Содержание:

Введение

В век информационных технологий огромную роль играет интернет, а любое путешествие по просторам интернета невозможно без поисковых систем, позволяющих комфортно просматривать любимые веб страницы. Поисковая система – это аппаратно-программный комплекс, который предназначен для осуществления функции поиска в интернете, и реагирующий на пользовательский запрос который обычно задают в виде какой-либо текстовой фразы (или точнее поискового запроса), выдачей ссылочного списка на информационные источники, осуществляющейся по релевантности.

В этой работе будет проведён анализ и сравнение наиболее популярных поисковых систем.

3

1 глава.

Принципы работы поисковой системы

Все большие системы поиска имеют свою структуру, которая весьма отличается от других. Но все-таки можно выделить общие для всех поисковиков основные элементы:

1.

Модуль индексирования.

Данный компонент состоит из трех программ-роботов:

1.1 Spider (по англ. паук) – программа которая предназначена для того чтобы скачивать веб-страницы. «Паук» скачивает определенную страницу, одновременно извлекая из нее все ссылки. Скачивается код html практически с каждой страницы. Для этого роботы используют HTTP-протоколы

«Паук» функционирует следующим образом. Робот передает запрос на сервер “get/path/document” и иные команды запроса HTTP. В ответ программа-робот получает поток текста, который содержит информацию служебного вида и, естественно, сам документ.

Извлекаются все ссылки из тэгов. Вместе с ними обрабатывают редиректы. Любая скачанная страница сохраняется в таком формате:

URL скаченной страницы;

дата, когда осуществлялось скачивание страницы;

заголовок http-ответа сервера;

html-код, «тела» страницы.

1.2 Crawler («путешествующий» паук). Данная программа автоматически заходит на все ссылки, которые найдены на странице, а также выделяет их. Его задача – определиться, куда в дальнейшем должен заходить паук, основываясь на этих ссылках или исходя из заданного списка адресов.

Crawler, исследуя найденные ссылки, ищет новые документы, еще не ставшие известными поисковой системе.

1.3 Indexer (робот-индексатор) – это программа, анализирующая страницы, которые скачали пауки.

Индексатор полностью разбирает страницу на составные элементы и проводит их анализ, применяя свои морфологические и лексические виды алгоритмов.

Анализ проводится над разнообразными частями страницы, такими как заголовки, текст, ссылки, стилевые и структурные особенности, теги html и так далее.

Таким образом, модуль индексирования дает возможность проходить по ссылкам заданного количества ресурсов, скачивать страницы, извлекать ссылочную массу на новые страницы из полученных документов и делать подробный их анализ.

2.База данных

База данных (или индекс поисковика) - комплекс хранения данных, массив информации в котором сохраняются определенным образом переделанные параметры каждого обработанного модулем индексации и скачанного документа.

3.Поисковый сервер

Это самый важный элемент всей системы, потому что от алгоритмов, лежащих в основе ее функциональности, прямо зависит скорость и, конечно же, качество поиска.

Поисковый сервер работает следующим образом:

Запрос, который идет от пользователя подвергается морфологическому анализу. Информационное окружение любого документа, имеющегося в базе, генерируется (оно и будет в дальнейшем отображаться как сниппет, то есть информационное поле текста соответствующего данному запросу).

Полученные данные передают как входные параметры специализированному модулю ранжирования. Они обрабатываются по всем документам, и в итоге для каждого такого документа рассчитывается свой рейтинг, который характеризует релевантность такого документа запросу пользователя, и иных составляющих.

В зависимости от условий заданных пользователем этот рейтинг вполне может быть подкорректирован дополнительными.

Затем генерируется сам сниппет, т.е. для любого найденного документа из соответствующей таблицы извлекают заголовок, аннотацию, наиболее отвечающую запросу, и ссылка на этот документ, при этом найденные словоформы и слова подсвечивают.

Результаты полученного поиска передаются осуществившему его человеку в виде страницы, на которую выдают поисковые результаты (SERP).

Все эти элементы тесно связаны между собой и функционируют, взаимодействуя, образовывая отчетливый, но достаточно непростой механизм

функционирования поисковой системы, требующий громадных затрат ресурсов.

2 глава.

Немного истории

Самым первым посковиком был поисковик Арчи – это первая в мире программа, индексирующая Интернет, благодаря чему и считается «дедушкой всех поисковиков». Он был изобретен Аланом Эмтеджем, студентом-компьютерщиком из Барбадоса, обучавшимся в университете МакГилл. До появления Арчи единственным способом найти что-то в Интернете или на FTP-серверах было спросить у кого-нибудь или получить email с указанием того, куда идти за информацией.С помощью Роберта Кальяу Тим Бернерс-Ли пишет первый www-вебсервер. Он выходит онлайн через компьютерную сеть под названием Интернет летом 1991 года. Чтобы строить свои базы данных по сайтам, поисковые машины должны регулярно ходить по вебу. В 1993 году Мэтью Грей представил миру программу World Wide Web Wanderer. Изначально он хотел померять размеры Интернета и создал этот бот, чтобы посчитать активные веб-серверы. Вскоре он проапгрейдил его так, чтобы тот смог собирать URL-ы сайтов. Полученная им база данных стала называться Wandex(созвучно с Яндекс не спроста).

В 1996 году компания Netscape хотела заключить эксклюзивную сделку с одной из поисковых систем, сделав её поисковой системой по умолчанию на веб-браузере Netscape. Это вызвало настолько большой интерес, что Netscape заключила контракт сразу с пятью крупнейшими поисковыми системами (Yahoo!, Magellan, Lycos, Infoseek и Excite). За 5 млн долларов США в год они предлагались по очереди на поисковой странице Netscape.

Студенты последнего курса компьютерного направления Стэнфордского университета Лари Пейдж и Сергей Брин начинают сотрудничество по разработке поисковой системы BackRub. Этот проект произведет революцию в веб-поиске, потому что BackRub будет учитывать ссылки на сайт и ранжировать сайты в соответствии с ними.В 1997 году Пейдж и Брин переименовали BackRub в Google. Они произвели это имя от числительного «гугол», которое в математике используется для обозначения чисел с сотней нулей. Такое название отразило их миссию – организовать кажущееся бесконечным количество информации в вебе.

Google взял на вооружение идею продажи ключевых слов в 1998 году, тогда это была маленькая компания, обеспечивавшая работу поисковой системы по

7

адресу goto.com. Этот шаг ознаменовал для поисковых систем переход от соревнований друг с другом к одному из самых выгодных коммерческих предприятий в Интернете. Поисковые системы стали продавать первые места в результатах поиска отдельным компаниям)

На сегодняшний день поисковые системы являются сложнейшими механизмами, представляющие собой не только инструмент для нахождения любой необходимой информации, но и довольно увлекательные сферы для бизнеса. Работа с помощью поисковых систем позволяет многим пользователям глобальной сети осуществлять быстрый поиск нужной информации в кратчайшие сроки.

Для поиска информации с помощью поисковой системы пользователь формулирует поисковый запрос. Работа поисковой системы заключается в том, чтобы по запросу пользователя найти документы, содержащие либо указанные ключевые слова, либо слова, как-либо связанные с ключевыми словами. При этом поисковая система генерирует страницу результатов поиска. Такая поисковая выдача может содержать различные типы результатов, например: веб-страницы, изображения, аудиофайлы. Некоторые поисковые системы также извлекают информацию из подходящих баз данных и каталогов ресурсов в Интернете.

Поисковая система тем лучше, чем больше документов, подходящих запросу пользователя, она будет возвращать. Результаты поиска могут становиться менее релевантными из-за особенностей алгоритмов или вследствие человеческого фактора. Самой популярной поисковой системой в мире является Google(google.com) 90.2%, на втором месте Bing (bing.com) 3,23%, на третьем китайская поисковая система Baidu (baidu.com) 2,2%, далее идут Yahoo! (yahoo.com) 2,09% и Российская Yandex(yandex.ru) 0,80% (рисунок 1).

8

В русскоязычном сегменте ситуация немного отличается. В лидерах находится Yandex 50.44%, на втором месте Google 46.06%, третьи Mail.ru 2.18%, далее bing, Yahoo, DuckDuckGo и Baidu (все вместе около 1.3%) (рисунок 2).

9

3 глава.

Рассмотрим самые популярные поисковые системы:

Google.

Название происходит от числа гугол — единица со 100 нулями (намёк на то, что поисковик может найти ответ на гугол вопросов и найти гугол вещей).В общем-то, поначалу было предложение назвать поиск GooglePlex (в правильном написании Googolplex — десятка в степени гугол), но оно показалось слишком длинным и остановились на упомянутом термине (почему-то написанном в результате не правильно, но менять уже не стали, т.к. был почти сразу же приобретен домен Google.com).

Историю развития этой компании можно начинать отсчитывать с 1996 года, хотя официально поисковая система начала работать только с осени 1998 года. Имеющиеся в то время поисковики с трудом справлялись со своей задачей. Результаты поисковой выдачи имели очень низкую корреляцию с тем, что хотел увидеть в ответ на свой запрос пользователь. Дело в том, что тогда основным фактором, по которому осуществлялось определение релевантности и ранжирование документов в выдаче, была частота использования слов из запроса пользователя в документе. Понятно, что такой критерий отбора очень легко поддается накрутке со стороны вебмастеров простым увеличением "тошноты" текстов. Сложно поверить, сколько времени уже прошло с тех пор, как появился текстовый спам, с которым поисковики только сейчас начали всерьез бороться и искоренять (ибо появились другие факторы, позволяющие существенно снизить важность частоты вхождения ключей в тексте при ранжировании). Ну, вот. Ларри Пейдж с детства на примере своих родителей, вращавшихся в научных кругах, видел и понимал, что авторитет того или иного ученого во многом зависит от того, в скольких научных работах на него ссылаются, как на первоисточник или как на авторитетного специалиста. Чем больше ссылок, тем авторитетнее имя

ученого.Тогда у Пейджа возникла идея перенести эту систему ранжирования на поиск в интернете. Ученых он ассоциировал с отдельными документами

(не сайтами, а именно отдельными вебстраницами), а ссылки в интернете

существовали аж с момента изобретения всемирной паутины WWW Тимом Бернерсом-Ли в далеком 1989 году (кстати, спустя пару лет, именно Тим основал консорциум W3C и разработал язык Html).

В результате появился фактор ранжирования, который учитывается поисковиками до сих пор — PageRank. Термин этот составной. Rank — означает ранжирование, а вот Page может означать либо веб страницу в английской вариации написания, либо то, что данный параметр ранжирования придумал никто иной как Пейдж.

PageRank совершил революцию и позволил поднять качество поиска будущего Google на недосягаемую высоту. Он позволял учитывать при ранжировании документов не только количество, но и качество ведущих на ту или иную вебстраницу ссылок. Ну, а качество ссылки, соответственно, зависело от количества входящих бэклинков на страницу-донора (донором в поисковой оптимизации принято называть того, с кого ведет линк, а акцептором — того, на кого он проставлен).

В конце концов детище Брина и Пейджа стало служить средством поиска для всех пользователей Стэнфордского университета. Создатели поисковой системы просили своих первых пользователей высказывать свои впечатления и замечания по работе поиска, пытались их учесть и доработать (фактически это был этап альфа тестирования). Поиск стал доступен в 1997 году по адресу google.stanford.edu. Из Стэндфорда была подана заявка на лицензирование технологии поиска с использованием PageRank.

На сегодняшний день, система Google является лидером среди поисковых систем мира. Google - это не только поиск, но и еще более 50 сервисов, включая очень популярный браузер Google Chrome. Помимо просто поиска, здесь можно сравнивать цены на товары в интернет-магазинах, читать новости и многое другое. Есть и служба блокировки назойливой интернет-рекламы, а так же, с помощью веб кэша, можно "оживить" казалось бы давно "мёртвые" сайты. По мнению многих специалистов, на сегодняшний день Google Chrome самый быстрый браузер в мире. Что касается оценки пользователей, то претензий к скорости работы не было

выявлено, браузер открывает страницы практически мгновенно.

PageRank, используемый в Google, в основном основан на link popularity (индекс цитирования). На данный момент база данных Google насчитывает более миллиарда проиндексированных страниц. Google – одна из немногих поисковых систем, которая глубоко индексирует ваш сайт. Google использует индекс цитирования как наиболее весомый фактор в определении релевантности страницы. Поэтому большим и популярным сайтам проще

попасть на высокие позиции в результатах поиска. Это также защищает Google от спама. Так же анализируется плотность и частота ключевых слов, ключевые слова в ссылках, выделенный текст. В поиске разработанны языковые модели, позволяющие определять, какие сочетания слов следует искать в индексе.Для этого выполняется ряд действий – от интерпретации орфографических ошибок до определения типа введенного запроса на основе результатов последних исследований в области понимания естественного языка. Например, даже если у введенного вами слова несколько значений, Google Поиск определит верное. Это стало возможным благодаря специальной системе синонимов, которая создавалась пять лет и позволяет существенно увеличить качество результатов по более чем 30% запросов на разных языках.

Каждый год Google предоставляет сведения по триллионам поисковых запросов. 15% запросов, которые ежедневно обрабатывает система, никогда ранее не использовались. Разработать алгоритмы поиска, которые обеспечивали бы полезные результаты по таким запросам, не так просто. Для этого требуются регулярные проверки качества поиска, а также вложения средств. На сегодняшний день, благодаря дистрибьюторским соглашениям с Yahoo, AOL и Ask Jeeves, Google ежедневно обрабатывает до 80% всех поисковых запросов, сделанных в интернете.

Yandex.

В системе производится поиск с учетом морфологии русского языка, поиск с учетом расстояния, и тщательно разработанный алгоритм оценки релевантности. Реализован естественно-языковой запрос: теперь поисковик можно спрашивать «по-русски», задавая длинные вопросы. Поисковый робот позволяет предоставить возможности поиска по разным зонам текста, ограничение поиска на группу сайтов, поиск по ссылкам и изображениям, так же существует индекс цитирования. Работает фильтрация результатов поиска от мата и порнографии. У Яндкса имеются почтовые службы, новости, открытки и закладки, автоматическое объединение новостей в сюжеты и выделение главных тем дня. Плюс, стилизованный под Google Toolbar,

спартанский поисковик ya.ru. Поиск ведется не только по веб-страницам, но и по специализированным массивам данных – новостям и товарам. Находит документы не только в формате HTML.

За последние несколько лет технологии машинного обучения совершили колоссальный рывок, сегодня сделав возможным то, что еще вчера казалось фантастикой. Обладая достаточным количеством эталонных примеров, образцов для подражания, нейронные сети научились самостоятельно творить, писать музыку в стиле великих композиторов или создавать картины, подражая манере известных художников. В Яндекс.Поиске используются те же технологии, но в каком-то смысле задача здесь сложнее, потому что отсутствуют заранее готовые эталонные данные для обучения, примеры, на которые можно было бы настраиваться.

Чтобы научить поиск понимать самые разные запросы пользователей и находить на них хорошие ответы, эталонные данные для обучения приходится готовить самим. В Яндексе уже несколько лет сбором данных для машинного обучения занимаются асессоры, специально отобранные люди, которые в основном занимаются оценкой релевантности документов.

Асессор получает реальный пользовательский запрос, который случайно попал в базу, и документы, которые могли бы найтись по этому запросу, и оценивает, насколько тот или иной документ может быть хорошим ответом на запрос пользователя. Чтобы обучить поиск ориентироваться во множестве возможных трактовок и смыслов и понимать самые разные пользовательские запросы, нужно обрабатывать как можно больше реальных пользовательских запросов и собирать

для них оценки релевантности. На протяжении многих лет, из года в год, Яндекс постоянно увеличивал количество оценок в обучающей базе. Сам поиск развивался: появился поиск по картинкам, по видео, появлялись многие внутренние классификаторы и алгоритмы, все они работали на технологиях машинного обучения, всем им нужны были данные для настройки. Чтобы собирать всё больше и больше данных по всё большему количеству проектов, требовалось больше и больше людей. В какой-то момент, когда асессоров стало больше полутора тысячи и их всё равно не хватало, в Яндекс поняли, что нужно что-то менять, что технологии и области применения машинного обучения развиваются так быстро, что никакая, даже очень хорошо масштабируемая и быстрорастущая, но ограниченная, команда не способна будет удовлетворить постоянно растущие потребности в обучающих данных. Тогда Яндекс создали свою открытую краудсорсинговую платформу, на которой любой желающий может зарегистрироваться как исполнитель, находить для себя интересные задания и выполнять их за

вознаграждения; любой заказчик, которому нужны данные для машинного обучения, может зарегистрировать и размещать там свои задания. Назвали данную платформу Толокой – по названию старинной деревенской традиции: когда жители деревни собирались вместе, чтобы сообща сделать большое дело, такое, какое не под силу одному человеку. Так и на платформе Яндекс.Толока: за несколько лет ее существования собралось уже более миллиона толокеров, сообща они сделали более двух миллиардов оценок, которые пошли на обучение искусственного интеллекта.

Открытие Яндекс.Толока дало колоссальный рывок в масштабируемости и объемах собираемых данных для обучения. Если раньше собирали миллионы оценок силами только асессоров, то сейчас счет уже идет на миллиарды; если раньше асессоры принимали участие в десятках разных проектов, то сейчас в Яндекс.Толока открыто более полутора тысяч разных типов заданий и разных проектов. Среди заданий, которые могут выполнять толокеры, есть и оценка релевантности документов, и задания для развития карт и геопоиска (вооружившись мобильным приложением, толокеры ходят по самым разным регионам и проверяют актуальность данных об организациях для базы справочника), есть и любимые многими толокерами задания для настройки речевых технологий, которые особенно актуальны сейчас, потому что именно такой способ общения с машиной предпочитают самые новые пользователи Яндекс. Все требования для толокеров

открыты и доступны, их можно увидеть в инструкциях, соответствующих заданиям в Яндекс.Толока. Иногда никогда инструкций не дается, а просто собирается субъективное мнение большого количества людей.

3.Bing

Bing – поисковая система, которая разработана корпорацией Microsoft. Microsoft уже делал попытки запустить свою поисковую систему. Это были MSN Search — search.msn.com – с 1998 до 2006 года, Windows Live Search – search.live.com – работал до марта 2007 года, Live Search – live.com – до 1 июня 2009 года и по сути это новое название последней поисковой системы, с

обновленными возможностями, новым подходом, и новыми технологиями. Название поисковика Bing расшифровывается как: Bing is not Google (Bing это не Google). Поисковая система является второй в США (после Google) с 16 процентами и второй в мире с 3.23 процентами доли поиска.

Bing имеет классический для поисковиков интерфейс с возможностью поиска по категориям: картинки, видео, карты, новости. Здесь же можно воспользоваться инструментами Microsoft Office в режиме онлайн, перейти на сайт корпорации, зарегистрировать учетную запись и ящик на Outlook или Hotmail, завести свое облачное хранилище.Поисковая система работает на русском языке, определяя автоматически местонахождение пользователя. При этом не существует адреса Bing.ru, поисковая система работает только на домене com.

Алгоритм ранжирования поисковой системы Bing более чем на 90% основан на машинном обучении. Взаимодействие с пользователем – один из самых сильных сигналов в Bing. Поэтому активное участие аудитории увеличивает шансы на высокие позиции в этой поисковой системы. Для оценки сайта поисковик использует термин, известный как «пого-стик». Он обозначает ситуацию, когда пользователь вводит запрос поисковую в строку, нажимает на результат в выдаче, а затем щелкает по кнопке «Назад», чтобы вернуться к поиску. Если процент

людей, которые переходят на сайт и сразу возвращаются к выдаче, высок, то поисковая система будет рассматривать этот сигнал как свидетельство плохого взаимодействия. Однако если люди задерживаются на сайте, у него есть все шансы на то, чтобы выйти в топ. Bing считает клики по результату в выдаче, плюсует к ним данные о взаимодействии с пользователем и использует этот микс сигналов для улучшения результатов поиска. К примеру если при равных условиях сайт на первой позиции получает меньшее количество кликов, а второй большее, то в скором времени, при сохранении тенденции, они поменяются местами в результатах поиска.

Bing придает большое значение анкорным текстам. Поэтому, если пользователь хочет, чтобы ссылка учитывалась при ранжировании, он должен использовать анкорный текст.

Bing не так хорош, как Google в сопоставлении ключевых слов. Поэтому ему нужно прописывать ключевые фразы в точном вхождении. Но с оптимизацией нельзя перебарщивать. В противном случае будет легко нарваться на Google Penguin (алгоритм поисковый системы, направленный на устранение веб-спама).

15

4.Mail.ru.

Mail.ru - русскоязычный интернет-портал, принадлежащий технологической компании Mail.ru Group. Второй по популярности русскоязычный поисковик в рунете. Объединяет главную страницу сайта и тематические проекты, служит единой «точкой входа» для принадлежащих компании интернет-служб — почты, поиска, социальной сети «Мой мир», облачного сервиса, мессенджеров «Агент Mail.ru» и ICQ.

Первые поисковые технологии в компании Mail.ru начали разрабатываться в 2004 году под руководством Михаила Костина, прежнего руководителя системы Апорт. Инвестиции в проект составили около 700 тыс. долл. Сам домен gogo.ru был приобретён компанией Mail.Ru ещё в 2000 году. Результатом работы стал открытый в 2007 году сайт GoGo.ru. Поисковик имел первый на тот момент в рунете поиск по

видео, а также поиск по картинкам. К отличительным особенностям поисковика можно отнести русскоязычный поиск по видеороликам, а также по базе данных проекта "Ответы mail.ru". Также разработчики предусмотрели возможность тематической фильтрации результатов текстового поиска. Выдачу можно ограничить информационными, коммерческими и пользовательскими (форумы и блоги) источниками. Поисковик мог исправлять опечатки и реализовывать поиск по ключевым словам. По итогам Российского семинара по оценке методов информационного поиска (РОМИП), формула текстового ранжирования GoGo заняла первое место. Тем не менее, нужно отметить, что GoGo так и не стал основным поисковиком на главной странице Mail.ru. В течение многих лет в поисковой строке на главной странице Мейл.ру использовался сторонний движок: в 2004—2006 и 2010—2013 годы использовался поиск Google, 2007—2009 годах — решение от Яндекса. В 2009 году Яндекс разорвал контракт с Mail.ru об использовании технологий последнего на главной странице портала. Причиной стал отказ размещать логотип Яндекса на поисковике Mail.ru. В течение 8 месяцев портал работал полностью на собственном движке. В августе 2010 года был заключен контракт с Google, о чём стало известно широкой публике только в декабре того же года. Согласно условиям контракта, на собственный поисковик Mail.ru приходится 40 % выдачи, остальные 60 % выдачи — на Google.26 января 2011 года был запущен новый AJAX-интерфейс поиска по картинкам. Вместо постраничной выдачи теперь

новые картинки подгружаются внизу страницы по мере надобности.

В марте 2011 был запущен «социальный поиск» — теперь страницы в поисковой выдаче будут сопровождаться информерами с количеством рекомендаций пользователей соцсетей. В июне того же года появился «Поиск по обсуждениям» (go.mail.ru/realtime), позволяющий в режиме реального времени отслеживать обновление информации на новостных лентах, блогах и микроблогах. В феврале 2012 года был запущен собственный независимый поисковый движок в режиме бета-тестирования по адресу o.go.mail.ru.

11 ноября 2012 года появились инструменты для вебмастеров. В качестве планов было заявлено создание собственной системы контекстной рекламы.

В ноябре 2012 от анонимных источников внутри компании стало известно, что Mail.ru откажется от услуг Google в пользу собственной поисковой технологии в

связи с будущим выходом на международный рынок (под брендом my.com). Также появились данные о создании собственной системы контекстной рекламы. Официально о переходе на собственные поисковые разработки Мэйл.ру объявила 1 июля 2013 года. Месячная аудитория сервиса на тот момент составляла 39,5 млн человек с долей около 10 % на рынке. Число сотрудников проекта Go.mail.ru выросло с момента запуска с 15 до 200 чел, а количество проиндексированных документов составляло 10 млрд.

В ноябре 2013 в Google Play появилась новая версия поискового приложения от компании Mail.ru, позволяющего переходить с главного экрана в любые социальные сети и содержащего быстрый доступ к поиску по картинам, видео и новостям. Android-приложение превратилось в мини-браузер, заточенный под эффективный поиск нужной информации. Утилита также научилась распознавать поисковые запросы, заданные не текстом, а голосом. Разработчики также отмечают, что создали специальный виджет, который можно поместить на главный экран смартфона или планшета на базе системы Google Android. Подразумевается, что это позволит ещё сильнее сократить время, затрачиваемое на поиск.

В декабре 2013 внедрена технология «ручного» механизма ранжирования, благодаря чему веб-мастера могут самостоятельно добавлять запрос и документ в индекс Поиска Mail.ru. Это позволяет сайту органически «встроиться» в ранжирование и влиять на выдачу естественным образом. Таким образом, механизм ранжирования становится «ручным»: теперь качество сайтов оценивают не алгоритмы, а люди. Вот что по этому поводу

сообщил Андрей Калинин:

"Сначала хотел бы уточнить: в индекс сайты попадают обычным способом. Мы же решили добавить для веб-мастеров, создающих хорошие и достойные сайты, возможность гарантированного попадания ресурсов на верхние позиции выдачи. До сих пор веб-мастерам приходилось не только делать хорошие сайты, но и заниматься поисковой оптимизацией. Но, во-первых, не всем интересно заниматься оптимизацией; а, во-вторых, лучшая оптимизация ещё не гарантирует, что контент сайта окажется максимально качественным. Теперь сами веб-мастера могут указать, какие запросы они считают «своими», и по каким правильно показывать их страницу на первой позиции."

В январе 2014 добавлен поиск по описанию и контенту мобильных приложений в AppStore и Google Play на основе технологий российского стартапа Osmino. Аналогичный поиск в других поисковых системах и магазинах приложений доступен только по описанию приложени. С 1 июля 2013 года сервис использует собственные поисковые технологии, которые разрабатывались командой инженеров Mail.ru.

4.Baidu

Baidu - китайская компания, предоставляющая веб сервисы, основным из которых является поисковая система с таким же названием — лидер среди китайских поисковых систем. Занимает, примерно, 1,5 % глобального рынка поисковиков. В индексе Байду содержится свыше 740 млн веб-страниц, 80 млн изображений и 10 млн медиафайлов. Baidu так же имеет энциклопедию Baidu, которая обогнала Китайскую википедию.

Основана в 2000 году, основатели — Робин Ли и Эрик Сю, получившие высшее образование в США. $1,2 млн стартового капитала привлекли от американских венчурных компаний. Название компании взяли из поэмы времен династии Сун, буквально оно означает «сто раз». Через год получили от венчурных компаний ещё $10 млн инвестиций. В 2004 году Baidu стала

лидирующей поисковой системой в Китае.

Самым важным этапом в процессе формирования компании стал правильный подбор сотрудников. Руководство Baidu тщательно подошло к этому процессу. В большинстве своем это такие же, как и Робин Ли, граждане КНР, получившие образование в вузах США и поработавшие в крупных ИТ-компаниях. При этом кадры в Baidu подобраны настолько идеально, что Google несколько раз пыталась увести некоторых ключевых специалистов. Через год компании удалось укрепить

собственные позиции, фактически монополизировав аудиторию в Китае. В 2005 году прибыль компании составляла $13,2 млн.

На данный момент Baidu контролирует 80% китайского рынка, что составляет около 465 млн пользователей. Огромную часть трафика (около 40%) в Baidu обеспечивает сервис по поиску прямых ссылок на скачивание аудиотреков. Такое наплевательское отношение к авторским правам вызвало волну осуждения со стороны звукозаписывающих компаний со всего мира, но Робин Ли проигнорировал его, заметив, что борьба с незаконно размещенным контентом не является целью поисковой системы. Вполне вероятно, что молодую компанию уничтожили бы при помощи исков, но китайское правительство не выдает своих граждан.

Высокий доход связан, в первую очередь, с оригинальным способом монетизации. Baidu сделала платной возможность попадания сайта в топ поиска. Стоимость зависит от тематики ресурса и ключевых фраз. При этом сервис оставался не очень избирательным в выборе клиентов, что привело к серии громких скандалов.

В 2007 году в китайских новостях прозвучала информации об обилии в Baidu рекламных объявлений шарлатанов, выдающих себя за дипломированных врачей. Меламиновый скандал в 2008 году чуть не уничтожил компанию. Суть в том, что в Китае началось следствие по поводу добавления вредных для здоровья веществ в детское питание. Часть молочных компаний, на которые было открыто уголовное дело, попыталась скрыть свою причастность путем удаления из поисковых запросов упоминаний о себе в теме следствия. Baidu ответила согласием на их предложение, и вскоре сама компания оказалась под ударом. Акции за один день опустились в цене на 60 пунктов, бренду был нанесен колоссальный ущерб. В таких условиях Робину Ли пришлось принести публичные извинения и начать кадровую чистку. После увольнения виновных, руководство компании решило более тщательно подойти к своей рекламной платформе.

Для решения проблемы был создан сервис «Гнездо феникса», при помощи которого проводится проверка рекламодателей, оказывается помощь в создании собственных объявлений, а также отслеживается трафик. Новая

рекламная платформа вскоре значительно укрепила финансовое состояние

китайского поисковика, позволив достигнуть высоких показателей прибыли. В 2014 году Baidu поставил собственный рекорд, увеличив доход на 53% и достигнув

отметки в $7,9 млрд за год.

Считается что Байду на порядок хуже индексирует сайты, расположенные на иностранных серверах. Официальная позиция такая: Baidu индексирует, а затем ранжирует сайты по многочисленным параметрам, в том числе по скорости ответа сервера. А для многих не секрет, что в Китае для иностранного трафика выделен небольшой канал. Поэтому если хотите попасть в индекс быстрее и быть выше — перенесите сайт в Китай, или как минимум в Гонконг. Обзаведитесь ICP-лицензией для сайта. Официально: ее наличие не влияет на индексацию. Но на практике ICP помогает Baidu быстрее решить хороший сайт или плохой (если одобрен правительством = хороший), и ставит его быстрее/выше.

Вкратце можно сказать, что как и с другими поисковиками, оценивается множество параметров: уникальность и качество контента, количество внешних ссылок и рейтинг ссылающихся сайтов, «плотность» ключевых слов и так далее. Пожалуй, только с плотностью дело тут обстоит несколько свободнее: если в Google/Яндекс за большое количество ключевиков на «сантиметр» текста можно получить «банхаммером», то Baidu на это смотрит проще. А самую главную роль в ранжировании играет совокупность ссылок, ведущих на ваш ресурс с других сайтов и количество внешних ссылок. Сразу стоит отметить, что поисковой маркетинги платная реклама в Байду никак не влияют на индексацию/позиционирование. Но так как у Baidu нет никакой совести и рекламные площади находятся сверху выдачи, снизу, сбоку и посередине (еще совсем недавно они почти никак не отличались от обычных результатов, но сейчас у них другой фон), то этот момент очень важно не упускать. Можно вложить в поисковую оптимизацию сотни тысяч юаней и оказаться на первом «органическом» месте, но это будет пятое (а недавно и десятое) место сверху от платных результатов. Некоторые эксперты говорят, что 80% пользователей не знают, что это проплаченные результаты. А те, кто знает, по инерции кликают в середину или в нижние результаты. Но не тут-то было, даже нижние результаты можно купить.

Вообще за Baidu давно сложилась репутация с одной стороны активного проправительственного цензора, а с другой монополиста в поиске, чья выдача продана более чем наполовину. Сейчас по этому поводу негодуют в основном «гики», но и среди обычных интернет-пользователей начинает расти недовольство. Байду старается потихоньку исправлять репутацию, но конкуренты не упускают возможности как-то на этом факте сыграть.

Китайскому поисковику пока не удалось составить конкуренцию Google в других странах. Несмотря на правильно выбранную стратегию, направленную на изначальное завоевание азиатского рынка, Baidu мешают факторы, за счет которых было достигнуто доминирование в Китае. Оказалось, что пользователям не очень нравится ограничение контента, ссылки на скачивание песен запрещены в других странах, а переводы поисковика на другие языки выполнены недостаточно грамотно. Кроме того, Google и другие поисковые системы уже обеспечили свое доминирование на этих рынках, и поэтому захватить аудиторию достаточно сложно.

В своих многочисленных интервью Робин Ли продолжает настаивать на том, что цель Baidu — стать мировым лидером среди поисковых систем. На деле все выглядит не настолько позитивно.

4 Глава.

Сравнение работы представленных поисковых систем.

1.Для начала создадим общий для поисковых систем запрос, понятный для всех, на английском языке.

2. По результатам поиска выясним количество полученных ссылок.

3. По полученным результатам конкретизируем наш запрос.

4. Проанализируем первые пятьдесят предложенных ссылок.

5. Определим положение ссылки с нужным результатом.

6. Сделаем тоже самое на русском языке.

7. Сравним результаты обоих запросов

8. Выставим оценку по пятибалльной шкале.

Для начала введём одинаковый запрос на английском языке для всех поисковых систем, результат в таблице 1.

Запрос на английском:

Таблица 1.

Название ПС	Google	Baidu	Яndex	Bing	mail.ru
Общее количество результатов	142000	17	9000	23200	Не указывается в поисковике
Положение нужного результата	4	Отсутствует	1	37	1

Уточним запрос и получим новый результат в таблице 2.

22

Уточнённый запрос на английском:

Таблица 2.

Название ПС	Google	Baidu	Яndex	Bing	mail.ru
Общее количество результатов	30590	1530	1000000	39400	Не указывается в поисковике
Положение нужного результата	3	Отсутствует в первых 50 результатах	1	39	52

По результатам запроса на английском языке Baidu не справился совсем. У поисковой системы малое количество ссылок в результатах поиска, а также предоставлена совершенно отличная от запрашиваемой информация. Поэтому он получает 1 балл. На “троечку” справился Bing, результаты были близкими но не совсем то что нужно. Так же на “троечку” справился и mail.ru. из за большого расхождения в результатах после уточнения. Такие расхождения с

первоначальным результатом скорее всего связанны с появлением слова “website” в запросе, так как предложения о создании сайтов и рекламы других сайтов откинули искомую ссылку аж на пятую страницу. Отлично справился поиском Google, на твёрдую “пятёрку”. Реклама отсутствовала, а выше стояли

только ссылки на информационные сайты. Ну а лучше всех, по моему мнению, справился Яндекс. Он так же получает пять баллов. Интересующая информация находилась на первом месте в результатах поиска в обоих случаях.

Теперь попробуем тот же запрос на русском языке для всех поисковых систем, результат в таблице 3.

Запрос на русском:

23

Таблица 3.

Название ПС	Google	Baidu	Яndex	Bing	mail.ru
Общее количество результатов	1950000	17300000	127000	39400	Не указывается в поисковике
Положение нужного результата	3	Отсутствует в первых 50 результатах	2	42	1

Уточним запрос и получим новый результат в таблице 4:

Таблица 4.

Название ПС	Google	Baidu	Яndex	Bing	mail.ru

Общее количество результатов	150000	43700000	12000000	529000	Не указывается в поисковике
Положение нужного результата	1	Отсутствует в первых 50 результатах	1	1	1

Как и с английским языком так и с русским китайская поисковая система показала себя с худшей стороны. Как и ранее с английским языком, поисковик показал огромное количество рекламы. Более менее полезная информация по запросу начинается поле седьмой страницы полученных результатов на английском, а с русским языком дела обстоят намного хуже. На “хорошо” справился американский поисковик от Microsoft. На пять баллов

сработали Google, mail.ru и Яндекс. Mail.ru и Яндекс справились с задачей лучше всех, оба поисковика сразу предлагали искомый вариант.

По результатам сравнения получились следующие результаты:

5 место поисковик Baidu – 2 балла

4 место поисковик Bing – 7 баллов

3 место поисковик Mail.ru – 8 баллов

2 место поисковик Google – 10 баллов

1 место поисковик Яндекс – 10 баллов

24

Яндекс справился лучше всех с поиском нужной для меня информации, почти всегда предлагая мне интересующий меня сайт, как на английском, так и на русском языках. поэтому первое место я отдаю ему.

Поиск проводился через браузер опера, с включенным VPN, на новом компьютере.

25

Заключение.

По результатам даже нашего небольшого сравнения работ поисковых систем очевидно, что Робину Ли, с его поисковой системой, предстоит ещё очень большой путь для причисления к мировым лидерам среди поисковых систем. Может она и не плоха на китайском рынке, но конкуренцию в других странах она заметно уступает.

Mail.ru тоже неприятно удивил, своим разбросом результатов из за добавления всего лишь одного двух слов в поисковой запрос, что заставляет усомнится в релевантности его результатов.

Порадовал Bing (хоть и стоит ниже чем Mail.ru). Пусть не всегда искомый результат находился близко к первым местам, однако результаты всё же были близкими, поисковик как будто “ходил вокруг да около”, не совсем уверенно предлагая разные варианты. Видимо это по тому что им не часто пользуются на территории России (как и Baidu).

Но всё же отдельной строкой стоит именно Google, а не победивший Яндекс. Да наш поисковик показал себя с лучшей стороны, однако запрос касался России. Google же, практически, ни в чём не уступал ему. Если провести такое же сравнение касательно Соединённых штатов, то думаю что Яндекс уступил бы не только Google, но и Bing.

Но всё же поисковые системы - это бизнес, продвижение товара, мнений, мировоззрений. Хоть и считается что самые популярные поисковики довольно открыты, но со стороны государства ведётся постоянный контроль за их релевантностью. Будь то Google, Яndex и уж тем более Baidu. Есть и свободные от контроля со стороны власти поисковики но они “обитают” в так называемом Даркнете, являющимся одним из ответвлений интернета, но о них почти ничего не известно основной массе пользователей глобальной сети интернет.

Список литературы

1. Кузьмин А.В. Золотарева Н.Н. Поиск в Интернете - Санкт - Петербург.: Издательство НиТ, 2011г.
2. Д. Н. Колисниченко Поисковые системы и продвижение сайтов в Интернете Издательство Вильямс 2007г.
3. https://ru.wikipedia.org
4. www.baidu.com - Поисковая система Baidu.
5. www.seop.ru - Search engine optimization project, рейтинг основных поисков
6. Экслер, А.Б., "Самоучитель работы в Интернете" - Москва.: NT Press, 2010г.