Математические структуры и моделирование
2002, вып. 9, с. 1–11

УДК 681.518

СБОР ИНФОРМАЦИИ О ДОСТУПНЫХ РЕСУРСАХ ИНТЕРНЕТ

И.А. Земсков

We first study the generic search engine architecture, then review all problems of collecting information about Web's resources and finally describe three building conception of search engine's module for collecting information about Web's resources.

Введение

Согласно последним исследованиям, в Интернете уже опубликовано более двух миллиардов страниц, и их число экспоненциально увеличивается [22,25]. Таким образом, с каждым годом, месяцем, днем становится труднее находить «дорогу» к нужной информации. По этой причине становятся все более актуальными научные исследования в области информационного поиска в Интернет. Результаты этих исследований [2,4] нацелены прежде всего на создание новых поисковых систем [26,27] или на усовершенствование алгоритмов работы уже существующих средств информационного поиска.

Целью данной статьи является рассмотрение такого аспекта работы информационно-поисковых систем в Интернет, как технологии сбора информации о содержимом Веб-ресурсов. Для того чтобы обсуждать технологии, предложенные различными исследователями, нужно удостовериться в том, что мы правильно понимаем общие принципы функционирования поисковых систем в общем и процесса сбора информации о Веб в частности. Поэтому вначале мы рассмотрим некоторые общие понятия и получим краткое описание процессов, протекающих в поисковых системах. Далее мы опишем основные проблемы, с которыми сталкивается процесс сбора информации, и будем опираться на них при рассмотрении различных концепций поведения, направленных на получение «представления» о Веб-ресурсах.

1. Основные компоненты поисковых средств

Множество работ по проблемам поиска в Интернете описывают схему организации поисковых средств [17,19,21,23], но для целей данной статьи мы приведем эти схемы к общему виду. Результат можно видеть на рисунке 1.

Основные компоненты поисковых средств Интернет (1): модуль сбора информации о доступных ресурсах, репозиторий собранной информации, модуль

E-mail: zemskov@univer.omsk.su
Омский государственный университет

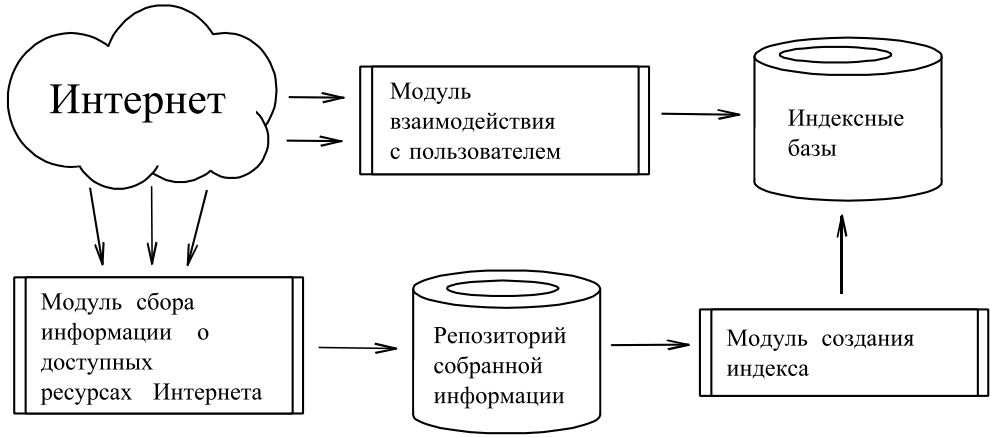

Рис. 1. Обобщенная схема организации поисковых систем Интернет

создания индексов, сами индексы (индексные базы) и модуль взаимодействия с пользователем.

Постараемся описать каждый элемент схемы, не выходя при этом за рамки обобщения.

Интернет. Следует отметить, что в Интернет реализовано и функционирует несколько технологий организации информационных ресурсов (например протоколы: NNTP, WAIS, Gopher, HTTP, FTP). Самый популярный источник информационных ресурсов – это Веб (WWW реализовано с использованием протокола HTTP). В пользу этого утверждения говорят факты увеличения как пользователей ресурсов этого типа, так и рост количества самих ресурсов. Однако многие поисковые системы позволяют своим пользователям производить поиск не только по источникам Веб. Например, стало популярным производить поиск по сообщениям новостных групп (реализация новостных групп основана на использовании протокола NNTP).

В дальнейшем изложении мы сознательно ограничим себя информационными источниками из Веб. Сузим тем самым область охвата поисковых систем. Однако это сужение позволит нам досконально разобраться с одной средой (под средой понимается множество информационных источников, реализованных в рамках одной технологи), и в дальнейшем попытаться перенести свои наработки на другие информационные среды.

Ограничившись средой Веб, необходимо разобраться с тем, что же мы будем понимать под словами «информационный ресурс» в данном контексте. Для начала надо вспомнить, что в этой среде есть несколько типов ресурсов, имеющих свою вполне определенную «природу». Основными типами ресурсов являются следующие: форматированный текст (для форматирования используется HTML код), графические изображения (форматы: JPEG, GIF, PNG, SWF), аудиофайлы(форматы: WAV, MIDI, MP3, RA), видео. Таким образом, помня о том, что ресурс любого из этих типов может стать объектом поиска, мы будем понимать под словами «информационный ресурс» некий файл, имеющий внутреннюю структуру согласно спецификации одного из объявленных типов и

находящийся на специальном сервере (информационный источник) в Интернет. Сервер должен обеспечивать доступ к этому файлу по протоколу HTTP.

Модуль сбора информации о доступных ресурсах. На входе этот модуль имеет все множество информационных ресурсов Веб (помним о нашем ограничении на Интернет). Функция модуля заключается в том, чтобы, используя некий внутренний алгоритм, собрать в одном месте информацию обо всем многообразии доступных информационных ресурсов. Таким образом, на выходе этого модуля получаем репозиторий собранной информации, т.е. некое представление о Веб, готовое к удобному использованию.

Репозиторий собранной информации. Как было указано ранее, репозиторий – это некое представление о Веб. Далее мы поясним данное утверждение.

Учитывая огромные размеры хранимой в Вебе информации разного типа, становится понятным желание избежать простого складирования на своих дисках копий тех данных, которые есть в Веб. Тем самым мы приходим к решению о целесообразности создания «некоего представления» о Веб, описывающего существующие ресурсы Веб без потребности хранить их у себя в полном объеме.

Из всего выше сказанного вытекает, что репозиторий – это пассивный элемент схемы, т.к. сам он никаких действий по отношению к другим элементам схемы не выполняет. Его основная функция заключается в накоплении в нужном виде того, что ему передаёт на хранение модуль сбора информации. Другой его функцией является предоставление накопленных сведений как модулю сбора информации, так и модулю создания индексов.

Модуль создания индексов. Выполнение поиска на данных, накопленных в репозитории, является неэффективным. На этом факте основано решение об использовании индексных структур. Таким образом, на вход рассматриваемого модуля поступает информация, накопленная в репозитории. Он использует её для построения индексных структур. Количество и состав индексов зависит от цели поставленной перед разработчиками поисковой системы. То есть на выходе этого модуля мы получаем набор из индексных баз (минимум одну базу).

Индексы. Индексы сами по себе также являются пассивным элементом схемы, т.к. не производят действий по отношению к другим элементам схемы. Основная функция индексов заключается в специальной организации данных, с помощью которых можно с приемлемой эффективностью производить поиск нужной информации среди всего массива накопленной информации о Веб.

Модуль взаимодействия с пользователем. На вход данного модуля поступают запросы пользователей. Причем реализация языка запросов пользователей преимущественно стремится к естественному языку. Таким образом, основная функция данного модуля заключается в переводе запросов пользователя с языка, понятного пользователю, на язык, понятный машине, работающей с индексом. На выходе этого модуля получается список ссылок на информационные ресурсы Веб.

2. Проблемы

Перед разработчиками поисковых систем встает большое количество проблем, их рассмотрению посвящено множество работ [11, 13, 15, 17, 21]. В данном параграфе делается попытка обобщения проблем, которые стоят перед всеми исследователями и непосредственно влияют на архитектурные решения при разработке модуля сбора информации о доступных ресурсах. В дальнейшем такое обобщение позволит нам при рассмотрении вопросов разработки рассматриваемого модуля глубже понимать суть применяемых стратегий и концепций.

Возникающие перед исследователями проблемы можно разделить на три класса. В первый класс мы отнесём проблемы, связанные с «природой» информационных ресурсов. В этот класс попали следующие проблемы [15, 17, 21]:

- **Большой объем информационных ресурсов:** размер накопленных человечеством в рамках Веб данных поистине огромен. С этим фактом трудно спорить, т.к. исследователи в данной области давно оперируют сотнями гигабайт в своих работах [2]. Также нельзя выпускать из виду факт экспоненциального роста объемов накапливаемой в рамках Веб информации.
- **Разное время существования информационных ресурсов:** документы или файлы могут быть легко добавлены и так же легко удалены в Веб. Для большинства других членов сети Интернет эти манипуляции с файлами могут остаться незамеченными, но в целях обеспечения эффективного поиска видится необходимым четкое отслеживание производимых изменений в содержимом информационных ресурсов. Тем самым, например, можно заметно уменьшить шанс столкновения в результатах поиска со «сломанной» ссылкой, т.е. ссылкой на уже несуществующий ресурс.
- **Разнородность информационных ресурсов:** данные информационных ресурсов в Веб разнотипны. Они создаются в различных форматах, имеют различную медиаприроду (текст, звук, изображение), а также различаются по применяемым естественным языкам.
- **Динамичность изменения содержимого информационных ресурсов:** для пояснения этой проблемы достаточно вспомнить страницы сайтов с расположенным на них разделом «Новости» или «Объявления». В зависимости от активности владельца ресурса информация в этих разделах может меняться от «очень часто» (раз в 10 минут или еще чаще) до «очень редко» (раз в год или еще реже). Независимо от скорости изменения содержимого страницы её адрес для посетителей сайта остается прежним, благодаря чему становится большой вероятность из поисковой системы посетителю попасть на страницу, которая уже не содержит нужную информацию.
- **Различное качество и уровень полезности информационных ресурсов:** здесь под словом «качество» понимается широкий круг проблем, начиная с того, что в процесс создания многих ресурсов не привлекаются профессиональные программисты, дизайнеры, редакторы, и заканчивая тем, что в сети очень много дублирующих друг друга ресурсов. Отдель-

ным блоком можно выделить правовые и морально-этические вопросы, касающиеся качества публикуемой информации. Еще один блок вопросов ставит нас перед философской дилеммой целесообразности владения информацией о некотором доступном ресурсе. Своими корнями дилемма уходит к проблеме большого объема доступных в Веб ресурсов.

- **«Скрытность» информационных ресурсов:** в последнее время появились работы, в которых рассматривается проблема организации поиска по так называемому скрытому Веб [4]. Под словами «скрытый Веб» здесь понимаются информационные ресурсы для получения доступа, к которым нужно пройти сложную процедуру регистрации или сформировать с помощью предлагаемой разработчиками ресурса формы некий запрос. Однако на этапе регистрации и при составлении запросов возникают большие сложности. Например, они начинаются уже с того, что заранее не известны ни предлагаемые формы, ни смысл их полей ввода, и заканчиваются тем, что нужно научиться понимать ответы систем на посланные нами запросы. А этап регистрации вдобавок может состоять из нескольких стадий, что еще более усложняет задачу. Одним словом, разработка системы автоматического сбора информации о таких ресурсах сродни разработке системы искусственного интеллекта.
- **Доступность информационных ресурсов:** суть этой проблемы заключается в различном качестве коммуникационных связей между распределенными по всему миру узлам и сегментам сети Интернет. Бывают моменты, когда узел по причинам плохого уровня связи с Интернет становится труднодоступен остальным членам сети. Однако для нас является нежелательным отказ от учета информации, предоставленной этим узлом (источником).

Во второй класс отнесём проблемы, касающиеся нагрузки на различные элементы, участвующие во взаимодействии:

- **Минимизация нагрузки на информационный источник:** независимо от выбираемой стратегии сбора информации о доступных информационных ресурсах информационный источник (место хранения информационных ресурсов) будет нести потери вычислительной мощности и неэффективной нагрузки на его аппаратные ресурсы (например создание нагрузки на дисковые накопители). Это, в свою очередь, может вызвать вполне законное негодование со стороны владельцев информационного источника.
- **Минимизация нагрузки на каналы связи:** здесь следует вспомнить о том, что сбор информации о доступных ресурсах подразумевает перекачивание определенных объемов информации с источника информации на узел сбора по существующим коммуникационным линиям. Если принять во внимание упоминавшиеся большие объемы хранимой информации, то станет понятно, что перекачка со всех источников создаст некоторую (довольно заметную) нагрузку на каналы Интернет. Что вполне законно отражается на плате за использование этих самых каналов связи. Еще нужно помнить о том, что каналы связи не имеют способности по первому же

запросу увеличивать беспредельно свою «ширину», т.е. они имеют вполне определенную максимальную пропускную способность.

- **Оптимизация нагрузки на модули сбора и накопления информации:** описывая данную проблему, следует еще раз напомнить тот факт, что модуль сбора информации должен будет обрабатывать большие объемы данных, приводя их в надлежащий вид, пригодный для, возможно, своей последующей работы и работы модуля создания индексов. Т.е. без должной оптимизации всех нюансов (алгоритмов, структур данных и т.п.), протекающих в модуле процессов, мы невольно обрекаем себя на бесполезную трату дорогостоящих ресурсов.

К третьему классу относятся все проблемы, касающиеся внедрения технологий, положенных в основу модуля сбора информации о доступных ресурсах. Весь список проблем мы приводить не будем, однако отметим, что большинство из них вызвано большой разобщенностью владельцев информационных ресурсов и разработчиков поисковых систем.

3. Концепции

В прошлом параграфе мы кратко рассмотрели проблемы, стоящие перед разработчиками модуля сбора информации о доступных ресурсах. Вполне резонно предположить, что различные группы разработчиков по-разному ставят приоритеты при решении этих проблем. Однако при всем кажущемся многообразии подходов к решению этих проблем все существующие стратегии реализации модуля сбора укладываются в три конкурирующие концепции. В рамках каждой из концепций общие проблемы, описанные в предыдущем параграфе, принимают новый вид, новое осмысление. В проблемах появляется некоторая конкретика, касающаяся направления их решения. Как следствие выбора одной из концепций становится постановка нового круга проблем, на решение которых исследователи бросают все свои усилия. Так происходит во всех отраслях знаний, а область исследования проблем поиска информации в Интернет не стала исключением. Но есть косвенные факты, которые заставляют задуматься о правильности выбираемых направлений приложения усилий. Например, сейчас модными стали публикации об исследовании объемов «охвата» Веб поисковыми системами [25]. Эти исследования показывают, что даже самые «большие» поисковые системы охватывают своим поиском лишь малый процент доступных ресурсов (конкретные цифры объемов на текущий момент можно посмотреть в приведенной ссылке). С учетом непрерывного роста объемов Веб еще более актуальным становится вопрос о правильности выбранной разработчиками «больших» поисковых систем концепции сбора информации о доступных ресурсах Веб. Другим косвенным фактом может служить само существование нескольких концепций.

После всего выше сказанного видится актуальным применение широко известных научных методов исследования, например имитационного моделирования, для исследования эффективности и перспективности существующих, а также вновь разрабатываемых в рамках каждой концепции технологий сбора

информации о доступных ресурсах Веб.

Далее мы кратко опишем все три концепции. Рассмотрение начнем с самой популярной и наиболее проработанной концепции:

Концепция роботов. Согласно этой концепции весь модуль сбора информации располагается и работает на аппаратном обеспечении разработчика поисковой системы. Другими словами, разрабатывается некий программный комплекс, который реализует модуль сбора информации для поисковой системы и оставляет без изменения программные технологии, лежащие в основе средств создания информационных источников (т.е. Веб-сервера). Основу реализаций программного комплекса составляет некоторый программный код, именуемый в литературе сетевым роботом, пауком, краулером и т.п. Т.к. единого мнения по этому вопросу нет, то мы договоримся далее в статье называть его роботом. Алгоритм работы робота заключается в рекурсивном «обходе» ресурсов Веб и извлечении из «обойденных» ресурсов ссылок (URL) на новые ресурсы. Его работа начинается с некоторого набора ссылок на ресурсы Веб и заканчивается при выполнении некоторого условия. Под словом «обход» здесь понимается скачивание ресурса к себе для последующей обработки. Таким образом, «обойденные» ресурсы - это ресурсы Веб, которые уже скачаны роботом к себе и о содержании которых составлено некоторое представление.

Первым результатом данной концепции становится отодвигание на задний план или практически полное снятие с рассмотрения одной из проблем, описанной в прошлом параграфе, а именно проблемы разобщенности разработчиков поисковых систем и владельцев информационных ресурсов. Фактически проблема остается, но теперь разработчики поисковых средств остаются одни перед лицом других, не менее серьезных, проблем и пытаются преодолеть их только своими силами. А круг оставшихся проблем настолько широк и разнообразен, что вызывает у них логичное желание сконцентрировать свои усилия не на всех, а только на некоторых проблемах. Это выражается, например, в желании разрабатывать специализированных роботов, т.е. роботов, которые имеют свою стратегию обхода Веб и свои условия остановки работы, продиктованные его специализацией. Такие роботы имеют различный объем охвата доступных ресурсов Веб. Например, робот для создания представления о доступных музыкальных ресурсах Веб может с известной долей рвения отвергать ссылки на графические ресурсы и прекращать составление представления об информационном источнике (Веб-сервере) при достижении определенного уровня проникновения в его «глубины». Еще одним интересным примером специализированного робота может быть робот, направленный на сканирование скрытого Веб [4].

Создание специализированных роботов многим исследователям кажется некоторой панацеей. Зная некоторые особенности организации и существования ресурсов в какой-либо информационной области интересов, исследователи могут далеко продвинуться в решении проблемы максимального охвата при сборе информации о ресурсах в рамках выбранной тематики. Однако собранная информация будет касаться только одной области интересов! А как же быть с другими интересными тематиками? Создавать новых роботов?

В рамках данной концепции проблема охвата максимально возможного объ-

ема доступных ресурсов неразрывно связана с проблемами, попавшими во второй класс прошлого параграфа, а именно минимизации и оптимизации нагрузок на участвующие во взаимодействии аппаратные ресурсы. С ростом объемов Веб эти проблемы только обостряются. Даже самые простые подсчеты передаваемых от информационных источников к поисковым системам по каналам связи объемов данных показывают большие объемы финансовых расходов [25]. Но сами по себе большие финансовые расходы еще не повод для беспокойства. Беспокойство появляется после обнаружения некоторых подробностей алгоритмов работы модулей поисковой системы. А именно того факта, что в скачиваемой информации содержится большой объем html-кода, который выступает некоторой оболочкой для форматированного текста и который после стадии извлечения «полезной» информации попросту удаляется.

Дополнительные сведения о проблемах и предлагаемых решения, связанных с данной концепцией, можно почерпнуть из специальных обзоров [15, 21].

Нам остается отметить лишь тот факт, что при построении модели данной концепции нужно будет очень аккуратно подходить к её формализации, т.к. накоплен большой багаж наработок и предложений по улучшению определенных моментов в работе роботов [1, 2, 5–9, 11, 18, 20] и отказ от их учета в модели может повлечь за собой построение неадекватной модели.

Концепция сенсоров. Основной причиной возникновения данной концепции стала попытка найти более дешевый метод изменения представления о содержимом доступных информационных ресурсов вместе с изменениями самих ресурсов. Суть концепции сенсоров кроется в специальной доработке программного обеспечения, находящегося на стороне информационного источника, т.е. Веб-сервера. Доработка Веб-сервера заключается в создании некоего модуля, который доступными ему средствами пытается обнаружить новые ресурсы на данном сервере или пытается обнаружить изменения в уже найденных ресурсах. После обнаружения каких-либо изменений в состоянии информационного источника модуль сообщает об этих изменениях некоему «головному» серверу.

Одна из попыток реализаций данной концепции была осуществлена в [3]. Результатом этой работы стал модуль, который автоматически с определенной периодичностью исследовал содержимое каталогов сервера в поисках изменений с момента его последнего запуска. Текущее состояние сравнивалось с состоянием, сохраненным в специальных файлах. При обнаружении изменений создавался файл с так называемой мета-информацией, которая описывала суть произошедших изменений. Затем файл архивировался и отсылался модулем на специальный сервер. После этого роботы специального сервера на основе принятой мета-информации принимали решение о скачивании ресурса для последующей обработки.

Заметным недостатком данной реализации можно назвать то, что модуль работает только с серверами, на которых информация хранилась в виде файлов. А этот подход к созданию информационных Веб-ресурсов начинает уступать место т.н. динамическим сайтам, в которых содержимое страницы берется из базы данных и соединяется с шаблоном дизайна в момент запроса странички посетителем сайта.

В статье [24] предлагается еще один вариант сенсора. Но теперь сенсор не является активным по отношению к поиску изменений содержимого, т.е. предложенный вариант модуля можно назвать пассивный сенсор (прошлому варианту больше подходило название активный сенсор). Другими словами, в программное обеспечение Веб-сервера предлагается встроить модуль, который будет «следить» за поступающими запросами информационных ресурсов и «слушать» ответы сервера. Каждому запросу ставится в соответствие ответ сервера, и далее эта пара ищется в уже накопленной базе мета-описаний ресурсов. При обнаружении изменений предлагается рассмотреть два варианта поведения: в первом случае предлагается послать основному серверу некое мета-описание найденных изменений, а во втором предлагается сразу провести предварительную обработку найденного ресурса по очищению от «мусора» (это в большей степени относится к страницам в формате HTML). Такая очистка (например от лишних конструкций языка разметки) может дать уменьшение нагрузки на каналы связи. Но даже в случае передачи одного мета-описания мы можем получить выигрыш, т.к. нет надобности постоянных повторных обходов, как это наблюдается в концепции роботов.

Большим недостатком данного предложения является то, что оно существует пока только на бумаге, т.е. еще не существует программной реализации, способной показать практические результаты от её применения. Однако уже сейчас можно предложить еще одно направление исследований касательно этого варианта сенсора. Оно заключается в том, чтобы изучить поведение сенсора на объектах самой сложной природы, а именно на сайтах, подпадающих под определение скрытого Веб. Например, видится интересным получение ответа на вопрос: сколько времени потребуется пассивному сенсору для помощи поисковой системе в составлении представления о самых полезных данных, хранимых «за формой» на одном из ресурсов (подразумевается то, что вопросы через форму будут задавать посетители сайта).

В заключение описания данной концепции хочется указать одну общую проблему для её исследователей. Проблема заключается в том, что концепция подразумевает вмешательство в программное обеспечение Веб-серверов. А это является огромным сдерживающим фактором на пути внедрения технологий, основанных на этой концепции.

Концепция мобильных роботов. Эта концепция по своей сути является гибридом двух рассмотренных ранее концепций. Её разработчики осуществили попытку объединить весь накопленный положительный опыт в рамках концепции роботов и заманчивую идею снижения нагрузки на каналы связи за счет снижения объема передаваемых по ним данным, описывающих содержимое ресурсов. В результате этой «попытки» появилась технология, которая предлагает модернизировать программное обеспечение Веб-сервера таким образом, чтобы оно приобрело способность принимать от специального сервера поисковой системы к себе некоторый код, описывающий поведение робота. В дальнейшем этот код должен будет отработать на принявшем его сервере. Результатом работы робота-кода становится представление о найденных информационных ресурсах на данном Веб-сервере. Это представление уже не содержит в себе такого

большого объема мусора, т.к. его основная чистка в найденных ресурсах теперь происходит не на стороне поисковой системы с роботом, а на самом источнике информации.

Однако и в этой бочке мёда есть большой черпак дегтя. Это можно почувствовать еще на стадии поверхностного ознакомления с данной концепцией. Для этого достаточно задать несколько «безобидных» вопросов. Например, как будет себя вести технология данной концепции по отношению к информационным ресурсам, имеющим способность очень часто «незначительно» менять свое содержимое? К сожалению, в этом вопросе данная концепция солидарна с концепцией роботов, а именно предлагается осуществлять периодический обход всех известных Веб-серверов. После такого ответа вопрос о нагрузке на принимающие к себе код робота Веб-сервера возникает сам собой. А он может усложнить и так тяжелый вопрос, связанный с возможностью внедрения «в массы» данной концепции.

За более подробной информацией можно обратиться к [14, 16].

Заключение

Рассмотрена общая структура поисковых систем для сети Интернет. Предпринята попытка сформулировать полный список общих проблем, с которыми сталкиваются разработчики модуля сбора информации о доступных ресурсах Веб. Опираясь на полученный список, были кратко рассмотрены основные концепции при разработке этого модуля. Попутно были обозначены возможные направления дальнейших исследований.

Литература

1. Najork M., Wiener J.L. *Breadth-First Search Crawling Yields High-Quality Pages.* – `http://www10.org/cdrom/papers/pdf/p208.pdf`
2. Melnik S., Raghavan S., Yang B., Garcia-Molina H. *Building a Distributed Full-Text Index for the Web.* – `http://www-db.stanford.edu/ rsram/pubs/www10/www10paper.pdf`
3. Brandman O., Cho J., Garcia-Molina H., Shivakumar N. *Crawler-Friendly Web Servers.* – `http://rose.cs.ucla.edu/ cho/papers/cho-server.pdf`
4. Raghavan S., Garcia-Molina H. *Crawling the Hidden Web.* – `http://dbpubs.stanford.edu/pub/2000-36`
5. Shkapenyuk V., Suel T. *Design and Implementation of a High-Performance Distributed Web Crawler.* – `http://cis.poly.edu/tr/tr-cis-2001-03.pdf`
6. Cho J., Garcia-Molina H., Page L. *Efficient Crawling Through URL Ordering.* – `http://rose.cs.ucla.edu/ cho/papers/cho-order.pdf`
7. Menczer F., Pant G., Srinivasan P., Ruiz M.E. *Evaluating Topic-Driven Web Crawlers.* – `http://dollar.biz.uiowa.edu/ fil/Papers/sigir-01.pdf`
8. Diligenti M., Coetzee F.M., Lawrence S., Giles C.L., Gori M. *Focused Crawling Using Context Graphs.* – `http://www.neci.nec.com/homepages/coetzee/focusCrawler.pdf`

9. Buyukkokten O., Garcia-Molina H., Paepcke A. *Focused Web Searching with PDAs.* – http://www-db.stanford.edu/ orkut/papers/pb2.pdf
10. Fox A., Brewer E.A. *Harvest, Yield, and Scalable Tolerant Systems.* – http://www.cs.ucsb.edu/ tve/cs290i-sp01/papers/fox99harvest.pdf
11. Najork M., Heydon A. *High-Performance Web Crawling.* – http://citeseer.nj.nec.com/najork01highperformance.html
12. Green J.W. *HYPERDOG - Up To Date Web Monitoring Through Metacomputers.* – http://www.cnds.jhu.edu/pub/papers/hyperdog.pdf
13. Brewington B.E., Cybenko G. *Keeping up with the changing Web.* – http://www.ece.eng.wayne.edu/ czxu/ece7995/reading/keep-up-change.pdf
14. Fiedler J., Hammer J. *Mobile Web crawling.* – http://www.cise.ufl.edu/tech-reports/tech-reports/tr98-abstracts.shtml
15. Arasu A., Cho J., Garcia-Molina H., Paepcke A., Raghavan S. *Searching the Web.* – http://rose.cs.ucla.edu/ cho/papers/cho-toit01.pdf
16. Bowman C. M., Danzig P.B., Hardy D.R., Manber U., Schwartz M.F. *The Harvest Information Discovery and Access System.* – http://citeseer.nj.nec.com/bowman95harvest.html
17. Lam S. *The Overview of Web Search Engines.* – http://citeseer.nj.nec.com/lam01overview.html
18. Rennie J., McCallum A.K. *Using Reinforcement Learning to Spider the Web Efficiently.* – http://www.ai.mit.edu/ jrennie/papers/icml99-text.pdf
19. Brin S., Page L. *The Anatomy of a Search Engine.* – http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm
20. Koch T., Ardo A., Brumer B., Lundbr S. *The building and maintenance of robot based internet search services – A review of current indexing and data collection methods.* – http://www.ub.lu.se/desire/radar/reports/D3.11v0.3/tot.html
21. Некрестьянов И.С., Пантелеева Н. *Системы текстового поиска для Веб.* – http://meta.math.spbu.ru/ nadejda/papers/web-ir/web-ir.html
22. Некрестьянов И.С. *Тематико-ориентированные методы информационного поиска* // Канд. дис., Санкт-Петербург, 2000. – http://meta.math.spbu.ru/ igor/thesis/thesis.html
23. Браславский П.И. *Методы повышения эффективности поиска научной информации (на материале Internet)* // Канд. дис.,. Екатеринбург, 2000.
24. Земсков И.А. *О концепции индексации информационных ресурсов сети Интернет.* – ftp://cmm.univer.omsk.su/pub/sbornik8/zemskov.zip
25. *Сайт обзоров поисковых систем.* – http://www.searchenginewatch.com/
26. *Поисковая система Google.* – http://www.google.com/
27. *Поисковая система Teoma.* – http://www.teoma.com/