Занятие 15. Лекция

«Поисковые службы Интернет. Поисковые серверы WWW.»

План лекции:

- Выписать определения
- Виды поисковых систем и поисковых запросов
- Схема поиска в сети

1. Поиск информации

Конец XX - начало XXI века, характеризуется огромными массивами постоянно растущей разнообразной информации, доступной и представляющей интерес для самых широких слоев социума. Более того, Интернет-технологии и программно-технические средства, также доступные большинству людей, позволяют осуществлять данный процесс в любое время, практически в любом месте по любым запросам.

Поиск - процесс, в ходе которого в той или иной последовательности производится соотнесение отыскиваемого с каждым объектом, хранящимся в массиве.

Цель любого поиска заключается в потребности, необходимости или желании находить различные виды информации, способствующие получению лицом, осуществляющим поиск, нужных ему сведений, знаний и т.д. для повышения собственного профессионального, культурного и любого иного уровня; создания новой информации и формирования новых знаний; принятия управленческих решений и т.п.

Существуют различные толкования термина "поиск информации" или "информационный поиск".

Термин "**информационный поиск**" (англ. "information retrieval") ввёл американский математик К. Муэрс. Он заметил, что побудительной причиной такого поиска является *информационная потребность*, выраженная в форме информационного запроса. К объектам информационного поиска К. Муэрс отнес документы, сведения об их наличии и (или) местонахождении, фактографическую информацию.

ИП различают следующим образом:

- в зависимости от цели - адресный (формально-механический) и семантический (тематический);
- от объекта поиска - документный и фактографический;
- от степени использования технических средств - ручной или автоматизированный.
- в зависимости от функциональной роли - доминирующие/второстепенные, центральные/периферические, устойчивые/ситуативные потребности.

Решать проблемы фактографического поиска первыми стали представители библиотек. Они разработали средства информационного поиска, получившие название "*справочно-поисковый аппарат*" (каталоги, библиографические указатели и др.). В профессиональной отечественной печати данный термин используется с 1970-х годов. Библиотекари определяют "***информационный поиск***" как нахождение в информационном *массиве документов*, соответствующих *информационному запросу пользователей*.

С точки зрения использования компьютерной техники "***информационный поиск***" - совокупность логических и технических операций, имеющих конечной целью нахождение документов, сведений о них, фактов, данных, релевантных запросу потребителя.

ИП производится при помощи **информационно-поисковых систем** (ИПС).

ИПС - это комплекс связанных друг с другом отдельных частей, предназначенный для выявления в каком-либо множестве элементов информации, отвечающих на предъявленный информационный запрос. Массив элементов информации, в котором производится ИП, называется **поисковым массивом**.

Наиболее эффективный метод поиска документов, содержащих научную информацию - прочитать каждый документ некоторой библиотеки. Но такой способ практически неосуществим, поскольку число документов обычно бывает слишком большим, чтобы все их можно было прочитывать при каждом информационном запросе. Поэтому приходится использовать другой, менее эффективный метод, при котором ИП производится не по самим текстам документов, а по кратким характеристикам содержания или определенным внешним признакам документов. Для этого каждый документ снабжается поисковым образом документа (ПОД) - характеристикой, в которой кратко выражается основное смысловое содержание документа. В виде такой же краткой характеристики - поискового предписания или поискового образа запроса (ПОЗ) - должен быть сформулирован и информационный запрос. Благодаря этому процедура ИП может быть сведена к простому сопоставлению ПОД с заданным ПОЗ. Если ПОД в необходимой и достаточной степени совпадает с ПОЗ, считается, что этот документ отвечает на информационный запрос. Такое сопоставление оправдано лишь тогда, когда поисковый образ и поисковое предписание формулируются в терминах одного и того же языка, и притом такого, в котором каждая фраза допускает одно и только одно толкование.

Существует три основных типа информационно-поисковых задач:

• ретроспективный информационный поиск, т.е. отыскание письменных документов (всех или части), в которых содержатся сведения по определенному вопросу;

• срочное оповещение отдельных специалистов (абонентов) о публикациях, представляющих для них потенциальный интерес. Данный тип информационного поиска называется избирательным (адресным) распределением информации (ИРИ). Он производится по постоянным информационным запросам (так называемым «профилям интересов»), которые формулируются самими потребителями. Это особый случай ИП;

• поиск имен специалистов, располагающих информацией по определенному вопросу.

Понятие и функции поисковой системы

Поисковая система - это программно-аппаратный комплекс, предназначенный для осуществления поиска в сети Интернет и реагирующий на запрос пользователя, задаваемый в виде текстовой фразы (поискового запроса), выдачей списка ссылок на источники информации, в порядке релевантности (в соответствии запросу).

Наиболее крупные международные поисковые системы: «Google», «Yahoo», «MSN». В русском Интернете это – «Яндекс», «Рамблер», «Апорт».

Рассмотрим подробнее понятие поискового запроса на примере поисковой системы «Яндекс». Поисковый запрос должен быть сформулирован пользователем в соответствии с тем, что он хочет найти, максимально кратко и просто.

Поисковый запрос (или поисковая фраза) — чаще всего это слово, словосочетание или целое предложение, которое вводит посетитель поисковой системы при обращении к ней. Запрос может содержать проблему, название товара или услуги, вопрос, информацию о которых посетитель хочет получить от поисковой системы.

Поисковые системы делятся на следующие виды:

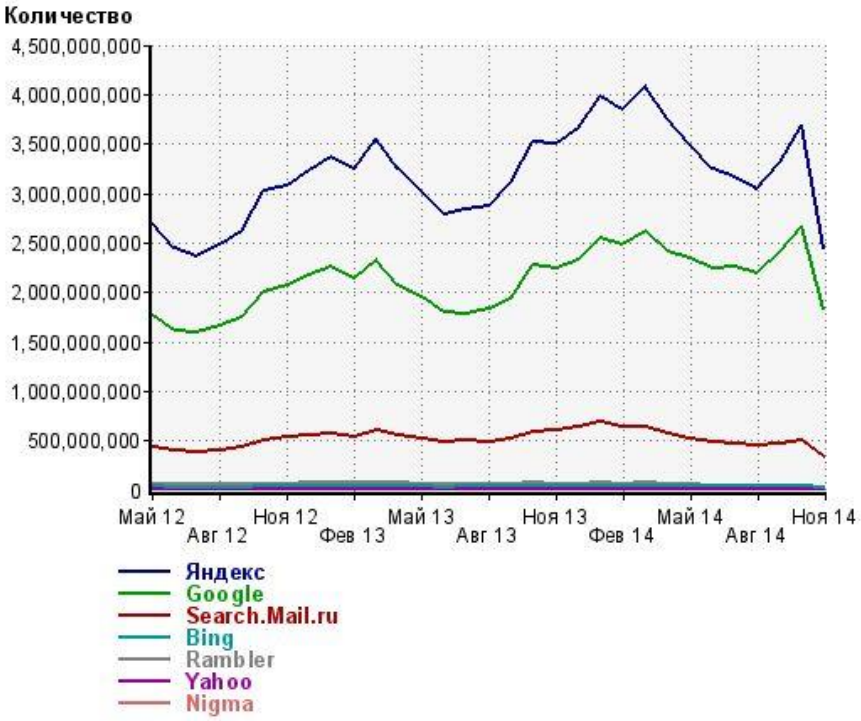
- **Национальные поисковые системы**

Поисковые системы разрабатываемые изначально для поиска сайтов внутри конкретной страны, т.е. для внутреннего рынка. Большинство из них постепенно вышли за рамки свого государства, но при этом не перешли в разряд транснациональных.

Пример национальных поисковых систем: Yandex (rus), Mail.ru (rus), Спутник (государственная поисковая система в России), Cade (br), Alcanseek (cn), Alexa (us), Anzwers (au), ...

- **Транснациональные поисковые системы**

Поисковые системы, осуществляющие поиск ответа на запрос пользователя по сайтам всех стран, независимо от их доменной зоны и страны нахождения.



Виды поисковых запросов

- **Информационные запросы**
  Цель подобных запросов — найти информацию о товаре, компании, событии.
  Например: «что такое кукумбер», «как лечить больное горло», «поисковая система», «самые богатые люди мира», ...
- **Транзакционные запросы**
  Цель запросов — совершить какое-либо действие, например: купить, заказать, скачать, зарегистрироваться и пр., т.е. поиску подвергается сайт, на котором это действие можно совершить. *Например:* «тест-драйв ниссан мурано», «купить детский велосипед», «доставка пиццы», ...
- **Навигационные запросы**
  Цель запроса — найти вполне конкретный сайт.
  Например: «сайт дом 2», «в контакте», «ургу», ...
- **Общие запросы**
  Запросы вида: «холодильник», «тойота», «детская одежда» и пр., которые по сути являются очень общими и не содержат конкретики. Подобные запросы чаще всего задают люди, находящиеся на самой ранней стадии готовности к покупке, т.е. когда они только начинают изучать предметную область.

Первоочередная задача любой поисковой системы – доставлять людям именно ту информацию, которую они ищут.

Основные характеристики поисковой системы

- *Полнота*

Полнота - одна из основных характеристик поисковой системы, представляющая собой отношение количества найденных по запросу документов к общему числу документов в сети Интернет, удовлетворяющих данному запросу. К примеру, если в Интернете имеется 100 страниц, содержащих словосочетание «как выбрать автомобиль», а по соответствующему запросу было найдено всего 60 из них, то полнота поиска будет 0,6.

- *Точность*

Точность - еще одна основная характеристика поисковой машины, которая определяется степенью соответствия найденных документов запросу пользователя. Например, если по запросу «как выбрать автомобиль» находится 100 документов, в 50 из них содержится словосочетание «как выбрать автомобиль», а в остальных просто наличествуют эти слова («как правильно выбрать магнитолу и установить в автомобиль»), то точность поиска считается равной 50/100 (=0,5).

- *Актуальность*

Актуальность - не менее важная составляющая поиска, которая характеризуется временем, проходящим с момента публикации документов в сети Интернет, до занесения их в индексную базу поисковой системы.

- *Скорость поиска*

Скорость поиска тесно связана с его устойчивостью к нагрузкам. Например, по данным ООО «Рамблер Интернет Холдинг», на сегодняшний день в рабочие часы к поисковой машине Рамблер приходит около 60 запросов в секунду. Такая загруженность требует сокращения времени обработки отдельного запроса. Здесь интересы пользователя и поисковой системы совпадают: посетитель желает получить результаты как можно быстрее, а поисковая машина должна отрабатывать запрос максимально оперативно, чтобы не тормозить вычисление следующих запросов.

- *Наглядность*

Наглядность представления результатов является важным компонентом удобного поиска. По большинству запросов поисковая машина находит сотни, а то и тысячи документов. Вследствие нечеткости составления запросов или неточности поиска, даже первые страницы выдачи не всегда содержат только нужную информацию. Это означает, что пользователю зачастую приходится производить свой собственный поиск внутри найденного списка.

2. Краткая история развития поисковых систем

В начальный период развития Интернет, число его пользователей было невелико, а объем доступной информации сравнительно небольшим. В большинстве своем, доступ к сети Интернет имели лишь сотрудники

научно-исследовательской сферы. В это время задача поиска информации в Интернете не была столь актуальной, как в настоящее время.

Одним из первых способов организации доступа к информационным ресурсам сети стало создание открытых каталогов сайтов, ссылки на ресурсы в которых группировались согласно тематике. Первым таким проектом стал сайт Yahoo.com, открывшийся весной **1994 года**. После того, как количество сайтов в каталоге Yahoo значительно увеличилось, была добавлена возможность поиска нужной информации по каталогу. В полном смысле это еще не было поисковой системой, так как поисковая область была ограничена только ресурсами, присутствующими в каталоге, а не всеми Интернет ресурсами.

Каталоги ссылок широко использовались ранее, однако практически полностью утратили свою популярность в настоящее время. Так как даже современные, огромные по своему объему каталоги, содержат информацию лишь о ничтожно малой части сети Интернет. Самый большой каталог сети DMOZ (его еще называют Open Directory Project) содержит информацию о 5 миллионах ресурсов, тогда как база поисковой системы Google состоит из более чем 8 миллиардов документов.

Первой полноценной поисковой системой стал проект WebCrawler, вышедший в свет в 1994 году.

В 1995 году появились поисковые системы Lycos и AltaVista. Последняя долгие годы была лидером в области поиска информации в сети Интернет.

В 1997 году Сергей Брин и Ларри Пейдж создали поисковую машину Google в рамках исследовательского проекта вСтэндфордском университете. В настоящий момент Google - самая популярная поисковая система в мире!

. В сентябре 1997 года была официально анонсирована поисковая система Yandex, являющаяся самой популярной в русскоязычном Интернете

В настоящее время существуют три основные поисковые системы (международные) – Google, Yahoo и MSN, имеющие собственные базы и алгоритмы поиска. Большинство остальных поисковых систем (коих насчитывается большое количество) использует в том или ином виде результаты трех перечисленных. Например, поиск AOL (search.aol.com) использует базу Google, а AltaVista, Lycos и AllTheWeb – базу Yahoo.

3. Интернет-поисковые системы

Для получения информации в среде Интернета создаются специальные поисковые системы. Как правило, они общедоступны и обслуживают пользователей в любой точке планеты, где имеется возможность работы с Интернетом. Непосредственно для поиска используются поисковые машины, число которых в мире исчисляется несколькими сотнями. Они ориентируются на определенные типы запросов или их сочетание (библиографический, адресный, фактографический, тематический и др.). Кроме того, бывают полнотекстовые, смешанные и другие поисковые машины.

Для проведения поиска в Интернете (в WWW) функционирует множество сайтов и поисковых систем, поэтому необходимо не только ориентироваться в таких системах, но и уметь осуществлять в них эффективный поиск, то есть использовать соответствующие технологии.

"**Технология поиска** (англ. "Search Technology") означает совокупность правил и процедур.

Поисковые системы характеризуются также временем выполнения поиска, интерфейсом, предоставляемым пользователю и видом отображаемых результатов.

При выборе поисковых систем обращают внимание на такие их параметры, как охват и глубина. Под *охватом* понимается объём базы поисковой машины, измеряемый тремя показателями: общим объёмом проиндексированной информации, количеством уникальных серверов и количеством уникальных документов. Под глубиной понимается - существует ли ограничение на количество страниц или на глубину вложенности директорий на одном сервере.

Каждая поисковая машина имеет свои алгоритмы сортировки результатов поиска. Чем ближе к началу списка, полученного в результате проведения поиска, оказывается нужный документ, тем выше релевантность и лучше работает поисковая машина.

Поиск информации в интернет

Эффективный доступ к информации в Интернете обеспечивают такие ***зарубежные поисковые системы*** (машины), как Альта-Виста (AltaVista), "Lycos", "Yahoo", "Google", "OpenText", "Wais", "WebCrawler" и др. Их адреса в Интернете: www.altavista.com, www.yahoo.com, www.gogle.com, www.opentext.com,

К отечественным поисковым машинам относятся: Апорт ("Aport" АО Агама), Rambler (фирма Stack Ltd.), Яндех ("Yandex" фирма CompTek Int), "Русская машина поиска", "Новый русский поиск", и др. Их адреса в Интернете: www.aport.ru, www.rambler.ru, www.yandex.ru, search.interrussia.com, www.openweb.ru соответственно) и др.

Все эти поисковые машины позволяют по ключевым словам, тематическим рубрикам и даже отдельным буквам оперативно находить в сети, например, все или почти все тексты, где эти слова присутствуют. При этом пользователю сообщаются адреса сайтов, где найденные ИР постоянно присутствуют. Однако ни одна из них не имеет подавляющих преимуществ перед другими. Для проведения надежного поиска по сложным запросам специалисты рекомендуют использовать последовательно или параллельно (одновременно) различные ИПС.

Полнотекстовая поисковая машина индексирует все слова видимого пользователю текста. Наличие морфологии дает возможность находить искомые слова во всех склонениях или спряжениях. Кроме этого, в языке HTML существуют тэги, которые также могут обрабатываться поисковой машиной (заголовки, ссылки, подписи к картинкам и т.д.). Некоторые машины умеют искать словосочетания или слова на заданном расстоянии, что часто бывает важно для получения разумного результата.

При проведении поиска поисковые серверы обычно используют данные, хранящиеся в веб-страницах в тегах метаданных: (title), (meta name=”keywords”) и (meta name=”description”). Формируя свои страницы, следует отражать в этих тегах сведения о назначении сайта и его тематике.

Полноту и точность ответа пользователь получает в зависимости от точности сформулированного им запроса. В результате поиска ему обычно предоставляется гораздо больше информации, чем ему необходимо, часть которой может вообще не иметь отношение к сформированному запросу. Легко заметить, что многое зависит не только от грамотно сформулированного запроса, но и от возможностей поисковых систем, которые весьма различны. При этом достаточно ярко проявляется "*лесной синдром*" (из-за леса не видно дров), заключающийся в том, что в полученных данных можно пропустить главные, необходимые сведения. Очевидно, никакие меры не являются исчерпывающими в условиях постоянного расширения среды и появления новых разнообразных ИР, что подтверждает трудности поиска в WWW.

Простые запросы в виде отдельных достаточно распространенных терминов приводят к извлечению тысяч (сотен тысяч) документов, абсолютное большинство которых пользователю не требуется (*информационный шум*).

4. Поиск в сети

Приемы работы, используемые при работе с теми или другими поисковыми инструментами, практически одинаковы. Перед тем как перейти к их обсуждению, рассмотрим следующие понятия:

1. Интерфейс поискового инструмента представлен в виде страницы с гиперссылками, строкой подачи запроса (строкой поиска) и инструментами активизации запроса.
2. **Индекс поисковой системы** – это информационная база, содержащая результат анализа веб-страниц, составленная по определенным правилам.
3. **Запрос** – это ключевое слово или фраза, которую вводит пользователь в строку поиска. Для формирования различных запросов используются специальные символы ("", , ~), математические символы (*, +, ?).

Схема поиска информации в сети