Данная работа посвящена описанию поисковых систем, которые осуществляют поиск информации.

В информационно-поисковой системе должен храниться весь необходимый информационный массив, из которого по требованиям пользователей выдается нужная информация. Поиск информации по требованию пользователя осуществляется либо автоматически, либо вручную (как в библиотеках, когда с запросом к работнику справочного фонда обращается читатель, а работник пользуется системой каталогов).

Во втором случае используются ЭВМ, снабженные специальными программными средствами, анализирующими процессы запросов, поиска и выдачи нужных документов. Таким образом, информационно-поисковые системы (ИПС) реализуют вопросно-ответное отношение, что сближает задачи, стоящие перед создателями таких систем, с теми задачами, которые решают создатели человеко-машинных систем.

Поиск информации является одной из наиболее распространенных и одновременно наиболее сложных задач, с которыми приходится сталкиваться в Сети любому пользователю. Однако если для рядового члена сетевого сообщества знание методов эффективного информационного поиска является желательным, но далеко не обязательным качеством, то для работников высокоинтеллектуальной сферы умение быстро ориентироваться в ресурсах Интернет и находить требуемые источники сегодня относится уже к числу базовых квалификационных навыков.

Цель работы – описать и дать характеристику информационно-поисковым системам.

Данная цель решается с помощью раскрытия следующих основных задач:

1) описать принципы работы поисковых машин;

2) дать характеристику глобальным поисковым системам;

3) описать стратегию и методику профессионального поиска информации.

1. Сущность поисковых машин

Задача поисковых машин - обеспечивать детальное разыскание информации в электронной вселенной, что может быть достигнуто только за счет учета (индексирования) всего содержания максимально возможного числа web-страниц. В отличие от справочников, все они функционируют в автоматизированном режиме и имеют одинаковый принцип действия. Поисковые системы состоят из двух базовых компонентов. Первый компонент представляет собой программу-робот, задача которого путешествовать с сервера на сервер, находить там новые или изменившиеся документы и скачивать их на главный компьютер системы. При этом робот, просматривая содержимое документа, находит новые ссылки, как на другие документы данного сервера, так и на внешние сайты. Программа самостоятельно направляется по указанным ссылкам, находит новые документы и ссылки в них, после чего процесс повторяется вновь, напоминая хорошо известный в библиографии “метод снежного кома”.

Выявленные документы обрабатываются (индексируются) вторым компонентом поисковой системы. При этом, как правило, учитывается все содержание страницы, включая текст, иллюстрации, аудио и видео файлы и пр. Индексации подвергаются все слова в документе, что как раз и дает возможность использовать поисковые системы для детального поиска по самой узкой тематике. Образуемые гигантские индексные файлы, хранящие информацию о том какое слово, сколько раз, в каком документе и на каком сервере употребляется и составляют базу данных, к которой происходит обращение пользователей, вводящих в строку запроса сочетания ключевых слов.

Выдача результатов осуществляется с помощью специального модуля, который производит интеллектуальное ранжирование результатов. При этом берется в расчет местоположение термина в документе (название, заголовок, основной текст), частота его повторения, процентное соотношение искомого термина к остальному тексту страницы, а также число и авторитетность внешних ссылок на данную страницу с других сайтов.

Основные параметры поисковых машин

К основным параметрам поисковых систем относятся:

- объем индексных файлов (число проиндексированных серверов и отдельных документов);

- степень оперативности обновления базы данных за счет включения сведений о новых материалах и удаления устаревших;

- возможности для составления запроса;

- интеллектуальность системы ранжирования результатов поиска;

- наличие дополнительных сервисных функций, облегчающих работу пользователя.

Первая величина, являющаяся ключевой, устанавливает широту охвата материала и определяется числом проиндексированных документов. Сейчас эта цифра для лидеров мирового сетевого поиска колеблется в пределах от 1 до 3 с лишним миллиардов.

Учитывая тот факт, что в среднем интернетовский адрес сохраняет актуальность до полугода, после чего документ или меняет местоположение или убирается с сервера, большое значение имеет уровень оперативности обновления данных, характеризующий степень соответствия индексного файла поисковой системы реальному местоположению документов на сайтах. В настоящее время этот параметр колеблется от двух недель до полутора месяцев.

Возможности поискового механизма выражать запрос максимально точно в значительной степени предопределяют долю релевантных документов в перечне полученных результатов. Каждая машина имеет свою собственную лексику, которая по-разному позволяет детализировать поисковое предписание.

Все поисковые машины обладают модулем ранжирования результатов поиска. Создание таких модулей - целая область программирования, в которой конкурируют сложнейшие алгоритмы, созданные разными компаниями. Перечень факторов, принимаемых во внимание при определении места документа в перечне ссылок необычайно широк: от местоположения слова на странице до рейтинга (авторитета) страниц, имеющих ссылки на найденный документ.

Не последнюю роль играет и простота интерфейса, наличие дополнительных сервисных функций, как например, возможность перевода текста документа на иностранный язык, способность выделять все документы с определенного сайта, сужение критериев в ходе поиска, нахождение документов "по образцу" и т.д.

По этим параметрам среди внушительного числа поисковых систем выделяются несколько наиболее признанных, позволяющих выявлять информацию с высокой степенью полноты и надежности. К наиболее авторитетным поисковым системам всемирного масштаба в настоящее время относятся Google (www.google.com), AlltheWeb (www.alltheweb.com) и Alta Vista (www.altavista.com).

Практически все всемирно известные справочники и поисковые системы в настоящее время превратились во внушительные информационные корпорации с многомиллионными доходами. Заработав авторитет наиболее посещаемых мест в Сети, они предоставляют свои страницы для размещения рекламной информации, доходы от которой и составляют основу их бюджета. Постепенно поисковые сервера превращаются в многофункциональные порталы, в которых поисковый сервис остается главной приманкой для пользователей, но далеко не единственной и даже не основной из предоставляемых услуг. Помимо разыскания информации, такие сервера обычно предоставляют пользователям бесплатную электронную почту, возможность бесплатно размещать собственные страницы, сведения о погоде, текущих новостях, биржевые котировки, карты местности и т.д.

2. Глобальные поисковые системы

2.1. Поисковая система Google

Поисковая система, запущенная в 1998 году и являющаяся ныне единоличным лидером среди глобальных поисковых систем по всем значимым параметрам. Главное достоинство Google - объем его индексного файла, который составляет на сегодня более 3 миллиардов web-страниц и статей из групп новостей по интересам. В сутки программы-роботы системы индексируют порядка трех миллионов новых и обновленных страниц, при том, что актуализация базы производится каждые 28 дней.

Второе несомненное преимущество Google - его способность индексировать документы не только в виде HTML-файлов, но также документы в форматах PDF, RTF, PS, DOC, XLS, PPT, WP5 и ряде других. При этом Google позволяет моментально конвертировать страницы в указанных форматах в обычный HTML-файл, что освобождает пользователя от необходимости иметь специальное программное обеспечение для доступа к файлу.

Следующим важнейшим достоинством является специально разработанный модуль ранжирования результатов - PageRank. Он основан на алгоритме, согласно которому вначале устанавливается структура ссылок во всей Сети, а затем каждая отдельная страница ранжируется в соответствии с числом и значимостью ссылок на нее с других страниц. При этом авторитетность внешних ссылок более важна, чем их количество. Подобный алгоритм позволяет существенно повысить релевантность ссылок в следствии чего Google отличает высокая степень соответствия найденной информации интересам пользователя. Этот результат

достигается, в частности, еще и за счет специальной подсистемы защиты пользователя от сайтов, которые продвигаются с помощью различных недобросовестных методов.

Google отличается высокой степенью комфорта для пользователя. Несмотря на то, что это глобальная поисковая система, пользователи из неанглоязычных стран автоматически переадресовываются на интерфейс на их родном языке. Русскоязычный интерфейс, в частности, находится по адресу www.google.com.ru. Длительность процесса в большинстве случаев не превышает одной секунды, несмотря на огромный объем индексного файла системы.

Интерфейс первой страницы Google - на сегодня в Сети у него нет достойных конкурентов.

Методика поиска с помощью Google предельно проста. В поисковую строку водится запрос на естественном языке - неважно на русском, английском или любом другом. Язык запросов не допускает усечения терминов знаком "*", поэтому все возможные варианты слов (library, libraries, librarians) пользователю следует вводить самостоятельно. Все термины запроса по умолчанию объединяются условием AND (И) - перед ними нет нужды ставить знак "+". Таким образом в список результатов попадают лишь страницы, содержащие все введенные ключевые слова. Для поиска по точной фразе традиционно используются кавычки: так запрос "Кто убил кошку у мадам Поласухер?" прямо приведет к ссылке на полный текст "Собачьего сердца" Михаила Булгакова. Поисковый механизм игнорирует стоп-слова (предлоги, союзы, артикли), однако если какое-либо из таких слов существенно перед ним необходимо поставить "+", давая понять системе, что в данном случае термин даже из одной буквы является значимым (например: Александр +I).

Google имеет в своем арсенале множество опций для максимальной конкретизации запроса. Все они доступны через меню Advanced Search "Расширенный поиск". Помимо уже описанных возможностей добавляются фильтры, ограничивающие язык документа, его формат (к примеру, "только документы в PDF"), время опубликования ("последние три месяца"), месту термина в самом документе ("в заголовке страницы") или расположение страницы в определенном домене или даже сайте.

Выдаваемые в результате поиска ссылки на документы, помимо традиционных сведений о заглавии документа, контексте искомых слов и данных о размере

содержат функцию Cached "Сохранено", позволяющую полностью восстановить весь документ, если по каким-либо причинам сайт на котором он расположен недоступен. Еще одна функция Similar pages "Похожие страницы" позволяет получить перечень страниц, содержание которых схоже с указанным источником. Это функция, впрочем, пока выполнена без особого успеха.

Помимо поиска текстовых материалов Google обладает лучшими на сегодня возможностями поиска иллюстраций с помощью режима "Поиск изображений" (Images, "Картинки"). В его базе данных учтено более 390 тысяч иллюстраций, разыскание которых ведется аналогично поиску текстовых документов с возможностью ограничения определенным размером, форматом или цветностью графических файлов - все через опцию "Расширенный поиск изображений".

При использовании Google следует, однако, помнить, что при работе с файлами большого объема он индексирует не весь источник, а лишь его первые 101 Кб. (для PDF-файлов лимит ограничен 120 Кб.) поэтому индексация документа пока не всегда гарантирует возможность его нахождения по любому фрагменту текста.

В качестве собственного справочника ресурсов Интернет Google использует усовершенствованный массив Open Directory Project, что порой позволяет сочетать достоинства обоих поисковых инструментов.

Дополнительным платным сервисом Google является поиск труднодоступной информации непосредственно человеком. Стоимость этого вида обслуживания - $2.50 за ответ.

2.2. Поисковая система AlltheWeb

Поисковая система, существующая с 1997 года, расположена в Европе (Норвегия) и изначально была ориентирована преимущественно на европейские сайты. В настоящее время отражает более 2.1 миллиарда документов, среди которых весомое место занимают русскоязычные материалы. Обновление базы производится раз в две недели, среднее время индексации - 5 недель. AlltheWeb способен индексировать PDF-файлы, которые обрабатываются без ограничения их размера.

Весь Web - интерфейс главной страницы.

Помимо текстовых разысканий в WWW, AlltheWeb поддерживает поиск в группах новостей по интересам, файлов на FTP-серверах, иллюстраций, видео фрагментов и

MP3 файлов.

Система оперирует традиционным языком запросов, включающим знаки "+", "-", кавычки для поиска цитат, круглые скобки для поиска по условию ИЛИ. Более сложные ограничения задаются специфическими операторами системы. Однако для формирования многоуровневых запросов лучше обратится к Advanced Search ("Расширенному поиску"), который позволяет легко применить фильтры, в числе которых ограничения по местоположению термина в документе, определенному домену или сайту, времени опубликования, языку и даже объему исходной страницы.

Среди сервисных функций AlltheWeb возможность автоматически объединять результаты в "тематические" папки. Специальная функция позволяет каждому пользователю создать свой собственный интерфейс системы с указанием приоритетов при сортировке результатов и дополнительными услугами, как например, отправка найденных ссылок по электронной почте.

Скорость работы AlltheWeb весьма высока и полностью оправдывает название поискового механизма, именуемого Fast Search (Быстрый поиск).

2.3. Поисковая система Alta Vista

Система введена в эксплуатацию в 1996 году и в течении трех лет являлась признанным лидером сетевого поиска по объему индексного файла, эффективности ранжирования результатов и сервисным функциям. Именно в ней был впервые опробован ставший ныне традиционным язык запросов: знаки "+" и "-", усечение с помощью знака "*" и кавычки для поиска по точной фразе. С форме углубленного запроса были впервые эффективно использованы булевые операторы и оператор расстояния - NEAR.

На данный момент Alta Vista содержит сведения всего об 1 миллиарде web-страниц и статей из телеконференций. Декларируется полное обновление базы каждые три месяца. На сегодня этих показателей уже явно недостаточно, чтобы сохранить лидирующие позиции в мировом поисковом сервисе.

Перечень результатов в AltaVista. Под строкой с запросом приведены термины,

ассоциируемые с искомыми ключевыми словами.

Единственным сервисом Alta Vista, сохраняющим актуальность является поиск мультимедийный файлов, в особенности аудио и видео. Поисковый механизм при

этом опирается на имена мультимедийных файлов и подписи к иллюстрациям. До некоторой степени может быть полезна и функция перевода найденных страниц на ведущие европейские, а также китайский, корейский и японский языки.

Помимо перечисленный глобальных поисковых систем в некоторых случаях (более по инерции) продолжают использоваться устаревшие поисковые сервисы, среди которых наиболее заметны Lycos (www.lycos.com), InfoSeek (www.infoseek.com) и Excite (www.excite.com). Малый объем их индексных файлов в настоящее время не позволяет полагаться на предоставляемые ими сведения. "Молодые" поисковые системы, запущенные в 2000 и 2001 годах, такие как WiseNut (www.wisenut.com) и Teoma (www.teoma.com), несмотря на внушительный объем проиндексированных документов пока не представляют особого интереса. Ни одна из них, в частности, не способна осуществлять поиск документов на русском языке.

3. Информационно-поисковая система по законодательству – Garant

Программа Garant содержит в себе все Российское законодательство, а так же бухгалтерские законопроекты и системы налогов и документы для руководителя Менеджера. Все это тоже может пригодиться адвокату. С помощью этой программы юрист сможет быстро отыскать необходимый ему закон или документ, что значительно увеличит его работоспособность.

Эта программа имеет очень хорошо отлаженную поисковую систему, «По реквизитам» и «По ситуации». При использовании поиска «По реквизитам», Garant выдает небольшую табличку, в которой можно указать часть искомого документа, например:

- выбрать тип документа, допустим «кодекс» и программа выдаст все кодексы нашего законодательства;

- выбрать орган, допустим «орган власти города Москвы» и Garant выдаст все документы и законы выпущенные относительного этого органа;

- выбрать раздел документа, допустим «законодательство о приватизации» и программа опять же выдаст все что у нее есть по этому поводу;

- можно просто указать сроки издания закона, например с 1 января 98 года, по 1 февраля 98 года, и все документы выпущенные в этот период предстанут перед вами, а можно просто указать с какого-нибудь определенного числа и по сей день, так же можно указать дату регистрации в министерстве юстиции документа;

- если вам нужно найти какой-нибудь определенный закон или документ, шифр которого вам известен, то можно его набрать и Garant вам тут же его предоставит;

- программа Garant так же позволяет искать документ по контексту, это когда вам известно только несколько слов из закона который вам нужно отыскать или найти все законы содержащие эту фразу;

- так же можно указать статус документа, т. е. можно указать искать во всех документах или в действующих или в уже утративших силу;

- при поиске по контексту можно указать тип поиска, т. е. искать только в названии документа или икать и самом документе.

При поиске «По ситуации», вам просто необходимо набрать ключевое слово, и программа выдаст все документы которые у нее есть отсортированные по этому ключу, а вам просто нужно выбрать необходимый вам документ которые нашел для вас Garant по введенному ключу.

Из этой программы никогда ничего не удаляется, а только вносятся новые законы и они будут написаны черным цветом, а старые которые уже утратили силу просто будут написаны желтым цветом, это его способность тоже может сильно пригодиться адвокату, видь адвокаты могут пересматривать дела заключенных которые сидят в тюрьме по закону который уже утратил силу. Здесь вступает обратная сторона закона и чтоб выпустить невинного заключенного юрист должен ознакомиться с нынешним законодательством и с прошлым, а для этого ему необходимо поднять старые законы, а программа Garant ему предоставит и новые законодательства и старые, что тоже значительно ускорит работу юриста.

Также программа Garant содержит и много других возможностей для облегченной работы с ней, но они более специфические и редко используются.

4. Стратегия и методика профессионального информационного поиска

Приступая к информационному поиску в Интернет следует всегда помнить несколько основных моментов. Прежде всего никакие средства навигации - справочники или поисковые машины не охватывают всего текущего информационного массива Интернет. По некоторым оценкам даже такие признанные лидеры сетевого поиска как Google или AlltheWeb отражают не более трети совокупного содержания Сети. Причина этого - постоянный колоссальный прирост объемов информации в Интернет, который, несмотря на все усилия

навигационных служб, содержит огромное число белых пятен.

Помимо быстрого роста и изменения местоположения документов, большинство поисковых систем имеют внутренние ограничения на отражение материалов одного сайта и на объем индексируемой части страницы. Программы-роботы зачастую не идут в глубь сервера дальше определенной директории, что также сокращает число отраженных материалов.

В тоже время некоторые серверы имеют собственную систему поиска, которая отражает весь их информационный массив. Выявив такие сервера с помощью справочников, можно провести более детальное их обследование, использовав локальный поисковый механизм. Например, при поиске сведений о конкретном виде креветки, искусственно разводимой человеком, весьма рациональным будет найти и просмотреть сервера, посвященные в целом аквакультуре, отрасли, занимающейся выращиванием морепродуктов в искусственных теплых водоемах, а при выявлении данных о конкретном заболевании - сервера учреждений, ведущих исследования в данной области.

Таким образом, для достижения наиболее полных результатов следует применять справочники и поисковые системы в сочетании друг с другом.

Существует также ряд общих требований к поисковой деятельности, соблюдение которых повышает эффективность и экономит время, затрачиваемое на разыскание данных.

1. Для поиска материалов по крайне узкой специфической тематике стоит начинать с мета-машин, дабы сразу получить представление о том насколько богато данная проблематика представлена в Интернет.

2. Для получения более полных результатов по сложному запросу (например, там где есть ограничения не только по содержанию документа, но и по дате обновления или местоположению документов) поиск рекомендуется проводить отдельно в каждой поисковой машине. Поисковые системы имеют сильный разнос в отражении документов и их последовательное использование в значительной степени расширяет охват материала.

3. При разыскании документов об отдельной стране или на конкретном языке следует отдать предпочтение национальным/региональным поисковым средствам. Так, например, при поиске материалов на испанском языке, стоит обращаться не к глобальным, а к испанским поисковым системам, например, Trovator

(#"_Toc57438713">Заключение

Причина сложностей, возникающих при информационном поиске в Интернет определяется двумя главными факторами. Во-первых, число источников в Сети чрезвычайно велико. В конце 2001 года самые приблизительные подсчеты указывали ориентировочную цифру в 7,5 миллиардов документов, расположенных на серверах по всему миру. Во-вторых, массив информации в Сети не только колоссален по объему, но еще и крайне динамичен. За те полминуты, что вы потратили на чтение первых строк этого раздела в виртуальной вселенной появилось порядка сотни новых или измененных документов, десятки были перемещены на новые адреса, а единицы навсегда прекратили свое существование.

В отличии от стабильного и контролируемого фонда документов в библиотеке, в Сети мы имеем дело с гигантским и непрерывно меняющимся информационным массивом, поиск данных в котором является весьма и весьма сложным процессом.

Навыками информационных разысканий в той или иной степени обладают большинство пользователей глобальных компьютерных сетей. И дилетанты и профессионалы зачастую пользуются одними и теми же инструментами. Однако результаты разысканий и затраченное на них время различаются в очень значительной степени.

Поисковые системы (search engines) распространены в гораздо большем количестве, нежели электронные справочники и число их, составляющее сегодня нескольких десятков, продолжает неуклонно увеличиваться. Профессиональная работа с ними требует специальных навыков, поскольку простой ввод искомого термина в поисковую строку скорее всего приведет к получению списка из сотен тысяч документов, содержащих данное понятие, что практически равносильно нулевому результату.