#### Введение

Основные протоколы, используемые в Интернет (в дальнейшем также Сеть), не обеспечены достаточными встроенными функциями поиска, не говоря уже о миллионах серверах, находящихся в ней. Протокол HTTP, используемый в Интернет, хорош лишь в отношении навигации, которая рассматривается только как средство просмотра страниц, но не их поиска. То же самое относится и к протоколу FTP, который даже более примитивен, чем HTTP. Из-за быстрого роста информации, доступной в Сети, навигационные методы просмотра быстро достигают предела их функциональных возможностей, не говоря уже о пределе их эффективности. Не указывая конкретных цифр, можно сказать, что нужную информацию уже не представляется возможным получить сразу, так как в Сети сейчас находятся миллиарды документов и все они в распоряжении пользователей Интернет, к тому же сегодня их количество возрастает согласно экспоненциальной зависимости. Количество изменений, которым эта информация подвергнута, огромно и, самое главное, они произошли за очень короткий период времени. Основная проблема заключается в том, что единой полной функциональной системы обновления и занесения подобного объема информации, одновременно доступного всем пользователям Интернет во всем мире, никогда не было. Для того, чтобы структурировать информацию, накопленную в сети Интернет, и обеспечить ее пользователей удобными средствами поиска необходимых им данных, были созданы поисковые системы.

#### Поисковые системы

Поисковые cистемы обычно состоят из трех компонент:

* агент (паук или кроулер), который перемещается по Сети и собирает информацию;
* база данных, которая содержит всю информацию, собираемую пауками;
* поисковый механизм, который люди используют как интерфейс для взаимодействия с базой данных.

**2.1. Как работают механизмы поиска**

Cредства поиска и структурирования, иногда называемые поисковыми механизмами, используются для того, чтобы помочь людям найти информацию, в которой они нуждаются. Cредства поиска типа агентов, пауков, кроулеров и роботов используются для сбора информации о документах, находящихся в Сети Интернет. Это специальные программы, которые занимаются поиском страниц в Сети, извлекают гипертекстовые ссылки на этих страницах и автоматически индексируют информацию, которую они находят для построения базы данных. Каждый поисковый механизм имеет собственный набор правил, определяющих, как cобирать документы. Некоторые следуют за каждой ссылкой на каждой найденной странице и затем, в свою очередь, исследуют каждую ссылку на каждой из новых страниц, и так далее. Некоторые игнорируют ссылки, которые ведут к графическим и звуковым файлам, файлам мультипликации; другие игнорируют cсылки к ресурсам типа баз данных WAIS; другие проинструктирова-ны, что нужно просматривать прежде всего наиболее популярные страницы.

* Агенты - самые "интеллектуальные" из поисковых средств. Они могут делать больше, чем просто искать: они могут выполнять даже транзакции от Вашего имени. Уже сейчас они могут искать cайты специфической тематики и возвращать списки cайтов, отсортированных по их посещаемости. Агенты могут обрабатывать содержание документов, находить и индексировать другие виды ресурсов, не только страницы. Они могут также быть запрограммированы для извлечения информации из уже существующих баз данных. Независимо от информации, которую агенты индексируют, они передают ее обратно базе данных поискового механизма.
* Общий поиск информации в Сети осуществляют программы, известные как пауки. Пауки сообщают о содержании найденного документа, индексируют его и извлекают итоговую информацию. Также они просматривают заголовки, некоторые ссылки и посылают проиндексированную информацию базе данных поискового механизма.
* Кроулеры просматривают заголовки и возращают только первую ссылку.
* Роботы могут быть запрограммированы так, чтобы переходить по различным cсылкам различной глубины вложенности, выполнять индексацию и даже проверять ссылки в документе. Из-за их природы они могут застревать в циклах, поэтому, проходя по ссылкам, им нужны значительные ресурсы Сети. Однако, имеются методы, предназначенные для того, чтобы запретить роботам поиск по сайтам, владельцы которых не желают, чтобы они были проиндексированы.

Агенты извлекают и индексируют различные виды информации. Некоторые, например, индексируют каждое отдельное слово во встречающемся документе, в то время как другие индексируют только наиболее важных 100 слов в каждом, индексируют размер документа и число слов в нем, название, заголовки и подзаголовки и так далее. Вид построенного индекса определяет, какой поиск может быть сделан поисковым механизмом и как полученная информация будет интерпретирована.

Агенты могут также перемещаться по Интернет и находить информацию, после чего помещать ее в базу данных поискового механизма. Администраторы поисковых систем могут определить, какие сайты или типы сайтов агенты должны посетить и проиндексировать. Проиндексированная информация отсылается базе данных поискового механизма так же, как было описано выше.

Люди могут помещать информацию прямо в индекс, заполняя особую форму для того раздела, в который они хотели бы поместить свою информацию. Эти данные передаются базе данных.

Когда кто-либо хочет найти информацию, доступную в Интернет, он посещает страницу поисковой системы и заполняет форму, детализирующую информацию, которая ему необходима. Здесь могут использоваться ключевые слова, даты и другие критерии. Критерии в форме поиска должны соответствовать критериям, используемым агентами при индексации информации, которую они нашли при перемещении по Сети.

База данных отыскивает предмет запроса, основанный на информации, указанной в заполненной форме, и выводит соответствующие документы, подготовленные базой данных. Чтобы определить порядок, в котором список документов будет показан, база данных применяет алгоритм ранжирования. В идеальном случае, документы, наиболее релевантные пользовательскому запросу будут помещены первыми в списке. Различные поисковые системы используют различные алгоритмы ранжирования, однако основные принципы определения релевантности следующие:

1. Количество слов запроса в текстовом содержимом документа (т.е. в html-коде).
2. Тэги, в которых эти слова располагаются.
3. Местоположение искомых слов в документе.
4. Удельный вес слов, относительно которых определяется релевантность, в общем количестве слов документа.

База данных выводит ранжированный подобным образом список документов с HTML и возвращает его человеку, сделавшему запрос. Различные поисковые механизмы также выбирают различные способы показа полученного списка - некоторые показывают только ссылки; другие выводят cсылки c первыми несколькими предложениями, содержащимися в документе или заголовок документа вместе с ccылкой.

Когда Вы щелкаете на ссылке к одному из документов, который вас интересует, этот документ запрашивается у того сервера, на котором он находится.

**2.2 Сравнительный обзор поисковых систем**

**2.2.1 Поисковая система Рамблер**

Поисковая система содержит информацию о более чем 12 миллионах документов, расположенных на серверах России и стран СНГ. Рамблер обрабатывает ежесуточно не менее 500 тысяч поисковых запросов (в среднем - 5 запросов в секунду), сканируя 48 тысяч web-серверов и используя несколько одновременно работающих программ-роботов.

**Поисковые слова**

Запрос может состоять из одного или нескольких слов, разделенных пробелами. Могут быть использованы как русские, так и английские слова и словосочетания. По умолчанию находятся только те документы, в которых встретились все введенные Вами слова. Чтобы найти документы, содержащие хотя бы одно слово из запроса, используйте логическую связку **Or** (см. ниже) или выберите на странице детального запроса: "Слова запроса: любое". Чтобы исключить документы, содержащие те или иные слова, укажите на странице детального запроса: "Исключить документы, содержащие следующие слова ...".

**Регистр**

Все равно, с какой буквы написаны слова запроса: с большой или с маленькой. И при построении индекса, и при поиске по запросу все заглавные (большие) буквы "понижаются".

**Логические связки**

Слова запроса могут быть соединены логическими связками **And** ("и"), **Or** ("или"). Вместо связок (или в сочетании с ними) могут использоваться также символы **'&'**, **'|'**.

**Логические группы**

Части запроса могут быть сгруппированы с помощью круглых скобок ( ). Возможна многократная вложенность скобок в сочетании с логическими операторами.

**Словоформы**

Рамблер умеет искать слова во всех формах (например, аминокислота, аминокислоты, аминокислотой и т. д.). Чтобы слово находилось во всех формах, перед ним надо поставить служебный символ **'#'**. В меню детального запроса такой режим может быть включен для всех слов: "Расширение запроса: все формы слов". Служебный символ **'@'** перед словом позволяет находить не только само это слово, но и однокоренные слова. В меню детального запроса символу **'@'** соответствует режим "Расширение запроса: все однокоренные".

**Усечение слов**

По умолчанию наша система ищет слова запроса так, как Вы их ввели, чтобы уменьшить "шум" в найденных документах. Если Вы не помните, как пишется слово, или хотите расширить запрос, можно использовать метасимволы **'\*'** и **'?'** для обозначения произвольной части слова и произвольного символа.

**Поиск в части документа**

Ограничить поиск частями документов, такими как название документа, его заголовок, URL и т.п., можно через меню детального запроса "Искать в...".

**Язык документов**

Можно ограничить поиск документами только на русском или только на английском языке. Для этого надо выбрать соответствующий режим в меню детального запроса "Язык документа...". По умолчанию поиск выполняется по документам на всех языках.

**Сортировка результатов**

По умолчанию найденные документы сортируются по релевантности (соответствию запросу). Однако Вы можете потребовать, чтобы вместо этого в начало списка были помещены самые свежие (или, наоборот, самые старые документы). Для этого надо выбрать соответствующую установку в меню "Сортировать по..." на странице детального запроса.

Вы можете также ограничить поиск документами, созданными в определенный период времени: для этого необходимо на странице детального запроса указать "От даты ... до даты ...".

**Расстояние между словами**

Можно потребовать, чтобы Рамблер возвращал только те документы, где слова из запроса находятся на минимальном расстоянии друг от друга. Режим "Ограничить расстояние между словами" может быть включен в детальном запросе. Все перечисленные выше правила могут быть использованы совместно друг с другом в необходимой Вам последовательности.

**Выдача результатов**

По умолчанию результаты поиска выдаются порциями по 15 документов. Меню "Выдавать по..." на странице детального запроса позволяет увеличить это число до 30 или 50. Меню "Форма вывода..." позволяет получать описания документов с увеличенной или уменьшенной подробностью.

**2.2.2. Поисковая система Яндекс**

Яндекс ежедневно просматривает сотни тысяч Web-страниц в поисках изменений или новых ссылок. Коллекция ссылок постоянно растет.

Яндекс не требует от вас знания специальных команд для поиска. Достаточно набрать вопрос ("где найти дешевые компьютеры" или "нужны телефоны москвы и московской области"), и вы получите результат - список страниц, где встречаются эти слова.

Независимо от того, в какой форме вы употребили слово в запросе, поиск учитывает все его формы по правилам русского языка. Например, если задан запрос 'идти', то в результате поиска будут найдены ссылки на документы, содержащие слова 'идти', 'идет', 'шел', 'шла' и т.д. На запрос 'окно' будет выдана информация, содержащая и слово 'окон', а на запрос 'отзывали' - документы, содержащие слово 'отозвали'.

При этом поиск не ограничен лишь словами или фразами. Яндекс отыщет по названию web-страницу компании или файл с нужной картинкой.

**2.2.3. Поисковая система Апорт**

**Примеры простых запросов**

Обычно запрос представляет из себя просто одно или несколько слов, например:

*микропроцессоры компании Intel*

По такому запросу находятся документы, в которых встречаются все слова запроса. Есть, правда, ограниченное число слов (союзы, предлоги и т.п.), которые в запросе игнорируются, так как не несут сами по себе смысловой нагрузки. Например, по запросу:

*яблоки на снегу*

будут найдены все документы, в которых встречаются одновременно два слова: "яблоко" и "снег". Где в пределах документа расположены слова, в какой грамматической форме они находятся — не важно

Стоит еще раз подчеркнуть важное и очень полезное свойство Апорта: независимо от того, в какой грамматической форме вы пишите в запросе слово, оно находится в документах во всех своих формах. Например, по запросу:

*человек шел*

будут найдены среди прочих и документы, содержащие текст "люди идут". Распознавание всех форм работает для обычных слов русского языка. Для экзотических слов, неологизмов и т.п. оно не проходит. В этом случае может пригодиться оператор "\*" (звездочка). Например, вы хотите найти все, касающееся деятельности президента России, в том числе и документы, содержащие слово "ельцинизм". Воспользуйтесь запросом:

*ельцин\**

Он позволит вам найти то, что вы хотите (а также документы со словами Ельцин*ище*, ельцин*цы*, ельцин*енок* и т.п), поскольку звездочка заменяет собой любое число любых букв

**Поиск по адресам (по URL)**

Вы можете искать документы не только по всему русскоязычному Интернету, но и по его части. Самый простой случай — поиск по определенному серверу. Например

*url=www.intel.ru собака*

По данному запросу будут найдены все документы на сервере www.intel.ru, содержащие слово "собака". Возможно, вам интересно, а что будет, если написать просто:

*url=www.intel.ru*

В этом случае вы получите список всех документов, расположенных на указанном вами сервере

Вы можете ограничивать поиск и сильнее — одним из каталогов сервера. Например:

*url=www.intel.ru/sobaki/ сенбернар*

По данному запросу документы, содержащие слово "сенбернар", будут искаться только в каталоге /sobaki (и его подкаталогах) московского сервера корпорации Intel.

**2.2.4. List.ru**

На главной странице в верхней ее части расположены ссылки на наиболее популярные проекты. Ниже, под логотипом каталога, поисковая форма. В правой колонке и нижней части страницы - блоки самых актуальных новостей. Список ссылок на основные категории каталога занимает центральную часть. Цифры рядом с названием категории показывают количество сайтов, содержащихся в ней. Записанные мелким шрифтом заголовки отсылают при нажатиии на подкатегории раздела. Щелкнув по любому из названий, Вы попадете в соответствующую рубрику и под логотипом List.ru увидите полный путь до нее, начиная с главной страницы. Каждый промежуточный уровень структуры доступен по отдельной ссылке.Поиск в каталоге реализован таким образом, что в результате запроса могут быть найдены как отдельные сайты, так и рубрики. Поиск будет успешным URL, название, описание, ключевые слова. Допускается использованием языка запросов Яндекс. Расположенная рядом с поисковой формой ссылка "Структура каталога" открывает в отдельном окне полный рубрикатор каталога. Реализована возможность перехода из рубрикатора в любую выбранную подкатегорию. Более детальное тематическое деление текущей рубрики представлено списком ссылок. Помеченные символом @ приведут в подкатегории, структурно входящие в другие разделы, но содержащие близкую по содержанию информацию. Если Вы хорошо представляете, в какой рубрике содержатся нужные ресурсы, лучше перейти в нужную подкатегорию, воспользовавшись любым из предоставляемых средств навигации по каталогу. В противном случае можно искать их в полном списке. Каталог организован таким образом, что все сайты, содержащиеся на нижних уровнях структуры, представлены и в рубриках.Показываемый ниже список ресурсов упорядочен в алфавитном порядке, но, выбрав соответствующую сортировку ("По времени добавления" или "По переходам"), можно просмотреть их по порядку добавления в каталог (начиная с самых "свежих") или в зависимости от популярности среди посетителей каталога. Ссылка с названия сайта открывает в отдельном окне его зарегистрированную в данной рубрике страницу. Пометки RUS и ENG означают наличие на сайте страниц, соответственно на русском и английском языках. В красных скобках (Заходов: ххх.хх) показано среднесуточное количество переходов на ресурс со страниц List.ru.

**2.2.5. Поисковая система AltaVista**

Индексирование в этой системе осуществляется при помощи робота. При этом робот имеет следующие приоритеты:

* слова содержащиеся в теге <title> имеют высший приоритет; ключевые фразы в <Meta> тэгах;
* ключевые фразы, находящиеся в начале странички;
* ключевые фразы в ALT - ссылках
* ключевые фразы по количеству вхождений\присутствия слов\фраз;

Если тэгов на странице нет, использует первые 30 слов, которые индексирует и показывает вместо описания (tag description)

Наиболее интересная возможность AltaVista - это расширенный поиск. Здесь стоит сразу оговориться, что, в отличие от многих других систем AltaVista поддерживает одноместный оператор NOT. Кроме этого, имеется еще и оператор NEAR, который реализует возможность контекстного поиска, когда термины должны располагаться рядом в тексте документа. AltaVista разрешает поиск по ключевым фразам, при этом она имеет довольно большой фразеологический словарь. Кроме всего прочего, при поиске в AltaVista можно задать имя поля, где должно встретиться слово: гипертекстовая ссылка, applet, название образа, заголовок и ряд других полей. К сожалению, подробно процедура ранжирования в документации по системе не описана, но видно, что ранжирование применяется как при простом поиске, так и при расширенном запросе. Реально эту систему можно отнести к системе с расширенным булевым поиском.

**3.Примеры запросов**

В системе Апорт был сделан запрос «вакуум-испарительный кристаллизатор».Система сообщила следующие результаты (первые три):

Апорт

1

Тамбовский государственный технический университет

http://www.tstu.ru

О университете - история; список факультетов, специальностей; персоналии ведущих ученых и др. Рассказ о Тамбовском региональном центре информатизации высшей школы: устав, перспективные проекты, результаты. О Тамбове и области: история, культура, ссылки на страницы образовательных учреждений, туризм.

Похожие сайты 36 0000

http://www.tstu.ru/win/katalog/okp/36.htm, 04.11.1998 (проверено 08.12.2000), 104 kb, Windows 1251 ... 36 1428 4 - комбинированные 36 1429 5 - прочие 36 1430 5 Кристаллизаторы / 36 1431 0 - с вращающимися барабанами 36 1432 6 - емкостные 36 1433 ... сосудов и аппаратов 36 1600 9 Фильтры жидкостные 36 1610 3 Вакуум-фильтры / 36 1611 9 - барабанные наружные 36 1612 4 - барабанные наружные ...

Реконструкция текста

Остальные найденные на сайте страницы (1)

2

Институт тепло- и массообмена им. А.В. Лыкова АН Беларуси

http://www.itmo.by

Информация об институте и его лабораториях. Новости и события. Анонсы проводимых конференций и семинаров. Материалы "Инженерно-физического журнала".

Похожие сайты IFZh, 1984

http://www.itmo.by/jepter/SODER/461984r/contr46.html, (проверено 10.11.2000), 41 kb, koi8-r ... Майоров В. А. , Васильев Л. Л. Физические особенности испарительного жидкостного охлаждения пористого металлокерамического тепловыделяющего элемента ... ...

... Бодров В. В. Определение тепловых параметров тонкостенных кристаллизаторов барабанного типа ...622 Новиков Н. П. , Игнатенко К. И. , Михайлова Г. ... примеси в кремнии n типа при высокотемпературном нагреве в вакууме ...693 Булгак И. А. , Скоропанов А. С. , Вечер А. А. , Проценко Г. ... ...

... А. А. Эффективность геттерирования при массопереносе в вакууме ...949 Агеенко И. С. , Ильясов С. Г. , Красников В. В. , Тюрев Е. ...

Реконструкция текста

Остальные найденные на сайте страницы (1)

3

WWW server of Tambov state technical university

http://www.tixm.tambov.su

36 0000

http://www.tixm.tambov.su/win/katalog/okp/36.htm, 04.11.1998 (проверено 18.12.2000), 104 kb, Windows 1251 ... 36 1428 4 - комбинированные 36 1429 5 - прочие 36 1430 5 Кристаллизаторы / 36 1431 0 - с вращающимися барабанами 36 1432 6 - емкостные 36 1433 ... сосудов и аппаратов 36 1600 9 Фильтры жидкостные 36 1610 3 Вакуум-фильтры / 36 1611 9 - барабанные наружные 36 1612 4 - барабанные наружные ...

Реконструкция текста

Остальные найденные на сайте страницы (1)

В системе Яндекс был сделан то т же запрос, система ответила, что ничего не найдено.

#### 3. Список использованной литературы

Павел Храмцов "Поиск и навигация в Internet".  
http://www.osp.ru/cw/1996/20/31.htm

Обучение Интернет-профессиям. Search engine Expert.  
http://searchengine.narod.ru/archiv/se\_2\_250500.htm

Андрей Аликберов "Несколько слов о том, как работают роботы поисковых машин".  
http://www.citforum.ru/internet/search/art\_1.shtml