Содержание:

ВВЕДЕНИЕ

На начальном этапе развития сети Интернет, количество его пользователей было относительно 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 – Интернет использовался лишь для передачи данных от пользователя к пользователю напрямую.

Все изменилось, когда появилась задача получить распределенный доступ к информационным ресурсам. Кроме того, количество информации, обрабатываемой и передаваемой через Интернет, росло в геометрической прогрессии. Именно тогда и возникла необходимость поиска нужной информации.

Сегодня сложно представить свою жизнь без поисковых систем. Количество информации, которая доступна в сети Интернет огромна. И чтобы найти необходимую – невозможно обойтись без одной из поисковых систем.

На сегодняшний день существует несколько крупных игроков среди так называемых «поисковиков». Каждый из них обладает своими положительными и отрицательными качествами. У каждого своя система ранжирования, поиска информации и так далее.

Из вышесказанного видно, что потребность максимально быстрого и эффективного поиска информации велика. Она актуальна, как для частных пользователей, так и для корпораций в рамках решения бизнес-вопросов. Для последних оптимизация особенно важна, так как возможность ускорения рабочих процессов сможет повысить производительность сотрудников в целом.

Именно поэтому выбранная мной тема курсовой работы столь актуальна сегодня.

Цель данной курсовой работы – провести анализ пяти самых популярных и узнаваемых поисковых систем сети Интернет в 2019 году, на основе которого создать сводную таблицу.

Для достижения этой цели необходимо решить ряд задач, а именно:

· изучить историю появления и развития поисковых систем,

· изучить теоретическую базу вопроса,

· определить пять наиболее крупных и популярных поисковых систем в 2019 году и изучить, как они устроены,

· выделить критерии оценки поисковых систем

· свести информацию по критериям оценки в единую таблицу.

Предметом курсовой работы являются пять выбранных поисковых систем.

Объектом курсовой работы становятся поисковые системы в целом.

Курсовая работа состоит из введения, двух глав – теоретической и практической (аналитической), заключения и списка используемой литературы. В первой главе мы рассмотрим историческую базу и теоретические аспекты поисковых систем, а во второй – сделаем выборку по самым крупным и популярным поисковым системам, проведем их анализ и сформируем сравнительную таблицу по выбранным критериям.

Глава 1

1.1. История становления и развития поисковых систем

Интернет – глобальная компьютерная сеть, которая охватывает весь мир. Сегодня Интернет имеет около 4 383 810 342 абонентов в более чем 150 странах мира. Увеличение размера сети с 2000 по 2019 составило 1114%. Интернет – является глобальной системой взаимосвязанных компьютерных сетей для связи устройств по всему миру. Если ранее сеть использовалась исключительно в качестве среды передачи файлов и сообщений электронной почты, то сегодня решаются более сложные задачи распределённого доступа к ресурсам. При низкой стоимости услуг пользователи могут получить доступ к коммерческим и некоммерческим информационным службам России, США, Канады, Австралии и многих европейских стран. В архивах свободного доступа сети Интернет можно найти информацию практически по всем сферам человеческой деятельности, начиная с новых научных

открытий до прогноза погоды на завтра.

Кроме того, Интернет предоставляет уникальные возможности дешевой, надежной и конфиденциальной глобальной связи по всему миру. Это оказывается очень удобным для фирм, имеющих свои филиалы по всему миру, транснациональных корпораций и структур управления. Обычно, использование инфраструктуры Интернет для международной связи обходится значительно дешевле прямой компьютерной связи через спутниковый канал или через телефон.

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. Основным отличием поисковой системы от своих предшественников является предоставление возможности пользователям осуществлять поиск по любым ключевым словам на любой веб-странице. Сегодня эта технология является стандартом поиска любой поисковой системы. Поисковая система «WebCrawler» стала первой системой, о которой было известно широкому кругу пользователей.

В 1995 году появились поисковые системы Lycos и AltaVista. В 1996 году AltaVista внедрила морфологическое расширение для русского языка и стала первой поисковой системой, которая была доступна русскоязычным пользователям Интернета. В этом же году были запущены первые отечественные поисковые системы – «Rambler.ru» и «Aport.ru». Появление первых отечественных поисковых систем ознаменовало новый этап развития Рунета, позволяя русскоязычным

пользователям осуществлять запрос на родном языке, а также оперативно реагировать на изменения, происходящие внутри Сети.

С запуском в 1997 году поисковой системы «Яндекс» отечественные поисковые машины начали конкурировать между собой, улучшая систему поиска и индексации сайтов, выдачи результатов, а также предлагая новые сервисы и услуги.

В 1997 году Сергей Брин и Ларри Пейдж создали поисковую машину Google в рамках исследовательского проекта в Стэндфордском университете. В настоящий момент Google – самая популярная поисковая система в мире, которая дала пользователям возможность осуществлять качественный поиск с учетом морфологии, ошибок при написании слов, а также повысить релевантность в результатах выдачи запросов. Сегодня компания Google обрабатывает более 40 миллиардов запросов в месяц, что соответствует 62,4 % всех поисковых запросов в мире.

1.2. Понятие поисковой системы, состав и принципы её работы

Поисковая система (поисковик) – сайт, где пользователь может найти интересующую его информацию по заданному ключевому запросу. Сайты и их страницы разбросаны в Internet без какого-либо порядка, без первой или последней страницы. «Читать» Интернет подряд — невозможно.

Сегодня существует множество поисковых систем, среди которых есть наиболее известные и популярные. В мировом масштабе на первом месте – Google, в русскоязычном пространстве Интернета, который еще называют Рунетом, наиболее посещаемый поисковик — Яндекс.

Далее обратим внимание на принцип работы поисковиков. Пользователь заходит на сайт поисковой системы, где ему необходимо ввести ключевую фразу, по которой он ищет необходимую информацию, в специальную форму, затем послать запрос путем нажатия кнопки поиск. После этого пользователь получает список текстовых ссылок на сайты, соответствующие этому запросу. Так выглядит принцип работы поисковика для пользователя.

Теперь необходимо изучить, как происходит процесс работы незаметный пользователю и внутреннее устройство поисковых систем.

Поисковая машина – это аппаратно-программный комплекс, который производит быстрый поиск информации по ключевой фразе внутри сервера или Интернет-ресурса. Основа поисковой машины у всех поисковых систем примерно одинаковая. Обычно, это поисковый бот, необходимый для индексации и поиска сайта, программное обеспечение, отвечающее за составление каталога запроса, и ранжирование результатов по релевантности поискового запроса. Конечно, крупные игроки среди поисковых систем всегда держат в тайне точное содержание своей поисковой машины. Ключевое отличие – это база проиндексированных сайтов, релевантность и учет морфологии языка запроса. Эти аспекты в своей совокупности определяют качество работы поисковых машин.

Классифицируется поисковая машина по области поиска информации:

1. Локальный поиск. Нужен, чтобы для реализации поиска информации по локальной сети, а также по одному сайту. Поисковый скрипт на сайте или внутренние серверы больших фирм – отличные примеры локального поиска.

2. Глобальный поиск. Нужен для поиска информации в группе сайтов, по сайтам региона или всей сети Интернет. Именно такой глобальный поиск и используют большие поисковые системы Google, Яндекс, Yahoo и аналогичные им.

Поисковые машины осуществляют свой поиск по сети Интернет в различных форматах – географическое положение, фотографии и картинки, музыка и аудиофайлы, личная информация и так далее. Это могут быть графические форматы (.gif, .png, .svg,) или мультимедийные (аудио и видео). Но именно поиск по текстовым документам является самым распространенным (web-страницы, документы в формате doc, rtf, txt и другие). Поиск по картинкам, фотографиям, видео- и аудиозаписям менее распространен, потому что это гораздо сложнее с точки зрения технологии. Системы типа Яндекс.Картинки осуществляли поиск не по непосредственно изображениям, а по альтернативным текстам, соответствующим этим изображениям. Каталог поиска картинок в Google составляется вручную. У такой технологии есть свои плюсы и минусы: релевантность запроса увеличивается, но обновление баз изображений замедляет.

Модуль индексирования поисковой машины включает в себя 3 вспомогательных программы (робота):

1. Spider (паук) – программа нужная для скачивания web-страниц. «Паук» осуществляет скачивание страницы и вынимает все внутренние ссылки с этой страницы. С каждой страницы скачивается html-код. Роботы задействуют протоколы HTTP для скачивания страниц. Принцип работы «паука» такой: робот передает запрос “get/path/document” и некоторые другие команды HTTP-запроса на сервер. В ответ робот получает текстовый поток, содержащий непосредственно сам документ и информацию служебного характера. Ссылки извлекаются из тэгов a, area, base, frame, frameset, и других. Вместе со ссылками, многими роботами также обрабатываются перенаправления (редиректы). Все страницы, которые были скачаны, сохраняются в таком формате:

- URL страницы
- Дата скачивания страницы
- http-заголовок ответа сервера
- html-код – так называемое тело страницы

1. Crawler («путешествующий» паук). Принцип его работы заключается в автоматических переходах по всем ссылкам страницы и их выделении. Задача Crawler – определить дальнейший путь паука, исходя из заранее заданного списка адресов или основываясь на ссылках. Далее он проходит по найденным ссылкам и выполняет поиск новых документов, пока еще незнакомых поисковой системе.
2. Indexer (робот-индексатор) – анализирует веб-страницы, скаченные программами Spider и Crawler. Индексатор раскладывает страницу на составные части и проводит их анализ, применяя собственные морфологические и лексические алгоритмы. Робот-интексатор анализирует разные элементы страницы. Это и заголовки, и текст, и структурные ссылки. А еще стилевые особенности, специальные служебные html-теги и так далее.

В итоге этой работы, модуль индексирования дает возможность проходить по ссылкам заданное множество ресурсов, скачивать встречающиеся страницы, извлекать ссылки на новые страницы из получаемых документов и производить их глубокий анализ.

База данных или индекс поисковой системы – система хранения данных, информационный массив, в котором хранятся особым образом преобразованные параметры всех скачанных и обработанных модулем индексирования документов.

Поисковый сервер – важнейший элемент системы. Качество и скорость потока самым прямым образом зависят от алгоритмов, лежащих в основе его функционирования.

Поисковый сервер работает следующим образом:

- Полученный от пользователя запрос подвергается морфологическому анализу. Генерируется информационное окружение каждого документа, содержащегося в базе (которое и будет впоследствии отображено в виде сниппета, то есть соответствующей запросу текстовой информации на странице выдачи результатов поиска).
- Полученные данные передаются в качестве входных параметров специальному модулю ранжирования. Происходит обработка данных по всем документам, в результате чего, для каждого документа рассчитывается собственный рейтинг, характеризующий релевантность запроса, введенного пользователем, и различных составляющих этого документа, хранящихся в индексе поисковой системы.
- В зависимости от выбора пользователя этот рейтинг может быть скорректирован дополнительными условиями (например, так называемый «расширенный поиск»).
- Далее генерируется сниппет, то есть, для каждого найденного документа из таблицы документов извлекаются заголовок, краткая аннотация, наиболее соответствующая запросу и ссылка на сам документ, причем найденные слова подсвечиваются.
- Полученные результаты поиска передаются пользователю в виде SERP (Search Engine Result Page) – страницы выдачи поисковых результатов.

Как видно, все эти компоненты тесно связаны друг с другом и работают во взаимодействии, образовывая четкий, достаточно сложный механизм работы поисковой системы, требующий огромных затрат ресурсов.

1.3. Обзор поисковых систем

Google — поисковая система, которая принадлежит корпорации Google Inc.

Именно она сейчас является лидером и самым популярным поисковиком в мире (84,65 %). Обрабатывая 41 млрд 345 млн запросов в один месяц (доля рынка 62,4 %), она индексирует более 8 миллиардов web-страниц, и способна находить

информацию на 191 языке (с 15.10.2009).

Google способна осуществлять поиск в документах RTF, PostScript, PDF, Microsoft Word, Microsoft Excel, Microsoft PowerPoint и других форматах.

Она начиналась, как учебный проект двух талантливых студентов Стендфорского университета. Их звали Лари Пейдж и Сергей Брин. Они смогли предложить новую поисковую систему, которая в настоящий момент стала одной из наиболее известных и влиятельных компаний во всемирной сети Интернет.

История названия поисковика представляет особенный интерес. В его основе наименование математической величины гугол (от англ. googol) — число, в десятичной системе счисления изображаемое единицей с сотней нулей. Идея создателей заключалась в организации миллиардов байтов информации, которая содержится в сети Интернет, и название Google как нельзя лучше смогло передать суть их детища.

Поисковая система Google представляет собой мощный механизм. Без таких поисковых систем найти информацию в глобальной сети Интернет было бы крайне невозможно. Как и все поисковые системы, Google использует специальный поисковый алгоритм для получения результатов поиска. Часть базовых характеристик своего алгоритма компания не скрывает, но конечно отличительные особенности своего алгоритма являются объектом конфиденциальной, строго закрытой, информации. Такая политика компании позволяет Google сохранять лидирующие позиции в сети Интернет и защищает систему от взлома.

Как и большая часть поисковых систем, Google использует программы-пауки для автоматического выбора всех документов, на которые есть ссылки в первом выбранном документе. В специальною строчку вписываются ключевые слова, и стартует поиск. По какому критерию и как Google классифицирует итоги поиска на своей странице – особенность данной поисковой системы. В ней используется алгоритм PageRank, который занимается сортировкой всех web-страниц по смысловому соответствию.

Факторы, от которых зависит работа PageRank:

- От частоты повторов и местоположения ключевых слов на web-странице – если заявленное в строке поиска слово или словосочетание встречается на сайте лишь 1 раз, то страница отмечается низким балом.

- От времени существования сайта – новые появляются в Интернете ежедневно, но лишь часть из них задерживается на длительный срок. Поэтому алгоритм отдает преимущество тем сайтам, которые успели зарекомендовать себя в течение длительного срока.
- От количества web-страниц, связанных с «главной страницей» - поисковик считывает сколько страниц относится к этому сайту, на основании чего определяет её рейтинг среди всех прочих.

Обмануть систему Google практически невозможно, потому что она воспринимает все ссылки на web-страницы как «голоса». Как следствие, самый оптимальный метод сделать так, чтобы ваш сайт оказался в топе и на первой странице поиска – это наполнить его наиболее разнообразной информацией, которая сможет привлечь много разной аудитории. Не последнюю роль играют и ссылки: чем их больше на вашей страничке, тем выше её оценит Google и в частности поисковый агент PageRank.

До 2011 года для части результатов поиска Google давал дополнительное поле для поиска, которое обладало функцией находить информацию внутри конкретного сайта. Сегодня такой опции не существует для пользователей, чем многие были недовольны, так как она была удобна и пользовалась популярностью.

Весной 2009 года была запущена поисковая технология «Википоиск». Она давала пользователю возможность самостоятельно настраивать результаты поиска под себя – он мог сам удалять результаты из полученного списка выдачи, и даже поднимать их на более высокие строки списка. Как и вышеописанная опция, «Википоиск» продержалась недолго – до осени того же года.

В компании Google развитие не останавливается ни на минуту: существует целый сегмент бесплатных сервисов от Google, которые часто не требуют даже установки дополнительного программного обеспечения на персональный компьютер пользователя. Особую популярность заслужили такие сервисы, как «Gmail» и «Gtalk». Оба проекта отлично работают, как в связке, так и по отдельности. «Gmail» - почтовый сервис, который умеет автоматически фильтровать спам, располагает большим объемом почтового ящика и имеет, удобный для многих, мобильный доступ. «Gtalk» – сервис, который дает возможность обмена сообщениями – как текстовыми, так и голосовыми, причем по вашему желанию и в окне браузера, и с помощью специального программного обеспечения.

Также очень популярны сервисы контекстной рекламы «AdSense» и «AdWords». Ими пользуются владельцы разных популярных сайтов, чтобы монетизировать их – заработок строится на посещаемости страниц. Кроме того, можно привлечь и новых посетителей.

Для простых пользователей есть сервисы, которые дают доступ к новостям и справочной информации самой разной направленности и тематики, обмену картинками и фотографиями и многим иным ресурсам.

Yahoo! — американский поисковик, который находится на второй по популярности позиции в мире (6.35 %). Компания Yahoo! Также предлагает рынку линейку сервисов, объединённых интернет-порталом «Yahoo! Directory». В него входит и один из старейших и наиболее популярных в интернете серис электронной почты под названием «Yahoo! Mail». Существует и версия почтового интерфейса, которая основана на AJAX (русскоязычный обзор нового интерфейса).

История Yahoo! началась в январе 1994 года с создания web-сайта под названием «Путеводитель Джерри по Всемирной Паутине». Это был каталог разных других сайтов. Его авторами были Джерри Янг и Дэвид Файло. Уже через 3 месяца создатели переименовали «Путеводитель» в известный сегодня Yahoo!.

Существуют две различные истории о происхождении названия компании. Первой версии придерживаются создатели – Джерри Янг и Дэвид Файло. Её суть в том, что слово Yahoo! было взято из романа Джонатана Свифта «Путешествия Гулливера». Оно обозначало расу грубых и тупых человекообразных существ (в русскоязычной версии звучит, как Йеху).

Есть вторая версия. Придерживаясь её Yahoo! – это аббревиатура, которая была образована от фразы «Yet Another Hierarchical Officious Oracle». В приблизительном переводе на русский она означает «Еще один иерархический неотесанный (неофициальный) прорицатель».

Но существует и третья версия происхождения названия. В Японии есть слово Yahhoo, которое обозначает неформальный вариант значения слова «Привет». Возможно название поисковой системы Yahoo! было заимствовано именно из этого источника. Можно отметить, что Yahoo уже существовало в качестве зарегистрированной торговой марки, под которой продавался соус для барбекю. Поэтому Джерри Янг и Дэвид Файло добавили к названию один восклицательный знак.

Уже 2 марта 1995 года Yahoo! стал корпорацией.

По данным статистики Alexa Internet, сайт Yahoo! сегодня находится на четвертой строке по посещаемости в сети Интернет в мире. Около 28% посещений – просмотр лишь одной только страницы.

Bing — поисковая система от международной корпорации Microsoft. Но ранее она имела совсем другие названия:

- MSN Search — с момента появления и до 11 сентября 2006;
- Windows Live Search — до 21 марта 2007;
- Live Search — до 1 июня 2009.

Сегодня Bing находится на третьей строчке рейтинга самых популярных поисковых систем. Но в отличии от своих конкурентов с первой и второй строки, Bing имеет ряд уникальных возможностей. Например, вместо того, чтобы пролистывать множество страниц с результатами поиска, с Bing их можно посмотреть на 1-ой странице. Кроме того, здесь существует динамическое корректирование объёма информации, отображаемой для каждого результата поиска – это может быть только название, а также сводка большого или малого размера).

В американской версии Bing есть определенные интересные новшества относительно поиска, среди них:

- темы оформления стартовой страницы, которые меняются каждый день, плюс есть информационные блоки;
- вывод уточняющих вариантов поисковых запросов по отдельным категориям;
- видео с запуском предварительного просмотра, это происходит автоматически;
- по каждому результату поиска предоставляются дополнительные данные;
- для поиска маршрутов есть отдельный встроенный сервис;
- дополнительные функции, которые делают поиск информации, изображений и видео более удобным.

Не смотря на отличные показатели точности поиска при вводе запросов на английском, важно отметить, что для русскоговорящих пользователей Bing практически бесполезен. В России и странах, поддерживающих русский язык, наиболее релевантный результат выдает «Яндекс» и «Google».

Поисковая система «Яндекс» к началу 2013 года сзаняла четвертую строчку в рейтинге популярнейших поисковиков планеты (после Google, китайского Baidu и Yahoo!) с 4,84 млрд поисковых запросов, причём она стала самым быстрорастущим из ТОП-5.

Поисковая система Yandex.ru была заявлена официально 23 сентября 1997 года, и вначале осуществляла развитие в рамках компании CompTek International. Образование отдетьной компании "Яндекс" произошло только в 2000 году.

«Яндекс» постоянно совершенствует свои поисковые алгоритмы. Это дает ему возможность всегда отвечать самым актуальным и продвинутым критериям поиска и быть на одном уровне с компанией «Google» хотя бы на российском рынке. Сегодня это именно так, если оценивать уровень освоения обоих поисковых систем аудиторией русскоговорящих стран. «Яндекс», как и «Google», работает на кластерной системе организации компьютерных вычислительных сетей. Каждый кластер отвечает за определённый сегмент сохранённой информации.

Сканирующие роботы поисковой системы бывают 2 видов:

1) основной сканирующий робот

2) быстрый робот – он регулярно сканирует сайты, где скорость и частота обновления информации крайне велика. Робот добавляет результаты поиска с этих сайтов в поисковую систему, что обеспечивает быстрое обновление её индекса.

Два вида апдейтов (обновления) поисковой системы:

1) Апдейт поисковой базы. В результатах поиска, собранных основным поисковым роботом, начинают появляться обновленные страницы разных сайтов. Это происходит обычно несколько раз за один месяц.

2) Апдейт программной части (движка) поисковой системы. Здесь смысл заключается в изменениях алгоритмов ранжирования документов в поисковой системе. Подобные обновления обычно получают собственные названия, их появление анонсируется.

Очень важный момент «Яндекса» заключается в том, что он учитывает морфологию русского языка. Он обладает системой определения словоформ, причем довольно сильной. Кроме того, «Яндекс» позволяет сузить запрос до предельно точного – это стало возможно благодаря использованию особых поисковых формул и геотаргетинга; имеет свой специальный алгоритм оценки релевантности –

точности результата запроса по отношению к самому запросу – который работает на очень высоком уровне. Плюс «Яндекс» отличается крайне высокой скоростью реакции на поисковые запросы при практически полном отсутствии перегрузки своих серверов.

С появлением алгоритма "Снежинск" «Яндекс» научилась определять регион сайтов, благодаря этому стало возможно выводить результат поиска по географии пользователя. В настоящий момент «Яндекс» по праву можно назвать наиболее точным поисковиком Рунета по географическому критерию.

Вместе с тем, интернет-портал «Яндекс» - это далеко не только сильнейшая поисковая система. Под маркой собрано огромное количество самых разнообразных удобных сервисов из разных сфер.

Так, с помощью «Яндекса» вы можете узнать свежие новости («Новости»), связаться и пообщаться с друзьями и коллегами («Блоги», «Почта»), заработать («Мой Круг», «Директ», «Рекламная сеть»), продать или приобрести различные товары («Маркет», «Авто»), а также получить море полезной информации: среди них карты, пробки и схемы метро, афиша мероприятий и программа телепередач, очень популярный сервис такси, сервис прогноза погоды и валютных котировок. Очень известна и востребована у русскоязычных пользователей платежная система «Яндекс Деньги», которая позволяет осуществлять электронные платежи с помощью web-интерфейса или Интернет-кошелька. Совсем недавно появился, стремительно набирающий популярность, сервис Яндекс.Еда.

«Rambler» – поисковая система, появившаяся еще в 1996 году, была разработана, так как создатели понимали, что иностранные поисковые системы часто крайне плохо работали с русским языком и web-страницами с несколькими кодировками, а глубина индексирования страниц Рунета, была очень низкой.

До 2011 года «Rambler» был первым по популярности поисковиком Рунета. И несмотря на то, что сегодня на первом месте «Яндекс» и «Google», поисковик «Rambler» занимает уверенную позицию в Рунете – на его долю приходится 20-25% русскоязычных поисковых запросов.

Название поисковика «Rambler» переводится с английского, как бродяга, странник или даже праздношатающийся человек. Всё это неплохо отражает деятельность компании. «Rambler» позволяет искать информацию на различных языках, среди которых русский, украинский, английский, казахский и многие другие. Поисковик умеет работать со словоформами, а также приводить полученные результаты

поиска в структуру по уровню релевантности.

В начале 2009 года был внедрен алгоритм вертикального поиска, в его основе технология XAG (eXtended AGgregator), с помощью которого появилась возможность отсортировать результаты поиска по темам, что на порядок упростило использование поисковика. Интересная особенность и несомненно преимущество алгоритма вертикального поиска заключается в том, что, если в найденном документе недостаточно информации, она может быть дополнена данными из другого документа. Это позволяет еще и очистить данные от повтора и спама. Так, при поиске вакансий по телефонному номеру компании «Rambler» определяет ее название с помощью чего способен увидеть на иных web-страницах дубликаты объявлений и сомнительные вакансии.

Отличительная особенность «Rambler» - обслуживание только сайтов, находящиеся в следующих доменах первого уровня: Российская Федерация: .ru, .su; Украина: .ua; Белоруссия: .by; Казахстан: .kz; Киргизия: .kg; Узбекистан: .uz; Грузия: .ge.

Как и вышеперечисленные компании, «Rambler» - не только поисковая система. В рамках компании было запущено большое количество проектов и сервисов с возможностями: посетить наиболее популярные сайты, послушать музыку и посмотреть видео, узнать новости и другую полезную информацию, завести новые знакомства и пообщаться.

Mail.ru — один из крупных игроков Рунета, который принадлежит инвестиционной группе Mail.ru Group. Его аудитория сегодня превышает 80 миллионов уникальных и активных пользователей в месяц. По данным Alexa, на октябрь 2018 года сайт портала занимает 36-е место в мировом рейтинге, а также пятое место в России.

В 1998 году разработчики американской компании DataArt, которые работали в российском подразделении в Санкт-Петербурге, создали новое программное обеспечение для почтового web-сервера, которое в дальнейшем предполагалось продавать западным компаниям. Есть версия, что собственники были вдохновлены примером роста и развития компании Yahoo!, которая объединила на одном сайте поисковую систему, сервис электронной почты и информационные блоки. Изначально сервис Mail.ru выложили в открытый доступ для тестирования в сегменте российских пользователей, но сервис так стремительно набирал известность и посещаемость, что остался самостоятельной единицей.

Очень популярны и дополнительные сервисы, и разделы компании, среди них такие тематические проекты, как:

«Авто Mail.ru» - сайт автомобильной тематики,

«Кино Mail.ru» - онлайн-кинотеатр с возможностью купить отдельный продукт или подписку,

«Дети Mail.ru» - тематический сайт о правильном и здоровом воспитании детей, о беременности, родах и семье в целом,

«Здоровье Mail.ru» - тематический медицинский портал,

«Леди Mail.ru» - женский тематический сайт, где обсуждаются темы моды, стиля, красоты, звезд шоу-бизнеса, отношений, психологии и многие другие.

«Новости Mail.ru» - агрегатор новостей российских изданий,

«Спорт Mail.ru» - агрегатор новостей спорта,

«Hi-Tech Mail.ru» - новости в сфере технологий и потребительской электроники,

«Недвижимость Mail.ru» - сервис поиска по объявлениям о покупке, продаже и сдаче в аренду жилой недвижимости.

GoGo.ru – автономный проект от разработчиков Mail.ru. Соответствующий домен был зарегистрирован еще в 2000-м году, но сама разработка началась только в 2006 году. А уже в 2007 году поисковая система была запущена. В начале было немало проблем: низкие охваты, странные результаты поиска. Работа над оптимизацией всех аспектов поисковой машины велась постоянно и в 2008 году GoGo.ru была способна осуществлять поис по более чем 2,5 млр. документов, 140 млн. изображений, 2 млн. видео-файлов, и так далее.

Главным отличием GoGo.ru от других поисковиков стала возможность:

- поиска по русскоязычным видеороликам и WAP-сайтам,

- анализа базы данных Ответы@Mail.ru

- а также набор функций для web-мастеров и владельцев собственных Интернет-проектов.

GoGo.ru осуществляет поиск по видео-хостерам: Video.Mail.ru, RuTube.ru, LiveInternet.ru, Teledu.ru, Smotri.com, Myvi.ru, Video.i.ua и некоторые другие.

А при поиске изображений робот GoGo.ru способен отличать фотографии от прочих картинок, благодаря особому встроенному фильтру.

Высокая эффективность графического поиска в GoGo.ru обеспечивается наличием XML-синдикации с ведущими фотохостерами Рунета.

Важно отметить, что поисковик способен работать со словоформами и синонимами. Причем словарь поисковой машины пополняется в полуавтоматическом режиме.

Таким образом, на рынке представлено довольно большое количество поисковых систем. Благодаря конкуренции между ними, мы можем наблюдать постоянное совершенствование «поисковиков»: улучшение алгоритмов поиска, оптимизацию времени выдачи поисковых запросов, расширение морфологии.

В следующей главе будут проанализированы доли рынка известных поисковых систем, наиболее популярные будут выбраны для дальнейшего анализа эффективности.

Глава 2

2.1. Статистика популярности поисковых систем

На сегодняшний день поисковые системы являются сложнейшими и громадными механизмами. В рамках данной курсовой работы были собраны актуальные статистические данные по популярным мировым поисковым системам.

Рейтинг популярных систем мира по данным исследовательской компании NetMarketShare в период с мая 2018 по май 2019 возглавляет Google (78,36%), на втором месте китайская поисковая система – Baidu (13,37%). Тройу лидеров замыкает Bing (4,49%). На четвертом и пятом местах расположились Yahoo! (2,18%) и единственный в рейтинге российский поисковик Yandex (0,79%). На Рисунке 1 наглядно представлено распределение долей.

Рисунок 1. Рейтинг поисковых систем мира по популярности

(май 2018 – май 2019)

В России наблюдается несколько иная картина. По данным российского онлайн-сервиса Liveinternet за период с марта по май 2019 выявлена следующая ситуация: на первом месте всё также Google (54,9%), на втором – Yandex (42,3%), на третьем – Search.Mail.ru (2,5%).

Далее Rambler и Bing с одинаковым значением – 0,1%.

Наглядное распределение на Рисунке 2.

Рисунок 2. Рейтинг поисковых систем России по популярности

(март 2019 – май 2019)

2.2. Анализ поисковых систем по критериям эффективности

В прошлом параграфе мы выявили ТОП-5 поисковых систем в России и в мире. В данной курсовой работе хотелось подробнее остановиться на популярных поисковых системах именно нашей страны.

Один из важнейших критериев качества поисковых систем – релевантность, которая включает в себя несколько показателей. Наиболее интересные из них – полнота и точность поиска. Точность определяется соотношением между найденными релевантными и нерелевантными документами, а полнота поиска - общим количеством найденных документов. Релевантным будем считать документ, который удовлетворяет запросу пользователя. Нерелевантным – тот, который не смог удовлетворить запрос пользователя.

Для анализа нам необходимо назначить весовые коэффициенты - параметры, которые отражают в сравнении с другими критериями относительную важность, значимость, «вес» данных критериев. Сумма всех весов должна быть равной 1, поэтому для точности поиска весовому коэффициенту даем значение, равное 0.8, для полноты поиска - 0.2. Оформим результаты в виде Таблицы 1.

Таблица 1

Весовые коэффициенты

Критерий	**Весовой коэффициент**
Точность поиска	0,8
Полнота поиска	0,2

Были сформулированы пятнадцать запросов на различные темы. Каждый запрос был выполнен в каждой из пяти исследуемых поисковых системах. Из полученных списков результатов была получена следующая информация:

1. Общее количество найденных документов (Д).

2. Количество релевантных документов различной ценности (РД)

Количество релевантных документов оценивается при просмотре текста первых 10 найденных документов. Также определяется ценность найденной информации (степень удовлетворения найденным документом информационных потребностей). Ценность информации оценивается по 3-х бальной шкале: 2 балла – информация имеет ценность, 1 балл – информация имеет частичную ценность, 0 баллов – информация не имеет ценности. Результаты выполнения запросов были сведены в Таблицу 2.

Таблица 2

Результаты выполнения запросов

	Bing			**Google**			**Mail.ru**			**Rambler**			**Yandex**		
№	Д	РД		Д	РД		Д	РД		Д	РД		Д	РД	
		2	1 0		2	1 0		2	1 0		2	1 0		2	1 0
1	111015	6	4 0	7 240 000	10	0 0	3661995	5	2 3	5266123	9	1 0	6071953	9	0 1

2	216054	8	2	0	9988671	10	0	0	872111	8	1	1	2575905	10	0	0	2026800	10	0	0
3	420554	7	2	1	2022025	10	0	0	7331185	9	1	0	6198330	8	0	2	7966970	8	1	1
4	620689	8	2	0	9640000	9	0	1	9040318	8	1	1	9040318	9	1	0	9586458	9	1	0
5	2002188	8	1	1	15669000	9	0	1	7330374	7	1	2	8318276	9	0	1	9977900	9	1	0
6	487775	7	0	3	2380000	9	0	1	2575905	9	0	1	4242846	10	0	0	6155744	9	0	1
7	746000	9	0	1	4301003	8	0	2	962075	10	0	0	2776226	9	0	1	4749756	8	1	1
8	196987	7	2	1	900347	8	0	2	246098	9	1	0	735288	8	0	2	829904	8	0	2
9	425696	9	1	0	1320500	8	1	1	983974	7	1	2	9045322	9	0	1	8275010	10	0	0
10	999548	10	0	0	16874000	9	1	0	1989016	9	1	0	7981997	10	0	0	9977900	9	1	0
11	102178	10	0	0	4008750	8	1	1	882097	10	0	0	4719405	10	0	0	9323589	9	1	0
12	7326587	7	0	3	1874000	9	0	1	734819	9	0	1	9098659	8	0	2	9147106	8	0	2
13	115644	8	1	1	5440060	8	0	2	789022	9	0	1	5545995	8	0	2	1218709	8	0	2
14	259300	8	1	1	3971000	9	0	1	882097	7	1	2	3201308	8	1	1	425696	8	1	1
15	541100	7	1	2	4982000	9	0	1	930344	9	0	1	5266123	10	0	0	832686	10	0	0

Для нахождения наиболее эффективной поисковой системы для начала вычислим средние арифметические значения показателей для каждой поисковой системы Д,

РД(0), РД(1) и РД(2).

Далее необходимо определить место каждой поисковой системы по критерию "Полнота поиска".

Для его определения места будем использовать среднее количество найденных документов Д. Наилучшей считается та система, которая нашла больше документов. Ей присваивается первое место, самой худшей – место N (где N – это количество всех исследуемых систем).

Коэффициент точности поиска P для каждой поисковой системы определим по формуле:

P = a/(a+b)

a – число релевантных документов, которые выдала поисковая система в ответ на запрос.

a = 0.5 * РД(1) + РД(2)

b – число документов, которые полностью не имеют ценность, b = РД(0).

Далее необходимо определить место каждой поисковой системы по заданному критерию "Точность поиска". Лучшей будет считаться система, которая имеет наибольшее значение коэффициента точности поиска P. Ей присваивается первое место, а самой худшей – место N (где N – это количество исследуемых систем).

Следующим шагом будет вычисление коэффициента поискового шума S по формуле:

S=1 – P

В заключении необходимо вычислить по следующей формуле рейтинг каждой исследуемой системы R:

i - номер критерия оценки поисковой системы,

m - это количество критериев оценки,

wi - весовой коэффициент для критерия оценки i,

qi - это место ПС по критерию оценки i,

N - общее количество исследуемых систем.

Для первичной обработки информации данные были сведены в Таблицу 3.

Таблица 3

Результаты сравнительного анализа поисковых систем

Критерий	**Bing**	**Google**	**Mail.ru**	**Rambler**	**Yandex**
Полнота поиска (Д)	971421	6040757,1	2614095,3	5600808,1	5771078,7
Место (полнота поиска)	5	1	4	3	2
Среднее количество пертинентных документов (РД2)	7,9333333	9,0666667	8,3333333	8,8666667	8,8
Среднее количество частично пертинентных документов (РД1)	1,1333333	0,0666667	0,6666667	0,2	0,4666667
Среднее количество непертинентных документов (РД0)	0,9333333	0,8666667	1	0,9333333	0,7333333
Коэффициент точности поиска (Р)	0,9010601	0,9130435	0,8965517	0,9057239	0,9249147

Место (точность поиска)	4	2	5	3	1
Коэффициент поискового шума (S)	0,0989399	0,0869565	0,1034483	0,0942761	0,0750853
Рейтинг (R)	2,3	6,1	3,8	4,8	5,6

На основании проведенного анализа можно сделать вывод, что по показателям «Полнота поиска» и «Среднее количество пертинентных документов» на первое место выходит «поисковик» от Google. А по показателям «Среднее количество непертинентных документов», «Коэффициент точности поиска» и «Коэффициент поискового шума» лидерство у поисковой системы российской компании «Яндекс». По мнению автора курсовой работы первые два коэффициента являются более важными, чем три последних коэффициента, поэтому по результатам анализа наиболее эффективной поисковой системой стоит считать «Google». Однако «Яндекс» показал очень близкие к лучшим показатели критериев оценки. Это говорит о том, что он создает достойную конкуренцию лидеру анализа.

ЗАКЛЮЧЕНИЕ

В наше время информация играет огромную роль во всех сферах жизни. А Интернет является самым популярным способом её поиска. Каждому из нас – для личных и профессиональных целей – важно быстро находить в этом огромном потоке информации действительно нужную. Без помощи поисковой системы это было бы невозможно. Благодаря удобству в обращении и хорошим техническим характеристикам, различные поисковые системы могут помочь в этом и новичку, и опытному пользователю.

Рост информационного потока и развитие информационных технологий не останавливается ни на минуту, следовательно, и в будущем нельзя будет обойтись без поисковых систем.

Острая конкуренция среди поисковых систем является гарантом совершенствования технологий поиска и его соответствия нашим нуждам.

Как показывает статистика, пользователи русскоязычной части Интернета предпочитают несколько поисковых машин:

- Google,

- Яндекс,

- Mail,

- Rambler,

- Bing.

В ходе работы были выделен основной критерий качества поисковых систем – релевантность, который включает в себя точность, актуальность, полноту и ценность полученной информации.

По итогам анализа, который является целью данной курсовой работы, наиболее эффективным по выделенным критериям стала поисковая система «Google». Очень близка к ней по важнейшим показателям поисковая система «Яндекс».

Поставленные в начале работы цели и задачи выполнены в рамках имеющейся в открытом доступе информации о предмете и объекте курсовой работы.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Search Engine Market Share. - URL: netmarketshare.com

2. Статистика сайта. Переходы из поисковых систем: [Электронный ресурс]. М. - URL: liveinternet.ru

3. Мировые информационные ресурсы[Текст]: Учебное пособие/ В.К.Иванов; под. ред .В. К.Иванова.- Тверь:Изд-во ин-та ТвГТУ, 2012. - 37с.