модели и $P = 0.655$ для модели вероятностного латентного семантического индексирования. Полученные данные демонстрируют преимущество второго подхода.

Четвертая глава посвящена разработке настраиваемого пользовательского интерфейса, состоящего из двух сценариев.

Существующие технологии никак не учитывает интересы конкретного пользователя. Обычно пользователь регулярно обращается к поисковой системе с запросами, относящимися к нескольким интересующим его темам. В системе, проводящей регистрацию пользователя, можно хранить его архив запросов, являющегося информационным ресурсом, и использовать архив для выявления устойчивых информационных потребностей этого пользователя.

Первый сценарий - выбор тем пользователя.

Алгоритм состоит и следующих шагов:

1) Формирование архива запросов пользователя. Запросы аккумулируются в виде архива запросов на том сервере распределенной поисковой системы, где зарегистрирован данный пользователь.

2) Определение тем документов, возвращаемых коллекцией в ответ на запросы пользователя. Ссылки на найденные документы передаются пользователю. Одновременно с

этим в архиве запоминаются описания ***nt*** наиболее важных тем, присутствующих в ***nd*** наиболее релевантных документах. *Построение описания тем документа* ***d***:

а) Определение ***nt*** факторов для документа ***d*** с наибольшими весами: $W(z,d) = P(z)P(d|z)$. Таким образом будут выбираться, с одной стороны ,факторы, которые хорошо отражают тематику всего индекса в целом, а с другой - хорошо отражающие тематику данного конкретного документа ***d***.

б) Построение описаний выбранных тем. В качестве описания темы z документа ***d*** выбираются $n\omega$ слов из этого документа с максимальными весами. Веса слов вычисляются следующим образом: $W(\omega, z) = P(z)P(\omega|z)$. Каждая из выбранных тем задается парой фактор z и документ ***d***. Все множество выделенных тем (по ***nt*** лучших тем для каждого из ***nd*** документов) нумеруются и образуют множество $\{\tau_1, \tau_2, \ldots, \tau_{ntxnd}\}$. Вес слова ω из темы τ обозначим как $W_\tau(\omega)$ и будем полагать $W_\tau(\omega) = W(\omega, z)$, где z - фактор из пары фактор-документ, соответствующей теме τ.

3) Кластеризация тем документов, возвращенных на запросы из архива запросов.

а) Оценка степени близости двух тем τ_i и τ_j. Для кластеризации множество тем необ ходимо определить функцию близости двух тем. В данной работе эта функция определяется следующим образом: $$sim_{\min}(\tau_i, \tau_j) = \sum_{\omega \in \tau_i \cap \tau_j} \min\{W_{\tau_i}(\omega), W_{\tau_j}(\omega)\}$$

где $\tau_i \cap \tau_j$ - множество слов, входящих в описание обеих тем.

б) Single-pass кластеризация.

При появлении первой темы формируется первый содержащий данную тему кла

стер $C_1 = \{\tau_1\}$. Отнесение темы к существующему кластеру или формирование но вого кластера происходит следующим образом. Пусть τ - новая тема, $C_1, \ldots, C_m$ - семейство построенных на данный момент кластеров $$sim_{avr}(\tau, C_i) = \max_{j=1,\ldots,m} \frac{\sum_{\tau' \in C_j} sim(\tau, \tau')}{|C_j|}.$$

Если $sim_{avr}(\tau, C_i) \geq threshold$ -порог, величина которого определяет число полученных кластеров, то тема τ включается в кластер C_i, в противном случае формируется новый кластер $C_{m+1} = \{\tau\}$.

По завершении кластеризации формируются описания кластеров, которые бу-

дут предъявляться для просмотра пользователю: $\omega(C) = U_{\tau \in \tau(C)} U_{\omega \in \tau(C)} \omega$. В описание кластера C включаются не более nud слов из $\omega(C)$. В качестве веса терма ω в кластере C берется среднее арифметическое весов этого терма во всех темах данного кластера, в которые это слово входит: $W_C(\omega) = \frac{\sum_{\tau \in C, \omega \in \tau} W_\tau(\omega)}{\left|\{\tau \in C : \omega \in \tau\}\right|}$, где $\tau(C)$- множества всех тем из кластера C.

4) Просмотр кластеров пользователем. Пользователь просматривает и помечает темы из предъявленных кластеров в соответствии со своими потребностями.

5) Построение описаний тем пользователя $\tau(U)$, выбранных на предыдущем шаге, производится в два этапа.

а) Сопоставление идентификатору $\tau(U)$ темы пользователя номера соответствующего кластера с помощью функции $map: \tau(U) \rightarrow \{C_1, C_2, \ldots\}$

б) Для описания $\tau(U)$ выбирается ntd слов из $\omega(map(\tau))$ с максимальными весами $W_{(map(\tau))}(\omega)$.

Второй сценарий – поиск.

Предыдущий сценарий выполняется регулярно при смене интересов данного пользователя, в этом - описывается работу пользователя в обычном режиме при поиске интересующей его информации.

1. Ввод запроса. Интерфейс пользователя предоставляет пользователю форму для введения запроса. Все термы запроса соединяются логическим оператором ***OR***.
2. Поиск по запросу. Протокол HTTP используется для передачи введенного запроса CGI-скрипту, который реализует поиск в коллекции.
3. Представление результатов поиска пользователю содержит: ссылку на документ; несколько первых строк документа; тематику документа.

Каждому документу приписывается одна из тем пользователя, предложенных самим пользователем в предыдущем сценарии. В данном случае для полученного d формируется описание его темы, наилучшим образом представленной в данном документе. Это описание задается некоторым набором терминов с весами $disc(d) = \left(W_{\tau_d}(\omega_1), \ldots, W_{\tau_d}(\omega_n)\right)$, где n - фиксированная длина описания документа, $W_{\tau_d}(\omega)$ - вес слова ω в описании темы τ_d, которая является лучшей темой в докумен-

те d. Для каждой темы пользователя $\tau \in \tau(U)$ имеется ее описание также в виде набора терминов с весами $disc(\tau) = \left(W_{map(\tau)}(\omega'_1),..., W_{map(\tau)}(\omega'_{nud})\right)$. Документу d сопоставляется τ, на которой достигается максимум следующей функции: $\sum_{\omega \in disc(\tau) \cap disc(d)} \min\left\{W_{map(\tau)}(\omega), W_{\tau_d}(\omega)\right\}$.

Результаты экспериментов показывают, что предложенный метод оценивания тем документов в терминах тем, заданных пользователем, обеспечивает достаточно высокую точность (**0.904**). В качестве функции близости для оценки близости тем и близости темы и документа следует использовать сумму минимумов весов общих слов.

В **заключении** обобщаются основные научные результаты, полученные в диссертационной работе.

Приложение 1 содержится блок-схема алгоритма агента.

Приложение 2 содержится блок-схема алгоритма брокера.

Приложение 3 приведены формы интерфейса пользователя.

Приложение 4 содержит структуру и листинг программной реализации агента.

Приложении 5 приведены структура и листинг программы брокера.

Приложении 6 содержится структура и листинг программы интерфейса пользователя.

Приложение 7 документы подтверждающие внедрение и использование результатов диссертационной работы

ЗАКЛЮЧЕНИЕ

В диссертации решена поставленная научная задача: разработаны принципы построения, архитектура, методы и алгоритмы поиска, повышающие релевантность и оперативность результатов в распределенных поисковых системах для сложных глобальных сред гипертекстовой информации типа Internet.

Основные результаты диссертационной работы заключаются в следующем:

1. Выявлены преимущества архитектуры распределенного поиска по отношению к централизованной архитектуре.
2. Разработан и реализован информационный агент и его архитектура для формирования тематических коллекций (индексов).
3. Предложено новое построение тематической фильтрации документов с использованием двух типов фильтров, рекомендующих документ для включения в коллекцию (индекс); фильтр ядра индекса, с использованием эффективного метода, ос-

нованного на основе семантического анализа документов, и фильтр запросов, который позволяет тематике коллекции не устаревать с течением времени.

4. Предложен новый метод вычисления значимости новых термов из запросов пользователей, поддерживающий интерес сообщества пользователей в пополнении коллекции новыми документами данной тематики.
5. Предложен новый метод построения и управления очередью ссылок для обхода Интернет на основе динамического оценивания вероятности того, что документ, на который указывает данная ссылка, релевантен заданной тематике, обеспечивающий высокую долю релевантных документов среди всей совокупности документов, просмотренных сетевым роботом.
6. Разработан и реализован алгоритм маршрутизации запросов пользователей, позволяющий существенно повысить точность релевантности поиска и максимизации доходов пользователя.
7. Модифицирован к особенностям распределенной поисковой системы и реализован новый способ оценивания весов термов в запросах. А для документов использован метод на основе вероятностного латентного семантического индексирования, что повышает точность поиска системы в целом.
8. Найдено более простое условие для проверки оптимальности найденного решения задачи маршрутизации запросов.
9. Разработан новый подход построения интерфейса пользователя, на основе нового метода выявления информационных потребностей пользователя, базирующегося на анализе тематической принадлежности документов, возвращаемых системой поиска в ответ на запросы пользователя, приводящий к повышению оперативность поиска информации.
10. Разработан метод оценивания релевантности документа, возвращаемого поисковой системой в ответ на запрос пользователя, информационным потребностям данного пользователя.
11. Выявлено преимущество использования метода вероятностного латентного семантического индексирования над частотными моделями.
12. Результаты исследований позволили придти к выводу о необходимости использования теоретико-игровых постановок задач в децентрализованных поисковых системах и выработки стандарта для владельцев компонентов системы.

Основные результаты диссертации опубликованы в следующих работах:

1. Рушди А. Амамра. Информационный агент для формирования тематической коллекции электронных документов // Сборник трудов Третьей Всероссийской конференции по Электронным Библиотекам. Электронные библиотеки: перспективные методы и технологии, электронные коллекции. - Петразаводск, 2001, с. 156-159.
2. Рушди А. Амамра., Шкодырев В.П. Брокер для системы распределенного поиска // Компьютерные инструменты в образовании. Электронная версия: URL: WWW.ipo.spb.ru/journal., 2001.
3. Рушди А. Амамра. Интеллектуальная система распределенного поиска в Интернет // Телематика 2000. Материалы международной научно-технической конференции 29 мая – 1 июня 2000 г.- СПб., 2000, с. 95-97.
4. Рушди А. Амамра. Методы распределенного поиска информации в Интернет // Компьютерные инструменты в образовании. – СПб., 2001, № 3-4 с. 111-119.
5. Рушди А. Амамра. Модель мультиагентного поиска релевантной информации в WWW // Интернет технологии и современное общество. Материалы Всероссийской объединенной конференции 20-24 ноября 2000 г.– СПб., с. 163-165.
6. Рушди А. Амамра. Мультиагентная система распределенного поиска и формирования тематических коллекций // Приборостроение, № 9, 2000, с.10-15.
7. Рушди А. Амамра. Шкодырев В.П. Интеллектуальная система поиска релевантной информации в распределенной базе данных // Высокие интеллектуальные технологии образования и науки. Материалы VIII Международной научно-методической конференции 15-16 февраля 2001 г. – СПб., 2001. с. 63-64
8. Рушди А. Амамра. Шкодырев В.П. Интеллектуальный интерфейс пользователя // Приборостроение № 1, 2002. с. 20-26
9. Рушди А. Амамра. Шкодыров В.П. Интеллектуальные программные агенты распределенных систем сбора данных // Материалы международной конференции по мягким вычислениям и измерениям, 27-30 июня. - СПб, 2000, с. 258-261.
10. Рушди А. Амамра. Шкодырев В.П. Интерфейс поисковых систем, обеспечивающий информационные потребности пользователя // Высокие интеллектуальные технологии образования и науки. Материалы IX Международной научно-методической конференции 15-16 февраля 2002 г. – СПб., 2001. с. 271-273.
11. Rushdi A. Hamamreh. A brocker for distributed search engine // International conference on telematics and web-based education. – St.-Petersburg, 18-21 June. p. 31-32.
12. Rushdi A. Hamamreh. Agent for generation of subject-specific collection of electronic documents // International conference on computing and Measurements, St.- Petersburg 25-27 June, 2001 p. 204-207.