Содержание:

Введение

Всемирная сеть очень востребованна. Любой пользователь Интернета может найти в нем много разной и интересной информации, а также использовать все широкие возможности сети.

В огромном пространстве под название Интернет есть очень много разнообразных ресурсов, которые предлагают развлечения, а также информацию с самым разным характером. И в этой всей суматохе вы можете разобраться, только воспользовавшись поисковыми системами.

В интернете появляются специальные поисковые средства. Несколько лет назад говорили: в Интернете ничего невозможно найти, но там есть всё. Но когда появились и быстро развились поисковые каталоги, поисковые машины, и всевозможные поисковые программы ситуация в корне поменялась, и сейчас в интернете информацию которая вам нужна, можно найти намного быстрее, чем в открытой книге, лежащей у вас в руках.

Наиболее популярным и используемым способом поиска в Интернете является использование поисковых систем. Что же такое поисковая система? Поисковая система — это специальный ресурс в Интернете, который выдает информацию пользователю в соответствии с его запросом. То есть этот ресурс собирает все данные в глобальной сети, все веб-проекты и при поступлении от пользователя определенного запроса выдает необходимую искомую информацию путем направления его, например, на тематический блог или сайт.

Первоочередная задача любой поисковой системы – доставлять людям именно ту информацию, которую они ищут. В Рунете системы появились еще с 1996 года – это Апорт и Рамблер. Годом позже в 1997 году образовался Яндекс, а еще годом позже в 1998 году появился еще один конкурент – Google. В настоящий момент наиболее популярные – это Яндекс и Google.

Получая результат, пользователь может использовать именно те материалы в интернете, которые он запрашивал. Сегодня рассмотрим основные системы, кратко их историю и принципы работы.

Глава I. Общие понятия

1.1 Определение

Поисковая система – это огромная база веб-документов, которая постоянно пополняется и расширяется. У каждой поисковой системы есть поисковые пауки, роботы – это специальные боты, которые обходят сайты, индексируют размещенный на них контент, а затем ранжируют по степени его качества и релевантности поисковым запросам пользователей.

По своей сути поисковая система – это каталог сайтов, справочник, основная функция которого – поиск информации по этому самому каталогу.

Рассмотрим, как работает поисковая система. Принцип работы поисковых систем очень сложный, но я попробую объяснить простыми словами. Поисковый робот (паук) обходит страницы сайта, скачивает их содержимое и извлекает ссылки. Далее начинает свою работу индексатор – это программа, которая анализирует все скачанные пауками материалы, опираясь на собственные алгоритмы работы. Таким образом, создается база данных поисковой системы, в которой хранятся все обработанные алгоритмом документы. Работа с поисковым запросом проводится следующим образом: анализируется введенный пользователем запрос; результаты анализы передаются специальному модулю ранжирования; обрабатываются данные всех документов, выбираются самые релевантные введенному запросу; генерируется сниппет – заголовок, описание, слова из запроса подсвечиваются полужирным; результаты поиска представляются пользователю в виде страницы выдачи.

1.2 История развития

В первые годы развития Интернета, численность его пользователей было небольшим, а количество информации, доступной пользователю, прилично маленьким. В основном в те годы выход в интернет имели зачастую сотрудники научно-исследовательской сферы. Но и надобность поиска информации в Интернете не столь уж актуальной, как на сегодняшний день.

Создание открытых каталогов сайтов стало первым способом организации доступа к информационным ресурсам сети, в них по тематике группировались ссылки на ресурсы. Первым подобным проектом был сайт Yahoo.com, его открыли весной 1994 года. После увеличения количества сайтов в каталоге Yahoo, нужную информацию стало возможным искать по каталогу. В полном смысле это еще не представляло поисковую систему, потому что область поиска была ограничена непосредственно только ресурсами, которые присутствовали в каталоге, а не во всех ресурсах интернета.

Каталоги ссылок были распространены и ранее, но в настоящее время почти полностью потеряли свою популярность. Потому что даже в самых огромных современных каталогах, есть информация только о мельчайшей части интернета. В сети один из самых больших каталогов DMOZ (он ещё называется Open Directory Project) имеет информацию о 5 миллионах ресурсов, а если брать базу поисковой системы Google, то она состоит более чем из 8 миллиардов документов.

Первая полноценная поисковая система была «WebCrawler», которая вышла в мир в 1994 году. Главное отличие этой поисковой системы от последователей заключается в предоставлении пользователю возможности осуществлять поиск на любой веб-странице, по любым ключевым словам. В настоящее время такая технология есть стандарт поиска любой поисковой системы. Таким образом, поисковая система «WebCrawler» стала первой системой, о которой знали не только ученые, но и широкий круг обычных пользователей.

В 1995 году появились поисковые системы Lycos и AltaVista. В 1996 году AltaVista стала доступна русскоязычным пользователям, запустив морфологическое расширение для русского языка. В этом же году запущены такие отечественные поисковые системы как – «Rambler.ru» и «Aport.ru». Появились первые отечественные поисковые системы, и Рунет (интернет на русском языке) вышел на новый уровень, позволяя всем русскоязычным пользователям осуществлять запросы на русском языке, и оперативно реагировать на любые изменения, которые происходят внутри Сети.

После того как в 1997 году запустили поисковую систему «Яндекс», очень сильно между собой начали конкурировать отечественные поисковые машины, они улучшают систему выдачи результатов, поиска и индексации сайтов, а стали предлагать новые сервисы и услуги.

Сергей Брин и Ларри Пейдж в 1997 году, в рамках исследовательского проекта в Стэндфордском университете, создали поисковую машину Google. В настоящее время Google - самая популярная поисковая система в мире, именно она дала возможность пользователю осуществлять с учетом морфологии качественный и быстрый поиск, ошибок при написании слов, и в результатах выдачи запросов очень сильно повысила релевантность. На данный момент компания Google обрабатывает более 40 миллиардов запросов в месяц, это соответствует около 62,4 % из всех поисковых запросов в мире.

1.3 Цель 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 Концепция 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 сер*вер:** 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. Действующие поисковые системы

2.1 Принцип работы Google

В наше время поисковые системы, в частности Google, напоминают «витрину» Интернета и являются наиболее важным каналом распространения информации в цифровом маркетинге. С помощью глобальной рыночной доли, которая составляет более 65% по данным за январь 2016 года, Google явно доминирует в поисковой индустрии. Хотя компания официально не раскрывает степень своего роста, к 2012 году было подтверждено, что их инфраструктура обслуживает около 3 миллиардов поисковых запросов в день.

Google.com глобально занял звание сайта номер 1 в Alexa Top 500 Global Sites. Учитывая эти цифры, владельцам собственных веб-страниц особенно важно иметь хорошую видимость своих сайтов поисковой системой.

Чем нужнее становится Google для современного маркетинга, тем важнее понимать функции поиска и алгоритмы обновлений, которые оказывают непосредственное влияние на ранжирование результатов. Moz предполагает, что Google изменяет свои алгоритмы по 600 раз за год. Многие из этих изменений и связанные с ними факторы ранжирования держатся в секрете. И только о крупных обновлениях объявляют публично.

Своим появлением поисковые системы напрочь изменили привычный для нас способ сбора информации. Интересует ли вас обновление данных фондового рынка или вы хотите найти лучший ресторан в районе, либо пишете академический отчет об Эрнесте Хемингуэе — поисковик даст ответ на все запросы. В 80 годы ответы на вопросы потребовали бы посещения местной библиотеки. Теперь же все решается в течении миллисекунды с использованием алгоритмических полномочий поисковика.

В этом отношении главная цель поисковой системы заключается в том, чтобы максимально быстро найти уместную и актуальную информацию, как ответ на введенные поисковые термины, также называемые ключевыми словами. Поэтому центральным аспектом для любой поисковой системы, желающей выдать действительно полезный результат, является понятие цели поиска, того, как именно люди ищут.

Результат работы Google можно сравнить с интернет-каталогом, отобранным с помощью рейтинговой системы на основе алгоритмов. Более конкретно алгоритм поиска можно описать как «нахождение элемента с заданными свойствами среди списка элементов».

Процессы Google

Давайте теперь подробнее рассмотрим привлеченные процессы сканирования, индексирования и позиционирования.

А) Сканирование

Сканирование может быть описано, как автоматизированный процесс систематического изучения общедоступных страниц в Интернете. Проще говоря, во

время этого процесса Google обнаруживает новые или обновленные страницы и добавляет их в свою базу. Для облегчения работы он использует специальную программу. «Googlebots» (можно встретить альтернативные названия: «боты» или «роботы») посещают список URL-адресов, полученных в процессе прошлого сканирования и дополненных данными карты сайта, которую предоставляют веб-мастера и анализируют их содержание. При обнаружении ссылок на другие страницы во время посещения сайта, боты также добавляют их в свой список и устанавливают систематические связи. Процесс сканирования происходит на регулярной основе в целях выявления изменений, изъятия «мертвых» ссылок и установления новых взаимосвязей. И это при том, что только по данным на сентябрь 2014 года насчитывается около миллиарда веб-сайтов. Можете себе представить сложность такой задачи? Тем ни менее, боты не посещают абсолютно каждый сайт. Чтобы попасть в список проверяемых, веб-ресурс должен быть рассмотрен, как достаточно важный.

Б) Индексация

Индексация — процесс сохранения полученной информации в базе данных в соответствии с различными факторами для последующего извлечения информации. Ключевые слова на странице, их расположение, мета-теги и ссылки представляют особый интерес для индексации Google.

Для того чтобы эффективно хранить информацию о миллиардах страниц в базе данных поисковой системы, Google использует крупные центры обработки данных в Европе, Азии, Северной и Южной Америке. В этих центрах, как было подсчитано, на основе энергопотребления Google в 2010 году, работает около 900,000 серверов.

Основная цель процесса индексации: быстро реагировать на поисковой запрос пользователя. Его как раз мы и будем обсуждать на следующей стадии.

В) Обработка

Когда пользователь вводит запрос, Google производит в базе данных поиск, подходящий под условия и алгоритмически определяет актуальность содержания, что выводит к определенному рейтингу среди найденных сайтов. Логично, что результаты, которые считаются более релевантными для пользователя поисковой системы, намеренно получают более высокий ранг, чем результаты, которые имеют меньше шансов обеспечить адекватный ответ.

Хотя Google и не выпустил официальных данных об этом, компания подтверждает, что использует более 200 факторов для определения релевантности и значимости конкретной страницы.

Естественно, всем веб-разработчикам важно знать, каковы факторы ранжирования, которые влияют на позицию страницы в поисковой выдаче. Иногда Google дает определенные намеки, объявив важные изменения в обновлениях своих алгоритмов.

Все вышеописанные процессы сканирования, индексирования и позиционирования можно изобразить с помощью такой схемы:

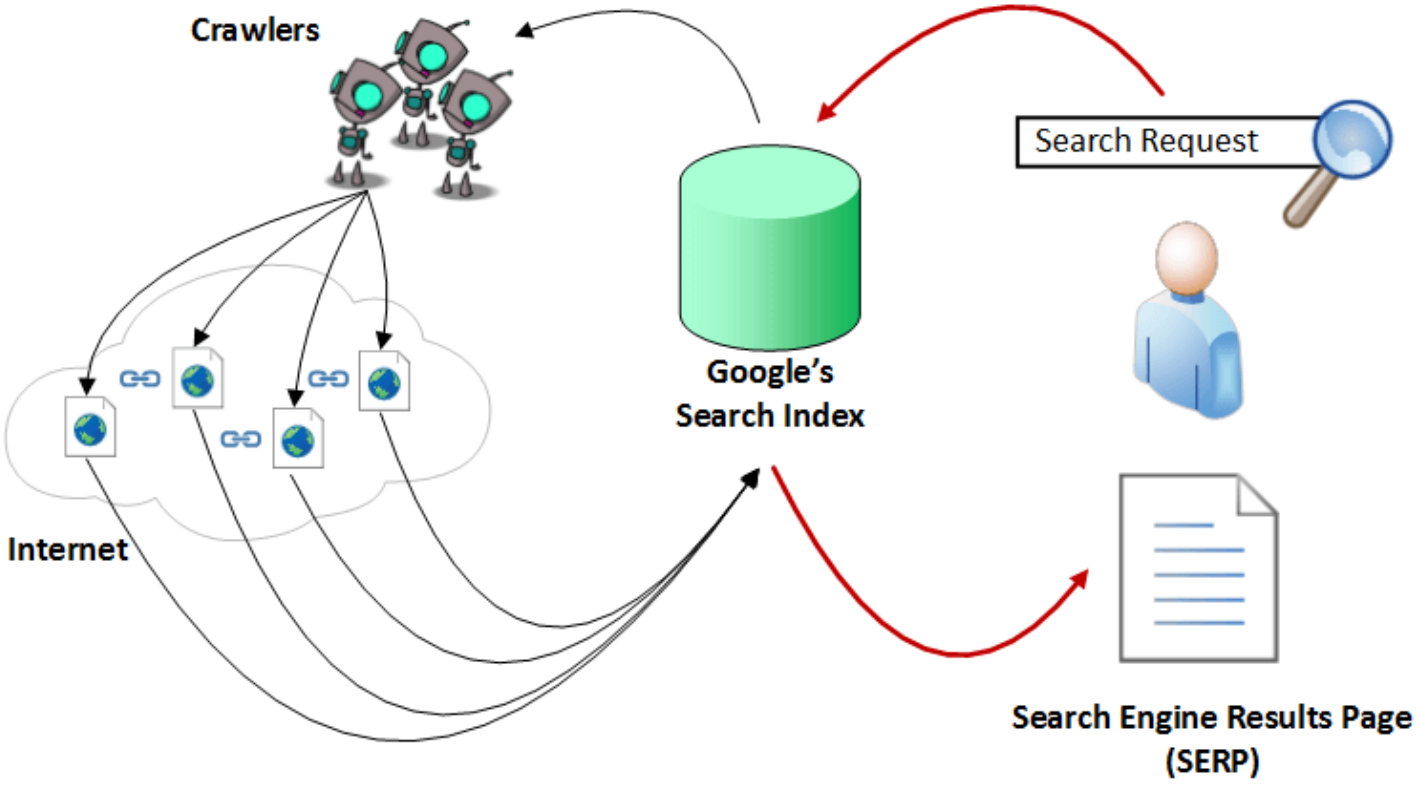

Рис. 1

Алгоритм ранжирования Google сложнее, чем алгоритм Яндекса. Продвигать сайты в Google, особенно на начальном этапе, немного сложнее. Раскрутка молодого сайта в Google затруднительна, так как на новые веб-ресурсы накладывается фильтр (так называемая «песочница»). Google при ранжировании использует порядка 200 факторов, оптимизатор может повлиять лишь на некоторые.

С другой стороны, поисковая система Google выглядит стабильнее своих конкурентов в плане смены алгоритма и апдейтов. Информация, только что

размещенная на сайте, может в считанные минуты попасть в основную выдачу. Поисковые роботы Google в три раза быстрее, чем роботы других поисковых систем. Фильтры (критерии «нормальности» сайта) почти не меняются с момента начала их внедрения.

Контент и ссылки – вот два фактора, на которые может повлиять оптимизатор при продвижении сайта в поисковой системе Google.

Релевантность контента относительно поискового запроса повышается следующим образом: простановка ключевых слов в заголовках (тегах title и h1 – h6). В title прописывается единственная ключевая фраза без лишних слов. Ключевые слова в начале html-кода страницы сайта так же увеличивает релевантность текста.

Внешние ссылки Google учитывает по нескольким параметрам: количество, авторитетность сайта-донора (т.е. насколько поисковая система доверяет сайту), тематичность. Сквозные ссылки (ссылки, ведущие со всех страниц сайта-донора, устанавливаются, например, в шаблоне сайта) в глазах Google обладают большим весом, нежели 10 ссылок (с этого же сайта-донора).

Сайт-акцептором называют сайт А, на который стоит ссылка с сайта В, а сайтом-донором – сайт В, который размещает ссылку на сайт А.

Перед продвижением сайта в Google следует:

- В случае нового сайта сообщить поисковой системе по адресу: https://www.google.com/webmasters/tools/submit-url/
- С помощью страницы «инструменты для веб-мастеров» https://www.google.com/webmasters/tools/home?hl=ru подтвердить права на сайт, создать файл sitemap.xml и добавить ссылку на карту сайта вида http://www.site.ru/sitemap.xml.
- Проверить код на валидность
- Проверить работоспособность всех ссылок на сайте, при необходимости исправить ошибки.

Это позволит поисковому роботу Google полнее и точнее проиндексировать сайт и выделить заслуженное место на страницах своей выдачи.

Понятие **Google PageRank** является одним из ключевых моментов в работе поисковой машины Google. Наряду с другими параметрами, влияющими на выдачу (сортировку) сайтов в результатах поиска, знание модели PageRank необходимо

как для понимания процесса поиска, так и для использования оптимизаторами при продвижении своих сайтов в поисковой системе.

PageRank (далее просто PR) это числовая величина — мера “важности” страницы в поисковой системе Google. Зависит от числа внешних ссылок на данную страницу и от их веса (важности). Другими словами от количества и качества ссылающихся страниц. А если говорить математическим языком, то PR – это алгоритм расчёта авторитетности страницы, используемый поисковой системой Google. PR не является основным, но является одним из вспомогательных факторов при ранжировании сайтов в результатах поиска.

Следует отметить, что при расчете PR Google учитывает не все ссылки, а отфильтровывает ссылки с сайтов, специально предназначенных для скопления ссылок. Некоторые ссылки могут не только не учитываться, но и отрицательно сказаться на ранжировании ссылающегося сайта (такой эффект называется **поисковой пессимизацией**). Основной формулой для расчета PR является сложная формула, ее можно посмотреть в открытых источниках.

2.2 Принцип работы Яндекса

Поисковик Яндекс знает несколько триллионов URL. И каждый день он изучает по паре миллиардов из них. Делают это специальные роботы-пауки, краулеры. Они заходят на страницу, анализируют содержимое, делают копию и отправляют на сервер. А затем уходят по ссылкам на другие страницы. Так происходит знакомство поисковика с сайтом. Далее следует этап индексикации. Если произвести нехитрые математические расчеты, то можно выявить, что пауки Яндекса обойдут все известные страницы приблизительно за 2 года. Но это будет неверно, так как количество урлов постоянно увеличивается. Вывод - работа по созданию поисковой базы бесконечна.

Этап 1. Поиск новых страниц

Вопреки заблуждению, поисковые системы выдают информацию не о страницах, находящихся в интернете, а о страницах, находящихся в базе данных поисковой машины. То есть, если сайт неизвестен Яндексу, то и в выдаче он не появится.

Задача поисковика на этом этапе заключается в поиске всех возможных адресов страниц в интернете. Выполняет эту работу так называемый робот «паук».

Интернет это ссылки, ссылки и еще раз ссылки и этот «паук» просто переходит по всевозможным ссылкам, записывая в свою базу адреса всех найденных страниц.

Попал на главную страницу сайта, на ней нашел ссылки на страницы рубрик, на страницах рубрик нашел ссылки на страницы со статьями, карточками товаров, ссылки на файлы или другой информацией. На каких-то из посещенных страниц одного сайта, он нашел ссылки на другие сайты – поисковая система переходит по ним и сканирует все, что нашла там.

Прекрасно помогают роботам для ориентирования файлы Robots.txt и карты сайта Sitemap.xml, их надо обязательно сделать, особенно, если сайт имеет много страниц.

Задача робота создать адресный справочник по типу – Город, Улица, Дом, Квартира.

Этап 2. Индексация

Как уже определили – в поисковую выдачу попадает информация не с сайтов, находящихся в интернете, а информация из базы данных поисковой системы. И следующая программка поисковика как раз занимается добавлением информации в базу. Она путешествует по всем известным адресам сайтов и страниц, копируя их содержимое на склады поисковой системы.

Называется этот процесс индексация – попадание информации в индекс поисковой системы.

Первый и второй процессы протекают непрерывно и, зачастую, одновременно. Постоянно пополняется база адресов страниц и база информации с этих страниц.

Кстати, в процессе индексации поисковые системы оценивают качество страниц, и информация некоторых из них не попадает в индекс. Как бы поисковик знает об их существовании, но по каким-то причинам считает их бесполезными для пользователя, поэтому не добавляет в выдачу – зачастую это не уникальный контент или служебные страницы.

Как часто происходит процесс индексации? В первую очередь это зависит от типов сайтов. Веб-ресурс первого типа очень часто меняет содержимое своих страниц. То есть, когда к этим страницам каждый раз приходит поисковый робот, они каждый раз содержат другой контент. По ним ничего в следующий раз уже не получится найти, поэтому такие сайты не включаются в индекс. Второй тип сайтов —

хранилища данных, на страницах которых периодически добавляются ссылки на документы для скачивания. Контент такого сайта обычно не меняется, поэтому его робот посещает крайне редко. Другие сайты зависят от частоты обновления материала. Имеется в виду следующее — чем быстрее появляется новый контент на сайте, тем чаще приходит поисковый робот. И приоритет отдается в первую очередь наиболее важным веб-ресурсам (новостной сайт на порядок важнее, чем любой блог, к примеру).

Индексирование позволяет выполнить первую функцию поисковой системы — сбор информации на новых страницах в сети Интернет. Но у Яндекса есть и вторая функция — поиск ответа на запрос пользователя в уже подготовленной поисковой базе.

Этап 3. Определение релевантности и ранжирование

Если то, что мы обсудили в предыдущих пунктах, работает непрерывно и независимо от внешних факторов (действий человека), то третий этап в алгоритме работы поисковых систем начинает действовать только под воздействием человека. Когда в поисковике задается запрос, система начинает искать на него ответ в наполненной базе знаний по критериям, заданным человеком в этом запросе (как узнать самые популярные запросы в Яндексе).

Сначала, система делает выборку, определяя все релевантные запросу страницы из известных (Релевантные – значит соответствующие, подходящие. Как проверить релевантность страниц сайта я писал тут). Например, для запроса «купить холодильник Норд» релевантными будут страницы содержащие слова «купить», «холодильник», «Норд». Все страницы, содержащие одно или несколько из этих слов, попадут в выдачу поисковой системы.

Процессом обработки запроса и выдачей релевантных ответов в Яндексе занимается компьютерная система «Метапоиск». Для своей работы сначала она собирает всю вводную информацию: из какого региона был осуществлен запрос, к какому классу относится, есть ли ошибки в запросе и т.д. После такой обработки метапоиск проверяет, есть ли в базе точно такие же запросы с такими же параметрами. Если ответ положительный, то система показывает пользователю заранее сохраненные результаты. Если же такого вопроса в базе не существует, метапоиск обращается поисковой базе, в которой содержатся данные индекса.

И вот здесь происходят удивительные вещи. Представьте себе, что существует один супермощный компьютер, который хранит в себе весь обработанный

поисковыми роботами Интернет. Пользователь задает запрос и в ячейках памяти начинается поиск всех документов, причастных к запросу. Ответ найден и все довольны. Но возьмем другой случай, когда появляется очень много запросов, содержащих в своем теле одинаковые слова. Система должна каждый раз пройтись по одним и тем же ячейкам памяти, что может увеличить время на обработку данных в разы. Соответственно, увеличивается время, что может привести к потери пользователя — он обратится за помощью к другой поисковой системе.

Чтобы таких задержек не было, все копии в индексе сайтов распределены по разным компьютерам. После передачи запроса, метапоиск дает команду таким серверам искать свой кусочек с текстом. После чего, все данные от этих машин возвращаются в центральный компьютер, он объединяет все полученные результаты и выдает пользователю первую десятку самых лучших ответов. С такой технологией сразу убивается два зайца: в несколько раз уменьшается время поиска (ответ получается за доли секунды) и благодаря увеличению площадок дублируется информация (данные не теряются из-за внезапных поломок). Сами компьютеры с дублирующей информацией составляют дата-центр — это комната с серверами.

Когда пользователь поисковой системы задает свой запрос, в 20-ти случаях из 100 получаются неоднозначные цели в вопросе. Например, если он пишет в строке поиска слово «Наполеон», то еще не известно, какой ответ ожидает — рецепт торта или биография великого полководца. Или фраза «Братья Гримм» — сказки, фильмы, музыкальная группа. Чтобы такой возможный веер целей сузить до конкретных ответов в Яндексе существует специальная технология Спектр. Она учитывает потребности пользователей, используя статистику поисковых запросов. Из всех вопросов, заданных в Яндексе посетителями, Спектр выделяет в них различные объекты (имена людей, названия книг, модели машин и т.д.) Эти объекты распределены по некоторым категориям. На сегодняшний момент таких категорий насчитывается более 60-ти. С помощью них поисковая система имеет в своей базе разные значения слов в запросах пользователей. Интересно, что эти категории периодически проверяются (анализ происходит пару раз в неделю), что позволяет Яндексу более точно давать ответы на поставленные вопросы.

На базе технологии Спектр Яндекс организовал диалоговые подсказки. Они появляются под поисковой строкой, в которой пользователь набирает свой неоднозначный запрос. В этой строке отражены категории, к которым может относится объект вопроса. От выбора пользователем такой категории зависят

дальнейшие результаты поиска. От 15 до 30% всех пользователей поисковой системы Яндекс желают получить только местную информацию (данные того региона, в котором они живут). Например, о новых фильмах в кинотеатрах своего города.

В задачу Яндекса входит не только поиск всех возможных вариантов ответа, но и подбор самых лучших (релевантных) по порядку. Ведь пользователь не будет рыться во всех ссылках, которые ему предоставит в качестве результата поисков Яндекс. Процесс упорядочивания результатов поиска называется ранжированием. То есть именно ранжирование определяет качество предлагаемых ответов.

Есть правила, по которым Яндекс определяет релевантные страницы:

- понижение в позициях на странице с результатами ждут сайты, которые ухудшают качество поиска. Обычно это такие веб-ресурсы, владельцы которых пытаются обмануть поисковую систему. К примеру, это сайты со страницами, на которых находится бессмысленный или невидимый текст. Конечно, он видим и понятен поисковому роботу, но не посетителю, читающему этот документ. Или сайты, которые при переходе на ссылке в зоне выдачи сразу переводят пользователя совсем на другой сайт.
- не попадают в выдачу результатов или сильно понижаются в ранжировании сайты, содержащие в себе эротический контент. Это связано с тем, что часто такие веб-ресурсы используют агрессивные методы продвижения.
- зараженные вирусами сайты не понижаются в выдаче и не исключаются с результатов поиска — в этом случае пользователь информируется об опасности с помощью специального значка. Это связано с тем, что Яндекс предполагает, что на таких веб-ресурсах могут находиться важные документы по запросу посетителя поисковой системы.

Есть также некоторые алгоритмы, которые являются основополагающими для любой поисковой системы:

— Алгоритм прямого поиска.

Что это такое – вы помните, что читали замечательную историю в одной из книг. И вы начинаете по очереди искать. Взяли одну книгу – полистали – не нашли, взяли другую... Принцип понятен, но этот способ чрезвычайно долгий. Это тоже понятно.

— Алгоритм обратного поиска.

Для этого алгоритма создается из каждой страницы твоего блога – создается текстовый файл. В этом файле перечисляются в алфавитном порядке ВСЕ слова, которые ты использовал. Даже позиция этого слова в тексте указывается (координаты в тексте).

Это достаточно быстрый способ, но уже поиск происходит с какой-то погрешностью.

Здесь главное понимать, что алгоритм этот ищет не в интернете, не поиском по блогу. А в отдельно взятом текстовом файле, который создан был когда-то давно. Когда робот заходил к тебе. И эти файлы (обратные индексы) хранятся на серверах Яндекса.

Так, это были базовые алгоритмы поиска. Т.е. как Яндекс просто находит нужные документы. Но ведь документов Яндекс знает не один и даже не 100, а по последним данным из моих источников – Яндекс знает порядка 11 млрд. документов (10 727 736 489 страниц) .

И среди всего этого количества нужно выбрать документы, подходящие под запрос. И что еще важнее – опять же - нужно как-то ранжировать их. Т.е. выстроить по степени важности, а точнее по степени полезности для читателя.

Для решения этого вопроса на помощь приходят математические модели:

- Булевская мат.модель – Если слово встречается в документе – документ считается найденным. Просто на совпадение и ничего сложного.

Но тут есть проблемы. Например, если ты как пользователь введешь какое-то популярное слово, а еще лучше предлог «в», который является самым распространенным словом в русском языке и встречается в КАЖДОМ документе – то тебе выдаст такое количество результатов, что ты даже не осознаешь такую цифру, сколько тебе документов нашлось. Поэтому появилась следующая мат модель.

- Векторная мат.модель – эта модель определяет «вес» документа. Уже не только совпадение встречается, но и это слово должно встречаться несколько раз. Причем чем больше слово встречается – тем выше релевантность (соответствие).

Именно векторную модель используют ВСЕ поисковики.

- Вероятностная модель – более сложная. Принцип такой: поисковик нашел сам эталон страницы. Например, вы ищете информацию об истории Яндекса. У Яндекса хранится какой-то эталон, допустим это будет моя предыдущая статья о Яндексе.

И все остальные документы он будет сравнивать с этой статьёй. И логика здесь такая: чем более страница твоего блога похож на мою статью – тем ВЕРОЯТНЕЕ тот факт, что твоя страница блога тоже будет полезна читателю и тоже рассказывает об истории Яндекса.

Также есть (кто бы мог подумать?) ручное участие в результатах поиска. Нужна эта релевантность еще и для оценки качества работы алгоритмов.

Для этого есть штаб – их называют Асессоры. Это специальные люди, которые руками просматривают поисковую выдачу. У них есть инструкция, как проверять сайты, как оценивать и т.п. И они руками определяют по порядку подходят твои страницы поисковым запросам или не подходит.

И вот от мнения асессоров зависит качество поисковых алгоритмов. Если все асессоры скажут, что поисковая выдача не соответствует запросам – значит неправильный алгоритм ранжирования и здесь вина только Яндекса.

Если асессоры говорят о том, что только один сайт не соответствует запросу – значит, сайт улетает куда-то далеко и понижается в выдаче. Точнее не весь сайт, а только одна статья, но это «не суть».

Конечно, асессоры не могут руками и глазами просмотреть и оценить ВСЕ статьи. И на помощь приходят другие параметры, по которым проходит ранжирование страниц.

Их очень много, например:

- вес страницы (вИЦ, PageRank, пузомерки в общем);
- авторитетность домена;
- релевантность текста запросу;
- релевантность текстов внешних ссылок запросу;
- а также множество других факторов ранжирования.

Асессоры вносят замечания, а люди, которые отвечают за настройку математической модели ранжирования уже, в свою очередь, редактируют формулу, в результате чего поисковик работает более качественно.

Основные критерии оценки работы формулы:

1. Точность выдачи поисковой системы — процент документов, соответствующих запросу (релевантных). Т.е. чем меньше страниц, не соответствующих запросу присутствует — тем лучше.

2. Полнота выдачи поисковой системы — это отношение релевантных веб-страниц по данному запросу к общему количеству релевантных документов, находящихся в коллекции (совокупности страниц, находящихся в поисковой системе). Например, если во всей коллекции релевантных страниц больше, чем в поисковой выдаче, то это означает неполноту выдачи. Это произошло из-за того, что некоторая часть релевантных веб-страниц попала под фильтр.

3. Актуальность выдачи поисковой системы — это соответствие веб-страницы тому, что написано в сниппете. Например, документ может сильно отличаться или вовсе не существовать, но в выдаче присутствовать.

Актуальность выдачи напрямую зависит от того, как часто сканирует поисковый робот документы из своей коллекции.

Сбор коллекции (индексация страниц сайта) осуществляется специальной программой — поисковым роботом.

Система кластеров

Основой работы поисковых систем как Google, так и Яндекс является система кластеров. Вся информация делится на определенные области, которые относятся к тому или иному кластеру. Индексация сайтов с целью получения данных о размещенной на них информации выполняется роботами-сканерами. Существуют следующие виды сканирующих роботов: основной робот-сканер и робот-сканер, отвечающий за сбор информации на ресурсах с частым обновлением содержания. Второй тип сканирующего робота предназначен для быстрого обновления списка проиндексированных ресурсов и значения их индексов в поисковой системе. Для наиболее полного обеспечения сбора информации в системе Яндекс применяются обновления базы поиска и обновления программного кода:

- База поисковой информации обновляется несколько раз в течение месяца, при этом на поисковые запросы выдается обновленная информация с сайтов. Такая информация добавляется с помощью основного робота-сканера.

- При обновлении программного кода или «движка» выявляются недостатки и изменяются алгоритмы, отвечающие за ранжирование ресурсов в поисковой системе. Как правило, перед выходом таких обновлений Яндекс публикует соответствующие анонсы.

Основная особенность системы Яндекс, делающая популярной ее среди русскоязычных пользователей, – это способность определять различные словоформы с учетом морфологических особенностей русского языка. При этом значения запроса с помощью геотаргетинга и формул поиска преобразуется в максимально точную формулировку. Кроме того, Яндекс отличается алгоритмом по определению релевантности индексируемых страниц (релевантностью называют соотношение содержания веб-страницы к содержанию поискового запроса). Также к положительным сторонам можно отнести высокую скорость ответной реакции на запросы и устойчивую, без перегрузок, работу серверов.

Большое значение для поисковой системы имеют динамические ссылки, наличие которых может привести к отказу от индексации ресурса поисковым роботом.

В процессе индексации Яндекс распознает текстовую информацию в документах с расширениями: .pdf, .rtf, .doc, .xls, .ppt. Последние два относятся к программам входящими в комплект Microsoft Office: Excel и PowerPoint.

При индексировании сайта поисковая система считывает данные из файла robots.txt, при этом поддерживается атрибут Allow и часть метатегов, а метатеги Revisit-After и Keywords игнорируются.

Так как сниппеты – краткие описания текстовых документов – составляются из фраз на искомой странице, то использование описания в теге не является обязательным, но может использоваться в отдельных случаях.

По заявлениям разработчиков кодировка индексируемых документов определяется автоматически, а значит, и метатег кодировки не имеет большого значения.

Поисковая система большое значение придает показателю последнего изменения информации (Last-Modified). Если сервер не будет передавать эту информацию, то процесс индексации данного ресурса будет происходить намного реже.

Пока что остается нерешенной проблема страниц, использующих фреймовые структуры, но она может быть обойдена с помощью скриптов, отправляющих пользователей поисковой системы в нужное место сайта.

Если у сайта существуют «зеркала» (например, http://www.site.ru, http://site.ru, https://www.site.ru, https://www.site.ru), необходимо принять соответствующие действия для исключения их из процесса индексации. Если индексацию «зеркал» избежать не удалось, можно «склеить» их путем внесения необходимой информации в robots.txt.

В случае попадания сайтов в Яндекс.Каталог система будет идентифицировать их как заслуживающих отдельного внимания, что может повлиять на продвижение сайтов. Также это способствует упрощению процедуры определения тематики сайта, что в свою очередь означает получение сайтом значимой внешней ссылки.

Команда поисковой системы Яндекс держит в секрете IP-адреса своих роботов. Но в лог-файлах отдельных сайтов можно встретить текстовые пометки, оставленные поисковыми роботами Яндекс.

Одними из самых интересных роботов-сканеров поисковой системы Яндекс можно назвать:

- Yandex/1.01.001 (compatible; Win16; I) – основной робот, занимающийся непосредственно индексацией сайтов;
- Yandex/1.01.001 (compatible; Win16; P) – робот-индексатор изображений;
- Yandex/1.01.001 (compatible; Win16; H) – робот, который выявляет «зеркала» индексируемых сайтов;
- Yandex/1.02.000 (compatible; Win16; F) – робот-индексатор пиктограмм ресурсов (favicons);
- Yandex/1.03.003 (compatible; Win16; D) – робот, который обращается к страницам, добавленным с помощью формы «Добавить URL»;
- Yandex/1.03.000 (compatible; Win16; M) – задействуется при переходе на страницу посредством ссылки «Найденные слова»;
- YaDirectBot/1.0 (compatible; Win16; I) – этот робот отвечает за индексацию страниц ресурсов, принимающих участие в рекламной сети Яндекс.

Из всех поисковых роботов самый важный так и называется – основной поисковый робот. От того, как он проиндексирует страницы сайта, будет зависеть значимость ресурса для поисковой системы.

Работа всех роботов происходит по индивидуальному расписанию, и если сайт проиндексирован одним из них, то это не значит, что скоро будет произведена индексация и другим.

В помощь основным созданы и роботы, которые периодически посещают сайты и устанавливают, насколько те доступны. К таким можно отнести роботов «Яндекс.Каталога» и рекламной сети Яндекс. [6]

Для поисковой системы Яндекс характерны следующие основные показатели внешней оптимизации:

- тИЦ – это общедоступный тематический индекс цитирования, он не оказывает прямого влияния на ранжирование и используется для определения позиций в тематической категории Яндекс.Каталога; применяется, когда необходима раскрутка сайта, тИЦ показывает, какое количество ссылок, в среднем, обращается к сайту.
- вИЦ, или взвешенный Индекс Цитирования, представляет собой алгоритм для подсчета количества внешних ссылок; значение его не разглашается и используется поисковой системой как определяющее при ранжировании сайтов в поисковой системе.
- Присутствие сайта в «Яндекс.Каталоге».
- Общее число страниц сайта, принявших участие в индексации.
- Частота, с которой индексируется содержимое сайта.
- Наличие и отсутствие ссылок с сайта, присутствие сайта в поисковых фильтрах.

Индекс цитирования создает основу для тематического и взвешенного индекса цитирования, которые влияют на ранжирование сайта.

Индекс цитирования (ИЦ) — это указатель цитирований (количества ссылок на источник) между публикациями, позволяющий узнать, какие из более поздних документов ссылаются на более ранние работы, при этом, ИЦ может рассматриваться как для отдельных статей, так и для авторов (ученных).

В поисковой системе Яндекс, а также в других поисковых системах, под индексом цитирования подразумевается количество обратных ссылок, без учета ссылок со следующих ресурсов: немодерируемых каталогов, досок объявлений, сетевых конференций, страниц серверной статистики, XSS ссылки и другие, которые могут добавляться без контроля со стороны владельца ресурса.

Заключение

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.

Самой лучшей иностранной поисковой системой по последним данным является Google, так как 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 неправильно. Поэтому для удобства поиска и полноты информации следует пользоваться несколькими поисковиками вводя в них нужные запросы. Из многих Российских поисковиков выделяется Яндекс, для них характерно постоянное обновление баз данных что, обеспечивает именно актуальность и точность предоставляемой информации.

На сегодняшний день не существует уже чисто поисковых систем. Кроме функции самого поисковика разработчики дают пользователям возможности пользоваться услугами почты, электронными деньгами, системами общения в среде пользователей поисковой системы, а также другие приятные мелочи, заключающиеся в показе погоды, пробок и прочего что зависит от вкуса самих разработчиков. Поисковые системы превращаются в ЭКОсистемы предоставления услуг и сервисов, которые будут полезны пользователям – оптимизировать время и энергию для достижения необходимых задач. То, что зарождалось как «закладки», в итоге полностью меняет представление о возможностях интегрирования с интернетом.

Список использованной литературы

- Открытые источники интернета.