Содержание:

Введение

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.

Большинство пользователей поисковых систем никогда не задумывались (либо задумывались, но не нашли ответа) о принципе работы поисковых систем, о схеме обработки запросов пользователей, о том, из чего эти системы состоят и как функционируют…

Количество данных в сети стремительно растет, уже сейчас всемирная есть насчитывает более полутора миллиарда сайтов (http://www.internetlivestats.com/total-number-of-websites/).

В сети каждый день появляются множество новых документов, и, конечно же, в большинстве случаев они оставались бы не востребованными, ни кем не найдены, и все это огромное количество информации оказалось бы никому не доступным и не нужным. Появилась необходимость создавать такие средства, которые позволили бы просто и понятно ориентироваться в информационных ресурсах всемирных сетей, мгновенно и качественно находить нужную информацию.

Наиболее популярным и используемым способом поиска в Интернете является использование поисковых систем. Что же такое поисковая система? Поисковая система – портал, осуществляющий поиск, сбор и сортировку информации в сети Интернет. Поисковые системы это инструмент, позволяющий пользователю глобальной сети в кратчайшие сроки найти интересующую его информацию.

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

Получая результат, пользователь оценивает работу системы, руководствуясь несколькими основными параметрами. Нашел ли он то, что искал? Если не нашел, то сколько раз ему пришлось перефразировать запрос, чтобы найти искомое? Насколько актуальную информацию он смог найти? Насколько быстро обрабатывала запрос поисковая машина? Насколько удобно были представлены

результаты поиска? Был ли искомый результат первым или же сотым? Как много ненужного мусора было найдено наравне с полезной информацией? Найдется ли нужная информация, при обращении к поисковой системе, скажем, через неделю, или через месяц?

Я выбрал тему курсовой «Анализ поисковых систем в сети Интернет», потому что эта тема очень актуальна в наше время и близка мне, т.к. я часто пользуюсь возможностями поисковых систем и их всевозможными сервисами.

Цель работы: ознакомиться с информацией о поисковых системах всемирной сети Интернет.

Задачи: анализ работы механизмов поиска информации, ознакомление с краткой историей поисковых систем, изучение их основных характеристик, типов и функций.

Объект исследования: Сеть Интернет.

Предмет исследования: поисковые системы Google, Яндекс, Yahoo!.

В качестве источников информации были использованы материалы сети Интернет, а также данные с сайтов самих компаний-поисковых систем.

Глава I. Теоретическая часть

1.1 История поисковых систем

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[1].

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 2017 год компания Google обрабатывала

167 миллиардов запросов в месяц.

1.2 Понятия и задачи поисковых систем

Поисковая система – это сайт, к которому пользователь обращается посредствам ключевого слова и находит интересующую его информацию**.** Сегодня поисковая система лучший способ, чтобы быстро и качественно найти интересующую вас информацию[2].

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

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.

1.3 Состав и принципы работы поисковой системы

Поисковая машина – это аппаратно-программный комплекс, который осуществляет быстрый поиск внутри сервера или Интернет-ресурса необходимой информации. У всех поисковых систем основа поисковой машины примерна одинаковая. В основном, это программное обеспечение, отвечающее за ранжирование результатов по релевантности поискового запроса и составление каталога запроса, поисковый бот, который необходим для поиска сайта и индексации. Но некоторые крупные поисковые системы держат содержание своей поисковой машины в секрете. Основным отличием является учет и релевантность морфологии языка запроса, база проиндексированных сайтов. Все это в совокупности и определяет критерий качества работы поисковых машин[3].

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[4].

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[5].

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[6].

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.

1.4 Поисковые системы в настоящее время

Во всем мире самые известные поисковые системы это: Google, Bing, Yahoo, Baidu(лидер среди китайских поисковых систем).

Русскоязычные — в основном все «русскоязычные» поисковые системы находят тексты и индексируют на нескольких языках — украинском, татарском, английском, белорусском и др. От «всеязычных» систем они отличаются тем, что

практически всегда индексируют те ресурсы, которые расположены в доменных зонах, где на первом месте стоит русский язык и тем, что они своих роботов ограничивают русскоязычными сайтами другими способами. А всеязычные индексируют все документы подряд.

По данным на май 2018 года доминирующие места в мировом рейтинге поисковых систем стабильно занимает компания Google (рис. 1).

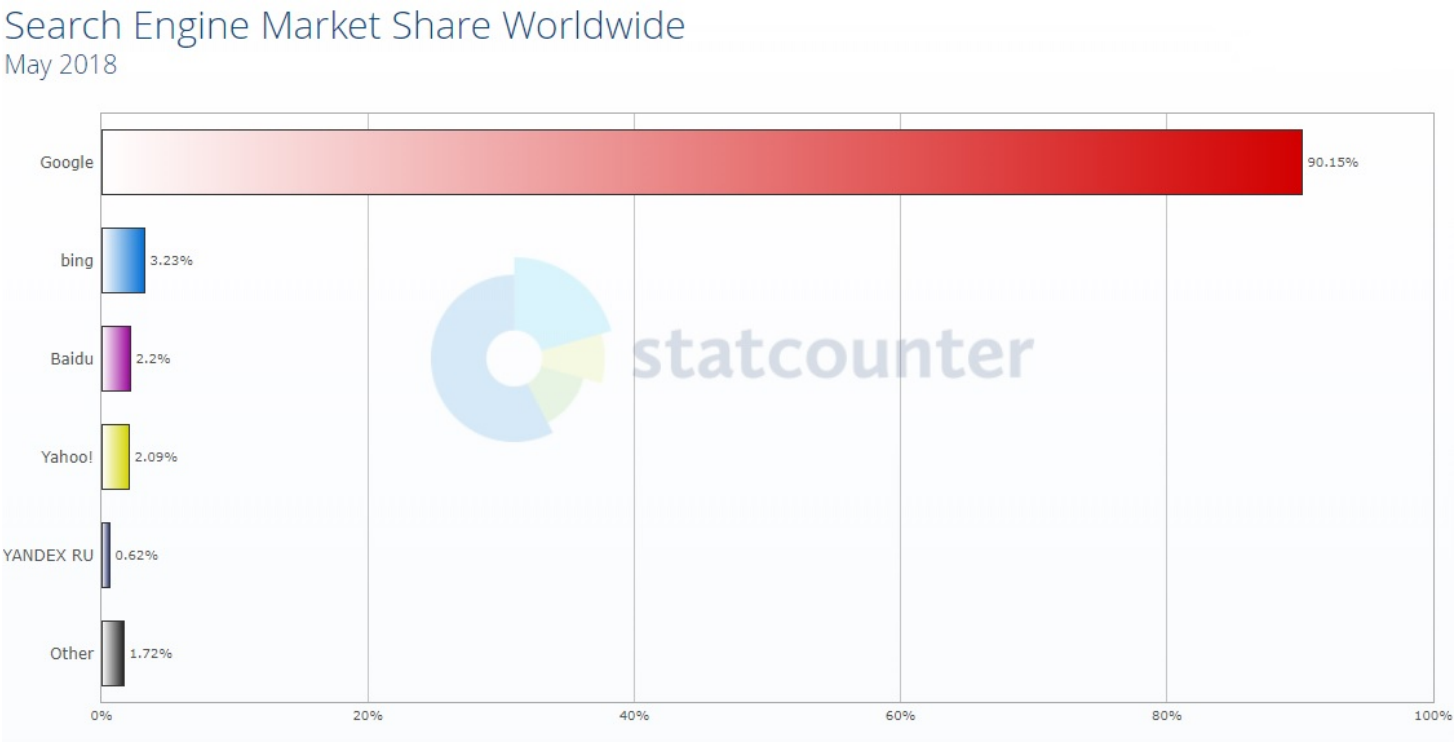

Рисунок 1. Рейтинг мировых поисковых систем (2018 год)

Несмотря на это, Google не собирается останавливаться, так в декабре 2018 года доля Google в России составила 53,5%, а Яндекса – 42,7% (рис. 2).

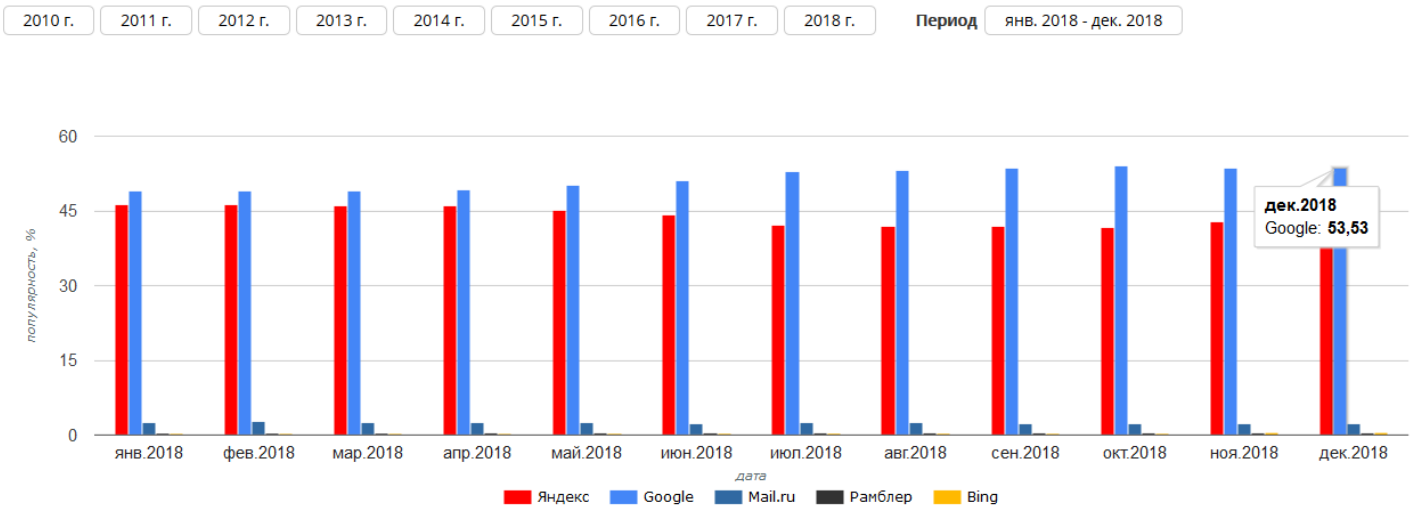

Рисунок 2. Объем трафика Yandex и Google (2018 год)

[7]

Глава II. Практическая часть

2.1 Принцип работы Google

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[8].

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. Так же необходимо заполнить мета тег description, где нужно кратко описать, какую информацию можно найти на странице. Пример тега: <meta name=”description” content=”Краткое описание страницы…”>.

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:

- Зарегистрировать сайт в Google Search Console https://search.google.com/search-console/welcome
- Указать в панели управления ссылку на файл sitemap.xml – файл, в котором указаны ссылки на страницы сайта в специальном формате
- Проверить код на валидность https://validator.w3.org/nu/
- Проверить работоспособность всех ссылок на сайте, при необходимости исправить ошибки
- Измерить скорость загрузки сайта и предпринять меры, если скорость низка https://developers.google.com/speed/pagespeed/insights/?hl=RU

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 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, отражающий какую долю веса может передать страница-донор на страницу-акцептор. Обычно его принимают равным 0.85, что означает, что страница может передать 85% веса (распределяется между 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 принципе, то же самое. Какое числовое значение у этого параметра знают только в Google, остальные из экспериментальных данных принимают его равным 0,85;

n – количество страниц, ссылающихся на страницу-акцептор (на которые не наложен фильтр);

Ti – i-ая ссылающаяся страница;

C(Ti) – количество ссылок на странице-доноре Ti .

Поскольку ссылающихся страниц может быть много, и общее количество страниц в поисковой системе Google достаточно велико, а также их количество постоянно растет, то представлять вес страницы в абсолютных значениях для вебмастеров было бы весьма неправильно. Для этого ввели понятие TLPR — ToolBar PageRank – значение PR, который имеет значение от нуля до 10 (шкала в Google Toolbar).

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

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

где base – основание логарифма, которое зависит от количества страниц в поисковой машине (возможно и от ряда других факторов). Некоторые принимают его равным 7;

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

Из вышесказанного неверно делать выводы, что нулевой TLPR означает нулевой реальный PageRank. По формуле PR видно, что даже при n=0 , мы получим минимальный PRmin =(1-d)=0,15. Это значение соответствует TLPR≈-1.

При таких (отрицательных) значениях тулбарного PR считается что PR=N/A (или еще не определен), однако он также оказывает влияние на распределение веса между ссылками-акцепторами. Также следует заметить, что тулбарное значение предназначено только для отображения вебмастерам в Google Toolbar и никак не влияет на позицию в выдаче. **На позицию в выдаче влияние оказывает реальный PR страницы.**

Исходя из принципов расчета **Google PageRank**, можно теперь легко рассчитать, с каких ссылок нужно ссылаться и сколько нужно ссылок, чтобы получить тот или иной PR.

Также можно прогнозировать PR. Один из важных выводов заключается в следующем: если у нового сайта более 10000 страниц (число страниц зависит от количества ссылок с них на другие страницы), они правильно перелинкованы и каждая ссылается на главную страницу, то главная страница получит хороший вес от этих ссылок. Учитывая, 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 хорошего PR без единой внешней ссылки с других сайтов.

Таким образом, существует множество способов повышения веса своих страниц, но главная идея — это качественные ссылки с других сайтов. Для этого можно использовать каталоги, социальные закладки, статьи, форумы, блоги и другие типы сайтов. Однако не следует глупо расставлять множество ссылок на других сайтах, так как 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 тулбарный PR равен нулю.

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

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

2.2 Принцип работы Яндекс

Основой работы поисковых систем как Google, так и Яндекс является система кластеров[9]. 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 в сервисе Яндекс.Вебмастер

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.Каталог и РСЯ (рекламной сети Яндекс).

Для поисковой системы Яндекс характерны следующие основные показатели внешней оптимизации:

- ИКС – индекс качества сайта, востребованность сайта аудиторией. Чем больше пользователей смогли удовлетворить свои запросы с помощью сайта, тем он лучше. При этом учитывается не только количество пользователей, но и степень их удовлетворенности и общий уровень доверия к сайту. На самом деле это всего лишь краткое описание ИКС. Алгоритм включает в себя сотни факторов ранжирования.
- вИЦ, или взвешенный Индекс Цитирования, представляет собой алгоритм для подсчета количества внешних ссылок; значение его не разглашается и используется поисковой системой при ранжировании сайтов в поисковой системе.
- Присутствие сайта в Яндекс.Каталог.
- 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 (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.

Стоит отметить, что в каталоге Апорт под ИЦ понимается взвешенный индекс цитируемости.

Рассчитывается этот индекс из ссылочного графа: если рассматривать ресурсы сети как вершины графа, а цитирование других ресурсов (ссылочные связи между сайтами) как связи вершин графа (ребра), тогда ссылочный граф можно представить в виде диаграммы, как показано на рисунке 3.1.

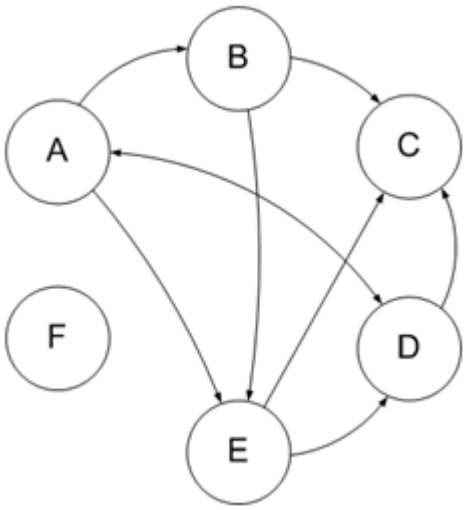

Рисунок 3. Ссылочный граф

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

ИЦ используется как один из факторов для ранжирования документов в поисковой выдаче, но не является главным.

Индекс цитируемости обычно рассматривается в качестве параметра значимости статьи, однако он не отражает структуру ссылок в каждой дисциплине (тематике), а также слабозначимые работы и труды с большой значимостью могут иметь одинаковый индекс цитируемости.

Поэтому был введен взвешенный индекс цитирования, который определяется не только количеством, но и качеством ссылающихся источников.

Введение ссылочного поиска и статической ссылочной популярности помогает поисковым системам справляться с примитивным текстовым спамом, который полностью разрушает традиционные статистические алгоритмы информационного поиска, полученные в свое время для контролируемых коллекций. ВИЦ является аналогом PageRank от Google.

Взвешенный индекс цитирования, как и другие ссылочные факторы ранжирования, рассчитывается из ссылочного графа.

Узнать вИЦ для своих страниц вы можете приблизительно, проверив их PageRank любым онлайн-сервисом проверки, однако, следует учесть, что в индексе Яндекса

присутствуют только русскоязычные документы, а из зарубежных лишь некоторые популярные, таким образом, урезая ссылочный граф по сравнению с Google.

2.3 Анализ Yahoo!

Yahoo![10] – это поисковая система, занимающая четвертую (на 2018 год) позицию по популярности во всем мире. Она представляет целый ряд сервисов, объединенных одним интернет порталом. Yahoo! Directory – портал, включающий в себя многофункциональную электронную почту. В 2004 году появилась новая версия почтового сервиса с обновленным интерфейсом. Данная разработка была создана на AJAX (отправка запросов на сервер без перезагрузки страницы). Электронная почта Yahoo! Mail одна из самых старых и наиболее популярных в сети Интернет. Следуя статистике Alexa Internet — весной 2012 года Yahoo! стал четвертым по посещаемости веб-сайтом, из них 28% посещений состоят из просмотра главной страницы.

Портал Yahoo продолжает предлагать своим пользователям неограниченные возможности благодаря постоянно совершенствующемуся алгоритму. Yahoo неоднократно и кардинально меняла принципы своей работы. Задача Yahoo – предоставление релевантных результатов своим пользователям в тех областях, где компьютерные алгоритмы «не оправдывают ожиданий» (речь идет о персонализированных результатах и мнениях).

Компания Yahoo ввела «социальный поиск», которому дали название My Web 2.0. Новый вид поисковой системы – социальная поисковая система, которая дополняет Интернет-поиск, позволяя пользователям получать ответы на интересующие вопросы не только в Интернет-ресурсах, но и непосредственно от знакомых и друзей. Технология, которой руководствуется «социальный поиск», называется My Rank.

My Rank обладает всеми преимуществами алгоритмического поиска и совмещает в себе многие достоинства, руководствуясь всего одной идеей: субъективное мнение по тем или иным вопросам. Технология My Rank позволяет получать ответы на интересующие вопросы не только от поисковиков, но и от определённых людей, оценивать эти мнения с целью нахождения оптимальных ответов, которые, по Вашему мнению, являются наиболее релевантными. Тем более, речь идет о людях, которые Вам знакомы, которые разделяют Ваши интересы, работают, возможно, в Вашей структуре и потенциально искали ответы на те же вопросы, что и Вы.

Совмещая возможности алгоритмического поиска с возможностью «войти в знакомое сообщество», технология My Rank способствует нахождению более релевантных ответов.

Все это становится реальностью благодаря предоставляемой возможности избирать, сохранять и делиться информацией с другими людьми, точно так же, как и получать информацию, с которой готовы поделиться другие люди. Социальный поиск привнес нечто новое в Интернет. Теперь поисковые результаты находятся в некоторой зависимости от мнения определенных людей.

На протяжении длительного периода времени Yahoo стремится стать уникальной концептуальной поисковой системой. Какая теория лежит за понятием «концептуальная модель поисковой системы»?
Компания Yahoo придерживается следующего мнения: все, что люди выражают сложной терминологией, можно заключить в простые понятия. Например, «Гавайи» и «Нью-Йорк» - абсолютно разные запросы, как по длине, так и по количеству слов, но в человеческом восприятии они совмещают в себе одно понятие. И, наоборот, человек воспринимает запрос «правоохранительные органы Нью-Йорка» как запрос, содержащий 2 разных понятия: «Нью-Йорк» и «правоохранительные органы».

Люди рассуждают о логической связи между понятиями. Например, понятия «правоохранительные органы» и «полиция» можно отнести к смежным областям. Пользователь, который вводит в поисковую строку одно из понятий, может заинтересоваться сайтами, которые относятся к смежному понятию, даже, если оно не содержит слов запроса[11].

До сих пор остается непонятным, какую технологию использует Yahoo, совершенствуя концептуальный поиск. Есть основания подозревать, что Yahoo предложит концептуальный поиск в виде отдельной поисковой системы, с использованием «социального поиска».
Тем не менее, стремление предложить пользователям точную информацию в соответствии с их индивидуальными потребностями, выглядит, по меньшей мере, утопично.

Yahoo уделяет первостепенное значение плотности ключевых слов. По некоторым оценкам, плотность ключевых слов в <title> составляет около 10% от требований алгоритма рассматриваемой поисковой системы. На первый взгляд, кажется, что алгоритм Yahoo представляет собой полную противоположность приоритетам,

которым уделяет внимание Google. Но на самом деле это не так.

Некоторые ассоциируют нынешний алгоритм Yahoo с алгоритмом Google двухлетней давности. С момента появления алгоритма Inktomi, поисковая система Yahoo стала уделять большее внимание обратным ссылкам, и все же это не является основополагающим компонентом работы алгоритма Yahoo, в отличие от алгоритма Google.

Оптимизируя под Yahoo, важно помнить, что алгоритм этой поисковой системы заинтересован в таких факторах, как контент, использование ключевых слов на странице, плотность ключевых слов на странице, жирный текст. Учитываются такие внешние факторы, как ссылочный текст, входящие ссылки и т.д. Yahoo предпочитает видеть ключевые слова в самих URL сайта или страниц, но отдает предпочтение жирному тексту, тексту, заключенному в <h1>.

2.4 Поведенческие факторы ранжирования. Рейтинг поисковых систем

В связи с увеличением количества сайтов в сети интернет поисковым системам приходится усложнять свои алгоритмы, чтобы пользователи могли найти именно ту информацию, которую они ищут и получить ее в максимально удобной форме. Так в недалеком прошлом были введены поведенческие факторы ранжирования.

Поведенческие факторы ранжирования – это совокупность всех действий пользователей с сайтом. Полный список этих факторов знают только создатели и разработчики поисковых систем. Вот список основных из них:

- Возврат в поисковую выдачу – фактор определяет, как часто пользователи, после посещения вашего сайта, возвращались к поиску.
- Показатель отказов – если за посетителем сайта не зафиксированы определенные действия и показатели, это считается отказом. Например: посетитель провел на сайте менее 15 секунд. Считать посещение отказом или нет определяет поисковая система по своим индивидуальным критериям.
- Время на сайте и глубина просмотра (количество страниц, просмотренных посетителем)
- Прямые заходы – посещение сайта напрямую, без помощи поисковой системы.
- Повторные посещения

Рейтинг популярности поисковых систем оценивается как доля поискового трафика, генерируемая данной поисковой

системой в Рунете.

Количество трафика поисковой системы (рис.4-5) оценивается по данным крупнейших в России сервисов интернет-статистики:

Яндекс.Метрика, SpyLog/Openstat, LiveInternet, Hotlog, Рейтинг@Mail.ru, а также на основании собственной статистики SEOAUDITOR

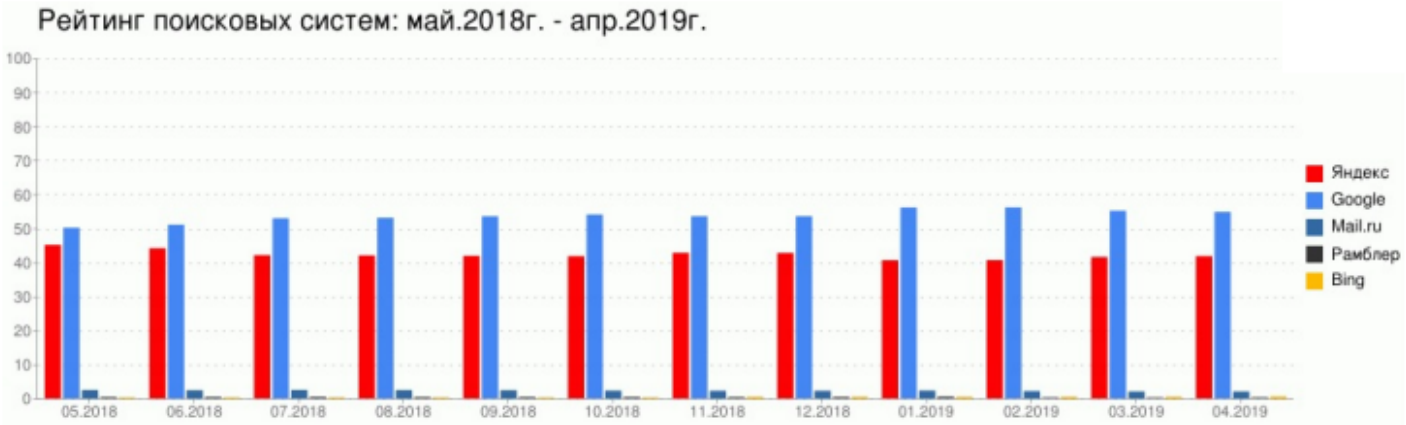

Рисунок 4. Рейтинг поисковых систем, 05.2018-04.2019

Популярность поисковых систем

	май. 2018	июн. 2018	июл. 2018	авг. 2018	сен. 2018	окт. 2018	ноя. 2018	дек. 2018	янв. 2019	фев. 2019	мар. 2019	апр. 2019
Яндекс	45.08%	44.11%	42.08%	42.00%	41.86%	41.77%	42.77%	42.71%	40.58%	40.61%	41.54%	41.78%
Google	50.14%	51.03%	52.93%	53.07%	53.51%	54.03%	53.52%	53.53%	56.09%	56.12%	55.13%	54.80%
Mail.ru	2.46%	2.40%	2.48%	2.44%	2.40%	2.29%	2.25%	2.25%	2.31%	2.16%	2.02%	2.02%
Рамблер	0.35%	0.38%	0.40%	0.35%	0.37%	0.37%	0.35%	0.38%	0.48%	0.24%	0.22%	0.26%
Bing	0.28%	0.29%	0.31%	0.30%	0.28%	0.28%	0.50%	0.51%	0.50%	0.50%	0.51%	0.57%
Yahoo!	0.19%	0.19%	0.20%	0.18%	0.17%	0.17%	0.17%	0.18%	0.17%	0.17%	0.17%	0.18%
Ask	0.02%	0.02%	0.02%	0.02%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
Nigma	0.02%	0.03%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
QIP	0.05%	0.05%	0.05%	0.04%	0.04%	0.04%	0.04%	0.04%	0.04%	0.03%	0.03%	0.04%

Рисунок 5. Популярность поисковых систем, 05.2018-04.2019

Можем сделать вывод, что, не смотря старания сервиса сделать поиск удобнее и точнее, поисковик Yahoo! в России не пользуется популярностью.

Заключение

В ходе выполнения работ мы с вами выяснили, что такое поисковые системы, из чего они состоят и как работают. Рассмотрели некоторые из известных факторов ранжирования веб сайтов.

Несмотря на то, что поисковые системы существуют уже более 20 лет, отмечается, что пользователи до сих пор вводят новые запросы, которые ранее были неизвестны поисковым системам, а так же количество информации в сети продолжает увеличиваться. Это значит, что поисковым системам есть куда стремиться и с течением времени поисковые машины будут совершенствоваться и внедрять новые прогрессивные технологии.

В работе нами были проанализированы три поисковика: Google, Яндекс, Yahoo!. Сделаны выводы, что поисковик Yahoo! постоянно старается стать совершеннее. Однако, в России он не пользуется популярностью. Самые популярные в России: Яндекс и Google. Именно на них надо ориентироваться при оптимизации сайта.

Также, отметим, что в связи с увеличением количества сайтов в сети интернет поисковым системам приходится усложнять свои алгоритмы, чтобы пользователи могли найти именно ту информацию, которую они ищут и получить ее в максимально удобной форме. Так в недалеком прошлом были введены поведенческие факторы ранжирования.

Список использованной литературы

1) Google – Поисковая система / [Электронный ресурс]. – Режим доступа: URL: https://www.google.ru/ (дата обращения 30.05.2019)

2) Yahoo – Поисковая система / [Электронный ресурс]. – Режим доступа: URL: https://www.yahoo.com/ (дата обращения 30.05.2019)

3) Yandex – Поисковая система / [Электронный ресурс]. – Режим доступа: URL: https://ya.ru/ (дата обращения 30.05.2019)

4) Поисковые системы / [Электронный ресурс]. – Режим доступа: URL: https://www.bsmu.by/downloads/kafedri/k_fiziki/2016-1/poisk_sistem.pdf (дата обращения 30.05.2019)

5) Wikipedia – Свободная энциклопедия / [Электронный ресурс]. – Режим доступа: URL: https://www.wikipedia.org/(дата обращения 30.05.2019)

6) Ашманов И.С. Оптимизация и продвижение в поисковых системах. 4-е изд. / Спб.: Питер, 2019

7) Байков В.Д. Интернет. Поиск информации. Продвижение сайтов/ Д.В Байков — СПб.: БХВ-Петербург, 2016.- 288 с.

8) Блог Яндекса для вебмастеров / [Электронный ресурс]. - Режим доступа: URL: https://webmaster.yandex.ru/blog/ (дата обращения 30.05.2019)

9) Вячеслав Тихонов. Поисковые системы в сети Интернет.-2000.- 24 февраля / [Электронный ресурс]. – Режим доступа: URL: http://citforum.ru/internet/search/searchsystems.shtml (дата обращения 30.05.2019)

10) Занзеров В.И.,Славинская Л.В.,Перинская Е.В. Поисковые системы.- 2012 г. / [Электронный ресурс]. – Режим доступа: URL: http://dpivi.ru/ HYPERLINK "http://dpivi.ru/156-poiskovye-sistemy.html"156 HYPERLINK "http://dpivi.ru/156-poiskovye-sistemy.html"-poiskovye-sistemy.html (дата обращения 30.05.2019)

11) Интернет-издание SEONEWS / [Электронный ресурс]. – режим доступа: URL: https://www.seonews.ru/ (дата обращения 30.05.2019)

12) Руководство по поисковой оптимизации для начинающих / [Электронный ресурс]. – Режим доступа: URL: https://support.google.com/webmasters/answer/7451184?hl=ru (дата обращения 30.05.2019)

13) Поисковые системы / [Электронный ресурс]. – Режим доступа: URL: http://inftis.narod.ru/is/is-n HYPERLINK "http://inftis.narod.ru/is/is-n8.htm"8 HYPERLINK "http://inftis.narod.ru/is/is-n8.htm".htm (дата обращения 30.05.2019)

14) Поисковые системы / [Электронный ресурс] . – Режим доступа: URL: http://uniofweb.ru/wiki/poiskovye_sistemy/ (дата обращения 30.05.2019)

15) Святослав Чернецкий. Поисковые системы в Интернете.-2012 г. / [Электронный ресурс] . – Режим доступа: URL: http://networkmy.ru/ HYPERLINK "http://networkmy.ru/636"636 (дата обращения 30.05.2019)

1. Вячеслав Тихонов. Поисковые системы в сети Интернет.-2000.- 24 февраля [Электронный ресурс]. URL: http://citforum.ru/internet/search/searchsystems.shtml (дата обращения 30.05.2019) ↑

2. Байков В.Д. Интернет. Поиск информации. Продвижение сайтов/ Д.В Байков — СПб.: БХВ-Петербург, 2016.- 288 с. ↑

3. Занзеров В.И.,Славинская Л.В.,Перинская Е.В. Поисковые системы.- 2012 г. [Электронный ресурс]. URL: http://dpivi.ru/ HYPERLINK "http://dpivi.ru/156-poiskovye-sistemy.html"156 HYPERLINK "http://dpivi.ru/156-poiskovye-sistemy.html"-poiskovye-sistemy.html (дата обращения 30.05.2019) ↑

4. Святослав Чернецкий. Поисковые системы в Интернете.-2012 г. [Электронный ресурс]. URL: http://networkmy.ru/ HYPERLINK "http://networkmy.ru/636"636 (дата обращения 30.05.2019) ↑

5. Поисковые системы [Электронный ресурс] URL: http://uniofweb.ru/wiki/poiskovye_sistemy/ (дата обращения 30.05.2019) ↑

6. https://www.bsmu.by/downloads/kafedri/k_fiziki/2016-1/poisk_sistem.pdf ↑

7. Интернет-издание SEONEWS / [Электронный ресурс]. – режим доступа: URL: https://www.seonews.ru/ (дата обращения 30.05.2019) ↑

8. google.com ↑

9. www.yandex.ru ↑

10. https://www.yahoo.com/ ↑

11. Поисковые системы [Электронный ресурс] URL: http://inftis.narod.ru/is/is-n HYPERLINK "http://inftis.narod.ru/is/is-n8.htm"8 HYPERLINK "http://inftis.narod.ru/is/is-n8.htm".htm (дата обращения 30.05.2019) ↑