**Информационно-поисковая система**

Рассмотрим постановку задачи поиска в общем виде. Для этого нам необходимо ответить на три вопроса: что искать, то есть, какие источники информации, где искать (места размещение этих источников) и как искать (какие инструменты для этого использовать).

**Каковы основные источники информации, представленные в Интернете**?

Это документы WWW, статьи в группах новостей и списках рассылки, файлы в библиотеках файлов, справочники адресной информации организаций и людей (электронная почта, адрес, телефон), статьи в тематических базах данных, энциклопедиях.

**Где эти источники информации размещаются**?

Это такие популярные ресурсы Интернет, как WWW, группы новостей, списки рассылки и FTP-серверы. Безусловно, можно искать нужные источники информации вручную, узнавать адреса из специализированных журналов по информатике и Интернету, использовать специальные бумажные справочники с классифицированными по категориям адресами. Однако для такого изменчивого пространства как Интернет необходимо научиться пользоваться специальными инструментами, цель которых - собирать данные об информационных ресурсах и предоставлять пользователям услугу быстрого поиска.

**ИПС (информационно-поисковая система)** - это система, обеспечивающая поиск и отбор необходимых данных в специальной базе с описаниями источников информации (**индексе**) на основе информационно-поискового языка и соответствующих правил поиска.

Главной задачей любой ИПС является поиск информации релевантной информационным потребностям пользователя. Очень важно в результате проведенного поиска ничего не потерять, то есть найти все документы, относящиеся к запросу, и не найти ничего лишнего. Поэтому вводится качественная характеристика процедуры поиска - релевантность.

**Релевантность** - это соответствие результатов поиска сформулированному запросу. Далее мы будем, в основном, рассматривать ИПС для всемирной паутины (WWW). Основными показателями ИПС для WWW являются пространственный масштаб и специализация. По пространственному масштабу ИПС можно разделить на локальные, глобальные, региональные и специализированные. Локальные поисковые системы могут быть разработаны для быстрого поиска страниц в масштабе отдельного сервера. Региональные ИПС описывают информационные ресурсы определенного региона, например, русскоязычные страницы в Интернете. Глобальные поисковые системы в отличие от локальных стремятся объять необъятное - по возможности наиболее полно описать ресурсы всего информационного пространства сети Интернет.

В общем случае, можно выделить следующие поисковые инструменты для WWW: каталоги, поисковые системы, метапоисковые системы.

**Каталог**

**Каталог** - поисковая система с классифицированным по темам списком аннотаций со ссылками на web-ресурсы. Классификация, как правило, проводится людьми. Поиск в каталоге очень удобен и проводится посредством последовательного уточнения тем. Тем не менее, каталоги поддерживают возможность быстрого поиска определенной категории или страницы по ключевым словам с помощью локальной поисковой машины. База данных ссылок (индекс) каталога обычно имеет ограниченный объем, заполняется вручную персоналом каталога. Некоторые каталоги используют автоматическое обновление индекса. Результат поиска в каталоге представляется в виде списка, состоящего из краткого описания (аннотации) документов с гипертекстовой ссылкой на первоисточник.

***Адреса популярных каталогов:***

*Зарубежные каталоги:*

Yahoo - **www.yahoo.com** Magellan - **www.mckinley.com**

*Российские каталоги:*

@Rus - **www.aport.ru** Weblist - **www.weblist.ru** Улитка - **www.ulitka.ru**

**Поисковая машина**

**Поисковая машина** - поисковая система с формируемой роботом базой данных, содержащей информацию об информационных ресурсах. Отличительной чертой поисковых машин является тот факт, что база данных, содержащая информацию об Web-страницах, статьях Usenet и т.д., формируется программой-роботом. Поиск в такой системе проводится по запросу, составляемому пользователем, состоящему из набора ключевых слов или фразы, заключенной в кавычки. Индекс формируется и поддерживается в актуальном состоянии роботами-индексировщиками. В описании документа чаще всего содержится несколько первых предложений или выдержки из текста документа с выделением ключевых слов. Как правило, указана дата обновления (проверки) документа, его размер в килобайтах, некоторые системы определяют язык документа и его кодировку (для русскоязычных документов). Что можно делать с полученными результатами? Если название и описание документа соответствует вашим требованиям, можно немедленно перейти к его первоисточнику по ссылке. Это удобнее делать в новом окне, чтобы иметь возможность далее анализировать результаты выдачи. Многие поисковые системы позволяют проводить поиск в найденных документах, причем вы можете уточнить ваш запрос введением дополнительных терминов. Если интеллектуальность системы высока, вам могут предложить услугу поиска похожих документов. Для этого вы выбираете особо понравившийся документ и указываете его системе в качестве образца для подражания. Однако, автоматизация определение похожести - весьма нетривиальная задача, и зачастую эта функция работает неадекватно вашим надеждам. Некоторые поисковики позволяют провести пересортировку результатов. Для экономии вашего времени можно сохранить результаты поиска в виде файла на локальном диске для последующего изучения в автономном режиме.

***Адреса наиболее популярных поисковых машин за рубежом и в России.***

*Зарубежные поисковые машины*:

Google - **www.google.com** Altavista - **www.altavista.com** Excite - **www.excite.com** HotBot - **www.hotbot.com** Nothern Light - **www.northernlight.com** Go (Infoseek) - **www.go.com** (**infoseek.com**) Fast - **www.alltheweb.com**

*Российские поисковые машины*:

Яndex - **www.yandex.ru** (или **www.ya.ru**)Рэмблер - **www.rambler.ru** Апорт - **www.aport.ru**

**Метапоисковая машина**

***Метапоисковая система.*** Обратите внимание на то, что различные поисковые системы описывают разное количество источников информации в Интернет. Поэтому нельзя ограничиваться поиском только в одной из указанных поисковых системах. Теперь познакомимся с инструментами поиска, которые не формируют собственный индекс, но умеют использовать возможности других поисковых систем. Это метапоисковые системы (поисковые службы) - системы, способные послать запросы пользователя одновременно нескольким поисковым серверам, затем объединить полученные результаты и представить их пользователю в виде документа со ссылками.

***Адреса известных метапоисковых систем:***

MetaCrawler - **www.metacrawler.com** SavvySearch - **www.savvysearch.com**

**Поиск источников информации**

Обсудим проблему поиска такого источника информации, как статьи в группах новостей. Инструментами поиска в данном случае могут являться рассмотренные поисковые машины WWW, которые индексируют не только пространство WWW, но и статьи в телеконференциях и имеют специальный режим поиска именно в этом ресурсе. Поиск в группах новостей поддерживает, например, поисковый сервер Altavistа. Следует отметить, что поисковые системы WWW весьма оперативно индексируют группы новостей и содержат информацию о статьях, реально существующих в сети. Для поиска в архивах новостей существую специализированные системы, самой известной из которых является система Deja (**www.deja.com**). Эта система позволяет проводить как поиск отдельных статей, содержащих введенный термин, так и поиск определенных групп новостей, посвященных обсуждению заданной темы. Можно зарегистрироваться в Deja и подписаться на определенные группы новостей.

Теперь рассмотрим инструменты, позволяющие проводить поиск файлов. Многие поисковые системы WWW стали оказывать услугу поиска мультимедийных файлов (Altavista, Aport). Для этого вовсе нет необходимости знать специальные операторы, а достаточно перейти с домашней страницы по ссылкам Картинки (Images), MP3/Audio или Video к специальному режиму поиска. Поиск проводится по возможному имени файла или по тексту в комментарии к ссылке на мультимедийный файл.

Что касается поиска программного обеспечения, во всемирной паутине существуют поисковые Web-серверы с коллекциями условно-бесплатного ПО, некоторые из них специализируются на поиск программного обеспечения для Интернета или для конкретной операционной системы. Эти системы в конечном итоге приведут вас к конкретному серверу, с которого и можно скачать искомый программный продукт. Следует упомянуть серверы Archie, также оказывающие услугу поиска файлов на FTP-серверах, однако пользоваться Web-серверами гораздо удобнее.

Рассмотрим поисковые инструменты для поиска адресной информации. Введем понятие Белого(White) и Желтого (Yellow) поиска.

**White-поиск** - поиск адресной информации по заранее известному собственному имени адресата (имя человека или организации)

**Yellow-поиск** - поиск собственного имени по дополнительным признакам (по роду деятельности, по географическому признаку), а затем поиск его адресной информации.

Обычно Yellow Pages системы фактически сразу включают в себя и White Pages - у найденного адресата сразу видны его телефон и почтовый адрес. Кроме того, некоторые Yellow Pages позволяют искать просто в алфавитном списке своих абонентов (white-поиск). С другой стороны, White pages также содержат элементы yellow-поиска - кроме задания собственного имени они обычно позволяют указать название города, штата и другие, сужающие поиск, данные (что необходимо в случае многих однофамильцев). Возможно, именно поэтому многие on-line телефонные справочники, выполняющие, фактически white-поиск, называют себя Yellow pages.

Здесь приведены адреса Web-систем для поиска адресной информации для людей и организаций.

***Поиск людей:***

* Поиск людей на Yahoo **(http://people.yahoo.com).**
* Система WhoWhere **(www.whowhere.com).**
* Система Bigfoot **(www.bigfoot.com).**

***Поиск организаций:*** раздел Желтые страницы (Yellow pages) на поисковых системах специализированные сервера **www.yellowpages.com** - для поиска в США и других странах.

Пользователям Internet уже хорошо известны названия таких сервисов и информационных служб, как Lycos, AltaVista, Yahoo, OpenText, InfoSeek, а без услуг этих систем сегодня практи чески нельзя найти что-либо полезное в море информационных ресурсов Сети. Но что собой представляют эти сервисы изнутри, как они устроены, почему результат поиска в терабайтных массивах информации осуществляется достаточно быстро и как устроено ранжирование документов при выдаче - все это обычно остается за кадром. Тем не менее без правильного планирования стратегии поиска, знакомства с основными положениями теории ИПС (Информационно-Поисковых Систем), насчитывающей уже двадцатилетнюю историю, трудно эффективно использовать даже такие скорострельные сервисы, как AltaVista или Lycos.

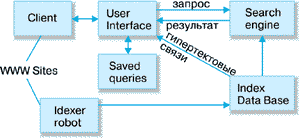
Информационно-поисковые системы появились на свет достаточно давно. Теории и практике построения таких систем посвящено множество статей, основная масса которых приходится на конец 70-х - начало 80-х годов. Среди отечественных источников следует выделить научно-технический сборник "Научно-техническая информация. Серия 2", который выходит до сих пор. На русском языке издана так же и "библия" по разработке ИПС - "Динамические библиотечно-информационные системы" Ж. Солтона [1], в которой рассмотрены основные принципы построения информационно-поисковых систем и моделирования процессов их функционирования. Таким образом, нельзя сказать, что с появлением Internet и бурным вхождением его в практику информационного обеспечения появилось нечто принципиально новое, чего не было раньше. Если быть точным, то ИПС в Internet - это признание того, что ни иерархическая модель Gopher, ни гипертекстовая модель World Wide Web еще не решают проблему поиска информации в больших объемах разнородных документов. И на сегодняшний день нет другого способа быстрого поиска данных, кроме поиска по ключевым словам. При использовании иерархической модели Gopher приходится довольно долго бродить по дереву каталогов, пока не встретишь нужную информацию. Эти каталоги должны кем-то поддерживаться, и при этом их тематическое разбиение должно совпадать с информационными потребностями пользователя. Учитывая анархичность Internet и огромное количество всевозможных интересов у пользователей Сети, понятно, что кому-то может и не повезти и в сети не будет каталога, отражающего конкретную предметную область. Именно по этой причине для множества серверов Gopher, называемого GopherSpace была разработана информационно-поисковая программа Veronica (Very Easy Rodent-Oriented Net-wide Index of Computerized Archives).

Аналогичное развитие событий наблюдается и в World Wide Web. Собственно еще в 1988 году в специальном выпуске журнала "Communication of the ACM" [2] среди прочих проблем разработки гипертекстовых систем и их использования Франк Халаз назвал в качестве первоочередной задачи для следующего поколения систем этого типа назвал проблему организации поиска информации в больших гипертекстовых сетях. До сих пор многие идеи, высказанные в той статье, не нашли еще своей реализации. Естественно, что система, предложенная Бернерсом-Ли [3] и получившая такое широкое распространение в Internet, должна была столкнуться с теми же проблемами, что и ее локальные предшественники. Реальное подтверждение этому было продемонстрировано на второй конференции по World Wide Web осенью 1994 года, на которой были представлены доклады о разработке информационно-поисковых систем для Web, а система World Wide Web Worm, разработанная Оливером МакБрайном из Университета Колорадо, получила приз как лучшее навигационное средство. Следует также отметить, что все-таки долгая жизнь суждена отнюдь не чудесным программам талантливых одиночек, а средствам, являющимся результатом планового и последовательного движения научных и производственных коллективов к поставленной цели. Рано или поздно этап исследований заканчивается, и наступает этап эксплуатации систем, а это уже совсем другой род деятельности. Именно такая судьба ожидала два других проекта, представленных на той же конференции: Lycos, поддерживаемый компанией Microsoft, и WebCrawler, ставший собственностью America On-line.

Разработка новых информационных систем для Web не завершена. Причем как на стадии написания коммерческих систем, так и на стадии исследований. За прошедшие два года снят только верхний слой возможных решений. Однако многие проблемы, которые ставит перед разработчиками ИПС Internet, не решены до сих пор. Именно этим обстоятельством и вызвано появление проектов типа AltaVista компании Digital [4], главной целью которого является разработка программных средств информационного поиска для Web и подбор архитектуры для информационного сервера Web.

**Архитектура современных ИПС для WWW**

Прежде чем описать проблемы построения информационно-поисковых систем Web и пути их решения рассмотрим типовую схему такой системы. В различных публикациях, посвященных конкретным системам, например [5,6], приводятся схемы, которые отличаются друг от друга только способом применения конкретных программных решений, а не принципом организации различных компонентов системы. Поэтому рассмотрим эту схему на примере, взятом из работы [6] (рис. 1).



**Рисунок 1.** Типовая схема информационно-поисковой системы.

**Client (клиент)** на этой схеме - это программа просмотра конкретного информационного ресурса. Наиболее популярны сегодня мультипротокольные программы типа Netscape Navigator. Такая программа обеспечивает просмотр документов WWW, Gopher, Wais, FTP-архивов, почтовых списков рассылки и групп новостей Usenet. В свою очередь все эти информационные ресурсы являются объектом поиска информационно-поисковой системы.

**User interface (пользовательский интерфейс)** - это не просто программа просмотра, в случае информационно-поисковой системы под этим словосочетанием понимают также способ общения пользователя с поисковым аппаратом: системой формирования запросов и просмотров результатов поиска.

**Search engine (поисковая машина)** - служит для трансляции запроса на информационно-поисковом языке (ИПЯ), в формальный запрос системы, поиска ссылок на информационные ресурсы Сети и выдачи результатов этого поиска пользователю.

**Index database (индекс базы данных)** - индекс, который является основным массивом данных ИПС и служит для поиска адреса информационного ресурса. Архитектура индекса устроена таким образом, чтобы поиск происходил максимально быстро и при этом можно было бы оценить ценность каждого из найденных информационных ресурсов сети.

**Queries (запросы пользователя)** - сохраняются в его (пользователя) личной базе данных. На отладку каждого запроса уходит достаточно много времени, и поэтому чрезвычайно важно запоминать запросы, на которые система дает хорошие ответы.

**Index robot (робот-индексировщик)** - служит для сканирования Internet и поддержания базы данных индекса в актуальном состоянии. Эта программа является основным источником информации о состоянии информационных ресурсов сети.

**WWW sites** - это весь Internet или точнее - информационные ресурсы, просмотр которых обеспечивается программами просмотра.

Рассмотрим теперь назначение и принципу построения каждого из этих компонентов более подробно и определим, в чем отличие данной системы от традиционной ИПС локального типа.

**Информационные ресурсы и их представление в ИПС**

Как видно из рис. 1, документальным массивом ИПС Internet является все множество документов шести основных типов: WWW-страницы, Gopher-файлы, документы Wais, записи архивов FTP, новости Usenet и статьи почтовых списков рассылки. Все это довольно разнородная информация, которая представлена в виде различных, никак несогласованных друг с другом форматов данных: тексты, графическая и аудиоинформация и вообще все, что имеется в указанных хранилищах. Естественно возникает вопрос - как информационно-поисковая система должна со всем этим работать?

В традиционных системах используется понятие поискового образа документа - ПОД. Обычно, этим термином обозначают нечто, заменяющее собой документ и использующееся при поиске вместо реального документа. Поисковый образ является результатом применения некоторой модели информационного массива документов к реальному массиву. Наиболее популярной моделью является векторная модель [7], в которой каждому документу приписывается список терминов, наиболее адекватно отражающих его смысл. Если быть более точным, то документу приписывается вектор размерности, равный числу терминов, которыми можно воспользоваться при поиске. При булевой векторной модели элемент вектора равен 1 или 0, в зависимости от наличия или отсутствия термина в ПОД. В более сложных моделях термины взвешиваются - элемент вектора равен не 1 или 0, а некоторому числу (весу), отражающему соответствие данного термина документу. Именно последняя модель стала наиболее популярной в ИПС Internet [4,6,7].

Вообще говоря, существуют и другие модели описания документов: вероятностная модель информационных потоков и поиска и модель поиска в нечетких множествах [7]. Не вдаваясь в подробности, имеет смысл обратить внимание на то, что пока только линейная модель применяется в системах Lycos, WebCrawler, AltaVista, OpenText и AliWeb. Однако ведутся исследования по применению и других моделей, результаты которых отражены в работах [4, 6]. Таким образом, первая задача, которую должна решить ИПС, - это приписывание списка ключевых слов документу или информационному ресурсу. Именно эта процедура и называется индексированием. Часто, однако, индексированием называют составление файла инвертированного списка, в котором каждому термину индексирования ставится в соответствие список документов в которых он встречается. Такая процедура является только частным случаем, а точнее, техническим аспектом создания поискового аппарата ИПС. Проблема, связанная с индексированием, заключается в том, что приписывание поискового образа документу или информационному ресурсу опирается на представление о словаре, из которого эти термины выбираются, как о фиксированной совокупности терминов. В традиционных системах существовало разбиение на системы с контролируемым словарем и системы со свободным словарем. Контролируемый словарь предполагал ведение некоторой лексической базы данных, добавление терминов в которую производилось администратором системы, и все новые документы могли быть заиндексированы только теми терминами, которые были в этой базе данных. Свободный словарь пополнялся автоматически по мере появления новых документов. Однако на момент актуализации словарь также фиксировался. Актуализация предполагала полную перезагрузку базы данных. В момент этого обновления перегружались сами документы, и обновлялся словарь, а после его обновления производилась переиндексация документов. Процедура актуализации занимала достаточно много времени и доступ к системе в момент ее актуализации закрывался.

Теперь представим себе возможность такой процедуры в анархичном Internet, где ресурсы появляются и исчезают ежедневно. При создании программы Veronica для GopherSpace предполагалось, что все серверы должны быть зарегистрированы, и таким образом велся учет наличия или отсутствия ресурса. Veronica раз в месяц проверяла наличие документов Gopher и обновляла свою базу данных ПОД для документов Gopher. В World Wide Web ничего подобного нет. Для решения этой задачи используются программы сканирования сети или роботы-индексировщики [8]. Разработка роботов - это довольно нетривиальная задача; существует опасность зацикливания робота или его попадания на виртуальные страницы. Робот просматривает сеть, находит новые ресурсы, приписывает им термины и помещает в базу данных индекса. Главный вопрос заключается в том, что за термины приписывать документам, откуда их брать, ведь ряд ресурсов вообще не является текстом. Сегодня роботы обычно используют для индексирования следующие источники для пополнения своих виртуальных словарей: гипертекстовые ссылки, заголовки, заглавия (H1,H2), аннотации, списки ключевых слов, полные тексты документов, а также сообщения администраторов о своих Web-страницах [9]. Для индексирования telnet, gopher, ftp, нетекстовой информации используются главным образом URL, для новостей Usenet и почтовых списков поля Subject и Keywords. Наибольший простор для построения ПОД дают HTML документы. Однако не следует думать, что все термины из перечисленных элементов документов попадают в их поисковые образы. Очень активно применяются списки запрещенных слов (stop-words), которые не могут быть употреблены для индексирования, общих слов (предлоги, союзы и т.п.). Таким образом даже то, что в OpenText, например, называется полнотекстовым индексированием реально является выбором слов из текста документа и сравнением с набором различных словарей, после которого термин попадает в ПОД, а потом и в индекс системы. Для того чтобы не раздувать словарей и индексов (индекс системы Lycos уже сегодня равен 4 Тбайт), применяется такое понятие, как вес термина [10]. Документ обычно индексируется через 40 - 100 наиболее "тяжелых" терминов.

**Индекс поиска**

После того как ресурсы заиндексированы и система составила массив ПОД, начинается построение поискового аппарата. Совершенно очевидно, что лобовой просмотр файла или файлов ПОД займет много времени, что абсолютно не приемлемо для интерактивной системы WWW. Для ускорения поиска строится индекс, которым в большинстве систем является набор связанных между собой файлов, ориентированных на быстрый поиск данных по запросу. Структура и состав индексов различных систем могут отличаться друг от друга и зависят от многих факторов: размер массива поисковых образов, информационно-поисковый язык, размещения различных компонентов системы и т.п. Рассмотрим структуру индекса на примере системы [6], для которой можно реализовывать не только примитивный булевый, но и контекстный и взвешенный поиск, а также ряд других возможностей, отсутствующие во многих поисковых системах Internet, например Yahoo. Индекс рассматриваемой системы состоит из таблицы идентификаторов страниц (page-ID), таблицы ключевых слов (Keyword-ID), таблицы модификации страниц, таблицы заголовков, таблицы гипертекстовых связей, инвертированного (IL) и прямого списка (FL).

Page-ID отображает идентификаторы страниц в их URL, Keyword-ID - каждое ключевое слов в уникальный идентификатор этого слова, таблица заголовков - идентификатор страницы в заголовок страницы, таблица гипертекстовых ссылок - идентификатор страниц в гипертекстовую ссылку на эту страницу. Инвертированный список ставит в соответствие каждому ключевому слову документа список пар - идентификатор страницы, позиция слова в странице. Прямой список - это массив поисковых образов страниц. Все эти файлы так или иначе используются при поиске, но главным среди них является файл инвертированного списка. Результат поиска в данном файле - это объединение и/или пересечение списков идентификаторов страниц. Результирующий список, который преобразовывается в список заголовков, снабженных гипертекстовыми ссылками возвращается пользователю в его программу просмотра Web. Для того чтобы быстро искать записи инвертированного списка, над ним надстраивается еще несколько файлов, например, файл буквенных пар с указанием записей инвертированного списка, начинающихся с этих пар. Кроме этого, применяется механизм прямого доступа к данным - хеширование. Для обновления индекса используется комбинация двух подходов. Первый можно назвать коррекцией индекса "на ходу" с помощью таблицы модификации страниц. Суть такого решения довольно проста: старая запись индекса ссылается на новую, которая и используется при поиске. Когда число таких ссылок становится достаточным для того, чтобы ощутить это при поиске, то происходит полное обновление индекса - его перезагрузка. Эффективность поиска в каждой конкретной ИПС определяется исключительно архитектурой индекса. Как правило, способ организации этих массивов является "секретом фирмы" и ее гордостью. Для того чтобы убедиться в этом, достаточно почитать материалы OpenText [11].

**Информационно-поисковый язык системы**

Индекс - это только часть поискового аппарата, скрытая от пользователя. Второй частью этого аппарата является информационно-поисковый язык (ИПЯ), позволяющий сформулировать запрос к системе в простой и наглядной форме. Уже давно осталась позади романтика создания ИПЯ, как естественного языка, - именно этот подход использовался в системе Wais на первых стадиях ее реализации. Если даже пользователю предлагается вводить запросы на естественном языке, то это еще не значит, что система будет осуществлять семантический разбор запроса пользователя. Проза жизни заключается в том, что обычно фраза разбивается на слова, из которых удаляются запрещенные и общие слова, иногда производится нормализация лексики, а затем все слова связываются либо логическим AND, либо OR. Таким образом, запрос типа:

**>Software that is used on Unix Platform**

будет преобразован в:

**>Unix AND Platform AND Software**

что будет означать примерно следующее: "Найди все документы, в которых слова Unix, Platform и Software встречаются одновременно".

Возможны и варианты. Так, в большинстве систем фраза "Unix Platform" будет опознана как ключевая фраза и не будет разделяться на отдельные слова. Другой подход заключается в вычислении степени близости между запросом и документом. Именно этот подход используется в Lycos. В этом случае в соответствии с векторной моделью представления документов и запросов вычисляется их мера близости. Сегодня известно около дюжины различных мер близости. Наиболее часто применяется косинус угла между поисковым образом документа и запросом пользователя. Обычно эти проценты соответствия документа запросу и выдаются в качестве справочной информации при списке найденных документов.

Наиболее развитым языком запросов из современных ИПС Internet обладает Alta Vista. Кроме обычного набора AND, OR, NOT эта система позволяет использовать еще и NEAR, позволяющий организовать контекстный поиск. Все документ в системе разбиты на поля, поэтому в запросе можно указать, в какой части документа пользователь надеется увидеть ключевое слово: ссылка, заглавие, аннотация и т.п. Можно также задавать поле ранжирования выдачи и критерий близости документов запросу.

**Интерфейс системы**

Важным фактором является вид представления информации в программе-интерфейсе. Различают два типа интерфейсных страниц: страницы запросов и страницы результатов поиска.

При составлении запроса к системе используют либо меню - ориентированный подход, либо командную строку. Первый позволяет ввести список терминов, обычно разделяемых пробелом, и выбрать тип логической связи между ними. Логическая связь распространяется на все термины. На приведенной на рис. 1 схеме имеется сохраненные запросы пользователя - в большинстве систем это просто фраза на ИПЯ, которую можно расширить за счет добавления новых терминов и логических операторов. Но это только один способ использования сохраненных запросов, называемый расширением или уточнением запроса. Для выполнения этой операции традиционная ИПС хранит не запрос как таковой, а результат поиска - список идентификаторов документов, который объединяется/пересекается со списком, полученным при поиске документов по новым терминам. К сожалению, сохранение списка идентификаторов найденных документов в WWW не практикуется, что было вызвано особенностью протоколов взаимодействия программы-клиента и сервера, не поддерживающих сеансовый режим работы.

Итак, результат поиска в базе данных ИПС - это список указателей на удовлетворяющие запросу документы. Различные системы представляют этот список по-разному. В некоторых выдается только список ссылок, а в таких, как Lycos, Alta Vista и Yahoo, дается еще и краткое описание, которое заимствуется либо из заголовков, либо из тела самого документа. Кроме этого, система сообщает, на сколько найденный документ соответствует запросу. В Yahoo, например, это количество терминов запроса, содержащихся в ПОД, в соответствии с которым ранжируется результат поиска. Система Lycos выдает меру соответствия документа запросу, по которой производится ранжирование.

При обзоре интерфейсов и средств поиска нельзя пройти мимо процедуры коррекции запросов по релевантности [7]. Релевантность - это мера соответствия найденного системой документа потребности пользователя. Различают формальную релевантность и реальную. Первую вычисляет система, и на основании чего ранжируется выборка найденных документов. Вторая - это оценка самим пользователем найденных документов. Некоторые системы имеют для этого специальное поле [6], где пользователь может отметить документ как релевантный. При следующей поисковой итерации запрос расширяется терминами этого документа, а результат снова ранжируется. Так происходит до тех пор, пока не наступит стабилизация, означающая, что ничего лучше, чем полученная выборка, от данной системы не добьешься.

Кроме ссылок на документы в списке, полученном пользователем, могут оказаться ссылки на части документов или на их поля. Это происходит при наличии ссылок типа http://host/path#mark или ссылок по схеме WAIS. Возможны ссылки и на скрипты, но обычно такие ссылки роботы пропускают, и система их не индексирует. Если с http-ссылками все более или менее понятно, то ссылки WAIS - это гораздо более сложные объекты. Дело в том, что WAIS реализует архитектуру распределенной информационно-поисковой системы, при которой одна ИПС, например Lycos, строит поисковый аппарат над поисковым аппаратом другой системы - WAIS. При этом серверы WAIS имеют свои собственные локальные базы данных. При загрузке документов в WAIS администратор может описать структуру документов, разбив их на поля, и хранить документы в виде одного файла. Индекс WAIS будет ссылаться на отдельные документы и их поля как на самостоятельные единицы хранения, программа просмотра ресурсов Internet в этом случае должна уметь работать с протоколом WAIS, чтобы получить доступ к этим документам.

**Заключение**

В обзорной статье были рассмотрены основные элементы информационно-поисковых систем и принципы их построения. Сегодня ИПС являются наиболее мощным механизмом поиска сетевых информационных ресурсов Internet. К сожалению, в российском секторе Internet пока не наблюдается активного изучения этой проблемы за исключением, может быть, проекта LIBWEB, финансируемого РФФИ и системы "Паук", которая работает недостаточно надежно. Наибольшим опытом разработки такого сорта систем безусловно обладает ВИНИТИ, но здесь работа сосредоточена пока на размещении своих собственных ресурсов в Сети, что принципиально отличается от информационно-поисковых систем Internet типа Lycos, OpenText, Alta Vista, Yahoo, InfoSeek и т.п. Казалось бы, что такая работа могла быть сосредоточена в рамках таких проектов, как Россия On-line компании SovamTeleport, но здесь мы пока наблюдаются ссылки на чужие поисковые машины. Развитие ИПС для Internet в США началось два года назад, учитывая отечественные реалии и темпы развития технологий Сети в России, можно надеяться, что у нас еще все впереди.