Донецкий национальный технический университет
Факультет компьютерных наук и технологий

Кафедра компьютерной инженерии

Курс «Интернет-технологии»

Лекция 3

Поисковые системы

Цололо С.А.,
к.т.н., доцент кафедры
компьютерной инженерии

Донецк, 2013

1 История веб-поиска. Предпосылки появления

Основные протоколы, используемые в Интернете, **не обеспечены достаточными встроенными функциями поиска**, не говоря уже о миллионах серверах

Протокол **HTTP 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.

Основная проблема заключается в том, что е**диной полной системы обновления и занесения** всего объема информации **никогда не было**.

2 История веб-поиска. Archie

Самый первый поисковый инструмент интернета назывался **Archie** (название произошло от искаженного слова archive).

Он был **создан в 1990 году** Аланом Эмтаджем, студентом Монреальского Университета.

Программа **скачивала списки файлов**, расположенные на публичных анонимных **FTP-сайтах**, создавая базы данных имен файлов, по которым можно было производить поиск.

3 История веб-поиска. World Wide Web Wanderer

В 1993 студент Мэтью Грей разработал **первого робота, который индексировал страницы интернета** – WWW Wanderer.

Первоначально программа позволяла пересчитывать веб-сервера, **измеряя масштабы веб-паутины**.

Wanderer **запускали ежемесячно в 1993-1995 гг.**

Позже Wanderer был использован для получения адресов ресурсов при формировании **первой базы данных веба**, который был назван Wandex (по мотивам «index»).

История веб-поиска. ALIWEB

В 1993 году Мартин Костер создал ALIWEB. Система позволяла **владельцам сайтов подавать заявки на индексацию** в поисковых машинах.

Фактически, ALIWEB был поисковой системой, основанной на автоматизированном сборе мета-данных для веба.

Финансирование поисковых систем становится прибыльным бизнесом. Инвесторы сочли, что **из интернета можно извлекать выгоду**, началось финансирование развития поисковых машин.

Разработка поисковиков стала прибыльным бизнесом.

История веб-поиска. Excite

В 1993 году шесть студентов Стэнфорда представили **Excite**.

Программа использовала статистический анализ слов в тексте, чтобы облегчить процесс поиска. В течение года Excite был усовершенствован и **вышел онлайн в декабре 1995 года**.

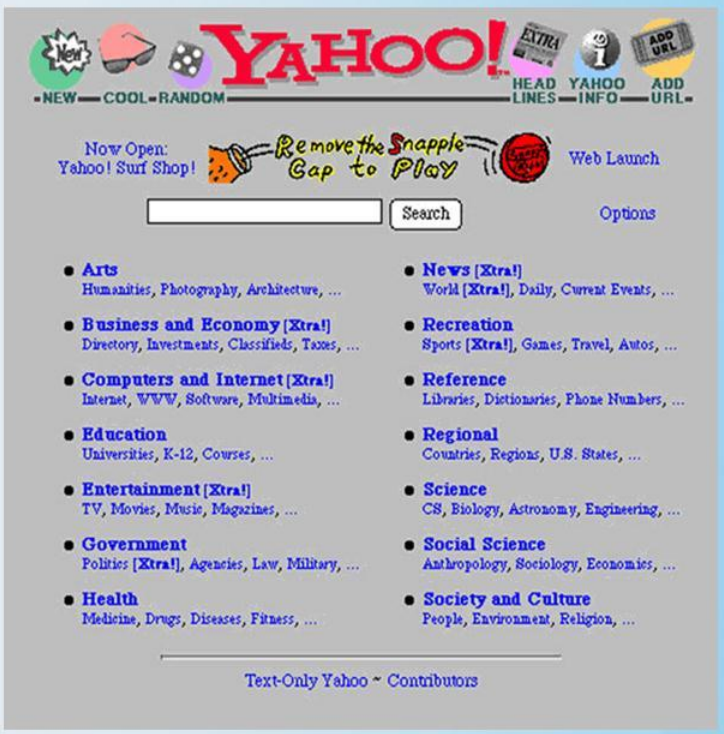
В 1999 году два аспиранта Стэнфордского университета, **Сергей Брин и Ларри Пейдж** пришли к руководству Excite и предложили купить разработанный ими поисковик **Google за $1 млн**, но получили отказ.

Это решение впоследствии называли **одной из крупнейших ошибок,** которые когда либо были сделаны в цифровой индустрии.

История веб-поиска. Yahoo!

Джерри Янг и Дэвид Фило создали Yahoo **в 1994 году.**

Проект начался с составления каталога их любимых веб-сайтов. Единственное, что отличало этот перечень от других, был **комментарий к каждой ссылке URL**.



Через год разработчики получили финансирование и создали **корпорацию Yahoo!**

К тому времени Yahoo уже был зарегистрированной торговой маркой соуса для барбекю, поэтому к названию был добавлен восклицательный знак.

7 История веб-поиска. Lycos

В 1994 году Lycos представил поисковую машину, **предлагающую наряду с результатами поиска ссылки на темы**, связанные с поисковым запросом.

В 1996 году это уже была **обширная поисковая система**, индексирующая более 60 миллионов документов, **самая крупная на тот момент**.

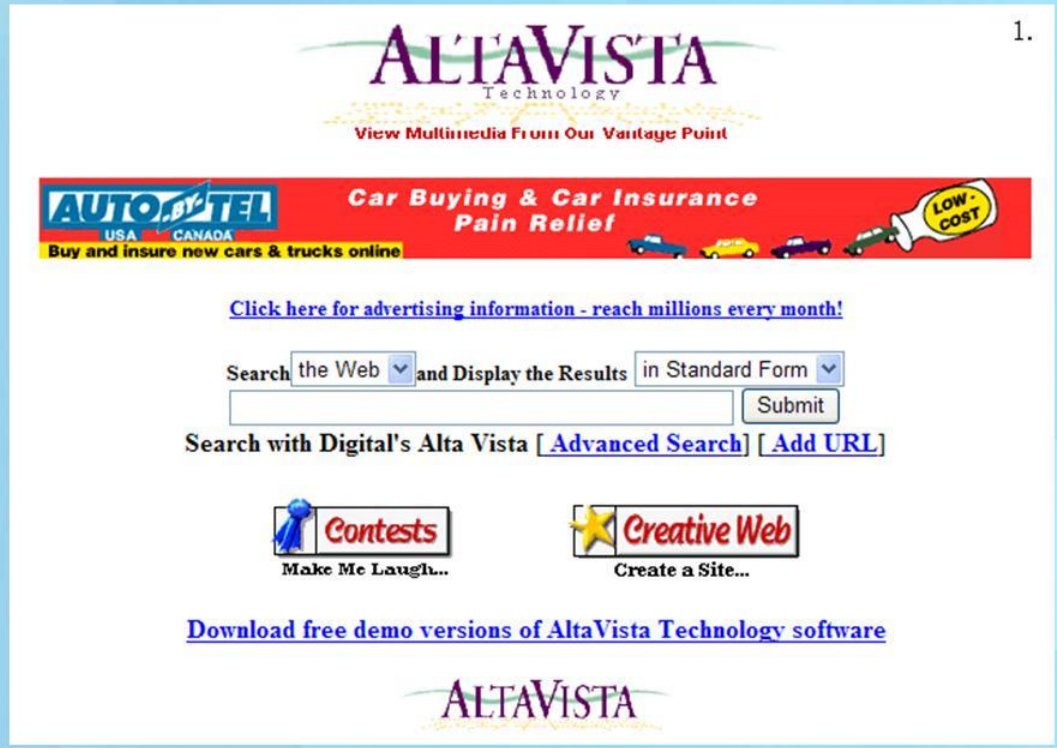
Позже компания вышла на IPO и превратилась в **один из первых в мире бизнес-проектов в интернете, приносивших доход**.

История веб-поиска. AltaVista

AltaVista начала работать **в 1995 году.**

Эта поисковая машина первой предложила **расширенную систему поиска** и принимала языковые запросы на так называемом «естественном языке».

Например, могла обработать запрос «Как пройти в библиотеку?», вместо «библиотека».



История веб-поиска. Google

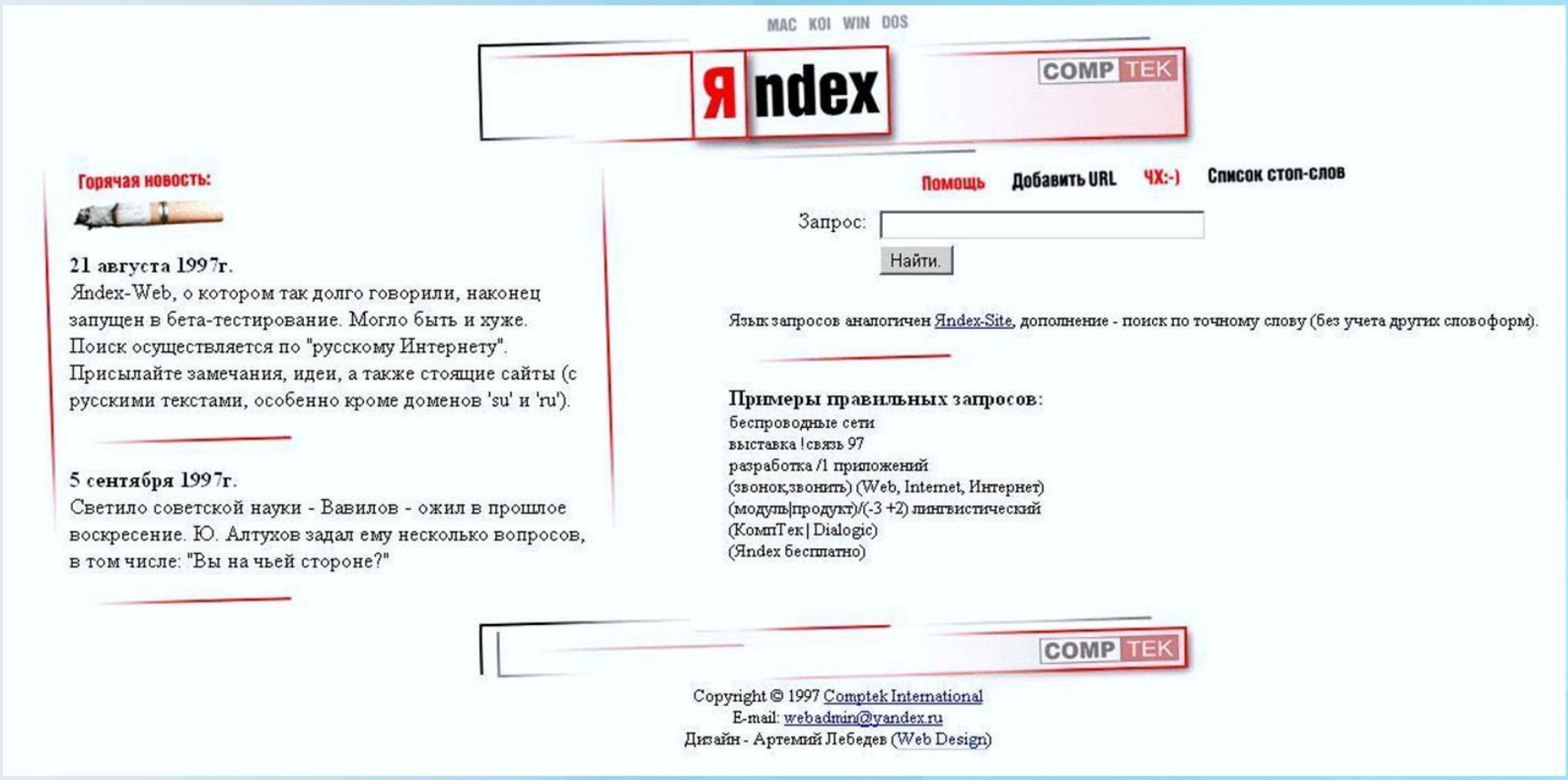
Google был запущен **в 1997 году** Сергеем Брином и Лари Пейджем как часть исследовательского проекта Стэнфордского университета.

При ранжировании результатов запроса Google учитывает **количество внешних ссылок на ресурс,** или **цитируемость**.

По одной из версий, которую принято считать официальной, **название поисковика произошло от намеренно искажённого создателями слова Googol** (Гугол), которое означает «десять в сотой степени» — 10^{100}.

История веб-поиска. Яндекс

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



Механизм веб-поиска

Поисковая система – программно-аппаратный комплекс с веб-интерфейсом, предоставляющий возможность поиска информации, которая размещается во всемирной паутине.

Основной программной частью поисковой системы является поисковая машина (**поисковый движок**) — комплекс программ, обеспечивающий полную функциональность поисковой системы и **обычно являющийся коммерческой тайной** компании-разработчика поисковой системы.

Механизм веб-поиска. Основные компоненты

Поисковые системы обычно состоят из **4-х компонент**:

1. **Поисковый агент**, который перемещается по сети и собирает информацию.

2. **База данных**, которая содержит всю информацию, собираемую пауками.

3. **Поисковая машина**, реализующая алгоритм поиска заданной информации и выдачу результата.

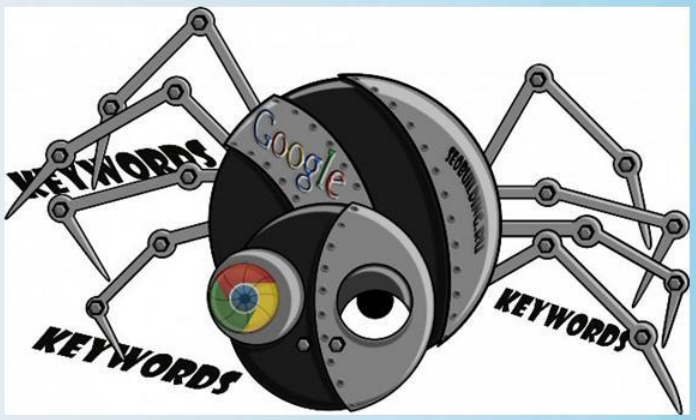
4. **Внешний интерфейс**, который используются для взаимодействия между поисковой машиной и пользователем.

Механизм веб-поиска. Агент

Поисковый агент состоит из нескольких элементов:

1. **Spider (основной паук)**. Скачивает веб-страницы, фактически работает аналогично браузеру. Паук не имеет никаких визуальных компонент.

2. **Crawler («путешествующий» паук)**. Основная задача – определять, куда дальше должен идти Spider, основываясь на ссылках или исходя из заранее заданного списка адресов.



3. **Indexer (индексатор)**. «Слепая» программа, которая анализирует веб-страницы, скачанные пауками.

Механизм веб-поиска. База данных

База данных – это хранилище всех данных, которые поисковая система скачивает и анализирует.

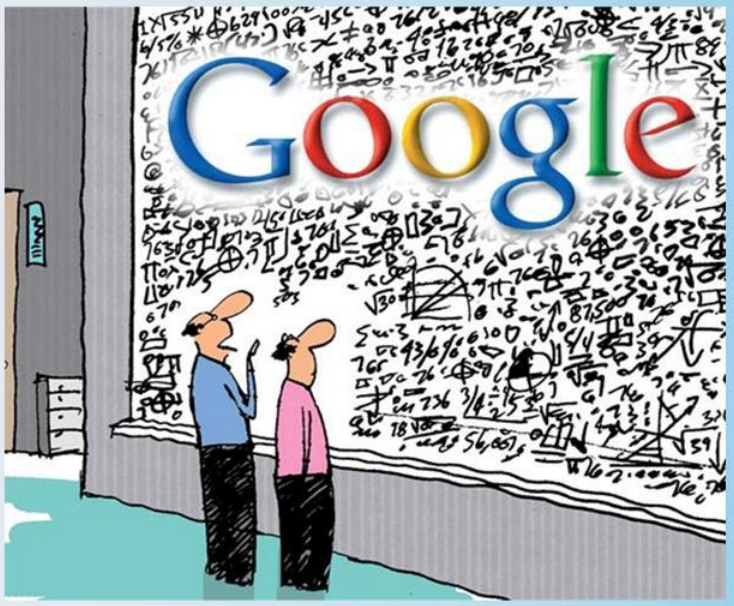
Хранение такого объема данных требует **огромных ресурсов**. В современных системах для этого используются целые датацентры, которую реализуют модель распределенной базы данных.

Так, по данным на конец 2011 года, в индексе Google находится около **40 млрд страниц**, количество уникальных URL – **более 1 квинтиллиона** (10^{18})

Механизм веб-поиска. Поисковая машина

Поисковая машина **отыскивает предмет запроса** в базе данных, основанный на информации, указанной в заполненной форме, и выводит соответствующие документы в результаты поиска.

Чтобы определить порядок, в котором список документов будет показан, поисковая машина применяет свой собственный **алгоритм ранжирования**.



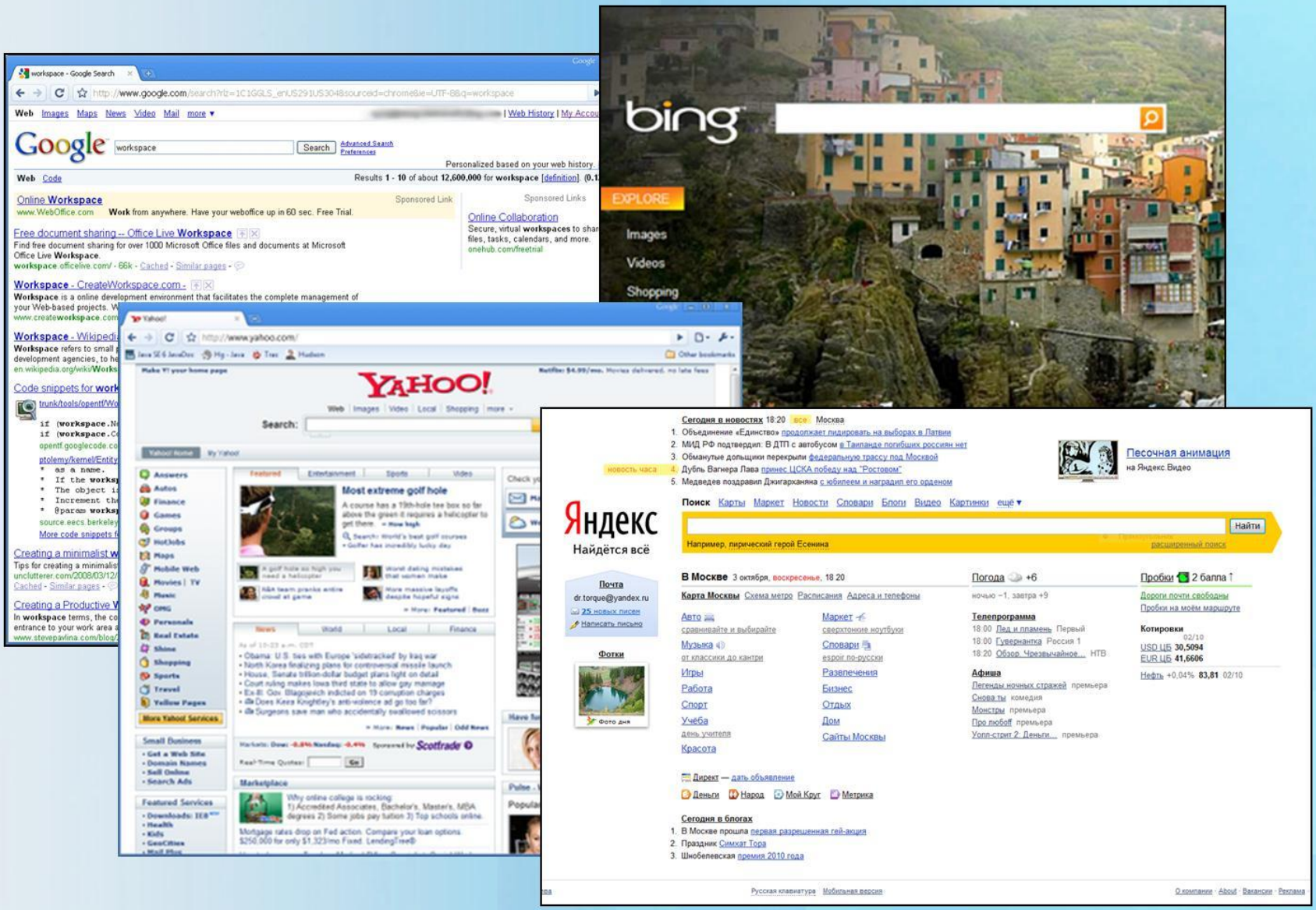
В идеальном случае, **документы, наиболее релевантные пользовательскому запросу, будут помещены первыми в списке**.

16 Механизм веб-поиска. Поисковая машина

Результаты поиска формируются на основании следующих критериев:

1. **Заголовок.** Присутствует ли ключевое слово в заголовке?
2. **Домен/адрес.** Присутствует ли ключевое слово в имени домена или в адресе страницы?
3. **Стиль.** Если место на странице, где ключевое слово использовано в жирных, курсивных фрагментах или текстовых заголовках?
4. **Плотность.** Как часто ключевое слово употреблено на странице?
5. **Метаданные.** Некоторые поисковые системы до сих пор читают мета ключевые слова и мета описания .
6. **Ссылки наружу.** На кого есть ссылки на странице и встречается ли ключевое слово в тесте ссылки?
7. **Внешние ссылки.** Кто еще в имеет ссылку на данный сайт?
8. **Ссылки внутри страницы.** На какие еще страницы данного сайта содержит ссылки эта страница?

17 Механизм веб-поиска. Внешний интерфейс



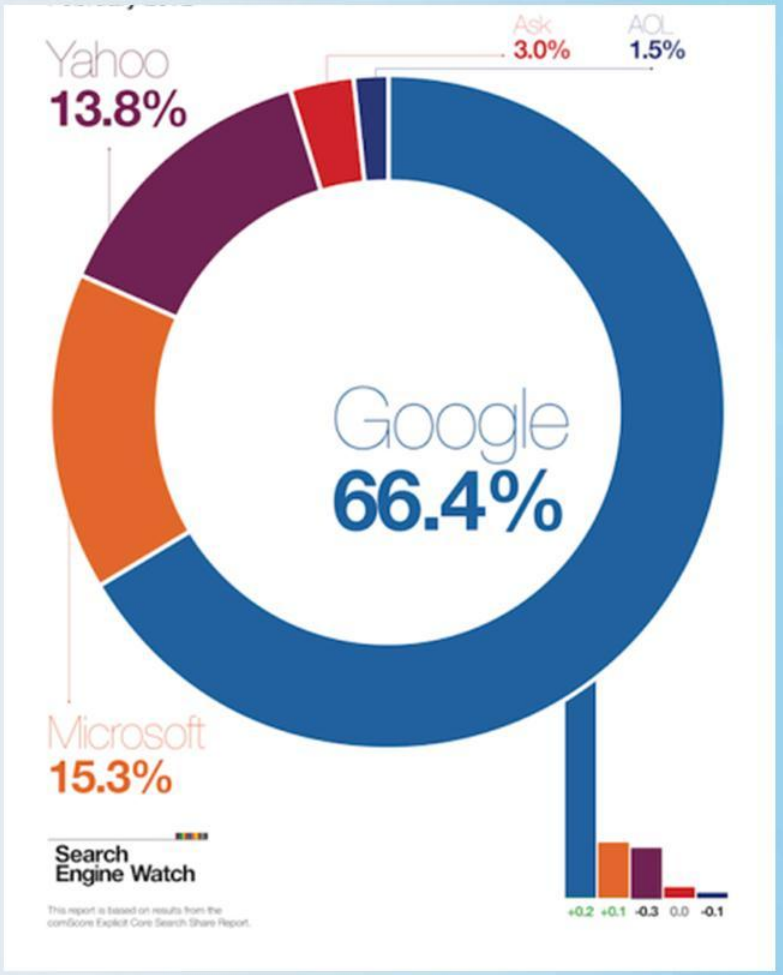
Особенности работы поисковых систем

1. **Алгоритм работы поисковой системы цикличен** - вся база данных поисковой системы обновляется не мгновенно, а периодически. У каждой поисковой системы свой цикл работы и обновления данных.

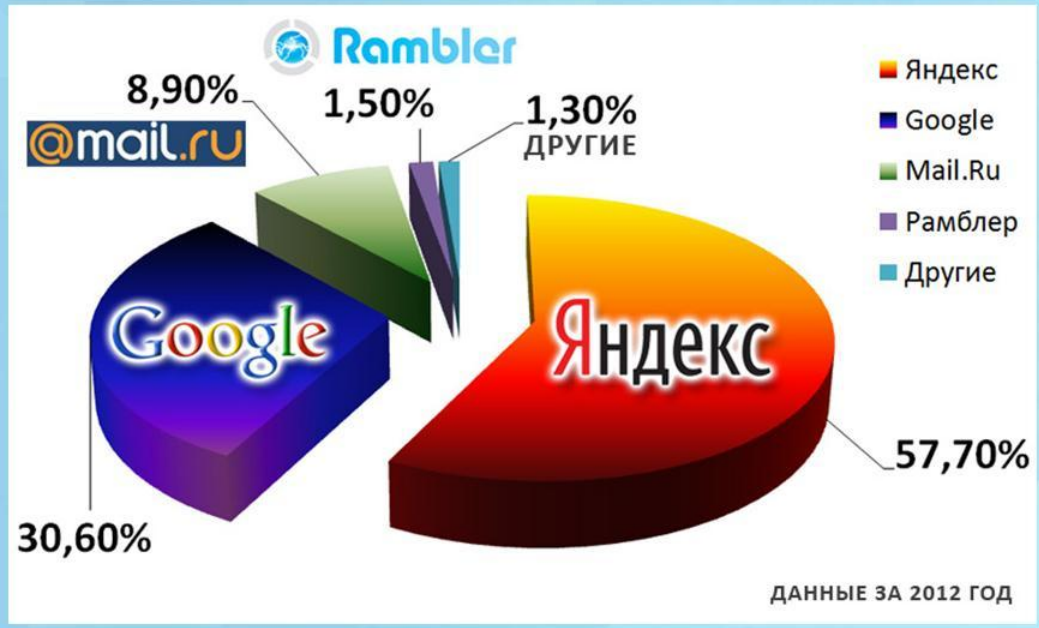
2. В алгоритм работы поисковой системы может быть **заложено более или менее частое сканирование** определенных сайтов: одни сайты будут просматриваться поисковым роботом чаще, другие - реже.

3. **Роботы** популярных поисковых систем **являются самонастраивающимися**. Чем чаще обновляется сайт, тем чаще его посещает работ поисковой системы.

Доли поисковых систем (2012)



Мир



Рунет

Обзор популярных поисковых систем

Google

Поисковая система Google была создана в результате реализации учебного проекта аспирантов Стэнфордского университета **Ларри Пейджа и Сергея Брина**.

В 1996 году они работали над поисковой системой **BackRub**, а страницы интернета поисковый робот BackRub начал индексировать в марте 1996 года.

Для преобразования собранных данных в уровень важности данной веб-страницы, Брин и Пейдж разработали алгоритм **PageRank**. Он стал основной и наиболее эффективной особенностью системы поиска.

В 1998 году на основе BackRub была создана система Google.

Обзор популярных поисковых систем

Google

Язык поисковых запросов Google (некоторые элементы):

«» (пробел) - логическое «И», даёт команду на поиск всех слов.

OR (или символ «|»**)** - логическое «ИЛИ» позволяет найти несколько вариантов сочетаний слов.

«+» - при обработке запроса обязательно учитывается слово, перед которым он стоит.

«-» (минус) - логическое «НЕ». Команда на исключение этого слова из результатов поиска

«» - двойные кавычки задает поиск на четкое соответствие.

«~» - поиск не только указанного слова, но и его синонимов

«*» - заменяет одно слово. Можно указать сколько может быть разных слов между искомыми.

«..» - две точки применяются при поиске чисел (диапазон «от - до»)

Обзор популярных поисковых систем

Google

Язык поисковых запросов Google (некоторые элементы)

filetype: оператор даёт возможность указать тип файла для поиска.

site: ограничивает поиск указанным доменом или сайтом.

movie: поиск информации о видеофильмах.

safesearch: безопасный поиск (без адалт контента)

allintitle: в результате поискового запроса будет выдан список страниц, у которых есть данные слова в заголовке.

allinurl: в результате запроса будет выдан список страниц, у которых есть данные слова в адресе страницы, в её URL.

allintext: используется для поиска слов только в тексте документов. Игнорирует ссылки, URL-адреса и названия.

related: выдаст страницы, похожие по тематике с указанной.

link: все страницы, которые ссылаются на страницу с указанным адресом.

Обзор популярных поисковых систем

Яндекс

Официально поисковая машина **Yandex.Ru** была анонсирована 23 сентября 1997 года на выставке Softool компанией CompTek (была основана еще в 1988 году).

Основными отличительными чертами Yandex.ru на тот момент были:

1. **Проверка уникальности документов** (исключение копий в разных кодировках).
2. **Учёт морфологии русского языка** (в том числе и поиск по точной словоформе).
3. **Поиск с учётом расстояния** (в том числе в пределах абзаца, точное словосочетание).
4. Тщательно разработанный **алгоритм оценки релевантности**, учитывающий не только количество слов запроса, найденных в тексте, но и «контрастность» слова (его относительную частоту для документа), расстояние между словами и положение слова.

Обзор популярных поисковых систем

Основные особенности:

1. У системы были три предшественника, а точнее три других названия: MSN Search, Windows Live Search, Live Search. Наконец, в 2009 году появился именно **bing.com**.
2. Сразу после появления у Bing появилась шуточная расшифровка – Bing Is Not Google. В Bing ежедневно меняется тема оформления стартовой страницы.
3. Очень **удобный поиск видео** – можно отбирать видео как по длине, так и по качеству изображения, так и по источнику.
4. Также не уступает ему по удобности и **поиск по картинкам** – там критериев ещё больше.
5. Есть возможность просмотра **результатов поиска на одной странице** без бесконечного листания.
6. **Интегрирован** с Facebook и Yahoo!.

Обзор популярных поисковых систем

В январе 1994 года аспиранты Стэнфордского университета Дэвид Файло и Джерри Янг создали веб-сайт, который назывался «**Путеводитель Джерри по Всемирной Паутине**». «Путеводитель» представлял собой каталог других сайтов.

Yahoo! стала одной из немногих крупных интернет-компаний, **выживших после «крушения дот-комов»** (2000-2001).

18 февраля 2004 года Yahoo! прекращает использование поисковой технологии Google и **переходит на свою собственную**.

В 2010 году компания возвращается к практике использования сторонней машины поиска. Только на этот раз – **bing разработки Microsoft** и только на территории США и Канады.