Содержание:

Введение

Интернет - глобальное информационное пространство, основанное на самых передовых технологиях, обладающее широким спектром информационных и коммуникационных ресурсов, содержащее колоссальные объемы данных. Появление Интернета принято связывать с 1969 г. Именно тогда в США начались работы по объединению в небольшие сети групп компьютеров. Это делалось с целью обеспечения сохранности информации в критических условиях. Уже в 1971 г. на основе этих разработок возникла электрическая почта. Успех этих начинаний и заложил основы Интернета в нынешнем виде.

Актуальность темы моего исследования является то-что наиболее популярным и используемым способом поиска в Интернете является использование поисковых систем, ресурсы Интернета давно превратились в незаменимый инструмент для повседневной работы людей многих профессий. Быстрый рост информации в сети сделали его океаном разнообразнейших данных, важность которых растет пропорционально их объему. По оценке экспертов объем информации, передаваемой по каналам Интернет, удваивается каждые полгода. Ежедневно в сети появляются миллионы новых документов, и естественно, что без систем поиска они в подавляющем своем большинстве остались бы не востребованными, вообще не были бы не кем найдены, и все то огромное количество информации оказалось бы никому не нужным. Возникла необходимость создания таких средств, которые позволили бы легко ориентироваться в информационных ресурсах глобальных сетей, быстро и надежно находить нужные сведения. В интернете появились специальные поисковые средства. Еще несколько лет назад бытовало такое мнение: в Интернете есть все, но найти там ничего невозможно. Однако с появлением и быстрым развитием поисковых каталогов, поисковых машин, и всевозможных поисковых программ ситуация изменилась, и теперь в Сети срочно понадобившуюся информацию иногда можно найти быстрее, чем в книге, лежащей на столе.

Целью моей курсовой работы является определить наиболее популярные Российские и Зарубежные поисковые системы. Разработать модель "идеальной"

поисковой системы

Задачей моей курсовой работы является раскрыть понятие поисковая система, максимально изучить, что же такое поисковая система, как работает поисковая система.

Глава I. Раскрытие понятия поисковая система

1.1 Что такое поисковая система

Поисковая система – это сайт, к которому пользователь обращается посредствам ключевого слова и находит интересующую его информацию. Другими словами поисковая система – портал, осуществляющий поиск, сбор и сортировку информации в сети Интернет, это инструмент, позволяющий пользователю глобальной сети в кратчайшие сроки найти интересующую его информацию. Сегодня поисковая система лучший способ, чтобы быстро и качественно найти интересующую вас информацию.

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

Рассмотрим, как работает поисковая система, что само по себе довольно просто. Пользователь, который зашел на сайт системы, должен ввести в поисковое окно, ключевую фразу, располагающуюся на сайте, по этой фразе система ищет информацию, и нажатием кнопки «поиск», послать запрос. После всего, пользователю будет выдан список текстовых ссылок на сайты, которые соответствуют данному запросу. В этом заключается весь принцип работы поисковой системы со стороны пользователя. Теперь рассмотрим внутреннее устройство и весь процесс работы системы, не заметный для пользователя.

Получая результат, пользователь оценивает работу системы, руководствуясь несколькими основными параметрами. Нашел ли он то, что искал? Если не нашел, то сколько раз ему пришлось перефразировать запрос, чтобы найти искомое? Насколько актуальную информацию он смог найти? Насколько быстро обрабатывала запрос поисковая машина? Насколько удобно были представлены результаты поиска? Был ли искомый результат первым или же сотым? Как много ненужного мусора было найдено наравне с полезной информацией? Найдется ли

нужная информация, при обращении к поисковой системе, скажем, через неделю, или через месяц?

1.2 Краткая история развития поисковых систем

Одним из первых способов организации доступа к информационным ресурсам сети стало создание каталогов сайтов, в которых ссылки на ресурсы группировались согласно тематике. Первым таким проектом стал сайт Yahoo, открывшийся в апреле 1994 года. После того, как число сайтов в каталоге Yahoo значительно увеличилось, была добавлена возможность поиска информации по каталогу. Это, конечно же, не было поисковой системой в полном смысле, так как область поиска была ограничена только ресурсами, присутствующими в каталоге, а не всеми ресурсами сети Интернет.
Каталоги ссылок широко использовались ранее, но практически утратили свою популярность в настоящее время. Причина этого очень проста – даже современные каталоги, содержащие огромное количество ресурсов, представляют информацию лишь об очень малой части сети Интернет. Самый большой каталог сети DMOZ (или Open Directory Project) содержит информацию о 5 миллионах ресурсов, в то время как база поисковой системы Google состоит из более чем 8 миллиардов документов.

Первой полноценной поисковой системой стал проект WebCrawler появившийся в 1994 году.

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

В 1997 году Сергей Брин и Лари Пейдж создали Google самую популярную на сегодняшний момент поисковую систему в мире.

23 сентября 1997 года была официально анонсирована поисковая система Yandex, самая популярная в русскоязычной части Интернет.

В настоящее время существует 3 основных международных поисковых системы – Google, Yahoo и MSN Search, имеющих собственные базы и алгоритмы поиска. Большинство остальных поисковых систем (коих можно насчитать очень много) использует в том или ином виде результаты 3 перечисленных. Например, поиск AOL (search.aol.com) и Mail.ru используют базу Google, а AltaVista, Lycos и AllTheWeb –

базу Yahoo.

В России основной поисковой системой является Яндекс, за ним идут Rambler, Google.ru, Aport, Mail.ru и KM.ru

1.3 Задачи поисковых систем

Все поисковые системы объединены несколькими основными задачами, такими как поиск новых сайтов, оценка сайта и максимально точный ответ пользователю на запрос. Главная задача любой поисковой системы, предоставить пользователь ту информацию, которую он ищет. Но, к сожалению нельзя научить пользователя производить «правильные» запросы к системе, т.е. запросы, которые соответствуют принципу работы поисковых систем. Вот почему разработчикам нужно создавать такие принципы работы и алгоритмы поисковых систем, которые бы позволяли пользователям находить искомую ими информацию.

Это значит, что поисковая система должна думать точно также как думает пользователь, когда ищет ту или иную информацию. Обращаясь к поисковой системе, пользователь надеется максимально просто и быстро найти интересующую его информацию. После получения результата, он оценивает работу системы, руководствуясь несколькими основными параметрами. Разработчики поисковых систем постоянно стараются совершенствовать алгоритмы и принципы поиска, пытаются всячески ускорить работу системы, добавляя новые функции и возможности, чтобы удовлетворить потребности пользователей

1.4 Состав и принципы работы поисковой системы

Поисковая машина – это аппаратно-программный комплекс, который осуществляет быстрый поиск внутри сервера или Интернет-ресурса необходимой информации. У всех поисковых систем основа поисковой машины примерна одинаковая. В основном, это программное обеспечение, отвечающее за ранжирование результатов по релевантности поискового запроса и составление каталога запроса, поисковый бот, который необходим для поиска сайта и индексации. Но некоторые крупные поисковые системы держат содержание своей поисковой машины в секрете. Основным отличием является учет и релевантность морфологии языка запроса, база проиндексированных сайтов. Все это в совокупности и определяет критерий качества работы поисковых машин.

Поисковые машины классифицируются по области поиска информации:

1. Локальный поиск. Он предназначен, чтобы осуществлять поиск информации по всемирной сети какой-либо ее части, например, по локальной сети, либо по одному или нескольким сайтам. Таким примером являются внутренние серверы крупных компаний или поисковый скрипт на сайте.

2. Глобальный поиск. Он предназначен для того, чтобы искать информацию по региональной части, по группе сайтов, либо в сети Интернет и т.д. Именно глобальным поиском пользуются такие крупные поисковые системы как Яндекс, Google, Yahoo и т.д.

Поисковые машины по сети интернет осуществляют различный поиск информации. Например, музыка, картинки, личная информация, географическое положение и т.д. Поисковая машина может работать с файлами различных форматов (например .html,.htm,.txt,.doc,.rtf, …), мультимедийного (видео, звука и другой информации) или графического (.gif, .png, .svg,) типа. Но самым распространенным поиском является поиск текстовых документов (документы в формате doc, rtf, txt, web-страницы и др.). Но с технологической точки зрения поиск по звукам, видео, изображениям является более сложным, поэтому он не реализован массово. Например, такие системы как Яндекс.Картинки ищут картинки по альтернативным текстам, соответствующим этим изображениям, а не по самим изображениям. А в компании Google каталог поиска картинок составляется вручную, это тормозит обновление баз изображений, но значительно увеличивает релевантность запроса.

Модуль индексирования: Модуль индексирования состоит из трех вспомогательных программ (роботов):

Spider (паук) – программа, которая предназначена для скачивания веб-страниц. «Spider» полностью обеспечивает скачивание страницы, и все внутренние ссылки извлекает с этой страницы. С каждой страницы скачивается html-код. Роботы используют протоколы HTTP для скачивания страниц. «Spider» работает следующим образом. Робот передает на сервер запрос «get/path/document» и несколько других команд HTTP-запроса. В ответ роботу приходит текстовый поток, который содержит сам документ и служебную информацию.

Ссылки извлекаются из тэгов frame, base, area, frameset, и др. Многие роботы, наряду со ссылками, обрабатывают редиректы (перенаправления). Все страницы сохраняются в таких форматах как:

- дата, когда страница была скачана
- тело страницы (html-код)
- URL страницы
- http-заголовок ответа сервера

Crawler («путешествующий» паук) – эта программа, автоматически проходит по всем ссылкам, которые нашла на странице. Выделяет все ссылки, присутствующие на странице. Его задача – состоит в том, чтобы исходя из заранее заданного списка адресов или основываясь на ссылках, определить, куда дальше должен идти паук. Crawler, осуществляет поиск новых документов, еще неизвестных поисковой системе, следуя по найденным ссылкам.

Indexer (робот - индексатор) - это программа, анализирующая веб-страницы, которые скачали пауки. Индексатор, применяя собственные лексические и морфологические алгоритмы, разбирает страницу на составные части и анализирует их. Разные элементы страницы подвергаются анализу, например, заголовки, текст, специальные служебные html-теги, ссылки структурные и стилевые особенности, и т.д.

Благодаря этому, модуль индексирования дает возможность извлекать ссылки на новые страницы из получаемых документов и производить полный анализ этих документов, обходить по ссылкам заданное множество ресурсов, скачивать встречающиеся страницы.

База данных: Индекс поисковой системы или база данных - это информационный массив, в котором хранятся преобразованные параметры всех документов скачанных и обработанных модулем индексирования.

Поисковый сервер: Поисковый сервер важнейший элемент всей системы, потому что скорость и качество поиска напрямую зависит от его алгоритмов, которые лежат в основе его функционирования.

Работает поисковый сервер следующим образом:

Запрос, который получен от пользователя подвергается морфологическому анализу. Генерируется информационное окружение каждого документа, содержащегося в базе (как раз оно и будет отображено в виде сниппета, т. е. текстовой информации соответствует запросу на странице выдачи результатов поиска).

Все полученные данные передаются специальному модулю ранжирования в качестве входных параметров. После чего по всем документам происходит обработка данных, далее подсчитывается собственный рейтинг для каждого документа, который характеризует релевантность разных составляющих данного документа, хранящихся в индексе поисковой системы запроса, введенного пользователем.

Этот рейтинг может быть составлен в зависимости от выбора пользователя дополнительными условиями (например, «расширенный поиск»).

Далее генерируется сниппет, т. е., из таблицы документов извлекаются краткая аннотация, наиболее соответствующая запросу, заголовок и ссылка на сам документ для каждого найденного документа, и еще подсвечиваются все найденные слова.

Пользователю результаты поиска, которые мы получили, передаются в виде SERP (Search Engine Result Page) – страницы выдачи поисковых результатов.

Все эти компоненты работают во взаимодействии и тесно связаны друг с другом, именно они образовывают тот самый довольно сложный механизм работы поисковой системы, который требует огромных затрат ресурсов.

Глава II. Анализ поисковых систем

2.1 Рейтинг основных мировых поисковых систем

Где-то с двухтысячного года самой крупной поисковой системой в мире считается Google. Однако не все страны и континенты пользуются одинаковыми поисковиками. Так, в странах Восточной Азии Гугл не в фаворитах.

В Китае популярны поисковики Soso и Baidu. Причём, последняя ПС ворвалась в десятку сайтов, лидеров по посещаемости, и продолжает там находится по сегодняшний день. Baidu — 8-ой поисковый сайт в мире по посещаемости.

В Тайване и Японии используют Yahoo! Taiwan и Yahoo! Japan.

В Южной Корее большинство жителей пользуются «отечественной» разработкой Naver.

В России Яндекс опережает Гугл.

В странах Ближнего Востока существуют поисковые системы, выдающие только «дозволенную» информацию с точки зрения религии. Это либо такие «молодые» системы, как Halalgoogling, либо уже знакомые нам Яху!, Гугл и Бинг с обусловленной системой фильтрации.

Самые крупные поисковые системы России на 2015 год

Яндекс — 50,65%

Google — 40,6%

Mail.ru — 6,4%

Рамблер — 1,7%

Bing — 0,7%

Самые крупные поисковые системы мира на 2015 год

Google — с суммарным процентом используемости в мире 66,41%;

Baidu — 12,33%;

Bing — 10,16%;

Yahoo! — 8,76%;

AOL — 0,7%;

Ask — 0,22%;

Конечно, эти списки не является окончательными, так как разные источники на основе своих критериев оценки формируют перечни популярных поисковиков, включая такие порталы, как: Infoseek, HotBot, Teoma, Exite, Galaxy, Microsoft MSN, AltaVista и др. Если говорить отдельно о такой поисковой системе, как Байду, то в китайском информационном пространстве в последнее время Baidu намного перегнал Google, Sina и Sohu.com, и на данный момент занимает 2 место в мире по числу обработанных запросов.

Система MCH для выдачи результатов поиска использует базы порталов Яху, Альтависты, Инктоми и др. Она тоже является одним из значимых ресурсов

интернета и ею широко пользуются в Бельгии, Дании, Англии, Японии и Новой Зеландии.

Яху насчитывает более 345 миллионов пользователей. Представительства компании (больше 30-ти) работают в тихоокеанском регионе, Европе, Азии и Северной Америке.

2.2 Обзор основных мировых поисковых систем

На сегодняшний день всемирная сеть Интернет насчитывает огромное множество поисковых систем во всех странах мира, из них всех можно выделить несколько самых крупных и пользующихся наибольшей популярностью среди пользователей:

2.2.1 Google

Лидер поисковых машин Интернета, Google занимает более 60 % мирового рынка, а значит, шесть из десяти находящихся в сети людей обращаются к его странице в поисках информации в Интернете. Сейчас регистрирует ежедневно около 50 миллионов поисковых запросов и индексирует более 8 миллиардов веб-страниц.

Была разработана в 2003 выпускниками Стэндфордского университета Сергеем Брином и Лари Пейджем, которые применили для ранжирования документов технологию PageRank, где одним из ключевых моментов является определение "авторитетности" конкретного документа на основе информации о документах, ссылающихся на него. Говоря общими словами, чем больше документов ссылается на данный документ и чем они авторитетнее, тем более авторитетным данный документ становится. Количественное значение авторитетности документа (другими словами, взвешенное количество ссылок или PageRank) относится к так называемым статическим факторам (то есть независящим от конкретного запроса) и учитывается при определении релевантности документа конкретному запросу как весовой коэффициент. Наряду с этим Google применил для определения релевантности документа не только текст самого документа, но и текст ссылок на него. Эта технология позволила ему обеспечить выдачу довольно релевантных результатов на фоне других поисковиков. Довольно быстро Google стал лидировать в различных опросах по такому показателю, как удовлетворенность пользователей результатами поиска.

Google осуществляет поиск по документам на более чем 35 языках, в том числе русском. В настоящее время многие порталы и специализированные сайты предоставляют услуги поиска информации в Интернете на базе Google, что делает задачу успешного позиционирования сайтов в Google еще более важной. Google проводит переиндексацию своей поисковой базы примерно раз в четыре недели. Во время этого усовершенствования, неофициально называемого Google dance, происходит обновление базы на основе информации, собранной роботами за время, прошедшее с предыдущего усовершенствования, и перерасчет значений PageRank документов. Также существует определенное количество документов с достаточно большим значением PageRank, информация о которых в поисковой базе обновляется ежедневно, однако значение PageRank пересчитывается только во время Google dance. Нормированное значение PageRank для конкретного документа, загруженного в браузер, можно узнать, скачав и установив Google ToolBar - специальную панель инструментов для работы с этим поисковиком. Не смотря на то, что в поисковике имеется форма для бесплатного добавления страницы в базу, Google предпочитает сам находить новые документы по ссылкам с уже известных и не будет индексировать добавленную через форму страницу, если в его базе не найдется ни одной страницы, ссылающейся на нее.

Преимуществами поисковой системы Google является:

- Очень мощная поисковая система, которая находится в постоянном развитии.
- База индексов этой системы обновляется раз в два дня, качество выдачи очень высокое, найти необходимый документ или информацию довольно легко.
- Система ориентирована в основном на ссылки, причем учитываются как входящие, так и исходящие ссылки с ресурса.
- Способна выдавать результаты на запросы по семантике языка программирования (исходный код поиска).

Недостатками поисковой системы является:

- Нередко встречаются ссылки на сайты с уже устаревшей информацией.
- Случается, что ссылки, которые находятся в результатах поиска, ведут на сайт, находящийся в стадии разработки.
- На запрос «фильм» и «фильмы» результаты поиска будут отличаться.
- Отсутствие возможности указать конкретную грамматическую форму слова, либо ударение также значительно усложняет процесс поиска информации.

2.2.2 Yahoo

Одна из самых первых Поисковых систем (создана Дэвидом Фило и Джерри Янгом в апреле 1994года) по сей день остается и самой популярной из них, традиционно сочетая поиск, как по ключевым словам, так и с помощью иерархического дерева разделов.

Нынешнее развитие Yahoo можно определить как движение в он-лайн, интерактивность. Yahoo быстро осваивает эту область Интернет-услуг, но возникает одна проблема: ядро Yahoo! не было на это рассчитано. Не была в 1994 году заложено в него "онлайновая" составляющая, ее "приклеил" Тим Кугл несколькими годами позже. Естественно возникает угроза хакерских атак через эту незащищенную область.

Одно из новшеств поисковой системы Yahoo - панель задач для браузера Firefox,. Этот инструмент помогает пользоваться поиском Yahoo, не заходя на официальный сайт, а лишь используя функциональные кнопки панели.

1 сентября 2007 года поисковик Yahoo, которому принадлежит более 200 миллионов адресов электронной почты по всему миру, анонсировал запуск новой системы поиска текстов, фотографий и других документов, содержащихся в письмах.

Необходимость такого нововведения возникла вслед за увеличением объёма хранимых данных, ведь некоторые пользователи создают целые почтовые архивы. Подгоняемый конкурентом Google и его почтовым сервисом Gmail, Yahoo для хранения почты предлагает отныне 1 гигабайт бесплатного места, или 2 гигабайта по годовому абонементу. "Как только вы получаете возможность хранить больше информации, вам необходимы и расширенные поисковые возможности", - объясняет Эрик Петерсон, аналитик компании Jupiter Research.

Пользователи поисковой системы Yahoo, в свою очередь, смогут теперь использовать возможности детализированного поиска слов в названии или непосредственно в тексте письма, а также в присоединенных документах, не открывая их. Результат поиска отражается в трёх строках с указанием всех атрибутов. На панели справа отображаются все похожие документы. Найденные фотографии выводятся на экран в уменьшенном виде, что значительно облегчает поиск. Система также учитывает орфографические ошибки, позволяя искать слова лишь по первым буквам.

Для начала Yahoo планирует предложить новую систему небольшому числу американских пользователей, а затем распространить её по всему миру. Со стороны клиентов это не потребует никаких дополнительных усилий. "Когда услуга станет, доступна, в левом верхнем углу страницы вашего почтового ящика появится соответствующий баннер", - обещает компания Yahoo.

По данным comScore Media Metrix на июль этого года, домену Yahoo принадлежит 219 миллионов адресов электронной почты, что составляет 31,5% мирового рынка, уступая лишь Microsoft с 221 миллионом пользователей сервиса Hotmail (35,5% рынка).

Преимуществами поисковой системы Yahoo является:

- Содержит ссылки, которые наиболее полно отвечают указанной в запросе тематике.
- Имеются интеллектуальные средства «отсечения» пустых, находящихся в разработке или чисто рекламных сайтов, далеких от искомой тематики.
- Всегда легко определить, в каком разделе находится нужная информация.
- В случае если на Yahoo нет результатов, сразу выводятся результаты с AltaVista.

Недостатками поисковой системы является:

- Возможна проблема с отсутствующими страницами, поскольку веб-мастера обычно забывают удалить свои сайты с поисковых систем, а на Yahoo нет механизма автоматического обновления.
- Чисто русские ресурсы не добавляются, потому что их просто некому смотреть и оценивать содержимое.
- Нет собственной поисковой машины.
- Ищет слова, заданные в критерии поиска только в названии и описании страницы

2.2.3 Baidu

Baidu – лидер среди китайских поисковых систем. По количеству обрабатываемых запросов поисковый сайт Байду стоит на 3 месте в мире (3 миллиарда 428 миллионов; с долей в глобальном поиске 5,2 %). Хотя компания работает только в единственной стране: Китае! Но точно, что этот рынок растет неистово быстро:

Уже в конце года в Китае свыше 170 млн. пользователей займутся поиском информации в Интернете. Аналитик J.P. Морган Дик Вей исходит в своем актуальном анализе из того, что это число вырастет в течение следующих трех, четырех лет до 100 млн. пользователей. Гигантский рынок с экстремально высокими доходами для Baidu. Сравнивают только прибыль, которую Google достигает в США с очень похожей бизнес-моделью.

К концу 2002 года количество китайских сайтов, индексируемых Baidu, было на 50% больше, чем у любого конкурента.

Число заблокированных результатов поиска у Baidu на 30% больше, чем у Google

Google оставила Baidu далеко позади, поскольку предлагает рекламодателям выход на международные рынки.

Преимуществами поисковой системы Baidu является:

- Предоставляет пользователям возможность сортировать результаты поиска: по дате, по алфавиту, по релевантности.
- При осуществлении поиск по ключевому слову, команда специалистов компании отслеживает наиболее релевантные на их взгляд сайты, вручную отбирают и классифицируют их, и вносят в определенные рубрики директории.
- Ранжирования узлов по популярности и сезонным изменениям.
- Помощь со стороны человека-редактора.

Недостатками поисковой системы является:

- Поисковая система полна спамом.
- Использует внешние данные для обработки поисковых запросов, поэтому на релевантность влияют: расположение ключевых слов, популярность ресурса и текст ведущих на сайт, и ведущих с сайта ссылок.

Ближе всего к идеалу находятся поисковые системы Google, Яndex, Rambler, Апорт. Отмечу также, что поисковая система MSN лидирует в системе ранжирования.

2.3 Обзор основных Российских поисковых систем

Основное отличие русскоязычных поисковых систем от иностранных одно - это то, что глобальные поисковые системы, поддерживающие поиск на русском языке, не поддерживают русскую морфологию. В русскоязычной части сети Интернет работают около двух десятков поисковых систем, но подавляющие большинство пользователей работает лишь с несколькими, подробно остановимся на самых крупных:

2.3.1 Yandex

Яндекс - На сегодня наиболее популярная поисковая система, ежемесячно к ней обращаются более 35 миллионов пользователей Русскоязычной части Интернета. Начала свою работу во второй половине 1997 года учитывая морфологию русского языка. История компании "Яндекс" началась в 1990 году с разработки поискового программного обеспечения в компании "Аркадия". За два года работ были созданы две информационно-поисковые системы - Международная Классификация Изобретений, 4 и 5 редакция, а также Классификатор Товаров и Услуг. Обе системы работали локально под DOS и позволяли проводить поиск, выбирая слова из заданного словаря, с использованием стандартных логических операторов. В1993 году "Аркадия" стала подразделением компании CompTek. В 1993-1994 годы программные технологии были существенно усовершенствованы благодаря сотрудничеству с лабораторией Ю. Д. Апресяна (Институт Проблем Передачи Информации РАН). В частности, словарь, обеспечивающий поиск с учетом морфологии русского языка, занимал всего 300Кб, то есть целиком грузился в оперативную память и работал очень быстро. С этого момента пользователь мог задавать в запросе любые формы слов.

Слово Яндекс придумал за несколько лет до этого один из основных и старейших разработчиков поискового механизма. "Яndex" означает "Языковой index", или, если по-английски, "Yandex" - "Yet Another indexer". За 4 года публичного существования Яndex возникли и другие толкования. Например, если в слове "Index" перевести с английского первую букву ("I" - "Я"), получится "Яndex".

В начале 1996 года был разработан алгоритм построения гипотез. Отныне морфологический разбор перестал быть привязан к словарю - если какого-либо слова в словаре нет, то находятся наиболее похожие на него словарные слова и по ним строится модель словоизменения. В это время Интернет в России только начинался. Еще через полгода стало очевидно, что ничто не отделяет CompTek от

создания собственной глобальной поисковой машины. Объем Рунета составлял тогда всего несколько гигабайт. Осенью 1997 года был открыт Yandex.Ru.

Помимо поисковой системы, сегодня Яндекс - огромный портал с целым набором широко используемых сервисов, такими как каталог, Яндекс. деньги, и другие. Официально поисковая машина Yandex.Ru была анонсирована 23 сентября 1997 года на выставке Softool. Основными отличительными чертами Yandex.Ru на тот момент были проверка уникальности документов (исключение копий в разных кодировках), а также ключевые свойства поискового ядра Яндекс, а именно: учет морфологии русского языка (в том числе и поиск по точной словоформе), поиск с учетом расстояния (в том числе в пределах абзаца, точное словосочетание), и тщательно разработанный алгоритм оценки релевантности (соответствия ответа запросу), учитывающий не только количество слов запроса, найденных в тексте, но и "контрастность" слова (его относительную частоту для данного документа), расстояние между словами, и положение слова в документе. Сегодня Яндекс имеет внутри мощный поисковый робот, позволяющий производить поиск по самым различным критериям.

Преимуществами поисковой системы Яндекс является:

- Постоянное развитие системы.
- Качество выдачи растет, все больше удобных сервисов предлагает компания: каталог, карты, новости, прогноз погоды, почта.
- Глубокий морфологический анализ обрабатываемых терминов.
- Обладает хорошим механизмом распознавания одного документа в нескольких кодировках или на зеркальных серверах.
- Оригинально сконструированный механизм выдачи результатов.
- Огромная индексная база.

Недостатками поисковой системы является:

- Разница в выдаче при наборе слова с большой (маленькой) буквы (иногда выдача меняется, иногда нет).
- Частое выпадение секторов поисковой базы - когда исчезают части сайтов из выдачи и восстанавливаются через 2-5 дней.
- Обновление индексов поисковой базы происходит недостаточно часто и регулярно.

2.3.2 Rambler

Rambler - Старейшая поисковая система российского Интернет, запущена в 1996 году, на сегодня - вторая по популярности с обращением более 25 миллионов посетителей в месяц. Помимо поисковой системы, сегодня Рамблер - один из крупнейших порталов Русскоязычной части Интернета с большим набором широко известных сервисов, таких как каталог Рамблер, Рамблер-почта, Рамблер-ICQ или Рамблер-ТВ. По сути сегодня Рамблер - больше, чем просто поисковая система и набор сервисов, это крупная медиагруппа. Поисковая машина "Рамблер" начала работу в октябре 1996 года, на стартовом этапе содержала всего 100 тысяч документов. "Рамблер" не был первой отечественной поисковой системой, однако в первый год своего существования (когда весь русский веб с приемлемой степенью правдоподобия индексировался "Рамблером", "Апортом", "Русской поисковой машиной", а также шведской и калифорнийской AltaVista) вынес основной груз поисковых запросов. Вторая версия "Рамблера" начала разрабатываться летом 2004 года, в марте нынешнего года приняла достаточно законченные очертания. В нее были введены функции, давно уже имевшиеся в конкурирующих системах. Она учитывает координаты слов, обучена строгой и нечеткой морфологии, связывает поиск с каталогом, в качестве которого используется Top100 (http://top100.rambler.ru/), группирует результаты поиска по сайтам, ищет по числам. Достаточно удачная архитектура продукта позволяет "Рамблер" иметь для поисковика количество серверов в 2 раза меньшее, чем у "Яндекса", и в 3 раза меньшее, чем у "Апорта".

Преимуществами поисковой системы Рамблер является:

- Система работает с большой скоростью поиска.
- Обновление поискового индекса происходит несколько раз в день.
- Поисковик всегда находит самые свежие документы и последние новости.
- Обладает близким к оптимальному выводом результатов поиска.
- Производит ранжирование результатов в зависимости от частоты употребления и местоположения искомых терминов.
- Один и тот же документ в различных кодировках показывается только один раз, а его конкретные адреса суммируются в списке, идущим за резюме.

Недостатками поисковой системы является:

- На величину индекса релевантности влияет время существования сайта в сети. Эта особенность позволяет пользователям находить ресурсы, которые

давно существуют, успешно развиваются, а не сайты-однодневки. Но такой подход значительно затрудняет попадание в выдачу новых сайтов, информация на которых подчас оказывается актуальной и, возможно, более важной для пользователя.
- Невозможность осуществления поиска по целой фразе указывая в запросах предельное расстояние искомых терминов друг от друга.

2.3.3 Апорт

Апорт – Третья популярности на сегодня поисковая система с обращением более 16 миллионов посетителей в месяц. Апорт позволяет пользователям осуществлять полнотекстовый поиск документов с учетом морфологии русского языка в запросах. Поисковая система построена на основании новейших достижений в области информационного поиска и использует уникальные алгоритмы сортировки найденных результатов. Разнообразные специализированные поиски (Знакомства, Товары, Новости, Рефераты, MP3 и др.) дают пользователям дополнительные возможности находить различную информацию в Сети. В поисковую машину интегрирован один из крупнейших в Русскоязычной части Интернет каталогов Интернет-ресурсов "Апорт-каталог".

Поисковая машина "Апорт" была впервые продемонстрирована в феврале 1996 года на пресс-конференции "Агамы" по поводу открытия "Русского клуба". Тогда она искала только по сайту russia.agama.com. Потом она начала искать по четырем, потом по шести серверам... Короче, день рождения и фактический старт системы сильно "размазались" по времени, а официальная презентация "Апорта" состоялась только 11 ноября 1997 года. К тому времени в его базе был проиндексирован первый миллион документов, расположенных на 10 тысячах серверов. Создателем системы выступила компания "Агама" - разработчик программного обеспечения для платформы Windows, главным из которых являлся корректор орфографии "Пропись". Лингвистические разработки "Агамы" использовались при создании поисковой машины, в которой, скажем, в отличие от "Рамблер", изначально учитывалась морфология слов и осуществлялась по желанию клиента проверка орфографии запроса.

Важнейшими свойствами первой версии "Апорта" являлся перевод запроса и результатов поиска на английский язык и обратно, а также реконструкция всех проиндексированных страниц из собственной базы (что означает возможность

просмотра страниц, уже несуществующих в оригинале).

Апорт 2004" стал первой российской поисковой машиной, практически реализовавший две базовых технологии американской поисковой машины Google. Первая - учет "ранга страницы" (Page Rank), который характеризует ее популярность (вычисляется по количеству ссылок на ресурс из внешнего Интернета: вес ссылки с популярного сайта выше, чем вес ссылки с менее популярного; ссылки, включающие слова запроса, имеют больший вес, чем, скажем, слово "здесь"). Вторая - обработка запроса, ориентируясь на HTML-код страницы. В "Апорт 2004" учитывается также вхождение слов запроса в URL. Среди недокументированных особенностей - больший приоритет сайтам, получившим высшую и элитную лигу в каталоге AtRus.

Можно отметить и то, что "Апорт" первым устроил поиск по новостным лентам (какие бы ложные сведения о приоритете "Яндекса" в этом сервисе не распускал в свое время Internet.ru). И, наконец, еще одно первенство "Апорта" - использование платной нулевой строки в выдаче (кстати, "Апорт" первым среди наших поисковиков начал покупать такой сервис у AltaVista, которая за небольшую плату выдавала его ссылку первой при запросе "Russian Search"). Однако в "Апорте" нельзя купить не нулевое, а просто более высокое место для своего сайта в результатах поиска. Пользователи "Апорта" (в отличие завсегдатаев "Яндекса") мало пользуются расширенным поиском (на 8000 загрузок простой страницы приходится 300 вызовов страницы "Расширенный поиск").

Организация масштабируемости в архитектуре "Апорт 2004" такова, что можно дробить поисковую базу "Апорта" на несколько отдельных баз, каждый маленький "Апорт" работает на своем компьютере. "Апорт 2004" считает, что весь Интернет поделен на фрагменты. После проведения поиска по этим фрагментам, пользователю интегрируется и выдается общий ответ. Добавлять новые маленькие "апортики" можно путем не очень сложной процедуры. В случаях аварий отдельных машин выдаются несколько отличные от штатных интегральные результаты, что мы можем время от времени наблюдать.

Преимуществами поисковой системы Апорт является:

- Содержит довольно удобный в пользовании каталог.
- Широкие возможности составления запроса.
- Автоматический перевод запроса с русского на английский язык и наоборот.

- Реконструкция проиндексированных страниц происходит из собственной базы. Это дает возможность просмотра уже несуществующих страниц.

Недостатками поисковой системы является:

- не всегда быстро находит то, что от него просишь.
- каталог не обновлялся уже очень давно.
- способен выделять один и тот же документ в различных кодировках и выдавать ссылку на него лишь один раз, перечисляя конкретные адреса в списке URL.
- не всегда корректная обработка названий страниц, из-за чего в результатах поиска часто указывается “документ без названия”, в то время как метки title на большинстве таких страниц содержат важные данные.

2.3.4 Mail.ru

Национальная почтовая служба Mail.ru – это не только поисковая система но и один из крупнейших порталов российского Интернета. Ежедневная аудитория Mail.ru - более 5 миллионов пользователей. Общее число регистраций со дня основания около 60 миллионов. Mail.ru - самый быстроразвивающийся российский Интернет-ресурс. Через почтовые ящики Mail.ru ежедневно проходит более 25 миллионов писем. Mail.ru занимает лидирующую позицию среди бесплатных почтовых сервисов, предоставляя своим пользователям почтовый ящик неограниченного размера с защитой от спама и вирусов, переводчиком, проверкой правописания, архивом для хранения фотографий и многое другое.

В 2003-м году программисты, работающие в питерском офисе американской софтверной компании DataArt, создали новое ПО для почтового веб-сервера, которое в дальнейшем предполагалось продавать западным компаниям. Чтобы протестировать сервис, его временно выложили в открытый доступ для российских пользователей, а сервис вдруг стал стремительно набирать популярность.

20 февраля 2005 года произошло слияние двух крупных игроков российского Интернет-рынка, компаний Port.ru и netBridge под брендом Port.ru. В результате объединения родилась компания, которая сразу заняла лидирующие позиции среди российских Интернет – холдингов по доле рынка и охвату аудитории.

2.4 Модель "идеальной" поисковой системы

Главный недостаток современных поисковых систем – это их централизация. А централизация означает, что вся информация хранится в одном месте, все работы и расчёты производятся в одном месте, все решения (результаты выдачи) принимаются в одном месте.

Итак, почему это недостаток, здесь несколько причин:

1) Полная централизация требует колоссальных ресурсов – это огромные базы данных, множество компьютеров и т.д. Учитывая темпы роста Интернета в ближайшем будущем придется применять просто невероятные мощности.

2) Только при управлении в одном центре можно достичь полной конфиденциальности. А так как по нашей концепции поисковая система должна быть открытой, то и необходимость в централизации отпадает полностью.

3) Поисковая система не всегда может правильно оценить конкретный ресурс. Правильнее самому обладателю сайта поручить выполнение ранжирования документов внутри сайта. И теперь, самое главное как уйти от централизации и устранить все эти минусы - это внедрение в каждый сайт своей мини поисковой системы. Эта мини поисковая система будет индексировать содержимое сайта по правилам самого обладателя сайта. Только веб-мастер будет решать, какие страницы его сайта по каким запросам более релевантные. А потом свои индексы уже будет отправлять на сервер поисковой системы.

Ещё одной из основных проблем при создании новой поисковой системы является учет мнения пользователей.

Попытка непосредственного выявления представлений пользователей об идеальной поисковой системе обычно не приводит к нужному результату: пользователи перечисляют все, что когда-либо видели или использовали в существующих системах. Не стоит ждать от пользователей навыков проектирования – они вряд ли смогут быстро описать, как должна выглядеть идеальная поисковая система.

Более продуктивным подходом к решению этой проблемы является анализ идеальной модели поисковой системы, которой оперируют пользователи. Идеальная модель – это совокупность представлений пользователя о целях, функциях, структуре, способах контроля и управления, возможных действиях с

системой, которые определяют его деятельность. Такой подход – от анализа представлений пользователей и построения идеальной модели к проектированию интерфейсов продукта - снижает риск того, что продукт не понравится пользователям, не будет принят и востребован ими.

В идеальной модели должны присутствовать следующие компоненты:

Primary nouns (электронное письмо, товар в Интернет-магазине, картинка, доступная для просмотра в Интернете) – это основные элементы, с которыми пользователь производит действия или манипуляции при работе с системой.

Сценарий использования - это описание представлений пользователей о взаимодействии с системой, разбитое на элементарные шаги. Сценарий использования иллюстрирует поведение пользователя при решении определенной задачи с помощью поисковой системы.

Диаграмма задач является графическим отображением представлений пользователей о перечне решаемых в системе задач.

Диаграмма навигации демонстрирует представления пользователей о порядке смены экранов, с которыми они сталкиваются при работе с системой, и содержании этих экранов. Диаграмма построена на основе сценариев использования системы и используется в процессе проектирования интерфейсов.

Проблема 1: Оптимизаторы не могут ясно понять, каким должен быть, «хороший» сайт в понимании поисковика и как сделать его таким, чтобы поисковик считал его наиболее релевантным по запросам.

Решение этой проблемы хорошо реализовано в поисковой системе MSN Search. В системе ранжированием занимается не только поисковик, но ему также помогает человек-редактор. Благодаря этому, при осуществлении поиск по ключевому слову, команда специалистов компании отслеживает наиболее частые запросы, вводимые в поисковую форму, и подбирает сайты, наиболее релевантные тематике запроса, а так же вручную отбирают и классифицируют их, и вносят в определенные рубрики директории. Что, например, в сравнении с самой популярной поисковой системой мира – Google, которая сама определяет релевантность Интернет-страниц (страница, на которую ссылаются чаще, более релевантна и значит более популярна) помогает избежать этой проблемы.

Проблема 2: Наличие доступных и понятно изложенных правил по специальному синтаксису каждой отдельной поисковой системы.

Изложение доступных и понятно изложенных правил по специальному синтаксису присутствует в следующих поисковых системах:

Яndex;

Google;

Апорт;

Ближе всего к идеалу находятся поисковые системы Google, Яndex, Rambler, Апорт. Отмечу также, что поисковая система MSN лидирует в системе ранжирования.

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

Основные характеристики поисковых систем:

- Полнота
- Точность
- Актуальность
- Скорость поиска
- Наглядность

В состав поисковой системы входят компоненты:

- Модуль индексирования
- База данных
- Поисковый сервер

Заключение

По итогам сделанной мной работы я могу заключить что; поисковые системы уже давно стали неотъемлемой частью Интернета. Поисковые системы сейчас – это огромные и сложные механизмы, представляющие собой не только инструмент поиска информации, но и заманчивые сферы для бизнеса.

По моему мнению, самой лучшей иностранной поисковой системой является Google, так как для меня основное значение имеет точность и полнота предоставляемых

данных. Но можно заключить также что, каждая поисковая система будь то Российская или зарубежная предоставляет различные возможности поиска, из различных баз данных, поэтому сказать точно какой именно лучше пользоваться было бы не правильно. Поэтому для удобства поиска и полноты информации следует пользоваться несколькими поисковиками вводя в них нужные запросы. По моему мнению, из многих Российских поисковиков выделяются Яндекс и Рамблер, для них характерно постоянное обновление баз данных что, обеспечивает именно актуальность и точность предоставляемой информации.

Подводя итог можно сказать что, как правило, несмотря на обилие поисковых систем, пользователь предпочитает обращаться к услугам лишь одной – двух из них (точно также как при обилии газет или новостных сайтов мы регулярно просматриваем лишь некоторые, привычные и любимые). Самой популярной поисковой системой в мире является Google. Но по оценкам аналитиков, на просторах бывшего СССР чаще используется Яндекс.

Список литературы

1. Экслер А.Б. Самоучитель работы в Интернете – Москва.: NT Press, 2009г.
2. Кузьмин А.В. Золотарева Н.Н. Поиск в Интернете – Санкт – Петербург.: Издательство НиТ, 2008г.
3. Гусев В.С. Яндекс. Эффективный поиск – Москва, Санкт – Петербург, Киев.: Диалектика,2009г.
4. Егоров А.Б. Поиск в Интернете – Санкт – Петербург.: НиТ, 2009г.
5. Гусев В.С. Поиск, Internet –Москва, Санкт – Петербург, Киев.: Диалектика, 2006г.
6. Гусев В.С. Google. Эффективный поиск – Москва, Санкт – Петербург, Киев.: Диалектика, 2009г.
7. www.citforum.ru – CIT forum, Поисковые системы в сети Интернет
8. www.ru.wikipedia.org – Википедия – свободная энциклопедия
9. www.clx.ru – Описание зарубежных поисковых систем
10. www.seop.ru – Search engine optimization project, рейтинг основных поисковых систем