Содержание:

Введение

В России 70,4% (84 млн.) населения пользуется глобальной сетью. Все население Земли составляет 7,3 миллиарда человек из них 3,5 миллиарда (почти 48%) пользователей Интернета. Интернет имеет большое значение и нужен почти всем. Пользователи сети могут найти в ней много полезной и развлекательной информации и использовать её по своему желанию.

Мой выбор темы для моей курсовой работы «Анализ поисковых систем в сети Интернет» обусловлен актуальностью темы, а так же тем, что я постоянно пользуюсь ресурсами Интернета: различными приложениями для телефона, ищу нужную мне информацию, смотрю ролики на YouTube. Даже моя работа имеет непосредственное отношение к глобальной сети: я работаю специалистом по внедрению облачной CRM системы в компании.

Ресурсы глобальной сети уже много лет уже давно имеет не только развлекательный контент, их использует большинство людей самых разных профессий, а некоторые даже имеют свой весьма успешный бизнес в Интернете. Объем данных, хранящихся в Интернете, вплотную приблизился к отметке в 1,1 зеттабайта. Такие данные приводит аналитическая компания IDC, проводившая исследование по заказу EMC Corporation. 3,5 миллиарда - количество пользователей интернета в мире (всё население Земли составляет 7,3 миллиарда человек).

В сети каждые несколько часов появляется много новой информации, и большинство этой информации, конечно же, никто не нашел, и поэтому она оказалась бы никому не нужной и недоступной. Возникли потребности создать такие средства, которые бы позволили ориентироваться в информационных ресурсах Интернета, быстро находили бы необходимую информацию, при этом были бы простыми и интуитивно понятными обычным пользователям.

В это время в интернете начинают появляться поисковые средства. Несколько лет назад говорили: в Интернете ничего невозможно найти, но там есть всё. Однако когда поисковые каталоги, программы, машины развились ситуация сильно поменялась. Сейчас в глобальной сети можно очень быстро найти нужную информацию, на порядок быстрее, чем скажем в бумажном источнике, например в

словаре или книге.

Самым актуальным и распространенным способом поиска в Интернете служит применение различных поисковых систем (лично мне удобнее всего пользоваться Google). Давайте разберемся, что можно назвать поисковой системой.

Поисковая система (англ. search engine) — это компьютерная система, предназначенная для поиска информации. Одно из наиболее известных применений поисковых систем — веб-сервисы для поиска текстовой или графической информации во Всемирной паутине. Существуют также системы, способные искать файлы на FTP-серверах, товары в интернет-магазинах, информацию в группах новостей Usenet. Проще говоря это портал, который ищет, собирает и сортирует информацию в Интернете, это инструмент, который позволяет в самые короткие сроки отыскать нужную информацию .Основная задача каждой поисковой системы – чтобы каждый человек нашел конкретно ту информацию, которая ему нужна.

Когда пользователь получает результат своего поиска, он как правило оценивает работу системы: Нашел он то, что искал или нет? Если нет, то сколько раз он перефразировал запрос? Актуальна ли та информация , которую он нашел? Как долго поисковая машина обрабатывала его запрос? Какой по счету был нужный результат первым (в реальности это бывает редко, так как в основном первые 1-5 мест занимает реклама) или ему пришлось долго пролистывать страницы? Каково отношение найденной нужной информации и ненужного мусора? Найдется ли, а если найдется то через какое время эта нужная информация? Когда я попользовался многими поисковыми системами и позадавал себе эти вопросы, я решил использовать Google.

Глава 1. Теория

1.1 Поисковая система - это

Поисковая система – сайт, к которому обращается пользователь и по ключевым словам находит нужную ему информацию. На сегодняшний день поисковые системы – это лучший способ найти в интернете нужную вам информацию максимально быстро и качественно.

Давайте разберемся, как работает поисковая система, что сделать достаточно несложно. Пользователь заходит на сайт системы, в специальное окно вводит ключевую фразу или слово и нажимает кнопку «поиск», по этой фразе или слову система начинает поиск информации (во многих поисковых системах на сегодняшний день есть автозаполнение, зачастую случается что в процессе написания фразы или слова уже видишь нужный тебе запрос).

После этого пользователю выдает список ссылок на сайты, которые соответствуют текущему запросу. Пожалуй это можно назвать принципом работы поисковой системы со стороны пользователя. Сейчас давайте рассмотрим внутреннее устройство и полностью процесс работы системы, который обычным пользователям незаметен.

1.2 История

Когда Интернет только появился, и начал развиваться, количество его пользователей было не очень большим, и доступной для пользователей информации было очень мало. В те годы пользовались и имели доступ к Интернету в основном работники научно-исследовательской сферы. Да и надо сказать необходимость поиска информации в Интернете не была настолько сильной как на сегодняшний день.

Создание открытых каталогов сайтов стало пробным способом организации доступа к информационным ресурсам Интернета, в этих каталогах группировались по тематике ссылки на разные ресурсы. Сайт Yahoo.com стал первым похожим проектом, он был открыт в апреле 1994 года (правда работала она на тот момент только на японском языке). После того , как количество сайтов в каталоге Yahoo подросло, нужную информацию стало возможным искать в каталоге. Это еще не была поисковая система в полном смысле этого слова, и каким мы знаем их на сегодняшний день, потому что область поиска была ограничена ресурсами, которые присутствовали в каталоге Yahoo, не во всех ресурсах Интернета.

В настоящее время каталоги потеряли свою популярность, потому что даже в наиболее больших современных каталогах имеется информация лишь о небольшой части всего интернета. Например: DMOZ (он ещё называется Open Directory Project) – один из самых огромных каталогов хранит информацию о 5 миллионах ресурсов, а база поисковой системы Google состоит более чем из 8 миллиардов документов.

«WebCrawler» вышла в мир в июне 1995 года, это была первая полноценная поисковая система. Эта поисковая система отличалась от последователей, главное отличие заключалось в предоставлении пользователю возможности искать информацию на любой веб-странице, по любым ключевым словам. Сейчас эта технология – стандарт поиска любой поисковой системы. Таким образом , «WebCrawler» стала первой поисковой системой, которой пользовались не только научные работники, а еще и широкий круг простых пользователей.

В 1996 году появились поисковые системы Lycos и AltaVista. Через год AltaVista стала доступна русскоязычным пользователям. В этом же году были запущены отечественные поисковые системы - Rambler.ru» и «Aport.ru».

Рунет (интернет на русском языке) вышел на новый уровень, когда появились первые отечественные поисковые системы, эти системы позволили всем русскоязычным пользователям осуществлять запросы на русском языке, и быстро узнавать об изменениях, которые происходят внутри глобальной сети.

В 1997 году была запущена поисковая система «Яндекс», после этого конкуренция между отечественными поисковыми машинами была очень сильной, поисковые системы начали улучшать систему поиска, индексации сайтов, выдачи результатов, начали предлагать новые услуги и сервисы.

Сергей Брин и Ларри Пейдж в 1997 году, в рамках исследовательского проекта в Стэндфордском университете, создали поисковую машину Google.

Сейчас 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 сегодняшний день поисковая машина Google обрабатывает более 60 миллиардов запросов в месяц, а это более 63% всех поисковых запросов в мире.

1.3 Задачи

Основные задачи поисковых систем:

-поиск ранее не известных сайтов

-анализ сайта

-максимально верный ответ пользователю

Основная задача каждой поисковой системы, найти пользователю ту, информацию, какую он ищет. И в то же время, к сожалению невозможно обучить пользователя производить “верные” запросы к системе, т.е. запросы, которые ответствуют принципу работы поисковых систем. Именно поэтому создателям надо делать такие принципы работы и алгоритмы поисковых систем, которые бы позволяли пользователям находить ту информацию, которую они ищут.

Это говорит о том, что поисковая система обязана думать также как думает пользователь, когда ищет информацию. Когда пользователь обращается к поисковой системе, он хочет максимально просто и быстро отыскать информацию, которая его интересует. Уже после получения результата, пользователь оценивает работу системы, при этом он руководствуется некоторыми главными параметрами. Конструкторы поисковых систем всегда стараются совершенствовать алгоритмы и принципы поиска, стремятся ускорить работу системы, добавляя новейшие функции и возможности, с целью удовлетворить потребности пользователей.

1.4 Принципы работы

Поисковая машина (движок) – это программная часть поисковой системы, которая используется для сбора, обработки и представления данных пользователю. Именно эта часть составляет основу поисковых систем, которая отличает одну систему от другой. Она осуществляет быстрый поиск внутри сервера или Интернет-ресурса нужной информации. У всех без исключения поисковых систем основа движка примерно одинакова. Чаще всего, это программное обеспечение, которое отвечает за ранжирование результатов по релевантности запроса и составление каталога, поисковый бот, он необходим для поиска сайта и индексации. Однако есть такие крупные поисковые системы, которые содержат содержание своей поисковой системы в секрете. Главным отличаем является учет и релевантность морфологии языка запроса, база проиндексированных сайтов. Это все в совокупности и определяет критерий качества работы поисковых машин.

Поисковые машины классифицируются по области поиска информации:

1. Глобальный поиск.

Он предназначен искать информацию по региональной части, или по группе сайтов, или в сети Интернет. Глобальным поиском пользуются большинство крупных поисковых систем , таких как Яндекс, Yahoo, Google и т.д.

1. Локальный поиск.

Он осуществляет поиск информации по глобальной сети какой-либо её части, к примеру, по локальной сети, или по нескольким сайтам. Таким примером являются снутренние серверы солидных фирм или скрипт поиска на сайте.

Поисковые машины осуществляют разнообразный поиск по сети интернет. К примеру, картинки, географическое положение, музыка, личная информация и др. Поисковая машина может работать с файлами самых разнообразных форматов (.html,.htm,.txt,.doc,.rtf, и др.), мультимедиа (звук, видео и др.) , графического типа (.gif,.png,.svg). Самым распространенным поиском считается поиск текстовых документов (документы в формате .doc,.rtf,.txt, web-страницы и др.). С технологической точки зрения поиск по картинкам, звукам, видео более сложен, посему он не реализован массово. Например, такая система как Яндекс.Картинки ищут картинки по альтернативным текстам, которые соответствуют этим изображениям, а не по самим изображениям. В Google каталог поиска картинок составлен вручную, что тормозит обновление баз изображений, и в то же время значительно увеличивает релевантность запроса.

Модуль индексирования состоит из трех вспомогательных программ-роботов:

Spider – это программа, предназначенная для скачивания веб-страниц. ”Spider” полностью обеспечивает скачивание страницы, и извлекает все внутренние ссылки из этой страницы. Html-код скачивается с каждой страницы. Роботы используют протоколы HTTP Для того, чтобы скачать страницу. “Spider” работает следующим образом : робот передает на сервер запрос “get/path/document” и несколько других http команд запроса. В ответ приходит текстовый поток, содержащий сам документ и служебную информацию.

Ссылки извлекаются из тегов Frame, Base, Area, Frameset и др. Почти все роботы, наряду со ссылками обрабатывают перенаправления (редиректы).

Все страницы сохраняются в формах:

- дата, когда страница была скачана
- тело страницы html-код

- URL страницы
- http-заголовок ответа сервера

Crawler – это программа, которая автоматически проходит по всем

ссылкам, которые находит на странице. Её задача в том, чтобы исходя из заведомо заданного списка адресов или основываясь на ссылках, определить, куда дальше должен идти Spider. Crawler осуществляет поиск более новых документов, которые еще не известны поисковой системе, следуя по найденным ссылкам.

Indexer – это программа, которая анализирует веб-страницы, которые скачали Spider и Crawler. Индексатор, используя свои логические и морфологические алгоритмы, разбирает страницу на составные части и анализирует их. Все элементы страницы подвергаются анализу, например заголовки, текст, html-теги, ссылки, структурные и стилевые особенности и др.

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

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

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

Рассмотрим, как работает поисковый сервер:

- Запрос, который получен от пользователя подвергается морфологическому анализу. Генерируется информационное окружение всех документов, содержащихся в базе (как раз оно и будет отображено в виде сниппета, т. е. текстовой информации соответствует запросу на странице выдачи результатов поиска).
- Все данные передаются специальному модулю ранжирования в качестве входных параметров. После чего по всем документам происходит обработка данных, потом 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.

1.5 Поисковые системы сейчас

Самые известные поисковые системы во всем мире это : Google, Baidu,

Bing, Yahoo!, Ask Network, AOL, Inc.

Самые известные поисковые системы в России : Google, Яндекс, Mail.ru, Рамбрер.

В целом основные “русскоязычные” поисковые системы находят и индексируют тексты на нескольких языках – украинском, татарском, английском, белорусском и др. От всеязычных систем их отличает то, что они почти всегда индексируют те ресурсы, которые расположены в доменных зонах, где на первом месте стоит русский язык, а также тем, что они своих роботов ограничивают русскоязычные сайты и другими способами. В отличае от них всеязычные поисковые системы индексируют все подряд документы.

Русскоязычные поисковые системы : Яндекс, nova.rambler, Mail.ru, Нигма.

По данным исследования, которое проводилось в 2016 году, первое место в мире занимает Google. На сегодняшний день на долю Google приходится 65,97 % рынка.

Второе место занимает Bing – 8,28% рынка и Baidu 7,54% рынка.

Поисковая система	Доля поисковых запросов
Google	75,97%
Bing	8,28%
Baidu	7,54%
Yahoo	6,56%
AOL	0,10%
Bing	8,28%

Рейтинг мировых поисковых систем (2016год)

Пятерка лучших поисковых систем Рунета :

1. Яндекс – 48,3%;
2. Google – 45,1%;
3. Mail – 5,7%;
4. Rambler – 0,4%;
5. Bing – 0,3%.

Глава 2. Практика

2.1 Google. Принцип работы

Алгоритм ранжирования Google намного сложнее, чем алгоритм Янедекса. В Google продвигать сайты, особенно на начальном этапе, немного сложнее. Раскрутка нового сайта в Google затруднительна, потому что на новые веб-ресурсы накладываеся фильтр (“песочница”). Google при ранжировании пользуется примерно 200 факторами, оптимизатор может повлиять лишь на некотрые.

С другой стороны, Google выглядит стабильнее своих конкурентов в плане смены алгоритма и апдейтов. Информация, недавно размещенная на сайте, может в считанные минуты попасть в основную выдачу. Поисковые роботы Google в три раза быстрее, чем поисковые роботы других поисковых систем. Критерии “нормальности” сайта (фильтры) почти не меняются с момента начала их внедрения.

Контент и ссылки – это 2 фактора, на которые может повлиять оптимизатор при продвижении сайта в системе Google.

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

Внешние ссылки Google учитывает по некотрым параметрам: кол-во, авторитетность сайта-донора (доверие поисковой системы сайту), тематичность. Сквозные ссылки (ссылки, которые ведут со всех страниц сайта-донора, устанавливается, например, в шаблоне сайта) в глазах Google обладают большим весом, нежели 10 ссылок (даже с этого сайта-донора).

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

Для продвижения сайта в Google нужно:

- 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.

Эти действия позволят поисковому роботу Google вернее и полнее проиндексировать сайт и выделить место на страницах своей выдачи.

Понятие Google PageRank является одним из важных моментов в работе поисковой машины Google. Наряду с другими параметрами, которые влияют на сортировку (выдачу) сайтов в результатах поиска, знание можели PageRank необходимо как для понимания процесса поиска, так и для использования оптимизаторами при продвижении своих сайтов в поисковой системе.

PageRank это числовая величина – так называемая, мера “важности” страницы в поисковой системе Google. Зависит от числа внешних ссылокна эту страницу и от их веса (важности). Другими словами от кол-ва и качества ссылающихся страниц. И если говорить математических языком, то PageRank- это алгоритм расчёта авторитетности страницы, используемый поисковой системой Google. PageRank не является основным, но является одним из вспомогательных факторов ранжировании сайтов в результатах поиска.

Следует отметить, что при расчете PageRank, Google учитывает не вообще все ссылки, а отфильтровывает ссылки с сайтов, специально предназначенных для скопления ссылок. Есть такие ссылки, которые могут не только не учитываться, но и негативно сказаться на ранжировании ссылающегося сайта (поисковая пессимизация).

Основной формулой для расчета PR является формула:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right)$$

где PR(Ti) – значение PageRank для страницы;

d – демпфирующий коэффициент, он отражает то, какую долю веса может передать страница-донор на страницу-акцептор. Обычно его считают равным 0.85, что означает, что страница может передать 85% веса (распределяется межу всеми акцепторами, на которые ссылается донор).

В других источниках d является вероятностью, с которой пользователь перейдет на один из акцепторов, а не закроет браузер, что, в целом, то же самое. Какое числовое значение у этого параметра известно только в Google, остальные из экспериментальных данных принимают его равным 0,85;

n- кол-во страниц, ссылающихся на траницу-акцептор (на которые наложен фильтр);

Ti – i-ая ссылающаяся страница;

C(Ti) – кол-во ссылок на странице-доноре Ti.

Поскольку ссылающихся страниц может быть очень много, и общее кол-во страниц в поисковой системе Google достаточно велико (около десятка биллионов штук), а также их количество постоянно растет, то представлять вес страницы в абсолютных значениях для Webмастеров было бы весьма неправильно. Для этого ввели такое понятие, как TLPR – ToolBar PageRank – значение PageRank, который имеет значение от 0 до 10 (шкала в Google Toolbar).

Для того, чтобы уложить все веса страниц между значениями от 0 до 10 используют логарифмическую шкалу. ToolBar PageRank определяется по формуле:

$$TLPR = \log_{base}(PR) \cdot a,$$

где base – основание логарифма, которое зависит от кол-ва страниц в поисковой машине (возможно и от ряда других факторов).

Принимается равным семи;

a – коэффициент приведения, который удовлетворяет $0<a\leq 1$

Из всего, сказанного выше не совсем верно делать выводы что нулевой TLPR означает реальный PageRank (далее PR) равен нулю. По формуле PR видно, что даже при n=0, мы получаем минимальный PRmin = (1-d) = 0,15. Это значение соответствует TLPR ≈ -1.

При таких значениях Toolbarного PR считается, что PR = N/A (либо он еще не определен), и в то же время от оказывает влияние на распределение веса между ссылками-акцепторами. Также следует заметить, что Toolbarное значение предназначено только для отображения Webмастерам в Goodle Toolbar и никак не влияет на позицию в выдаче. На неё влияет реальный PR страницы.

Исходя из принципов расчета Goodle PageRank, можно теперь несложно рассчитать, с каких ссылок надо ссылаться и сколько их нужно, чтобы получить тот или иной PR.

Также можно прогнозировать PR. Один из важных выводов : если у нового сайта более 10000 страниц (число страниц зависит от кол-ва ссылок с них на другие страницы), они правильно перелинкованы а также каждая ссылается на главную

страницу, то главная страница получит хороший вес от этих ссылок. Принимая во нимание, что минимальный PR равен 0,15 и в среднем на одной странице 10 ссылок, для такого сайта вычисляется по формуле PR:

$$PR = (1 - 0{,}85) + 0{,}85 \cdot 0{,}15 \cdot \frac{10000}{10} = 127{,}65$$

A ToolBar PageRank по формуле TBPR:

$$PR = (1 - 0{,}85) + 0{,}85 \cdot 0{,}15 \cdot \frac{10000}{10} = 127{,}65$$

Это пример очень хорошего PR без единой внешней ссылки с др. сайтов.

Таким образом, существует очень много способов повышения веса своих страниц, но основная идея – это качественные ссылки с других сайтов. Для этого можно использовать каталоги, социальные закладки, статьи, блоги, форумы а также другие виды сайтов. Однако не следует расставлять множество ссылок на других сайтах, потому, что помимо PageRank существует множество других ранков, влияющих на выдачу страницы в результатах поиска (например TrustRunk).

Отрицательного PR не бывает. Реальный PR минимум равен 0,15, минимальный Toolbarный PR равен нулю.

Ссылки на своем сайте на другие сайты ставить нужно, так как своими ссылками вы увеличиваете PR страниц-акцепторов и тем самым, по первой формуле, к вам возвращается еще более больший вес из большой системы ссылок. На значение PR влияет только количество и качество ссылающихся ресурсов.

С картинок PageRank “ перетекает “, только если они являются ссылками, по которым пользователь имеет возможность перейти на другой ресурс.

2.2 Яндекс. Принцип работы

Основа работы Яндекс и Google – это система кластеров. Таким образом, информация разделяется на области, относящиеся к тому или иному кластеру. Индексация разных сайтов выполняется с помощью роботов-сканеров целью которых является получить данные о информации, размещенной на них. Существуют два вида сканирующих роботов: робот-сканер основной и робот-

сканер, который отвечает за сбор информации на тех ресурсах, где содержание часто обновляется. Другой тип сканирующих роботов предназначен для того, чтобы быстро обновить список проиндексированных ресурсов, а также значения индексов этих ресурсов в поисковой системе. Для обеспечения сбора как можно более полной информации в Яндексе применяются обновления программного кода и обновления базы поиска:

- При обновлении «движка» (программного кода) выявляются различные недостатки и меняются алгоритмы, которые отвечают за ранжирование ресурсов в поисковой системе. Как правило, перед выходом подобных обновлений Яндекс делает соответствующие анонсы.
- База поисковой информации обновляется как правило два-три раза в месяц, при данном событии на поисковые запросы выдается уже более новая информация с этих самых сайтов. Подобная информация добавляется основным роботом-сканером.

Главное свойство системы Яндекс, делающая её распространенной среди русскоязычных пользователей – способность определять различные словоформы, причем с учетом морфологических особенностей русского языка. Причем значения запроса преобразуется в максимально точную формулировку благодаря геотаргетингу и формулам поиска. Кроме этого, Яндекс отличается особенным алгоритмом, который определяет релевантность индексируемых страниц (релевантность - это соотношение актуальности содержания поискового запроса к содержанию веб-страницы). Высокую скорость ответной реакции на запросы также можно отнести к положительным сторонам, как и устойчивую к перегрузкам, работу серверов.

Для поисковой системы Высокое значение имеют динамические ссылки, их присутствие в силах привести к отказу поисковым роботом от индексации ресурса.

При индексации Яндекс распознает различную текстовую информацию в документах с такими расширениями, как: .rtf, .pdf, .ppt., doc, .xls. Последние три относятся к программам входящим в комплект Microsoft Office: PowerPoint , Word и Excel.

При индексировании сайта поисковой системой, эта система считывает данные из файла который имеет название robots.txt, при этом игнорируются Revisit-After и Keywords метатеги, а поддерживается непосредственно такие вещи как атрибут Allow и часть метатегов.

Так как сниппеты (сниппет - сокращенное изложение текстовых документов) на искомой странице составляются из фраз, то применение описания в теге не всегда является обязательным, таким образом оно может использоваться в отдельных случаях.

Кодировка индексируемых документов, по заявлениям создателей автоматически определяется, а это значит, что и метатег кодировки не обладает важным значением.

Особо большое значение поисковая система придает показателю того, когда в последний раз менялась информация (Last-Modified - последнее изменение информации). В том случае, если сервер не будет передавать нужную информацию, тогда намного реже будет происходить процесс индексации данного ресурса.

Пока что проблема страниц, использующих фреймовые структуры, остается нерешенной, но её можно достаточно легко обойти при помощи скриптов, которые бы отправляли пользователей в нужное место сайта, непосредственно из поисковой системы.

Если у сайта пристутствуют «зеркала» (для примера, http://www.mypersonalsite.ru, http://mypersonalsite.ru, https://www. mypersonalsite.ru, https://www.mypersonalsite.ru), тогда совершенно необходимо как можно быстрее предпринять соответствующие мероприятия для того, что бы исключить их непосредственно из процесса индексации.

А вот если не удается избежать индексацию «зеркал», можно их «склеить» следующим образом: внести необходимую информацию в файл, который имеет название robots.txt.

В случае, когда сайт попадает в Яндекс.Каталог система их будет идентифицировать как те, которые заслуживают отдельного внимания, что в свою очередь может напрямую повлиять непосредственно на продвижение этого самого сайта. Также это сильно способствует собственно упрощению определения тематики сайта, а это в свою очередь означает, что сайт получит значимую внешнюю ссылку по сравнению с другими сайтами, которые не попали в этот самый каталог.

Владельцы поисковой системы Яндекс IP-адреса своих роботов держит в секрете. Но оставленные поисковыми роботами Яндекс текстовые пометки, можно

встретить в лог-файлах отдельных сайтов.

Самыме интересные роботы-сканеры поисковой системы Яндекс:

- Yandex/1.01.001 (compatible; Win16; I) – занимающийся только индексацией сайтов, основной робот ;
- Yandex/1.01.001 (compatible; Win16; P) – робот-индексатор различных изображений;
- Yandex/1.01.001 (compatible; Win16; H) – робот, который выявляет «зеркала» сайтов, которые подвергаются индексации;
- Yandex/1.02.000 (compatible; Win16; F) – favicons робот-индексатор пиктограмм ресурсов ;
- Yandex/1.03.003 (compatible; Win16; D) – робот, обращающийся к страницам, которые добавлены с помощью формы «Добавить URL»;
- Yandex/1.03.000 (compatible; Win16; M) – робот, задействующи йся при переходе на страницу посредством ссылки «Найденные слова»;
- YaDirectBot/1.0 (compatible; Win16; I) – робот, отвечающий за индексацию страниц ресурсов, он принимает участие в рекламной сети Яндекса.

Основной поисковый робот - самый важный из всех поисковых роботов. От того, каким образом он проиндексирует страницы сайта, напрямую будет зависеть значимость для поисковой системы данного ресурса.

Работа каждого робота происходит по индивидуальному расписанию, и в том случае, когда сайт проиндексирован одним из них, это не значит, что в ближайшее время другим будет также произведена индексация.

В помощь основным роботам созданы также и роботы, которые периодически посещают сайты и проверяют их доступность. К ним можно отнести роботов рекламной сети Яндекс и «Яндекс.Каталога».

Следующие основные показатели внешней оптимизации характерны для поисковой системы Яндекс:

- тИЦ - тематический индекс цитирования, не оказывающий прямого влияния на ранжирование , он используется для определения позиций в тематических категориях Яндекс.Каталога; применяется он тогда, когда раскрутка сайта необходима, также тИЦ может показать количество ссылок, в общем, обращающихся к сайту.

- вИЦ - взвешенный Индекс Цитирования, он представляет из себя алгоритм, который подсчитывает количество внешних ссылок; его значение используется поисковой системой как определяющее при ранжировании в поисковой системе и не разглашается .
- Фигурирует или нет сайт в «Яндекс.Каталоге».
- Общее число страниц, которые приняли участие в индексации сайта.
- Как часто индексируется содержимого сайта.
- Отсутствие и наличие ссылок с сайта, присутствует ли сайт в поисковых фильтрах.

Основой для тематического и взвешенного индекса цитирования, которые влияют на ранжирование сайта служит индекс цитирования сайта.

ИЦ **(Индекс цитирования)** - указатель цитирований (количество ссылок на первоисточник) между публикациями, он позволяет узнать, какие из более поздних работ ссылаются на более ранние документы. При всем при этом, ИЦ может рассматриваться как и для авторов (ученных) , так и для отдельных статей.

В поисковой системе Яндекс, и других поисковых системах, под индексом цитирования понимается количество обратных ссылок, без учета ссылок таких ресурсов как : страниц серверной статистики, досок объявлений, сетевых конференций, XSS ссылок, немодерируемых каталогов и других, которые способны добавляться без контроля со стороны владельца сайта.

Тут нужно сказать о том, что в каталоге Апорт под ИЦ понимается не индекс **цитирования**, а взвешенный индекс цитируемости.

Этот индекс можно рассчитать из ссылочного графа: если рассматривать ресурсы сети как вершины графа, а ссылочные связи между сайтами (цитирование других ресурсов) как связи вершин ребра (графа), в таком случае ссылочный граф представляется в виде диаграммы, как показано на рисунке ниже.

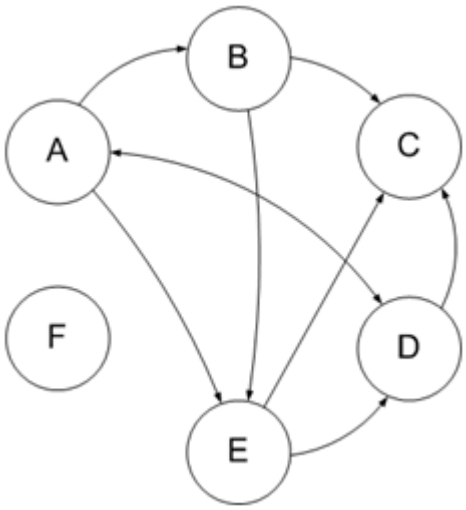

Рисунок – Ссылочный граф

На рисунке буквами A, B, …, F обозначены конкретные сайты в индексе поисковой системы, а стрелки изображают направление связей —двусторонние либо односторонние.

ИЦ используется для ранжирования документов в поисковой выдаче, но не является главным фактором.

Однако путать обычный индекс цитирования с взвешенным и тематическим, о которых будет написано несколько позже не стоит. Индекс цитируемости всегда непосредственно целое число и не имеет зависимости от тематик ссылающихся на них документов.

Таким образом ИЦ (Индекс цитируемости) обычно рассматривается в качестве одного из основных параметров значимости статьи, и в то же время он не отражает структуру ссылок в каждой тематике (дисциплине), получается труды с большой значимостью и слабозначимые работы могут иметь одинаковый индекс цитируемости.

Именно поэтому ввели взвешенный индекс цитирования, определяющийся не одним количеством, а еще и качеством ссылающихся источников.

Введение статической ссылочной популярности и ссылочного поиска помогло всем поисковым системам справляться с примитивным текстовым спамом, полностью разрушающим традиционные статистические алгоритмы информационного поиска, которые были получены в свое время для контролируемых коллекций. ВИЦ это аналог PageRank от поиска Google.

Из ссылочного графа рассчитывается как взвешенный индекс цитирования, как и другие ссылочные факторы ранжирования.

вИЦ для своих страниц можно узнать примерно, проверив их PageRank практически любым онлайн-сервисом проверки, и в то же время, стоит учесть, что в индексе Яндекса фигурируют в основном русскоязычные документы, зарубежных очень немногие, самые популярные, урезая ссылочный граф по сравнению с Google.

Тематический индекс цитирования был введен непосредственно для отражения авторитетности сайта в пределах своей тематики.

Когда определяется тематика сайта, сначала должно строится описание ресурса, который рассматривается (из структуры URL его страниц, названия категорий сайта и заголовков).

Потом высчитывается оценка близости между описаниями тематик заранее подготовленных (каталог) и описаниями ресурсов, где существует выбор наиболее близких тематик для них.

Таким образом тематическая близость двух документов выражает вероятность принадлежности обоих к одной и той же тематике. Этот показатель способен влиять на значение веса, передаваемого этой ссылкой.

Формула расчета тИЦ:

$$PF(v,t) = \frac{n_v}{N} \cdot \sum_{i \in P} \frac{PF(i,t) \cdot w(i)}{N(i)}$$

где PF(v,t) – тИЦ ресурса v;

P - количество ресурсов, ссылающихся на сайт v и имеют ту же тематику;

Nv - количество страниц на сайте v;

N - общее число страниц в индексе Яндекса (при этом, а nv/N - вероятность того, что пользователь читает сайт v);

w(i) – частота цитируемости ресурсом i сайта v;

N(i) – общее число ссылок на i-ом сайте.

При этом, PF(v,t) является нормализованной величиной.

Изначально тематический индекс цитирования отражал положение дел в Рунете, однако со временем индекс Яндекса существенно расширился на такие географические сегменты, как Украина, Беларусь, Казахстан и другие. В Яндексе появились более новые версии каталога для других регионов.

Таким образом, для того чтобы ранжировать сайты в каждом из региональных Яндекс.Каталогов, понадобилось ввести региональный тИЦ, учитывающий, помимо тематической, также географическую близость ссылок.

Соответственно, тИЦ обладает следующими свойствами:

1. тИЦ зависит от числа уникальных страниц на сайте и чем их соответственно больше, тем больше становится результирующий показатель.

2. А чем меньше исходящих ссылок на сайте-доноре, тем больше с него передается тИЦ.

3. тИЦ никаким образом не зависит от перелинковки.

4. Анкоры ссылок никак не участвуют в определении тематической близости двух сайтов.

1. При наличии у ресурса нескольких копий (зеркал), при их склейке результирующий тИЦ суммируется.

2.3 Поисковые машины

Потрясающе, однако эта необыкновенно знаменитая система, обслуживающая миллионы запросов каждый день, зародилась как обычная коллекция закладок, пополнением которой занимались всего лишь 2 человека, а именно Джерри Янг и Дэвид Фило. На настоящее время Yahoo!, это уже не только лишь каталог, а это целая группа разных сервисов, среди которых фигурируют такие как каталог Yahooligans - Yahoo! для детей, система персональных каналов My Yahoo!, бесплатный E-mail сервис, система "Shop with Yahoo!" (покупайте с Yahoo!), совместный с MTV проект MTV unfURLed и многое, многое другое.

Среди всех рассмотренных систем, Yahoo! является единственной чисто каталоговой, на Yahoo! Нет поисковой машины. И в то же время список категорий на Yahoo! является наиболее простым и полным в отличие от других каталогов, на

Yahoo! всегда несложно определить, в каком разделе находится необходимая информация. Главная странца Yahoo! грузится достаточно быстро, несмотря на то, что на ней очень много ссылок, но они все текстовые.

Центральная часть страницы, ясное дело, занята окном поиска и списком категорий. Графические Ссылки вверху страницы обеспечивают доступ к информации, такой как "More Yahoos", "что хорошего", "что нового". Рекомендуется посетить последнюю ссылку, потому что она приводит на страницу с большим количеством ссылок на различные каталоги и сервисы Yahoo. В нижней части главной страницы Yahoo! расположено огромное количество ссылок на наиболее популярные разделы каталога.

При вводе ключевых слов с главной страницы Yahoo, запрос обрабатывается по методу "Intelligent default", а именно Yahoo! ищет более подходящие результаты в областях: в категориях Yahoo; в Web-сайтах, зарегистрированных на Yahoo; на Altavista (запрос передается при отсутствии результатов); в новостях. Такой интеллектуальный поиск занимает достаточно долгий срок.

При задании критериев поиска для Yahoo! необходимо помнить о том, что Yahoo! ищет эти слова только в названии и описании страницы потому, что полнотекстового индекса на Yahoo! нет. В связи с этим не следует указывать при поиске чересчур много терминов либо синонимов - количество результатов с Yahoo! снизится или даже может быть нулевым. При вводе ключевых слов, необходимо выбрать область поиска - весь каталог Yahoo! или лишь его данный раздел. Это делается с помощью кнопок под полем ввода.

На странице с результатами поиска выводятся сперва удовлетворяющие критерию поиска категории, а уже потом сайты. Возле любой категории в скобках стоит число, которое является количеством сайтов в данной категории.

В ситуации если на Yahoo! нет результатов, тогда выводятся результаты с Altavista. Сверху и снизу страницы выводится небольшая табличка, которая позволяет одним нажатием кнопки мыши произвести поиск в категориях Yahoo!, на Altavista, в новостях а также событиях. Число результатов поиска на Yahoo!, вне всякого сомнения, невелико, зато можно сказать, что большинство из них являются релевантными.

В Yahoo! вероятна проблема с отсутствующими страницами, поскольку веб-мастера зачастую забывают удалять свои сайты из поисковых систем, а на Yahoo! нет механизма автоматического обновления. Для расширенного поиска Yahoo!

предоставляет не очень большой, зато весьма полезный набор инструментов. Для того, чтобы попасть на страничку расширенного поиска, следует перейти по ссылке "options" с главной страницы Yahoo!.

Среди инструментов расширенного поиска есть такие инструменты, как ограничение результатов по дате, поиск в Yahoo!, Usenet и среди E-mail адресов, использование логических операций над терминами, а так же поиск конкретной фразы. Также в Yahoo! присутствует возможность искать слова с произвольными окончаниями, указывать слова, которые должны либо не должны присутствовать в документе, и т.д. Только русские ресурсы в Yahoo! не добавляются, ибо в Yahoo! Inc. попросту некому смотреть и оценивать их содержимое. Однако те запросы, которые не дали результатов на Yahoo! передаются на Altavista, а там имеется отличный индекс русских ресурсов.

2.4 Каким образом осуществляется поиск

Как пишут сами создатели Yahoo!, их страница с результатами поиска существует для того, чтоб помочь пользователям находить то, что им нужно, в дружественном и удобном для работы интерфейсе.

Давайте рассмотрим более подробно разные разделы на странице с результатами поиска.

Inside Yahoo! (Внутренний Yahoo!) Это услуги или продукты Yahoo!, которые соответствует пользовательском критериям поиска. Например, если человек задал в запросе "лошадь" ("horse"), Inside Yahoo! покажет итоги поиска областями, где пользователь сможет найти разные типы информации, такие как изображения из Картинной галереи Yahoo!, элементы для продажи в Yahoo! Аукцион, факты о лошадях от Yahooligans!

Directory Category Matches (Категории директивных сделок): Эта область подсвечивает категории в Yahoo! каталоге, которые соответствуют пользовательскому запросу в поиска. В случае, если пользователь захочет увидеть совокупность сайтов по специфической теме, тогда ему следует щелкнуть по самой необходимой категории, и тогда пользователю представится список сайтов, который был собран редактором Yahoo! по этой теме.

В случае, если категорий существенно больше, чем может отображаться, то вверху справа появится ссылка "Next". Одно нажатие по данной ссылке позволит пользователю видеть как коммерческие, так и некоммерческие категории в Yahoo! каталог, которые в свою очередь соответствуют запросу поиска.

Sponsor Matches (Спонсорские сделки): Спонсорские сделки – это релевантные результаты поиска, за которые платят организации или предпринимателями и обеспечивается альтернативным (сторонним) средством доступа поискового сервера.

Web Matches (Сетевые сделки): Эти результаты показывают комбинации релевантных web-страниц и сайтов, обеспеченных сторонними средствами доступа поискового сервера и Yahoo! Каталог. Это является заданным по умолчанию стилем, в котором появляются результаты.

В случае, когда сайт, будучи перечисленным в результатах поиска, также перечислен в Yahoo! каталог, листинг результата поиска покзывает заголовок и описание, обеспеченному Yahoo! каталог. Кроме этого, пользователь будет видеть ссылку " More sites about", она находится внизу. Нажав на эту ссылку, человек сможет просмотреть совокупность сайтов по той же самой теме в Yahoo! Каталог.

В списки каталога включают сайты, которые прошли через специальную программу Yahoo!. Эти сайты заплатили Yahoo! для того, что-бы люди могли их рассматривать, они считаются включенными в Yahoo! Каталог.

2.5 Как осуществлять расширенный поиск

Расширенный поиск - особенность, которая поможет пользователю совершенствовать результаты поиска.

В поисковой системе Yahoo! возможен буквальный поиск (то есть поиск осуществляется именно по заданным словам) и расширенный поиск.

Расширенный поиск способен значительно увеличить точность результатов поиска, благодаря использованию дополнительного синтаксиса, для того, чтобы сосредоточить поиск. Человек может ввести большую часть следующих параметров поиска непосредственно в блок поиска, или выбрать их на странице расширенного поиска, на которую можно пройти по ссылке advanced search, которая находится справа от строки поиска.

Страница расширенного поиска представлена ниже.

Advanced Search

Find web pages

include all of the words:

include this exact phrase:

include at least one of these words:

exclude these words:

Search:

the Web Yahoo! Directory listings

<< Fewer options

More options

Language:

only show pages in

Country:

only show pages from

Date:

only show pages updated in the

Keyword Locations:

show pages where the keyword is

Domain:

show pages from the site or domain

e.g., yahoo.com, .org, .gov

Search by URL (Web Address)

Find web pages similar to

Find web pages that link to

Предлагаю рассмотреть данную страницу более подробно.

Include all of the words (Включите все слова) - эта опция позволяет найти те результаты поиска, которые включают в себя все слова, которые пользователь напечатали в блоке поиска. Это подобно вставке "AND" между словами или символом "+" перед словом.

Include this exact phrase (Включите эту точную фразу) - эта опция позволяет исследовать результаты, точно соответствующие словам, введенными пользователями. Это напоминает помещение цитат (" ") вокруг набора слов. (Например: вы ищете известное высказывание или цитату: "Я хочу домой").

Include at least one of these words (включите по крайней мере одно из этих слов) - это опция для поиска результатов по нескольким показателям, соответствующих одному или большему количеству слов, заданных для поиска. Это соответствует вставке "OR" между словами. (Например, если пользователь хочет найти информацию относительно катеров или лодок.)

Exclude these words (Исключите эти слова) - эта опция исключает заданные слова из поиска. В стандартном поиске это соответствует вставке "NOT" между разными словами или символом " " перед словом. (Например, вы ищете информацию о цветах, но не хотите, чтобы выдавалась информация о тюльпанах. Для этого нужно ввести "цветы" во "All of the words", а в "Exclude these words" введите "тюльпаны").

Search (поиск) - тут пользователю требуется выбрать, где он хочет искать информацию, в Сети или только лишь в Yahoo-каталоге.

More options (больше Вариантов) - использовать дополнительные опциями, которые появятся после нажатия этой кнопки. Предлагаю дать им краткое описание:

Language (язык) - позволяет выбирать, на каком языке будут отображаться сайты на странице с результатами.

Country (страна) - данная функция позволит показать результаты в зависимости от выбранной страны.

Date (дата) - ограничит результаты поиска лишь теми сайтами, которые были модифицированы в пределах прошедших 3, 6 или 12 месяцев.

Keyword Location (местоположение ключевых слов) - позволит пользователю самому выбрать условия для поиска - на странице, где-нибудь еще, в заголовке, в тексте, в URL или в ссылках на иные страницы.

Domain (домен, область поиска) - запрашивает, на каких конкретно доменах должен (или не должен) происходить поиск (например, с com, org, gov, net, biz, info, name).

Search by URL (поиск по URL) - пользователь может найти web-страницы, которые являются подобными или принадлежащими к специфическому узлу.

Заключение

В заключение хочу сказать, что все поисковые системы имеют свои плюсы и минусы, различных роботов и свои алгоритмы поиска и за последние годы поисковые системы все сильнее усовершенствуются. В основном сейчас актуальны Яндекс и Google, большую роль сыграло то, что именно эти компании разработали свои браузеры уже со встроенным поиском (например в Яндекс браузере можно ввести запрос в строку адреса и поиск будет производиться поисковой машиной Яндекса, аналогично браузер Google chrome подобный поиск производит уже своей поисковой машиной), так же эти компании зарабатывают хорошие деньги на рекламе в строке поиске и почтовых сервисах, что приносит ресурсы для развития этих поисковых систем. В заключении считаю нужным упомянуть такую поисковую систему как mail.ru, но не смотря на высокое количество пользователей почты @mail.ru, она несколько отстает от Яндекса и Google по количеству людей, которые пользуются этим ресурсом. Лично я пользуюсь поисковой системой Google, так как мне приходится искать много релевантной технической информации, и в то же время если нужно найти развлекательный контент многие люди предпочитают использовать Яндекс. Yahoo!, Rambler, Bing каталоги и поисковые системы на мой взгляд уже давно перестали пользоваться популярностью в Российской Федерации, и все же я уверен, что даже сейчас многие из них не утратили своей актуальности, и наверняка есть небольшой процент людей, которые их используют. Поисковые

системы сейчас совершенствуются очень высокими темпами, для нас уже не в новинку голосовой поиск, автозаполнение поисковой строки по самым популярным запросам, в наше время даже с помощью мобильного телефона можно зайти в поисковую систему и сделать запрос, например в Google. Даже всемирно известный видеохостинг youtube имеет свой каталог. На этом я закончу свой анализ поисковых систем в сети Интернет, но технологический и программный прогресс не стоит на месте, и возможно в скором времени будет разработано что-то новое такое, что перевернет наши представления о поисковых системах в целом.

Список использованной литературы

- 1. Поисковые системы Интернета: Яндекс, Google, Rambler, Yahoo. Состав, функции, принцип работы. https://www.seonews.ru/masterclasses/poiskovye-sistemy-interneta-yandeks-google-rambler-yahoo-sostav-funktsii/
  2. Информационно-поисковая система. Основные компоненты. http://megaobuchalka.ru/9/27414.html
  3. Интернет ресурсы, поисковые системы. https://sites.google.com/site/konspektlekcij/internet-navyki-raboty-v-seti/2-internet-resursy-poiskovye-sistemy
  4. Поисковые системы: состав, функции, принципы работы. https://studfiles.net/preview/5866029/page:20/
  5. Основные принципы поиска в информации в Internet. http://ecocyb.narod.ru/410-417/inrs5_7.htm
  6. Поисковая система. https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%B8%D1%81%D0%BA%D0%BE%D
  7. Анализ поисковых систем в сети Интернет. https://vuzlit.ru/971968/printsip_raboty_google
  8. Принцип работы Google. http://www.seodiplom.ru/3-poiskovie-sistemi/princip-raboti-google/
  9. Что такое PageRank. http://maksis.ru/chto-takoe-pagerank.html
  10. Особенности продвижения сайтов в поисковой системе Google. https://www.seone.ru/articles/osobennosti-prodvigenia-sajta-v-google/
  11. Применение алгоритма PageRank для определения весов web-страниц. https://moluch.ru/archive/102/23433/
  12. Поисковая система Yahoo! https://vuzlit.ru/995697/rasshirennyy_poisk