Рушди А. Амамра

МЕТОДЫ РАСПРЕДЕЛЕННОГО ПОИСКА ИНФОРМАЦИИ В ИНТЕРНЕТ

Этой статьей мы открываем новый цикл, связанный с публикацией результатов ведущихся исследований. Как правило, эти исследования связаны с подготовкой диссертаций в области компьютерных наук и содержат специальные вопросы, отражающие вклад автора в решение проблемы. Поэтому такие статьи будут публиковаться в двух версиях:

– *научно-популярной (в печатном и электронном варианте журнала)*
– *специальной (только в электронном варианте).*

Редакция считает, что проблематика многих исследований в области информатики может быть понятна и интересна школьникам и студентам младших курсов, ведь информатика – наука молодая. Замечательно, если окажется, что школьники придут в университеты, а студенты в аспирантуру, хорошо представляя себе, какие проблемы сейчас стоят и решаются в выбранной ими области деятельности.

ВВЕДЕНИЕ. ИНДЕКСИРОВАНИЕ И ПОИСК.

Традиционные методы поиска информации основывались на различных системах индексирования печатных изданий, внесения индексной информации в каталоги и на ручном поиске в каталоге. С появлением ЭВМ каталоги стали храниться в базах данных, обеспечивающих быстрый поиск, как локальный, так и удаленный. Одновременно появилась возможность хранить в электронной форме и сами документы. В результате решаются (или потенциально могут быть решены) проблемы автоматизации индексирования документов и полнотекстового поиска.

Ручное индексирование документов является весьма трудоемким и дорогостоящим процессом. Существуют стандарты, требующие для каждого документа заполнения формы, содержащей до 150 полей. Хранение документов в электронной форме позволяет автоматизировать процесс индексирования.

С помощью традиционного индексирования можно выполнять поиск по ограниченному набору атрибутов (автор, название, год издания и т.п.). Электронная форма документа позволяет сохранять в индексе информацию о всех вхождениях каждого слова в документ, о структуре документа (разбиение текста на главы и т.п.), о выделении отдельных частей текста специальными шрифтами. В результате становится возможен так называемый полнотекстовый поиск, при котором учитывается весь текст документа.

ИНТЕРНЕТ. ЗАДАЧА ПОИСКА ИНФОРМАЦИИ.

Интернет - новый и важный источник информации. Он характеризуется нижеследующими особенностями, связанными с информационным поиском (мы используем здесь различные источники, в некоторых случаях противоречащие друг другу, так что приводимые числовые величины можно рассматривать только как некоторые экспертные оценки):

– **Огромный объем информации**. На февраль 2000 года в Интернет было размещено более миллиарда страниц.

– **Динамичность информации.** По некоторым оценкам ежемесячно публикуется около 30 миллионов новых документов, причем ежемесячно изменяется до 40 процентов ранее опубликованной информации. Среднее время жизни WWW страницы около 24 дней.

– **Большой объем виртуальной информации**. Большая доля документов виртуальны в том смысле, что они формируются в ответ на некоторый запрос пользователя и не хранятся в явном виде. Как правило, это результаты поиска в различных базах данных.

– **Большое число используемых языков и форматов данных**. В Интернет используется более 100 естественных языков. Сами документы представляются в различных форматах, таких как html, xml, Word, PostScript, PDF и т.п.

Приведенные данные показывают, что без эффективных систем поиска информации польза от Internet будет мала.

КЛАССЫ ПОИСКОВЫХ СИСТЕМ

Говоря о системах поиска, необходимо упомянуть о двух классах поисковых систем – фактографических и документальных.

Фактографические системы дают точные и полные ответы на запрос пользователя. Это возможно только при весьма серьезных ограничениях и на представление информации в таких системах, и на язык запросов.

В настоящее время наиболее популярные системы такого типа представлены реляционными базами данных. Информация в реляционных базах данных хранится в виде совокупности таблиц заданных форматов. Язык запросов, например, SQL, позволяет пользователю описать новую таблицу, которую он хотел бы получить как ответ на свой запрос. При этом указываются связи между таблицей – результатом поиска и таблицами, хранящимися в базе данных. Система самостоятельно формирует алгоритм построения таблицы результата.

В документальных системах в ответ на свой запрос пользователь получает документы или некоторые их части. Запрос на поиск информации часто формулируется на естественном языке. В этом случае принципиально невозможно говорить о полном и точном ответе. Даже эксперты в конкретной области могут поспорить о степени релевантности некоторого данного документа заданному запросу.

В области документального поиска также имеются два конкурирующих подхода.

В рамках первого подхода документ представляется как некоторая структура, образованная словами, сгруппированными в предложения, абзацы, параграфы, главы и т.п. Кроме того, возможен учет тех частей текста, которые выделены специальным образом за счет использования шрифтов, тегов и т.п. При этом подходе не делается попыток распознать смысл отдельных слов, предложений и текста в целом. Все методы основаны лишь на сборе и учете статистики распределения слов в документе, в коллекции документов, в запросах пользователей. Условно такой подход можно назвать статистическим.

В рамках второго подхода (условно называемого семантическим) ставится задача раскрытия смысла отдельных слов, предложений и т.п. в контексте документа, коллекции документов. При этом широко используются различные словари, тезаурусы, составляемые вручную, что ограничивает применимость указанного подхода.

ОСНОВНЫЕ МОДЕЛИ ПРЕДСТАВЛЕНИЯ ДАННЫХ И ПОИСКА.

Наиболее употребительной моделью является булева модель. Заметим, что в рамках как данной, так и других моделей информационного поиска часто, вместо слова, используется более общее понятие «терм». Под термом здесь понимается отдельное слово, основа слова (корень), группы слов.

В рамках булевой модели документ d представляется своим словарем T(d). Запрос пользователя q в общем случае представляет собой выражение булевой алгебры, в которое входят термы, логические операторы (OR;AND;ANDNOT), скобки. Результат поиска R(q) может содержать огромное число документов. Пользователю это множество предъявляется в виде ранжированного списка.

Булева модель широко применяется в коммерческих поисковых системах в связи с возможностью эффективной реализации алгоритма поиска. При этом применяется технология так называемых инвертированных файлов. Для каждого терма формируется упорядоченный список документов, содержащих этот терм. Логические операции реализуются как операции над этими списками – слияние, пересечение, разность. Другой моделью, о которой только упомянем, является модель векторного пространства, латентное семантическое индексирование, вероятностное латентное семантическое индексирование.

ЦЕНТРАЛИЗОВАННЫЕ И ДЕЦЕНТРАЛИЗОВАННЫЕ СИСТЕМЫ ПОИСКА.

Системы поиска бывают централизованными и децентрализованными. На практике сейчас используются первые из них. Но централизованная архитектура имеет следующие фундаментальные недостатки, препятствующие выживанию таких систем в будущем:

– **Снижение доли проиндексированного Интернет.** Опубликованная в различных изданиях статистика показывает, что поисковые системы с централизованной архитектурой теряют свои позиции. Например, крупнейшая поисковая система Alta Vista индексирует самую большую долю Интернет из всех поисковых систем. Однако доля индексированного Интернет падает из года в год. Это связано с тем, что объем документов, опубликованных в Интернет, растет значительно быстрее, чем растут возможности централизованной системы по индексированию.

– **Низкое качество поиска**. Централизованная система стремится охватить информационные потребности всех возможных пользователей, в связи с чем в индекс включаются данные о всех доступных системе документах. В результате тематика проиндексированных документов меняется в очень широких пределах. С другой стороны, запросы пользователя, как правило, содержат не более двух ключевых слов. В ответ пользователь централизованной системы поиска получает огромное число документов, большая часть которых относится к категории мусора. Система не в состоянии выполнить качественное ранжирование результатов, и пользователь тонет в потоке нерелевантных документов.

СИСТЕМЫ С ДЕЦЕНТРАЛИЗОВАННОЙ АРХИТЕКТУРОЙ

Системы с децентрализованной архитектурой занимают пока значительно меньший сектор рынка информационного поиска, но будущее принадлежит им.

Основные компоненты системы поиска с распределенной архитектурой:

– **Тематические индексы.** В системе с децентрализованной архитектурой имеется заранее не ограниченное число индексов. Каждый индекс покрывает определенную тематическую область, обеспечивая и хранение информации о документах, и поиск. В отличие от централизованной системы, принадлежащей одному владельцу, различные тематические индексы могут принадлежать различным владельцам. Это, с одной стороны, приводит к конкуренции и возможному перекрытию тематик различных индексов, а с другой стороны, позволяет привлечь новые ресурсы для индексирования Интернет.

– **Тематические сетевые роботы.** В отличие от сетевого робота централизованной системы, сканирующего весь Интернет, тематический робот ориентирован на определенную тематику. Это позволяет использовать более интеллектуальные алгоритмы сканирования и повысить полноту представленной в индексе информации по заданной тематике.

– **Брокеры.** В распределенной системе запрос пользователя направляется брокеру, задачей которого является оценка тематической принадлежности запроса и выбор индексов, в которые следует переправить запрос для поиска (задача маршрутизации запроса пользователя).

Как и тематические индексы, брокеры также могут принадлежать различным владельцам, конкурирующим друг с другом. Разные брокеры могут специализироваться на поиске в различных более или менее широких группах тем.

– **Репозитарий.** В связи с открытостью системы с распределенной архитектурой должна иметься возможность владельцам новых тематических индексов регистрировать свои индексы для доступа к ним брокеров. В репозитарий заносится информация описательного, административного характера об индексе и, самое главное, описание содержимого индекса в форме, пригодной для чтения брокером. Именно на основе этой информации брокер принимает решение о соответствии того или иного индекса заданному запросу.

Приведенное выше описание основных компонент системы поиска с распределенной архитектурой показывает, что эти системы способны преодолеть те недостатки, которые были отмечены для систем с централизованной архитектурой.

Рассмотрим более подробно два основных компонента систем поиска с децентрализованной структурой.

ИНФОРМАЦИОННЫЙ АГЕНТ

Как уже отмечалось ранее, в поисковой системе с распределенной архитектурой различные тематические индексы могут принадлежать различным владельцам, что, с одной стороны, может привести к конкуренции и перекрыванию тематик различных индексов, а с другой стороны, независимость индексов способствует привлечению новых ресурсов в процесс индексирования Интернет, что необходимо для достаточно полного его охвата. Тематический индекс формируется автоматически с помощью тематического сетевого робота или, как его еще называют, информационного агента. От точности его работы зависит качество самого индекса и стоимость его построения и сопровождения.

Информационный агент – это программная компонента, обладающая способностью к самостоятельному принятию решений и проведению автономных действий, направленных на достижение цели, соответствующих интересам пользователя в сложной информационной среде (в контексте рассматриваемой задачи, такой целью является пополнение коллекций новыми релевантными ее тематике документами из WWW).

АГЕНТ ДЛЯ АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ КОЛЛЕКЦИИ ДОКУМЕНТОВ.

Задача агента состоит в пополнении индекса новыми ссылками на релевантные его тематике документы.

Известны два принципиально разных подхода к построению таких агентов. Первый из них – использование индексов существующих универсальных поисковых систем.

Этот подход достаточно широко применяется на практике и имеет положительные и отрицательные стороны.

Положительные стороны:

– **Повторное использование ранее полученных данных**. Сканирование Интернет – это дорогостоящий процесс, который приводит не только к большим затратам проводящей его организации, но и затрагивает интересы многих других сторон владельцев индексируемых сайтов, пользователей. Представляется весьма неразумным многократно сканировать Интернет ради предоставления доступа к одной и той же информации из большого числа конкурирующих поисковых систем. Идеальной была бы ситуация, при которой некоторая крупная международная организация выполняла регулярное сканирование всего Интернет и предоставляла бесплатный доступ к полученной необработанной информации всем желающим для последующей обработки и использования. Однако такое решение не масштабируемо, так как ресурсы этой гипотетической организации, сколь бы они ни были велики, всегда ограничены, и их рост не может поспевать за ростом объема информации, опубликованной в Интернет. На

практике часто используется индекс такой коммерческой поисковой системы, как Alta Vista в связи с тем, что ее индекс покрывает самую большую долю Интернет.

– **Новизна используемой информации.** Сетевые роботы универсальных поисковых систем стремятся (в идеале) индексировать все новые документы, не ограничивая себя фиксированной и, возможно, уже устаревшей тематикой. Индекс таких систем предоставляет представительную выборку относительно недавно опубликованных документов, что можно использовать при анализе тенденций в той или иной области.

– **Низкая стоимость получения информации.** Фактически на данный момент получение информации в виде ответов на автоматически генерируемые запросы от коммерческих поисковых систем бесплатно.

Отрицательные стороны:

– **Старение индекса.**

– **Закрытость методики получения используемой информации**. Алгоритмы сканирования Интернет и поиска в индексе являются коммерческой тайной, и, следовательно, индекс коммерческих универсальных систем представляется используемым его системам следующего уровня в виде черного ящика.

Вряд ли возможно делать какие-то объективные выводы о характере распределения информации в Интернет на основе косвенного (через систему поиска) анализа индекса коммерческой системы.

– **Ненадежность доступа к информации**. Коммерческие поисковые системы очевидно не заинтересованы в том, чтобы их индекс анализировался какими бы то ни было автоматическими системами. В любой момент эксперименты в этой области могут быть запрещены в связи с нарушением тех или иных прав коммерческих поисковых систем.

Второй подход к построению агента связан с обходом Интернет и основан на использовании ссылок на новые документы из ранее загруженных документов. Этот подход также широко используется, хотя и требует больших вычислительных ресурсов.

Положительные стороны:

– **Объективность**. В данном случае информация извлекается непосредственно из сети, что обеспечивает ее объективность.

– **Управляемость процесса получения информации**. В отличие от косвенного доступа к индексу коммерческой системы через формирование специальных запросов, доступ к информации в данном подходе более прозрачен. Имеется прямая и легко обнаруживаемая связь между алгоритмом сканирования Интернет и тематической направленностью загружаемых документов. Это позволяет подбирать параметры алгоритма, минимизирующие среднее отклонение тематики загружаемых документов от заданного тематического направления.

Отрицательные стороны:

– **Высокая стоимость**. В данном случае сетевой робот реально сканирует Интернет, что приводит к большим затратам даже при использовании ограниченного (заданной тематикой) поиска.

БРОКЕР ЗАПРОСОВ. ЗАДАЧИ, РЕШАЕМЫЕ БРОКЕРОМ ЗАПРОСОВ.

Важнейшая роль в распределенной системе отводится брокерам, решающим задачу маршрутизации запросов пользователей. В отличие от систем с централизованной архитектурой, которые сами занимаются своей рекламой ради привлечения новых пользователей, в системе с распределенной архитектурой отдельные тематические индексы невидимы для пользователей. Пользователь обращается к любому из брокеров системы, который и должен предоставить пользователю необходимые информационные услуги, привлекая для этого другие компоненты системы, прежде всего индексы, тематика которых в наибольшей степени соответствует информационным потребностям данного пользователя. Классическая задача, решаемая брокером – оптимальное распреде-

ление ресурсов, выделенных пользователем на выполнение поиска. Как правило, пользователь формулирует запрос, сопровождаемый информацией об ограничении на стоимость выполнения запроса. Каждый индекс имеет свои расценки на поиск, зависящие от качества и объема собранной в данном индексе информации, и от качества, обеспечиваемого данным индексом поиска. В связи с независимостью индексов, возможна ситуация, когда имеется несколько индексов, тематика которых близка тематике запроса. В этом случае брокер должен выполнить оценку числа релевантных документов, которые пользователь может получить от каждого индекса в ответ на его запрос, оценить точность поиска, предоставляемого этими индексами, учесть расценки индексов на поиск и стоимость доставки документов до пользователя. А прежде всего брокер должен оценить тематику самого запроса, который, как правило, может включать очень небольшое число терминов. Таким образом, сложно переоценить роль брокера в распределенной системе. От качества его работы зависит эффективность всей системы в целом.

Архитектура такой системы дает масштабируемое решение задачи сканирования Интернет и поиска только при привлечении очень большого (и все возрастающего вместе с ростом Интернет) числа независимых владельцев тематических индексов, обеспечивающих тематическое сканирование Интернет и предоставление услуг по поиску информации в рамках одной системы. Новые ресурсы (материальные, финансовые, интеллектуальные) будут вкладываться в развитие распределенной информационной системы только в том случае, если эта система обеспечивает справедливое для всех участников распределение дохода от работы системы как единого целого. Здесь роль распределителей доходов играют брокеры. Именно они перераспределяют запросы пользователей между различными тематически близкими индексами. В случае неоптимальной маршрутизации запросов пользователей, которая может проистекать от неэффективного алгоритма маршрутизации, возможны следующие последствия:

– **Неудовлетворенность пользователей качеством поиска**. Эта неудовлетворенность очевидна в случае переадресации запросов пользователей в индексы несоответствующей тематической направленности. Пользователь всегда может сравнить качество поиска, обеспечиваемое в данной системе и, например, в централизованной, и отказаться от использования некачественной услуги.

– **Отказ от участия в работе системы владельцев качественных тематических индексов**. В рамках распределенной системы владелец тематического индекса не рекламирует свою услугу, полагаясь на рекламу и привлекательность для пользователей всей системы в целом. При уменьшении числа пользователей системы владельцы качественных индексов могут принять решение о независимом существовании и самостоятельном рекламировании своей услуги. Очевидно, что с их выходом из системы качество ее работы ухудшится и число пользователей сократится.

АРХИТЕКТУРА БРОКЕРА

Прежде всего рассмотрим источники и типы информации, на основе которых брокер решает задачу маршрутизации запроса пользователя.

Пользователь предоставляет брокеру (через интерфейс пользователя) следующую информацию:

– запрос, состоящий из нескольких ключевых слов (термов), представленный в дизъюнктивной форме;

– общее число ссылок на документы, которые он желает получить в ответ на запрос;

– потери пользователя от получения и просмотра нерелевантных документов (штраф за получение нерелевантного документа);

– доход от получения релевантного документа.

Таким образом, информация, предоставляемая пользователем, позволяет

ставить и решать задачу оптимального распределения ресурсов, выделенных пользователем на поиск. Здесь ресурс – количество ссылок на документы, которые пользователь согласен просмотреть на предмет релевантности запросу. При этом загрузка каждого нерелевантного документа означает для пользователя прямые убытки – время, потраченное на загрузку документа и его чтение, стоимость трафика. Получение релевантного запросу документа эквивалентно для пользователя получению некоторого дохода, не зависящего от источника, из которого был получен сам документ и ссылка на него. Казалось бы, при задании общего числа ссылок, которые готов просмотреть пользователь, и величины штрафа за получение нерелевантного документа, задание величины дохода от получения релевантного документа является уже излишним. Однако задание данной величины необходимо в связи с тем, что пользователь еще вынужден платить индексам за поиск, который выполняется этими индексами в его (пользователя) интересах. При стоимости поиска, приходящейся на одну ссылку, превышающей доход от получения релевантного документа, пользователь должен отказаться от поиска в соответствующем индексе.

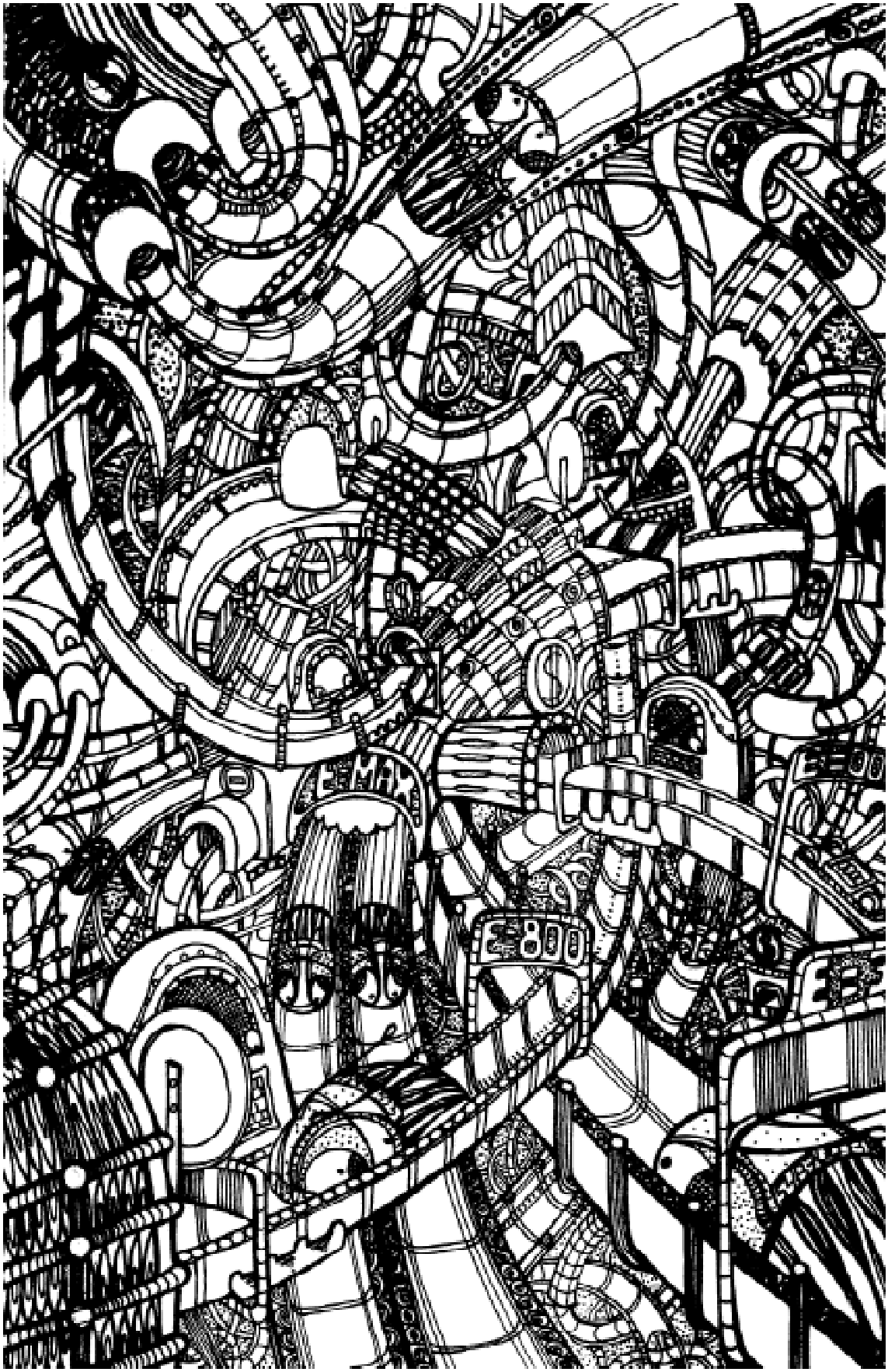
Итак, необходима информация о доступных тематических индексах. Эта информация должна поставляться и обновляться самими индексами (например, при их регистрации в системе). Традиционно для этой цели используется общедоступный репозитарий. **Репозитарий** хранит информацию о всех доступных (зарегистрированных) тематических индексах. Эта информация записывается в репозитарий самими индексами при их регистрации и регулярно обновляется. Любой брокер системы имеет доступ к репозитарию и использует его информацию для решения задачи маршрутизации запроса. В целях равномерного распределения нагрузки в системе имеется множество копий репозитария, обновление которых выполняется с помощью репликаций. О каждом индексе репозитарий хранит следующую информацию:

– описание индекса (некоторое множество термов с весами и некоторая статистика);
– стоимость ссылки на один документ, предоставляемой пользователю в ответ на его запрос (сюда входят и затраты на индексирование документов, поиск и т.п.)

Здесь уместно упомянуть одну проблему, которая не рассматривается в данной работе, но может представлять большой интерес при проектировании реальной распределенной системы. Эта проблема связана с достоверностью той информации, которую тематический индекс записывает в репозитарий. В соответствии с изложенной ранее архитектурой распределенной поисковой системы индекс получает доход, зависящий от числа направленных ему запросов. Маршрутизация запросов выполняется брокером, который в свою очередь учитывает информацию, переданную в репозитарий самим индексом. Содержание этой информации сложно контролировать и сопоставлять с реальным объемом данных, имеющимся у того или иного индекса. Например, очевидный подход – делать пробные запросы и сопоставлять результаты, полученные от индекса с его описанием, может использоваться для выявления недобросовестных владельцев индексов, но он требует больших временных и финансовых затрат и загружает систему избыточной работой. Данная проблема создает возможность для недобросовестных владельцев индексов включать в репозитарий ложные данные, подтверждающие возможность данного индекса ответить на любой вопрос по любой теме. При наличии такого индекса, брокер будет обязан пересылать ему все запросы всех пользователей. Данный индекс получит в короткие сроки значительную прибыль, но система в целом потеряет пользователей из-за низкого качества обслуживания.

ЗАКЛЮЧЕНИЕ.

Итак, мы представили читателю постановку проблемы и пути ее решения,



предложенные автором. В диссертационной работе на одноименную тему рассматриваются три следующие задачи, связанные с разработкой систем информационного поиска с децентрализованной архитектурой:

1. Разработка алгоритма для тематического сетевого робота (тематического информационного агента)
2. Разработка алгоритма для брокера, осуществляющего маршрутизацию запросов пользователя
3. Разработка настраиваемого пользовательского интерфейса.

Подробнее об алгоритме для брокера, осуществляющего маршрутизацию запросов пользователя, можно прочесть в электронной версии журнала.

Литература.

1. Lan Huang. Survey On Web Information Retrieval Technologies. Computer Science Department, State University of NewYork at Stony Brook, 2000.
2. Junghoo Cho and Lawrence Page. Efficient crawling through url ordering. Proceedings of the 8th International WWW conference, Canada, Toronto, May 1999, http://www7.scu.edu.au/programme/fullpapers/1919/com1919.htm
3. S. Haverkamp. Intelligent information Agents. JASIS, V49, N4, 1998, http://sourceforg.net/projects/openmuscat
4. Norbert Fuhr. A decision-theoretic approach to database selection in networked ir. In Workshop on Distributed IR, Germany, 1996.
5. Thomas Hoffman. Probabilistic latent semantic indexing. In Proc. of SIGIR'99, pp. 50–57, Berkeley, CA, USA, August 1999.
6. Steve Lawrence and C. Lee Giles. Accessibility of information on the web. Nature, 400:107–109, 1999.
7. A. Palet, Petrosijan, and W. Rosenstiel, editors. OASIS: Distributed Search System in the Internet. St. Petersburg State University Published Press, St. Petersburg, 1999.

Рушди А. Амамра,
аспирант Санкт-Петербургского Государственного Технического Университета.