Федеральное агентство по образованию

Смоленский государственный университет

**Технологии поиска информации в сети Интернет.**

Смоленск

2010

**Содержание:**

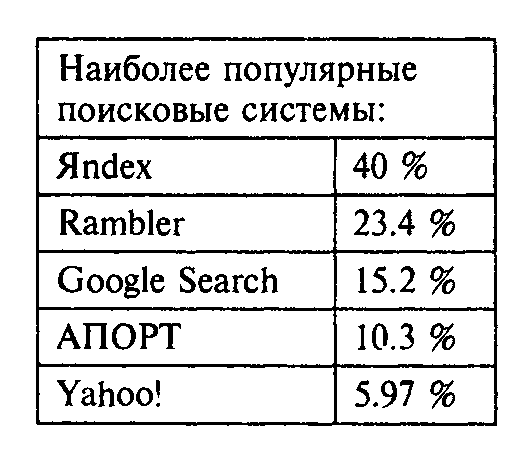
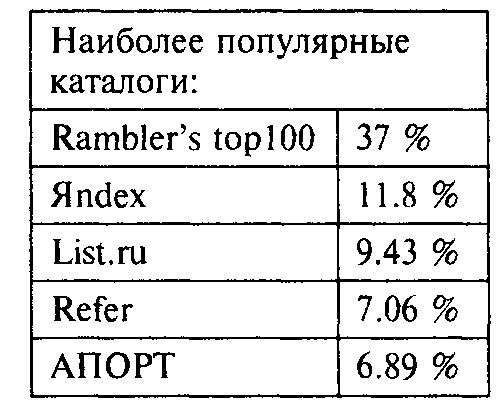
|  |  |
| --- | --- |
| Введение…………………………………………………………………………..  1.Классификация поисковых систем …………………………………………...  2.Приемы поиска………………………………………………………………....  2.1.Простой поиск………………………………………………………………..  2.2.Расширенный поиск………………………………………………………….  3.Тематический каталог Yahoo.…………………………………………………  4.Автоматический индекс Alta Vista……………………………………………  5.Поиск информации в конференциях (Usenet)………………………………..  6.Отечественные поисковые службы …………………………………………..  6.1.Rambler ……………………………………………………………………….  6.2.Яндекс………………………………………………………………………….  7.Поиск файлов …………………………………………………………………..  8.Источники специализированной информации ………………………………  9.Поиск отдельных людей в сети Internet ……………………………………...  10.Метапоиск……………………………………………………………………..  11.Эффективный поиск информации…………………………………………... Заключение………………………………………………………………………. Литература……………………………………………………………………….. | 3  3  6  6  8  10  12  13  14  15  15  16  17  18  19  19  21  22 |

**Введение**

Любой, кто пытался когда-нибудь найти нужную информацию в Интернете, наверняка согласится с тем, что этот процесс похож на поиск иголки в стоге сена. Глобальная сеть Internet содержит огромный объем информации, который стремительно увеличивается с каждым днем. В силу этого часто оказывается, что задача нахождения необходимой информации в этом информационном океане является чрезвычайно сложной и нужно уметь эффективно использовать различные поисковые системы (программы поиска информации в сети). Необходимо учесть, что "поиск - это искусство".

1. **Классификация поисковых систем**

Существует много разных поисковых систем принадлежащих разным компаниям. Прежде всего, выделяют так называемые тематические каталоги (например, Yahoo) и автоматические индексы (например, AltaVista), хотя необходимо иметь ввиду, что целый ряд поисковых систем занимает некоторое промежуточное положение между этими двумя "полюсами", то есть они содержат в себе элементы обоих этих классов. Каждая из поисковых систем имеет свою обширную базу данных об адресах (местоположении) различных Web-документов, и поиск ссылок на необходимую нам информацию происходит, не в самих Web-документах, а именно в этой базе данных.



Тематические каталоги и автоматические индексы различаются, прежде всего, по тому, как формируются и пополняются их базы данных: принимают ли в этом процессе участие люди, или все происходит совершенно автоматически.

Базы данных тематических каталогов составляются и систематически пополняются экспертами в соответствующих областях на основании новых Web-документов, обнаруженных в Internet специальными поисковыми программами. Тематический каталог представляет пользователю Internet некоторую древовидную структуру категорий (разделов и подразделов), на верхнем уровне которой собраны самые общие понятия, такие как Наука, Искусство, Бизнес и т. п., а элементы самого нижнего уровня представляют собой ссылки на отдельные Web-страницы и серверы вместе с кратким описанием их содержимого. По этому иерархическому каталогу можно путешествовать, начиная с более общих категорий (понятий) к более узким, специализированным.

Например, для нахождения информации о состоянии научных исследований по теории суперструн можно спуститься вниз по следующей "лестнице" понятий:

Science (Наука)

Physics (Физика)

Theoretical Physics (Теоретическая физика)

Theories (Теории)

String Theories (Теории струн)

В результате будет получен список сайтов, среди которых наибольший интерес представляет сайт Superstrings. Щелкнув мышью на гиперссылке Superstrings, мы попадаем на home page сайта, с соответствующим заголовком, на котором можно найти on-line учебник по теории суперструн, различные ссылки для дальнейшего чтения, глоссарий по суперструнам и т.д.

Главным достоинством тематических каталогов является большая ценность получаемой пользователем информации, что обеспечивается присутствием "человеческого фактора" в процессе анализа и сортировки новых Web-страниц. С другой стороны, тематические каталоги имеют существенный недостаток, связанный опять же с человеческим фактором, ибо из-за ограниченных возможностей человека их базы данных охватывают лишь небольшую часть всего информационного Web-пространства (менее 1 %). Таким образом, несмотря на всю полезность тематических каталогов, использование лишь поисковых систем этого вида часто оказывается явно недостаточным.

Сводная таблица избранных предметных каталогов

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | List.Ru | Апорт | Яндекс | Rambler | Yahoo! | About |
| Общая характеристика | 19 разделов верхнего уровня | 14 разделов верхнего уровня | 10 основных разделов, 7 комбинированных, дополнительная классификация | 56 разделов | 14 основных разделов | 36 разделов |
| Сортировка ресурсов внутри раздела | Алфавит, оценка гидов, популярность (посещаемость), дата | Алфавит, посещаемость, лига, оценка числа ссылок на данный ресурс, мнение пользователей | Алфавит, дата добавления, индекс цитируемости | по посещаемости | по алфавиту | по оплаченности ссылок |
| Булевские операторы | Используется язык поисковой машины Яndex | Используется язык поисковой машины Апорт | Используется язык поисковой машины Яndex | Используется язык поисковой машины Rambler | Нет | Нет |
| Поиск по фразе | " " | " " |
| Префиксы | +, - | +, - |
| Итеративный поиск (в результатах) | Есть поиск внутри категории | После входа щелкните More… |
| замена части слова | \* | \* (не всегда корректно) |

В отличие от тематических каталогов, базы данных для автоматических индексов создаются и пополняются полностью автоматически некоторыми специальными, внутренними поисковыми программами-роботами, которые в круглосуточном режиме просматривают Internet-узлы (сайты) в поисках вновь появившихся Web-документов. Из каждого такого документа робот извлекает все содержащиеся в нем новые ссылки и добавляет их в свою базу адресов, в результате чего у программы-робота возникает возможность просмотра еще некоторого количества новых для него Web-документов. В каждом новом Web-документе робот анализирует все входящие в него слова и в разделе базы данных, соответствующем каждому данному слову, запоминается адрес (URL) документа, где это слово встретилось. Таким образом, база данных, создаваемая автоматическим индексом, фактически хранит сведения о том, в каких Web-документах содержаться те или иные слова. В отличие от тематических каталогов, автоматические индексы охватывают до 25 % общего Web-пространства.

Автоматический индекс имеет отдельную поисковую систему для обеспечения интерфейса с пользователем. Эта система может, просматривая базу данных, по заданному набору ключевых слов находить и выдавать на экран пользовательского компьютера адреса и краткую информацию обо всех Web-страницах, которые содержат данный набор ключевых слов. Таким образом, автоматический индекс состоит из трех частей: программы-робота, собираемой этим роботом базы данных и интерфейса для поиска в этой базе данных. Именно с последней составляющей и работает пользователь. В силу такой организации, автоматический индекс не делает какой-либо классификации или оценивания информации.

Интерфейс автоматических индексов позволяет пользователю задать некоторый набор ключевых слов, которые с его точки зрения, являются характерными для искомых им документов, и позволяют, таким образом, найти достаточно ограниченное число потенциально относящихся к делу Web-страниц. В качестве таковых слов могут быть использованы некоторые специфические термины и их комбинации, достаточно редкие фамилии и т. д.

Успех поиска нужной информации в значительной степени определяется именно удачным выбором ключевых слов, ибо в противном случае поисковая система может выдать многие тысячи и миллионы ссылок на не относящиеся к делу Web-документы.

Следует иметь ввиду, что начинающего пользователя подстерегает много разных неожиданностей, иногда доходящих до анекдотических ситуаций.

Если для автоматических индексов поиск по ключевым словам является единственным средством нахождения необходимой информации, то в тематических каталогах (например, в Yahoo!) это средство является альтернативным способом поиска наряду с путешествием по системе (дереву) вложенных друг в друга категорий.

Некоторой разновидностью поисковых служб являются рейтинговые службы. Они предоставляют клиенту готовый список некоторых ссылок, к которым обращались наиболее часто другие пользователи сети Internet. Когда речь идет о темах, имеющих общественный интерес, таких как новости, музыка и т. д., такие рекомендации, полученные статистическим методом, являются весьма удобными и полезными. Такие услуги, в частности, обеспечивает отечественная служба Rambler.

1. **Приемы поиска** 
   1. **Простой поиск**

Каждая поисковая система (ПС ) предоставляет свои методы поиска и имеет свои особенности в правилах записи комбинаций ключевых слов. Однако есть общие элементы одинаково справедливые для большинства поисковых систем. Обычно все ПС допускают поиск Web-документов по ключевым словам, которые являются характерными для искомого документа. Выбор таких слов часто является нетривиальной задачей.

Как правило, указание одного ключевого слова является недостаточным, и тогда возникает вопрос о том, как задавать их соответствующие комбинации. Пусть, например, мы хотим получить информацию о российском лауреате нобелевской премии 2000 г. Жоресе Алферове. Указание в поле поиска только фамилии

Alferov

приведет к обнаружению огромного количества документов, большинство из которых никакого отношения к интересующему нас физику не имеет. Задавая же в поле поиска два ключевых слова

Zhorez Alferov,

разделенных символом пробела, мы должны иметь ввиду, что разные поисковые системы реагируют на этот символ по-разному. Большинство ПС рассматривают пробел как знак логической операции "или" (OR) (к ним относятся Yahoo, AltaVista и т. д.). Поэтому поиск по комбинации слов Zhorez Alferov не сократит, а, наоборот, увеличит количество документов, найденных поисковой системой: будут найдены все документы, где встречается или слово Zhorez, или слово Alferov, или оба эти слова вместе.

В таких случаях для более адекватного поиска можно использовать заключение ключевых слов в кавычки, которые определяют точную комбинацию заключенных в них символов:

"Zhorez Alferov".

Однако, при этом не будут найдены документы, в которые входит комбинация этих слов в другом порядке (Alferov Zhorez), ибо всё заключенное в кавычки является единым словосочетанием, и поисковая система ищет в документах абсолютно точное совпадение всех символов этого словосочетания.

Для того, чтобы найти только те Web-страницы, на которых одновременно присутствуют все наши ключевые слова, перед каждым из них нужно поставить знак плюс. Например,

+Zhorez +Alferov

приведет к нахождению Web-документов, где обязательно присутствует и слово Zhorez, и слово Alferov, причем не только в любом порядке, но и на любом расстоянии друг от друга (в последнем случае в найденном документе между этими словами может не быть никакой логической связи).

Знак минус перед словом исключает все документы, которые его содержат, что в ряде случаев может быть очень полезно. Например, если нам нужно найти разные работы по геометрической, но не волновой оптике, то можно задать такую комбинацию ключевых слов

+свет+оптика-волна

Необходимо иметь ввиду, что многие слова могут иметь одинаковую основу, но разные окончания (например, указание множественного числа в английском языке, падежного окончания в русском и т. д.). В связи с этим многие поисковые системы, получив запрос в виде последовательности ключевых слов, производят его нормализацию: отбрасывают из этого запроса все слова, состоящие менее чем из четырех символов (например, английский предлог of), а также отбрасывают различные окончания и суффиксы.

Например, чтобы не пропустить документ, в который входит фраза "В геометрической оптике световой луч...", поисковая система может приведенный выше запрос свести (нормализовать) к виду

+свет+оптик\*-волн\* (1)

Здесь символом звездочка (\*) обозначены любые комбинации символов после корня соответствующих слов.

Поскольку не все ПС проводят такую первоначальную обработку запроса клиента, целесообразно самим задавать запрос в форме (1), то есть использовать символ "звездочка" для указания возможных вариаций окончаний ключевых слов.

Необходимо иметь в виду некоторое различие при использовании больших (прописных) и малых (строчных) букв в ключевых словах. Обычно поиск по маленьким буквам учитывает совпадения и когда в тексте стоит большая буква, и когда маленькая. Написание же ключевых слов большими буквами означает поиск документов, содержащих слово именно в таком виде.

Очень полезным может быть поиск ключевых слов только в заголовках Web-страниц, поскольку важность заголовка документа обычно больше важности его текста. В разных ПС поиск ключевых слов лишь в заголовках Web-страниц, задается с помощью разных служебных слов: в AltaVista это слово title:, в Yahoo! — t: в Rambler — $ title:, и т. д.

* 1. **Расширенный поиск**

Приемы простого поиска не всегда позволяют с достаточной эффективностью найти требуемые нам Web-документы, в связи с чем поисковые системы предоставляют клиенту возможность использовать средства так называемого расширенного (advanced) поиска. Запросы для расширенного поиска строятся из ключевых слов, связанных знаками логических операций, и имеют вид аналогичный булевским выражениям, используемым в языках программирования.

При этом кроме стандартных логических операций OR ("или"), AND ("и") и NOT (отрицание) используется, (и является очень эффективной!) операция NEAR, определяющая степень близости друг к другу ключевых слов в Web-документе.

При выполнении сложных запросов важно понимать, в каком порядке выполняются приведенные выше операции, то есть каков их приоритет. В последовательности OR, AND, NOT, NEAR приоритет операций возрастает слева направо. Порядок выполнения отдельных элементов сложного запроса можно изменить с помощью круглых скобок, как это принято при записи булевских выражений в программировании.

В отличие от средств простого поиска, средства расширенного поиска го-раздо более стандартизованы для разных поисковых систем. Большим преимуществом расширенного поиска является также гибкость предлагаемых клиенту возможностей построения запроса, а существенным недостатком — значительно более медленная его работа по сравнению с простым поиском.

OR (операция "или") выполняет ту же самую функцию, что и знак пробела в запросах простого поиска в большинстве ПС. Например, запрос

bush OR modes

приводит к поиску документов, в которых есть или слово bush, или слово modes, или оба эти слова одновременно.

AND (операция "и")приводит к нахождению документов, в которых обязательно встречаются оба связанных этой операцией слова, но при этом они могут находиться на любом расстоянии друг от друга. В результате может оказаться, что в документе нет какой-либо логической связи между двумя заданными ключевыми словами. Например, в ответ на запрос

bush AND modes

поисковая система AltaVista выдает ссылки не только на соответствующие научные работы по нелинейной динамике, но и на отдельные главы известного фантастического романа Г. Уэльса "Борьба миров", которые, разумеется, не имеют никакого отношения к предмету поиска.

NOT — операция отрицания, позволяющая исключить те Web-документы, которые содержат ненужное словосочетание (таким образом, эта операция в сложном поиске играет ту же самую роль, что и знак "минус" в простом поиске).

NEAR (логическая операция, указывающая степень близости друг к другу ключевых слов) является одной из самых удобных команд расширенного поиска. В разных поисковых системах понятие близости слов разное. Например, AltaVista считает близкими слова, которые разделены не более чем десятью другими словами. Некоторые другие ПС допускают при записи операции NEAR указание степени близости в явном виде, то есть позволяют определить максимально допустимое число посторонних слов, которые могут стоять между двумя ключевыми словами, связанными этой операцией.

Запрос вида

bush NEAR modes

приведет к поиску документов, в которых два указанных ключевых слова стоят достаточно близко друг к другу, и стало быть, скорее всего связаны по смыслу. Например, в тексте Web-страницы может быть фраза: "Bushes of normal modes..".

Запрос

Жорес NEAR Алферов

приведет к поиску документов, в которых есть словосочетания Жорес Иванович Алферов, Алферов Жорес Иванович и т. д. Заметим, что при использовании операции NEAR порядок следования ключевых слов в документе указать невозможно, что, впрочем, является скорее не недостатком, а преимуществом, и это хорошо видно из только что приведенного запроса.

Как уже было сказано, с помощью вышеописанных логических операций можно задавать запрос на поиск в виде сложного логического выражения.

Несмотря на существование некоторых общих принципов организации поиска Web-документов, различные ПС могут весьма сильно отличаться друг от друга по предоставляемым клиенту возможностям, по внешнему виду своих заглавных страниц и т. д., в силу чего знакомиться с ними желательно при непосредственной работе в сети Internet.

В качестве примера ниже приведено очень краткое, схематическое описание двух наиболее популярных в настоящее время ПС Yahoo (тематический каталог) и AltaVista (автоматический индекс).

1. **Тематический каталог Yahoo**

Данная система появилась в Сети одной из первых, и сегодня Yahoo сотрудничает со многими производителями средств информационного поиска, а на различных ее серверах используется различное программное обеспечение. Вызов поисковой системы Yahoo можно осуществить с помощью адреса http://www.yahoo.com.

Заметим, что броузеры обычно допускают возможность не полного набора адреса. Например, в для вызова Yahoo можно набрать в поле адреса только слово Yahoo, а остальная часть адреса дописывается броузером автоматически.

Заглавная страница Yahoo предоставляет клиенту возможность обращаться к двум основным методам работы с этой поисковой системой — к поиску Web-документов по ключевым словам, поле набора которых находится слева от кнопки "Search" ("Поиск"), и к поиску с помощью иерархического дерева различных разделов (категорий), расположенного ниже.

Заметим, что и в том, и в другом случае поиск ссылки на нужный Web-документ происходит в одной и той же базе данных, но разными способами.

Открыв интересующий раздел тематического каталога щелчком мыши по его названию, увидим новую страницу Yahoo, которая имеет точно такую же структуру, как и заглавная страница: она содержит поле поиска по ключевым словам и список подразделов выбранного нами раздела, и т. д.

На заглавной странице Yahoo можно увидеть также рекламные объявления, различные дополнительные категории, например, новости, которые содержат, в частности, информацию о погоде и т. д.

Если общее число найденных Yahoo ссылок достаточно велико, то список этих ссылок разбивается на порции. По умолчанию такая порция содержит 20 ссылок, но это число можно изменить, если обратиться к странице опций поиска, кнопка запуска которой находится рядом с кнопкой поиска "Search". Там же можно переопределить роль пробела между ключевыми словами (считать ли его логической операцией OR или логической операцией AND), и т. д.

На странице выбора опций есть также переключатель, который позволяет включить или выключить режим поиска целых слов. В режиме Substrings Yahoo найдет все ссылки, в которых указанное нами слово стоит не только само по себе, но и является частью других слов, например, поиск слова "art" (искусство) сработает и на слове "department" (отдел, кафедра и т. д.). По умолчанию установлен режим Substrings, для того чтобы пользователь не был вынужден вводить каждое существительное в единственном и во множественном числе (т. е., с окончанием "-s"), и т. д. Напротив, в режиме Complete Words находятся заказанные нами ключевые слова только в том случае, если они ограничены с обеих сторон пробелами или знаками препинания.

Поисковая система Yahoo ищет ссылки на документы, соответствующие запросу, в нескольких различных базах данных, прежде всего это названия категорий и заголовки Web-сайтов. Если необходимые ссылки таким образом не были найдены, Yahoo автоматически организует поиск в базе данных индивидуальных Web-страниц. Клиент может заказать поиск в одной из этих баз данных щелчком мыши на кнопках, которые расположены ниже поля ввода ключевых слов. Результаты поиска ПС Yahoo упорядочивает в соответствии с несколькими критериями, которые определяют рейтинг данной ссылки. Более высокий рейтинг присваивается документам, в которых ключевые слова встречаются большее число раз, в которых они входят в заголовок или если ссылка соответствует более старшему разделу в иерархическом дереве категорий.

В конце страницы с результатами поиска Yahoo предлагает воспользоваться услугами нескольких других поисковых систем, если клиент не удовлетворен результатами этого поиска. Такие ссылки удобны тем, что система Yahoo, при обращении к ним, сама впишет наши ключевые слова в бланк запроса вызванной таким образом другой поисковой системы. Заметим в заключение, что Yahoo предоставляет около млн. ссылок на Web-страницы и это чуть более 0,1 % всего Web-пространства.

1. **Автоматический индекс Alta Vista**

Вызов этой поисковой системы можно осуществить по адресу: http://altavista.digital.com (для многих броузеров достаточно при этом набрать только слово altavista).

Alta Vista имеет одну из самых больших баз данных в классе автоматических индексов, и самые мощные и гибкие правила построения запросов. При этом Alta Vista имеет как систему простого поиска (simple search), так и систему расширенного поиска (advanced search), кнопка которой находится ниже поля ввода ключевых слов.

Предусмотрена возможность поиска информации и с помощью каталога категорий, устроенном аналогично каталогу системы Yahoo, причем, поиск по ключевым словам можно осуществить лишь в рамках определенной категории.

Поиск Web-документов по ключевым словам можно осуществлять в базе данных для WWW-страниц или в базе данных телеконференций (Usenet). Для выбора одной из этих баз данных — содержимого WWW или Usenet — служит выпадающий список, который можно вызвать с помощью кнопки, находящейся под панелью заголовка этой системы, на ее заглавной странице.

При использовании простого поиска системы Alta Vista необходимо учитывать его отличия от соответствующего средства ПС Yahoo. Действительно, по умолчанию Alta Vista, в отличие от Yahoo, ищет вхождения целых слов: заказанный термин должен стоять в Web-документе обособленно, а не быть частью других цепочек символов. Если же необходимо найти все вхождения данного ключевого слова, даже когда оно является частью других слов, необходимо использовать символ \* (этот символ может стоять только в конце ключевого слова и заменять не более пяти букв). Например, запрос вида Ада\* приведет к нахождению документов, в которых есть слова “Ада”, ”Адам”, ”Адажио” и т. д.

Поисковая система Alta Vista, как и большинство других поисковых систем, предлагает клиенту и ряд дополнительных возможностей, в частности, позволяет находить лишь те Web-документы, в которых заданные ключевые слова встречаются только в гипертекстовых ссылках, заголовках документов, в их URL-адресах (что дает возможность находить все Web-страницы, расположенные на сервере с данным адресом) и т. д. Для этого в Alta Vista используются специальные команды: Anchor, title и т. д. Такие возможности можно использовать как порознь, так и вместе друг с другом.

Бланк для расширенного поиска (advanced search) отличается от бланка простого запроса наличием двух полей. Второе из них — Results Ranking Criteria — полностью аналогично полю ввода бланка простого поиска, и в нем можно использовать те же специальные выражения, кавычки и знаки +, -, \*. Однако здесь это поле играет лишь вспомогательную роль, определяя порядок сортировки полученных результатов: документы, содержащие ключевые слова из поля Ranking будут стоять в списке первыми.

Ключевые же слова для собственно поиска должны вводиться в первом из полей — Selection Criteria. Запросы в этом поле строятся по описанным нами ранее правилам с использованием операций NEAR, NOT, AND, OR. Кроме того, внизу бланка поиска находится поле для ввода дат, позволяющих задать промежуток времени, в который были созданы или изменены интересующие нас документы.

Об особенностях поиска в Alta Vista документов на русском языке заметим, что искать документы по русским ключевым словам в каталоге Yahoo смысла не имеет, поскольку сотрудники этой службы на русском языке не говорят.

1. **Поиск информации в конференциях (Usenet)**

WWW-страницы являются самым масштабным и ценным, но не единствен-ным источником информации в Internet. Большой интерес представляет система телеконференций Usenet, в которой каждый день публикуется более 100 тысяч сообщений со всего мира.

Если клиент интересуется какой-либо определенной темой, он может просто подписаться на соответствующую группу телеконференций. Однако если он хочет не обсуждать свою проблему с живыми людьми, а просто выяснить, что известно по тому или иному вопросу, можно воспользоваться системами автоматического поиска в содержимом Usenet. Alta Vista ведет отдельный индекс по более чем 10 тысячам телеконференций, для поиска в котором нужно лишь перебросить переключатель поиска из значения "the Web" в значение "Usenet" (все правила составления запроса и установки опций в обоих этих случаях одинаковы). Результаты поиска представляют собой отсортированный список заголовков статей, полный текст любой из которых можно получить щелчком мыши по ее заголовку.

Во всем мире считается самым лучшим сервисом для поиска информации в конференциях Usenet сервер DejaNews ([http://www.dejanews.com](http://www.dejanews.com/)). Среди русскоязычных выделяется сервер RusNews ([http://news.corvis.ru](http://news.corvis.ru/)), являющийся также новостным сервером, содержащим более 2000 новостных групп. Также имеет место аналогичный российский сервер ТЕЛА-поиск (<http://tela.dux.ru/news.html>).

Совершенно уникальным источником информации являются документы с ответами на часто задаваемые вопросы (Frequently Asked Questions, FAQs). Такой жанр вопросов и ответов прекрасным образом вводит совершенно неподготовленного читателя в самую суть дела. Если первоначально списки ответов на вопросы существовали только для групп телеконференций, то сейчас документы с заголовком "FAQ" используются для клиентов самых разнообразных серверов и служб. В этом жанре пишутся статьи и обзоры, а фирмы и организации пропагандируют свои цели и устремления. Если клиенту нужно в сжатые сроки ознакомиться с новой областью науки, техники, культуры или политики, советуется начать с чтения раздела "FAQ" соответствующей телеконференции Usenet. И хотя таким образом нельзя приобрести слишком фундаментальные знания, зато можно быстро освоиться с терминологией, узнать о самых животрепещущих проблемах и нередко получить почти исчерпывающий список литературы.

**6. Отечественные поисковые службы**

Российский сектор Internet в настоящее время бурно развивается, и хотя отечественные поисковые системы еще отстают от соответствующих зарубежных систем по объему своих каталогов и индексов, по некоторым показателям (прежде всего, по применению новых технологий при обработке результатов поиска) они не уступают своим заграничным аналогам, а в ряде случаев и превосходят их.

Укажем, прежде всего, на автоматический индекс Aport 2000 и тематический каталог @Rus (Атрус), расположенных по адресам: http://www.aport.ru/ и http://www.atrus.ru/, соответственно. Эти две системы находятся в партнерских отношениях. Действительно, когда в результате проведенного поиска Aport 2000 выводит адрес некоторой Web-страницы, она может сопровождаться кратким описанием соответствующего Web-узла, взятым из каталога @Rus. С другой стороны, при наполнении каталога Атрус активно используются средства поисковой системы Aport. С точки зрения клиента, однако, эти поисковые системы являются все-таки независимыми средствами получения информации. Отметим некоторые их достоинства. Aport 2000 использует наиболее эффективную в настоящее время систему рейтингования по количеству ссылок, ведущих к данному ресурсу(по индексу цитирования). Поисковая система Атрус предоставляет клиенту удобный каталог-портал "Мой @Rus", который пользователь может настроить на быстрое получение наиболее нужной ему информации. Средства настройки этого каталога позволяют также отключить все лишнее на основной странице поисковой систе-мы, что делает работу с ним особенно быстрой и удобной.

**6.1. Rambler (**[**http://www.rambler.ru**](http://www.rambler.ru/)**)**

Поисковая система Rambler обладает одним из крупнейших индексов в Рос-сии, но основную популярность она приобрела в первую очередь как рейтинговая система. Она позволяет быстро выявить круг Web-узлов, поставляющих информацию на заданную тему, и оценить их популярность по количеству посещений разными клиентами Internet за последние сутки. Хотя число посещений данного Web-узла далеко не всегда свидетельствует об истинной ценности имеющейся на нем информации, в случае тем, представляющих общественный интерес, такому рейтингу популярности можно доверять.

По умолчанию находятся только те документы, в которых встретились все заданные нами ключевые слова, то есть пробел между словами воспринимается как логическая операция AND. Однако это значение пробела можно переопределить таким образом, чтобы он соответствовал логической операции OR (как это имеет место по умолчанию в Yahoo или Alta Vista). Для этого в бланке расширенного поиска нужно выбрать опцию “Слова запроса: любое”.

Чтобы исключить документы, содержащие те или иные слова, последние нужно указать на соответствующем поле бланка расширенного поиска.

**6.2. Яндекс (**[**http://www.yandex.ru**](http://www.yandex.ru)**)**

Поисковая система Яndex выделяется своими мощными средствами расши-ренного поиска, а также целым рядом технологических достижений, например, наличием интеллектуального механизма морфологического разбора слов, что особенно важно для русского языка. Независимо от того, в какой форме написали ключевое слово в запросе, Яndex будет учитывать все его формы. Например, если ключевым является слово идти, находятся ссылки на Web-документы, содержащие слова идти, идет, и даже шёл. Однако существует возможность поиска и по точной словоформе, для чего перед этой словоформой надо поставить восклицательный знак “!”.

Несколько набранных в запросе слов, разделенных пробелами, означает, что все они должны входить в одно предложение искомого документа (то есть пробел работает как знак логической операции AND).

Следует иметь ввиду, что в ПС Яndex операцию AND можно указать и в явной форме с помощью символа “&” (но не с помощью слова AND !). Удвоение же этого знака, т. е. использование символа “&&”, приводит к распространению действия операции AND на весь документ (т. е. связанные с помощью && слова должны обязательно присутствовать в пределах всего документа). Символом же операции OR в рассматриваемой поисковой системе служит знак “|” (но не само слово OR).

В ПС Яndex можно регулировать расстояние, на котором находятся друг от друга заданные ключевые слова в Web-документе. Например, запрос

физическое/(-2 4) образование

означает, что слово физическое может находиться как слева от слова образование (на расстоянии максимум двух слов от него), так и справа (на расстоянии максимум четырех слов от него).

Создавшая ПС Яndex компания CompTek бесплатно предоставляет для корпоративных клиентов (организаций) облегченную версию программы Яndex.Site, которая выполняет индексацию содержимого Web-узла. Это удобно владельцам тех Web-узлов, которые хотели бы организовать локальную систему для поиска информации в пределах своего собственного узла.

Поисковая система Яndex имеет очень хорошее описание в разделе “Помощь”, который настоятельно рекомендуется посмотреть перед использованием этой системы.

**7. Поиск файлов**

Для обмена файлами (это могут быть текстовые файлы или файлы с программами) в сети Internet существует специальный протокол FTP (File Transfer Protocol — протокол передачи файлов). Поскольку на многих серверах есть файлы, которыми их хозяева готовы поделиться с кем угодно, существует универсальное соглашение: зайдя почти на любой FTP-сервер, можно ввести вместо входного имени слово "anonymous", а вместо пароля — свой электронный адрес. Если у владельцев этого узла есть файлы, которые они распространяют без ограничений, мы получим к ним доступ. Такой способ обмена файлами называется "анонимным FTP". Программное обеспечение, которое можно получить таким образом, относится либо к бесплатным (freeware), либо к условно бесплатным (shareware) программам, или же представляет собой демонстрационные версии коммерческих программ с ограниченными возможностями.

Среди специальных систем поиска файлов в Интернете существуют аналоги уже рассмотренных ранее тематических каталогов (типа Yahoo) и автоматических индексов (типа Alta Vista). Разумеется, эти поисковые системы предоставляют клиенту не сами файлы, а лишь списки ссылок на них.

Одной из самых популярных поисковых служб типа тематических каталогов для поиска файлов является shareware.com по адресу http://www.shareware.com. Эта система классифицирует файлы только по одному признаку: для какой операционной системы они предназначены, но она хранит описания всех файлов, составленные людьми. Заглавная страница shareware.com устроена аналогично страницам уже рассмотренных нами поисковых систем. Она предлагает клиенту различные виды поиска, причем, ключевые слова могут содержать символ \*, соответствующий последовательности любых символов. Поиск производится как в именах, так и в описаниях файлов. Можно указать нижний временной порог поиска, чтобы получить ссылки на файлы, созданные не раньше заданной даты, и выбрать способ сортировки результатов—по дате или по алфавитному порядку имен файлов. Щелкнув на имени файла в списке результатов, получим ряд ссылок на узлы Internet, на которых хранятся копии этого файла с указанием надежности работы этих серверов и времени, необходимого для скачивания файла в зависимости от пропускной способности нашего канала.

В отличие от каталога shareware.com, поисковая система Archie является автоматическим индексом аналогичным Alta Vista. Список анонимных узлов для Archie приходится вести людям (его можно найти по адресу http://hoohoo.ncsa.uiuc.edu/ftp/). Поиск в базе данных сервера Archie производится с помощью ключевых слов, которые в данном случае представляют собой просто имена файлов или фрагменты имен.

Время ожидания связи с популярными FTP-узлами (такими, например, как богатое собрание программ для Windows (ftp://ftp.winsite.com)) может оказаться весьма большим, в силу чего в Интернете обычно имеются их точные копии или "зеркальные отражения" (mirrors), а поисковые системы выдают адреса всех этих зеркал.

Каталог, содержащий общедоступные файлы, почти всегда называется pub. В большинстве архивов в каждом каталоге имеется специальный файл с краткими—обычно не длиннее одной строки—описанием каждого файла этого каталога. Такой файл может называться 0index, 00index и т. п. (нолики приписываются к имени файла, чтобы он всегда попадал на первое место в отсортированном по алфавиту списке файлов).

**8. Источники специализированной информации**

Существуют информационные компании, обеспечивающие ин­формационное электронное обслуживание. Например, Knight-Ridder (KR) — это крупнейшая в мире информационная компания, предоставляющая доступ к своим службам в интерактивном режиме. Здесь объединились такие всемирно известные службы, как Dialog Information Service из США и Data Star из Европы. Используя Ин­тернет в качестве среды для распространения своих услуг, KR суме­ла автоматизировать ключевые моменты своей деятельности, улуч­шить обслуживание клиентов и, самое главное, расширить рынок своих услуг. Все это в конечном счете привело к тому, что KR стала мировым лидером в области доставки электронных документов и информационного сервиса.

Имеет собственную систему информационного поиска в Интер­нете фирма IBM — InfoMarket. В ней сочетаются средства получения информации и управления платежами с правами доступа.

Однако на начальной стадии освоения пространства Интернета не обойтись без самостоятельного поиска информации. Здесь весь­ма полезными могут оказаться адреса Web-узлов, хорошо зареко­мендовавшие себя источники деловой информации:

[www.kentis.com](http://www.kentis.com) — это сервер компании Kent Information System, сотрудники которой помогают профессионалам в области бухгалте­рии и финансов использовать компьютерные технологии, особенно Интернет;

<http://www.promotion.aha.ru> — on-line журнал по маркетингу в сети Интернет;

<http://www.inter.net.ru> — журнал «Интернет», регулярно публи­кующий материал по маркетингу и рекламе в сети Интернет;

<http://clickz.com> — журнал, полностью посвященный вопросам Интернет-маркетинга.

**9. Поиск отдельных людей в сети Internet**

Наиболее просто найти информацию о человеке, если у него имеется в сети Internet своя личная страница (personal home pages), на которой обычно имеется его фотография, e-mail и почтовый адрес, телефон и т. д. Нередко такая страница содержит краткую биографию автора, его хобби и т. д. Одна из крупнейших систем для поиска личных страниц называется Who’sWho и расположена по адресу http://web.city.ac.uk/citylive/pages.html.

Существуют также обширные каталоги электронных адресов различных людей, которые пользуются электронной почтой (e-mail). Лидерство здесь, по-видимому, принадлежит каталогу Four11 по адресу http://www.four11.com.

Отметим также каталог WhoWhere (http://www.whowhere.com), который производит поиск даже по схожести звучания или написания фамилий (например, “Kirsanov”, “Kirsanoff” и т. д.).

Отечественный каталог электронных адресов располагается по адресу http://www.botik.ru/~intermap/form.html.

Совершенно уникальным является всеамериканский адресный справочник по адресу http://www.databaseamerica.com, который выдает координаты любого из 90 млн. жителей США и любой из 10 млн. американских компаний.

**10. Метапоиск**

Базы данных разных поисковых систем в значительной мере не пересекаются. Поэтому для поиска достаточно редкой информации целесообразно обращаться не к одной, а к нескольким ПС. Однако правила оформления запросов для разных ПС, вообще говоря, отличаются друг от друга. Для того, чтобы не обращаться поочередно к разным поисковым системам и не думать о специфических правилах оформления запроса для каждой из них, были созданы так называемые метапоисковые системы.

Приняв заказ клиента, заданный с помощью ключевых слов в соответствие со своими собственными правилами его оформления, метапоисковая система сама пропишет его в бланках разных поисковых систем, разошлет эти бланки и будет ждать ответа. Когда все поисковые системы пришлют результаты поиска, метапоисковая программа сведет их в один документ и отправит пользователю. К таким метапоисковым системам относится MetaCrawler расположенный по адресу (http://metacrawler.cs.washington.edu:8080), который рассылает запрос на 9 различ-ных поисковых систем (в их число входит: Yahoo, Alta Vista, Lycos, Excite и т. д.). На случай разной интерпретации одних и тех же по смыслу опций в разных поисковых системах MetaCrawler предусматривает даже возможность проверки результатов поиска: прежде чем дать ссылку пользователю, он самостоятельно посмотрит на документ и проверит, соответствует ли он условиям запроса — так как их понимает MetaCrawler. Разумеется, этот режим проверки сильно задерживает получение результатов, но зато позволяет защититься как от неработоспособных ссылок, так и от бессмысленных результатов. Заметим, что на бланке запроса MetaCrawler можно задать время ожидания: в список будут включены только те результаты, которые успеют прийти с различных поисковых систем к этому моменту.

**11. Эффективный поиск информации**

После подробного изучения основных возможностей инструментов обратимся к проблеме эффективности поиска. Основными параметрами эффективности поиска являются:

- полнота поиска как отношение числа найденных документов к общему числу релевантных документов;

- точность поиска – отношение числа релевантных документов к общему числу полученных документов;

- актуальность ссылок на документы - существование найденных документов в сети в настоящий момент;

- скорость поиска.

В Internet существуют различные инструменты поиска, обладающие разными функциональными возможностями. Качество поиска, таким образом, зависит в первую очередь от параметров конкретной поисковой системы, например, от размеров индекса, от способа поиска (уточнение тем или поиск по запросу) и т.д. Далее, работая с конкретной поисковой системой, нужно иметь представление о методах составления запросов, знать необходимые операторы.

Таким образом, можно выделить следующие факторы, влияющие на эффективность поиска:

- свойства и возможности поисковой системы;

- качество формулировки запроса пользователем.

Теперь рассмотрим, как лучше подготовиться пользователю к составлению запроса. Прежде всего, необходимо провести всесторонний лексический анализ информации, которую необходимо найти. Затем желательно составить набор ключевых слов (при необходимости, на нескольких языках) в виде отдельных терминов и словосочетаний, специфичных для вашей предметной области.

Далее необходимо исследовать различные поисковые инструменты.

Действия:

- выбор поискового инструмента;

- точная формулировка запросов с использованием операторов, поддерживаемых данным поисковым инструментом;

- отправка тестовых запросов;

- анализ результатов поиска (по количеству и релевантности ссылок);

- при необходимости, корректировка запроса;

- повторный поиск;

- …

Исходя из вышеизложенного, можно выделить следующие приемы эффективного поиска:

- Поиск информации общего характера в поисковых системах-каталогах. - Поиск узкоспециальной информации в поисковых машинах. Для проведения более обширного поиска явно недостаточно использовать только системы-каталоги с ограниченным числом описанных ресурсов. Кроме того, узкоспециальная информация в каталогах может просто отсутствовать. Поэтому необходимо проводить поиск подобной информации в поисковых машинах, обладающих индексами большого объема.

- Использование операторов или бланка расширенного запроса для сужения области поиска. Для проведения качественного поиска необходимо ознакомиться с языком запросов конкретной поисковой машины. Эффективным и простым способом решения проблемы составления качественного запроса является использование режима расширенного поиска.

- Использование функции поиска среди найденных ресурсов. Большинство поисковых систем поддерживают возможность поиска внутри полученных результатов.

- Использование метапоисковых систем. Данные инструменты поиска отправляют ваш запрос сразу нескольким поисковым системам и от каждой системы получают несколько самых релевантных ссылок.

- Просмотр раздела Ссылки на специализированных сайтах. Авторы многих специализированных Web-узлов накапливают свои коллекции ссылок по тематике сайта.

- Поиск ответов на вопросы в группах новостей. При желании можно обратиться с конкретным вопросом о помощи в специализированную группу новостей.

- Подписка на специализированные списки рассылки и т.д.

## **Заключение**

В заключение, можно сказать, что единой оптимальной технологии поиска в Интернет не существует. В зависимости от специфики необходимой информации, для ее поиска должны использоваться соответствующие поисковые службы. Необходимо помнить, что чем грамотнее подобраны поисковые службы и составлен запрос на поиск информации, тем качественнее будут результаты поиска.

# **Литература**

1. Информатика и информационные технологии / Под ред. Романовой Ю.Д. М.: Эксмо, 2008
2. Степанов А.Н. Информатика. СПб.: Питер, 2002

## Майечак Б. Поиск информации в сети // Интернет для детей от 8 до 88. М.: Интерэксперт, 2002

1. Чечин Г. М., Положенцев Е. В., Нижникова С. В. Поиск информации в сети Internet. Ростов-на-Дону: РГУ, 2001 г.

5. Диканский Е.Ю. Осваиваем Internet: Практический курс по информационным и коммуникационным технологиям для начинающих пользователей. М.: Илекса, 2001

1. <http://www.dist-cons.ru/modules/searchinf/index.html>
2. <http://www.seonews.ru/masterclasses/detail/29812.php>
3. http://www.gdenet.ru/bibl/technology/transmission/5.1.html