

Содержание:

Введение

Всемирная сеть интернет в наше время является неотъемлемой частью в нашей жизни. Каждый может с легкость найти большое количество разной полезной информации, а также множество полезных функций и возможностей. Для меня решающими факторами в выборе темы «Анализ поисковых систем в сети Интернет», для моей курсовой работы, стали постоянно растущая актуальность на сегодняшний день, а также достаточная известность и открытость мне этой темы, так как я являюсь активным пользователем интернет сети.

Интернет давно перестал быть простым развлечением. На текущий момент интернет является незаменимым инструментом как для учеников и студентов, так и для рабочих людей разных профессий. А объемы информации в сети продолжает расти с огромной скоростью. По оценки разных экспертов объем данных, передаваемых по Интернету, увеличивается в два раза каждые полгода.

Каждый день в сети появляются миллионы новой информации. И что бы не потеряться в неиссякаемом потоке данных, были разработаны, поисковые системы, которые мы постоянно используем. На сегодняшний день сложно представить использование сети без таких ресурсов как Google, Яндекс, Mail.ru и других систем. Но всего лишь несколько лет назад бытовало мнение: что в Интернете безусловно много разной информации, но найти что-то слишком сложно или попросту невозможно. Однако с появлением и быстрым развитием поисковых каталогов, поисковых машин, и всевозможных поисковых программ ситуация изменилась, и теперь в Сети срочно понадобившуюся информацию иногда можно найти быстрее, чем в книге, лежащей на столе.

Так что же такое эта 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?

Глава 1. Теоретическая часть

1.1 Что такое поисковая система?

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

Рассмотрим, как работает поисковая система, что само по себе довольно просто. Пользователь, который зашел на сайт системы, должен ввести в поисковое окно, 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.

1.2 История развития поисковых систем.

Во времена, когда только начиналось развитие интернета, объём доступной информации был сравнительно мал, и пользователей сети было немного. На начальных стадиях развития сети, ее использовали сотрудники университетов и исследовательских лабораторий для обмена информацией между учреждениями. В то время поиск информации в сети интернет был не актуальным, в отличие от

нынешних дней.

Первым способом организации и систематизации доступа к информационным ресурсам стало создание каталогов сайтов. В них стали группировать ссылки согласно определенной тематике.

Первопроходцем области стал Yahoo, который появился в апреле 1994 года. С течением времени количество сайтов возросло, и Yahoo добавил опцию поиска по каталогам. Это не была поисковая система в прямом смысле, потому что область поиска ограничивалась ресурсами каталога.

Впоследствии каталоги распространились и стали использоваться повсеместно, но и интернет не стоял на месте, а продолжал развиваться. Вместе с ним развивались и методы поиска. На данный момент каталоги почти утратили популярность, это объясняется тем, что современный каталог, даже если он содержит огромное количество ресурсов, сможет обеспечить доступ лишь к малой части информации, которая содержится в сети.

В наши дни самым большим каталогом сети является Open Directory Project или DMOZ, включающий в себя информацию о 5 миллионах ресурсов, но это относительно мало, если сравнивать его, например, с поисковой системой Google, которая содержит около 8 миллиардов документов.

Полноценная же поисковая система вышла в свет только в 1994 году, ей стала система поиска WebCrawler.

Год спустя, в 1995 году появились проекты поисковых систем AltaVista и Lycos. Одна из них, в частности AltaVista, удерживала лидирующие позиции в области поиска в течение долгих лет.

По прошествии двух лет, в 1997 году, студенты Стэнфордского университета Сергей Брин и Ларри Пейдж разработали поисковую систему Google, являющуюся лидером в области поиска на сегодняшний день.

Этот год так же стал годом, когда было официально анонсировано о создании Российской поисковой системы Яндекс, которая и ныне лидирует в русскоязычном сегменте сети.

На данный момент существует только 3 поисковые системы, которые вышли на международный уровень, это MSN Search, Yahoo и Google. Эти системы имеют свои базы и поисковые алгоритмы. Большая часть остальных поисковиков используют их

результаты. Так Mail.ru использует базу Yandex, search.aol.com — Google, a Lycos, AltaVista и AllTheWeb — Yahoo.

Лидером поиска на русских просторах интернета на данный момент является Яндекс, за ним — Rambler, далее идут Google, Mail.ru, A port и KM.ru.

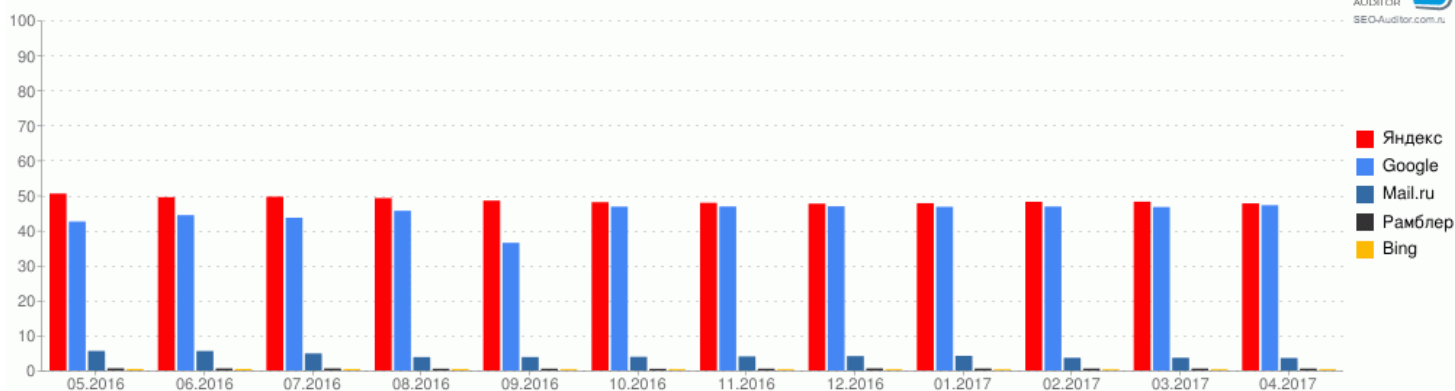
У поисковых систем разные алгоритмы работы, и чтобы занять хорошую позицию в поисковой выдаче и привлечь целевых посетителей, нужно знать особенности SEO оптимизации для разных поисковиков. Например оптимизация под Яндекс отличается от Google оптимизации.

1.3 Задачи поисковых систем.

Все поисковые системы объединены несколькими основными задачами, такими как поиск новых сайтов, оценка сайта и максимально точный ответ пользователю на запрос. Главная задача любой поисковой системы, предоставить пользователю ту информацию, которую он ищет. Но, к сожалению нельзя научить пользователя производить «правильные» запросы к системе, т.е. запросы, которые соответствуют принципу работы поисковых систем. Вот почему разработчикам нужно создавать такие принципы работы и алгоритмы поисковых систем, которые бы позволяли пользователям находить искомую ими информацию.

Это значит, что поисковая система должна думать точно также как думает пользователь, когда ищет ту или иную информацию. Обращаясь к поисковой системе, пользователь надеется максимально просто и быстро найти интересующую его информацию. После получения результата, он оценивает работу системы, руководствуясь несколькими основными параметрами. Разработчики поисковых систем постоянно стараются совершенствовать алгоритмы и принципы поиска, пытаются всячески ускорить работу системы, добавляя новые функции и возможности, чтобы удовлетворить потребности пользователей.

1.4 Рейтинг поисковых систем в России.



На сегодняшний день первое место в русском сегменте интернета, лидером среди поисковых систем является Яндекс и за ним с небольшим отставанием идет Google. Ниже представлена диаграмма статистики с сайта SEO-AUDITOR

Диаграмма 1 - Рейтинг Российских поисковых систем

Также для более точной информации ниже представлена информация в виде таблицы.

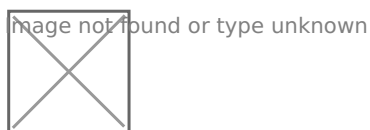


Таблица 1 - Рейтинг Российских поисковых систем

1.5 Мировой рейтинг поисковых систем.

По статистике в мире абсолютным лидером среди поисковых систем является Google. Его рейтинг составляет – 66,44% среди популярных поисковых систем.

Ниже представлена диаграмма с рейтингом поисковых систем в мире.

Диаграмма 2 - Мировой рейтинг поисковых систем за 2015г.

1.6 Обзор основных мировых поисковых систем

На сегодняшний день всемирная сеть Интернет насчитывает огромное множество поисковых систем во всех странах мира, из них всех можно выделить несколько самых крупных и пользующихся наибольшей популярностью среди пользователей:

1.6.1 Google

Лидер поисковых машин Интернета, Google занимает более 60 % мирового рынка, а значит, шесть из десяти находящихся в сети людей обращаются к его странице в поисках информации в Интернете. Сейчас регистрирует ежедневно около 50 миллионов поисковых запросов и индексирует более 8 миллиардов веб-страниц.

Была разработана в 1998 выпускниками Стэнфордского университета Сергеем Брином и Лари Пейджем, которые применили для ранжирования документов технологию PageRank, где одним из ключевых моментов является определение "авторитетности" конкретного документа на основе информации о документах, ссылающихся на него. Говоря общими словами, чем больше документов ссылается на данный документ и чем они авторитетнее, тем более авторитетным данный документ становится. Количественное значение авторитетности документа (другими словами, взвешенное количество ссылок или PageRank) относится к так называемым статическим факторам (то есть независящим от конкретного запроса) и учитывается при определении релевантности документа конкретному запросу как весовой коэффициент. Наряду с этим Google применил для определения релевантности документа не только текст самого документа, но и текст ссылок на него. Эта технология позволила ему обеспечить выдачу довольно релевантных результатов на фоне других поисковиков. Довольно быстро Google стал лидировать в различных опросах по такому показателю, как удовлетворенность пользователей результатами поиска.

Google осуществляет поиск по документам на более чем 35 языках, в том числе русском. В настоящее время многие порталы и специализированные сайты предоставляют услуги поиска информации в Интернете на базе Google, что делает задачу успешного позиционирования сайтов в Google еще более важной. Google проводит переиндексацию своей поисковой базы примерно раз в четыре недели. Во время этого усовершенствования, неофициально называемого Google dance, происходит обновление базы на основе информации, собранной роботами за время, прошедшее с предыдущего усовершенствования, и перерасчет значений PageRank документов. Также существует определенное количество документов с достаточно большим значением PageRank, информация о которых в поисковой базе обновляется ежедневно, однако значение PageRank пересчитывается только во время Google dance. Нормированное значение PageRank для конкретного документа, загруженного в браузер, можно узнать, скачав и установив Google

ToolBar - специальную панель инструментов для работы с этим поисковиком. Не смотря на то, что в поисковике имеется форма для бесплатного добавления страницы в базу, Google предпочитает сам находить новые документы по ссылкам с уже известных и не будет индексировать добавленную через форму страницу, если в его базе не найдется ни одной страницы, ссылающейся на нее.

1.6.2 Yahoo

Одна из самых первых Поисковых систем (создана Дэвидом Фило и Джерри Янгом в апреле 1994года) по сей день остается и самой популярной из них, традиционно сочетая поиск, как по ключевым словам, так и с помощью иерархического дерева разделов.

Нынешнее развитие Yahoo можно определить как движение в он-лайн, интерактивность. Yahoo быстро осваивает эту область Интернет-услуг, но возникает одна проблема: ядро Yahoo! не было на это рассчитано. Не была в 1994 году заложена в него "онлайновая" составляющая, ее "приклеил" Тим Кугл несколькими годами позже. Естественно возникает угроза хакерских атак через эту незащищенную область.

Одно из новшеств поисковой системы Yahoo - панель задач для браузера Firefox,. Этот инструмент помогает пользоваться поиском Yahoo, не заходя на официальный сайт, а лишь используя функциональные кнопки панели.

1 сентября 2005 года поисковик Yahoo, которому принадлежит более 200 миллионов адресов электронной почты по всему миру, анонсировал запуск новой системы поиска текстов, фотографий и других документов, содержащихся в письмах.

Необходимость такого нововведения возникла вслед за увеличением объёма хранимых данных, ведь некоторые пользователи создают целые почтовые архивы. Подгоняемый конкурентом Google и его почтовым сервисом Gmail, Yahoo для хранения почты предлагает отныне 1 гигабайт бесплатного места, или 2 гигабайта по годовому абонементу. "Как только вы получаете возможность хранить больше информации, вам необходимы и расширенные поисковые возможности", - объясняет Эрик Петерсон, аналитик компании Jupiter Research.

Пользователи поисковой системы Yahoo, в свою очередь, смогут теперь использовать возможности детализированного поиска слов в названии или непосредственно в тексте письма, а также в присоединенных документах, не открывая их. Результат поиска отражается в трёх строках с указанием всех атрибутов. На панели справа отображаются все похожие документы. Найденные фотографии выводятся на экран в уменьшенном виде, что значительно облегчает поиск. Система также учитывает орфографические ошибки, позволяя искать слова лишь по первым буквам.

Для начала Yahoo планирует предложить новую систему небольшому числу американских пользователей, а затем распространить её по всему миру. Со стороны клиентов это не потребует никаких дополнительных усилий. "Когда услуга станет доступна, в левом верхнем углу страницы вашего почтового ящика появится соответствующий баннер", - обещает компания Yahoo.

По данным comScore Media Metrix на июль этого года, домену Yahoo принадлежит 219 миллионов адресов электронной почты, что составляет 31,5% мирового рынка, уступая лишь Microsoft с 221 миллионом пользователей сервиса Hotmail (35,5% рынка).

1.7 Обзор основных Российских поисковых систем

Основное отличие русскоязычных поисковых систем от иностранных одно - это то, что глобальные поисковые системы, поддерживающие поиск на русском языке, не поддерживают русскую морфологию. В русскоязычной части сети Интернет работают около двух десятков поисковых систем, но подавляющее большинство пользователей работает лишь с несколькими, подробно остановимся на самых крупных:

1.7.1 Yandex

Яндекс - На сегодня наиболее популярная поисковая система, ежемесячно к ней обращаются более 35 миллионов пользователей Русскоязычной части Интернета. Начала свою работу во второй половине 1997 года учитывая морфологию русского языка. История компании "Яндекс" началась в 1990 году с разработки поискового программного обеспечения в компании "Аркадия". За два года работ были созданы две информационно-поисковые системы - Международная Классификация

Изобретений, 4 и 5 редакция, а также Классификатор Товаров и Услуг. Обе системы работали локально под DOS и позволяли проводить поиск, выбирая слова из заданного словаря, с использованием стандартных логических операторов. В 1993 году "Аркадия" стала подразделением компании ComrTek. В 1993-1994 годы программные технологии были существенно усовершенствованы благодаря сотрудничеству с лабораторией Ю. Д. Апресяна (Институт Проблем Передачи Информации РАН). В частности, словарь, обеспечивающий поиск с учетом морфологии русского языка, занимал всего 300Кб, то есть целиком грузился в оперативную память и работал очень быстро. С этого момента пользователь мог задавать в запросе любые формы слов.

Слово Яндекс придумал за несколько лет до этого один из основных и старейших разработчиков поискового механизма. "Yandex" означает "Языковой index", или, если по-английски, "Yandex" - "Yet Another indexer". За 4 года публичного существования Yandex возникли и другие толкования. Например, если в слове "Index" перевести с английского первую букву ("I" - "Я"), получится "Yandex".

В начале 1996 года был разработан алгоритм построения гипотез. Отныне морфологический разбор перестал быть привязан к словарю - если какого-либо слова в словаре нет, то находятся наиболее похожие на него словарные слова и по ним строится модель словоизменения. В это время Интернет в России только начинался. Еще через полгода стало очевидно, что ничто не отделяет ComrTek от создания собственной глобальной поисковой машины. Объем Рунета составлял тогда всего несколько гигабайт. Осенью 1997 года был открыт Yandex.Ru.

Помимо поисковой системы, сегодня Яндекс - огромный портал с целым набором широко используемых сервисов, такими как каталог, Яндекс. деньги, и другие. Официально поисковая машина Yandex.Ru была анонсирована 23 сентября 1997 года на выставке Softool. Основными отличительными чертами Yandex.Ru на тот момент были проверка уникальности документов (исключение копий в разных кодировках), а также ключевые свойства поискового ядра Яндекс, а именно: учет морфологии русского языка (в том числе и поиск по точной словоформе), поиск с учетом расстояния (в том числе в пределах абзаца, точное словосочетание), и тщательно разработанный алгоритм оценки релевантности (соответствия ответа запросу), учитывающий не только количество слов запроса, найденных в тексте, но и "контрастность" слова (его относительную частоту для данного документа), расстояние между словами, и положение слова в документе. Сегодня Яндекс имеет внутри мощный поисковый робот, позволяющий производить поиск по самым различным критериям.

1.7.2 Rambler

Rambler - Старейшая поисковая система российского Интернет, запущена в 1996 году, на сегодня - вторая по популярности с обращением более 25 миллионов посетителей в месяц. Помимо поисковой системы, сегодня Рамблер - один из крупнейших порталов Русскоязычной части Интернета с большим набором широко известных сервисов, таких как каталог Рамблер, Рамблер-почта, Рамблер-ICQ или Рамблер-ТВ. По сути сегодня Рамблер - больше, чем просто поисковая система и набор сервисов, это крупная медиагруппа. Поисковая машина "Рамблер" начала работу в октябре 1996 года, на стартовом этапе содержала всего 100 тысяч документов. "Рамблер" не был первой отечественной поисковой системой, однако в первый год своего существования (когда весь русский веб с приемлемой степенью правдоподобия индексировался "Рамблером", "Апортом", "Русской поисковой машиной", а также шведской и калифорнийской AltaVista) вынес основной груз поисковых запросов. Вторая версия "Рамблера" начала разрабатываться летом 2000 года, в марте нынешнего года приняла достаточно законченные очертания. В нее были введены функции, давно уже имевшиеся в конкурирующих системах. Она учитывает координаты слов, обучена строгой и нечеткой морфологии, связывает поиск с каталогом, в качестве которого используется Top100 (<http://top100.rambler.ru/>), группирует результаты поиска по сайтам, ищет по числам. Достаточно удачная архитектура продукта позволяет "Рамблер" иметь для поисковика количество серверов в 2 раза меньшее, чем у "Яндекса", и в 3 раза меньшее, чем у "Апорта".

1.7.3 Mail.ru

Национальная почтовая служба Mail.ru - это не только поисковая система но и один из крупнейших порталов российского Интернета. Ежедневная аудитория Mail.ru - более 5 миллионов пользователей. Общее число регистраций со дня основания около 60 миллионов. Mail.ru - самый быстроразвивающийся российский Интернет-ресурс. Через почтовые ящики Mail.ru ежедневно проходит более 25 миллионов писем. Mail.ru занимает лидирующую позицию среди бесплатных почтовых сервисов, предоставляя своим пользователям почтовый ящик неограниченного размера с защитой от спама и вирусов, переводчиком, проверкой правописания, архивом для хранения фотографий и многое другое.

В 1998-м году программисты, работающие в питерском офисе американской софтверной компании DataArt, создали новое ПО для почтового веб-сервера, которое в дальнейшем предполагалось продавать западным компаниям. Чтобы протестировать сервис, его временно выложили в открытый доступ для российских пользователей, а сервис вдруг стал стремительно набирать популярность.

20 февраля 2001 года произошло слияние двух крупных игроков российского Интернет-рынка, компаний Port.ru и netBridge под брендом Port.ru. В результате объединения родилась компания, которая сразу заняла лидирующие позиции среди российских Интернет – холдингов по доле рынка и охвату аудитории.

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

- Основные характеристики поисковых систем:

1. Полнота
2. Точность
3. Актуальность
4. Скорость поиска
5. Наглядность

В состав поисковой системы входят компоненты:

1. Модуль индексирования
2. База данных
3. Поисковый сервер

Подводя итог можно сказать что, как правило, несмотря на обилие поисковых систем, пользователь предпочитает обращаться к услугам лишь одной – двух из них (точно также как при обилии газет или новостных сайтов мы регулярно просматриваем лишь некоторые, привычные и любимые). Самой популярной поисковой системой в мире является Google. Но по оценкам аналитиков, на просторах бывшего СССР чаще используется Яндекс.

Глава 2. 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.

С другой стороны, поисковая система Google выглядит стабильнее своих конкурентов в плане смены алгоритма и апдейтов. Информация, только что размещенная на сайте, может в считанные минуты попасть в основную выдачу. Поисковые роботы Google в три раза быстрее, чем роботы других поисковых систем. Фильтры (критерии «нормальности» сайта) почти не меняются с момента начала их внедрения.

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

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

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

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

Перед продвижением сайта в Google следует:

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

- Проверить работоспособность всех ссылок на сайте, при необходимости исправить ошибки.

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

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

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

Следует отметить, что при расчете PR Google учитывает не все ссылки, а отфильтровывает ссылки с сайтов, специально предназначенных для скопления ссылок. Некоторые ссылки могут не только не учитываться, но и отрицательно сказаться на ранжировании ссылающегося сайта (такой эффект называется **поисковой пессимизацией**). [9]

Основной формулой для расчета PR является формула:

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

где $PR(T_i)$ – 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);

T_i – i -ая ссылающаяся страница;

$C(T_i)$ – количество ссылок на странице-доноре T_i .

Поскольку ссылающихся страниц может быть много, и общее количество страниц в поисковой системе Google достаточно велико (около десятка миллиардов штук), 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 $PR_{min} = (1-d) = 0,15$. Это значение соответствует $TLPR \approx -1$.

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.** [10]

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 Яндекса

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

- База поисковой информации обновляется несколько раз в течение месяца, при этом на поисковые запросы выдается обновленная информация с сайтов. Такая информация добавляется с помощью основного робота-сканера.
- При обновлении программного кода или «движка» выявляются недостатки и изменяются алгоритмы, отвечающие за ранжирование ресурсов в поисковой системе. Как правило, перед выходом таких обновлений Яндекс публикует соответствующие анонсы.

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

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

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

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

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

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

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

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

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

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

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

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

- Yandex/1.01.001 (compatible; Win16; I) – основной робот, занимающийся непосредственно индексацией сайтов;
- Yandex/1.01.001 (compatible; Win16; P) – робот-индексатор изображений;
- Yandex/1.01.001 (compatible; Win16; H) – робот, который выявляет «зеркала» индексируемых сайтов;
- Yandex/1.02.000 (compatible; Win16; F) – робот-индексатор пиктограмм ресурсов (favicons);

- Yandex/1.03.003 (compatible; Win16; D) – робот, который обращается к страницам, добавленным с помощью формы «Добавить URL»;
- Yandex/1.03.000 (compatible; Win16; M) – задействуется при переходе на страницу посредством ссылки «Найденные слова»;
- YaDirectBot/1.0 (compatible; Win16; I) – этот робот отвечает за индексацию страниц ресурсов, принимающих участие в рекламной сети Яндекс.

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

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

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

Для поисковой системы Яндекс характерны следующие основные показатели внешней оптимизации:

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

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

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

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

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.

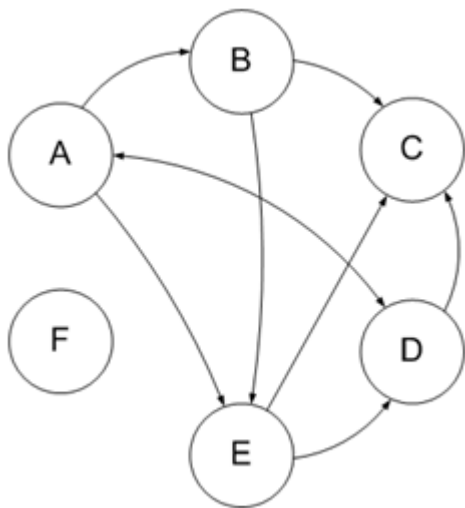


Рисунок №1 Ссылочный граф

На рисунке буквами А, В, ..., 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.

Не стоит путать обычный индекс цитирования с взвешенным и тематическим, о которых будет написано позже. Индекс цитируемости всегда целое число и не зависит от тематик ссылающихся документов. [7]

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.

Тематический индекс цитирования введен для отражения авторитетности сайта в своей тематике.

При определении тематики сайта сначала строится описание рассматриваемого ресурса (из названия категорий сайта, заголовков, структуры URL его страниц).

Далее вычисляется оценка близости между описаниями заранее подготовленных тематик (каталог) и описаниями ресурсов с выбором наиболее близких тематик для них.

Тематическая близость двух документов отражает вероятность принадлежности их обоим одной и той же тематике. Этот показатель может влиять на значение передаваемого ссылкой веса.

Расчет тИЦ основан на формуле:

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

где $PF(v, t)$ – ТИЦ ресурса v ;

P – количество ресурсов, которые ссылаются на сайт v и имеют ту же тематику;

n_v – количество страниц на рассматриваемом сайте v ;

N – общее число страниц в индексе Яндекса (при этом, n_v/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 ситуацию в Рунете, но со временем индекс Яндекса расширился на такие географические сегменты, как Беларусь, Украина и другие. В Яндексе появились новые версии каталога для дополнительных регионов. [8]

Соответственно, чтобы ранжировать сайты в каждом из региональных Яндекс.Каталогов, потребовалось ввести региональный ТИЦ, который учитывает, помимо тематической, географическую близость ссылок.

Таким образом, ТИЦ обладает следующими свойствами:

1. ТИЦ зависит от количества уникальных страниц на сайте и чем их больше, тем больше результирующий показатель.
2. Чем меньше исходящих ссылок на сайте-доноре, тем больше с него передается ТИЦ.
3. ТИЦ никак не зависит от перелинковки.
4. Анкоры ссылок не участвуют в определении тематической близости двух ресурсов.
 1. При наличии у сайта нескольких зеркал (копий), при их склейке результирующий ТИЦ суммируется.

2.3 Кратко о поисковой машине

Удивительно, но эта невероятно популярная система, обслуживающая миллионы запросов ежедневно, зародилась как простая коллекция закладок, которую пополняли всего 2 человека - Дэвид Фило и Джерри Янг. На сегодняшний день Yahoo!, это уже не просто каталог, это целая группа разнообразных сервисов, среди которых такие как каталог Yahoo!igans - 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 другое.

Среди всех рассмотренных систем, Yahoo! - единственная чисто каталоговая, на Yahoo! нет собственной поисковой машины. Зато список категорий на Yahoo! является наиболее полным и простым - в отличие от других каталогов, на Yahoo! всегда легко определить, в каком разделе находится нужная информация. Заглавная страничка Yahoo! грузится очень быстро - хотя на ней очень много ссылок, но все они текстовые.

Центральная часть страницы, конечно, занята окном поиска и списком категорий. Ссылки вверху страницы (графические) обеспечивают доступ к такой информации, как "что нового", "что хорошего", "More Yahoos". Последнюю ссылку рекомендуется посетить - она приводит на страницу с огромным количеством ссылок на разнообразные Yahoo! - каталоги и сервисы. В нижней части основной страницы Yahoo! расположено большое количество ссылок на наиболее популярные разделы Yahoo!.

При вводе ключевых слов с основной страницы Yahoo, запрос обрабатывается по методу "Intelligent default", то есть Yahoo! ищет наиболее подходящие результаты в таких областях: в категориях Yahoo; в Web-сайтах, зарегистрированных на Yahoo; на Altavista (запрос передается при отсутствии результатов); в новостях. Такой интеллектуальный поиск занимает довольно много времени.

При задании критериев поиска для Yahoo! нужно помнить, что Yahoo! ищет эти слова только в названии и описании страницы, поскольку полнотекстового индекса на Yahoo! нет. Поэтому не следует указывать при поиске слишком много терминов или синонимов - количество результатов с Yahoo! снизится или даже будет нулевым. При вводе ключевых слов со страницы каталога, нужно выбрать область поиска - весь каталог Yahoo! или только его текущий раздел. Это делается с помощью радио кнопок под полем ввода. поисковый информационный интернет

На странице с результатами поиска выводятся сначала удовлетворяющие критерию поиска категории, а потом сайты. Возле каждой категории в скобках стоит число - это количество сайтов в данной категории.

В случае если на Yahoo! нет результатов, сразу выводятся результаты с Altavista. Вверху и внизу страницы выводится маленькая табличка, с помощью которой можно одним нажатием кнопки мыши произвести поиск в категориях Yahoo!, на Altavista, в новостях и событиях. Количество результатов поиска на Yahoo!, естественно, невелико, зато большинство из них являются релевантными.

Возможна проблема с отсутствующими страницами, поскольку вебмастера обычно забывают удалить свои сайты с поисковых систем, а на Yahoo! нет механизма автоматического обновления. Для расширенного поиска Yahoo! предлагает не очень большой, но очень полезный набор инструментов. Чтобы попасть на страничку расширенного поиска, надо перейти по ссылке "options" с основной страницы Yahoo!.

Среди средств расширенного поиска - ограничение результатов по дате, поиск в Yahoo!, Usenet и среди E-mail адресов, использование логических операций над терминами и поиск конкретной фразы. Также присутствует возможность искать слова с произвольными окончаниями, указывать слова, которые должны или НЕ должны присутствовать в документе, и т.д. Чисто русские ресурсы в Yahoo! не добавляются, потому что в Yahoo! Inc. просто некому смотреть и оценивать их содержимое. Но те запросы, которые не дали результатов на Yahoo! передаются на Altavista, а там есть хороший индекс русских ресурсов.

2.4 Как осуществлять поиск

Как пишут сами разработчики Yahoo!, их страница с результатами поиска предназначена для того, чтобы помочь пользователям находить то, что они ищут, в дружелюбном и удобном для работы интерфейсе.

Рассмотрим более подробно различные разделы на странице с результатами поиска.

Inside Yahoo! (Внутренний Yahoo!) Это продукты или услуги Yahoo!, что соответствует пользовательскому критерию поиска. К примеру, если человек задал в запросе "лягушка" ("frogs"), Inside Yahoo! покажет результаты поиска областями,

где пользователь сможет найти различные типы информации, такие как изображения из Картинной галереи Yahoo!, элементы для продажи в Yahoo! Аукцион, факты о лягушках от Yahoo!igans!

Directory Category Matches (Категории директивных сделок): Эта область подсвечивает категории в Yahoo! Каталог, которые соответствуют пользовательскому запросу поиска. Если человек хочет увидеть совокупность сайтов по специфической теме, ему следует щелкнуть по самой необходимой категории, после чего пользователю представится наглядный список сайтов, который был собран редактором Yahoo! по заданной теме.

Если категорий больше, чем может отображаться, то справа вверху появится ссылка "Next". Щелчок по данной ссылке позволит пользователю видеть и коммерческие и некоммерческие категории в Yahoo! Каталог, которые соответствуют запросу поиска.

Sponsor Matches (Спонсорские сделки): Спонсорские сделки – релевантные результаты поиска, за которые платят предпринимателями или организациями и обеспечиваются 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 заданный по умолчанию стиль, в котором появляются результаты.

Когда сайт, перечисленный в результатах поиска, также перечислен в Yahoo! Каталог, листинг результата поиска покажет заголовок и описание, обеспеченному Yahoo! Каталог. Кроме того, пользователь будет видеть ссылку " More sites about", которая находится внизу. Кликая на эту ссылку, пользователь сможет просмотреть совокупность сайтов по той же самой теме в Yahoo! Каталог.

В списки каталога включают сайты, прошедшие через специальную программу Yahoo!. Эти сайты заплатили Yahoo! рассматривать и считать их для включения в Yahoo! Каталог.

2.5 Расширенный поиск

Расширенный поиск – это особенность, которая помогает вам совершенствовать ваши результаты поиска.

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

Расширенный поиск помогает увеличить точность результатов поиска, используя дополнительный синтаксис, чтобы сосредоточить поиск. Пользователь может ввести большинство следующих параметров поиска непосредственно в блок поиска, или же выбрать их на странице Расширенного поиска, на которую можно перейти по ссылке [advanced search](#), находящейся 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

Рассмотрим данную страницу более подробно.

Include all of the words (Включите все слова) – Эта опция позволяет найти результаты поиска, которые включают все слова, которые пользователь напечатал 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, которые точно соответствуют словам, которые пользователи ввели. Это подобно помещению цитат (" ") вокруг набора слов. (Пример: Вы ищете известное высказывание или цитату: "Я хочу домой").

Include at least one of these words (Включите по крайней мере одно из этих слов) – Эта опция для поиска результатов по нескольким показателям, которые соответствуют или одному или большему количеству слов, которых задаются для поиска. Это соответствует вставке "OR" между словами. (Например, если пользователь хочет найти информацию или относительно каноэ или относительно лодок.)

Exclude these words (Исключите эти слова) – Эта опция исключает заданные слова из поиска. В обычном поиске это соответствует вставке "NOT" между словами или символом " " перед словом. (Например, вы ищете информацию о цветах, но не хотите, чтобы выдавалась информация о розах. Для этого введите "цветы" во "All of the words", а в "Exclude these words" введите "розы").

Search (Поиск) – Здесь пользователю требуется выбрать, где он хочет искать информацию: в Сети или только в Yahoo-каталоге.

More options (Больше Вариантов) – Пользуясь дополнительными опциями, которые появляются при нажатии этой кнопки. Дадим им краткое описание:

Language (Язык) – Позволяет выбрать, на каком языке будут отображаться сайты на странице с результатами.

Country (Страна) – Данная функция позволяет показывать результаты в зависимости от выбранной страны.

Date (Дата) – Ограничивает результаты поиска теми сайтами, которые были модифицированы в пределах прошедших 3, 6, или 12 месяцев.

Keyword Location (Местоположение ключевых слов) – Позволяет пользователю самому выбрать условия поиска – на странице, где-нибудь, в заголовке, в тексте, в URL или в ссылках на другие страницы.

Domain (Домен, область поиска) – Запрашивает, на каких доменах должен (или не должен) происходить поиск (например, с com, org, gov, net, biz, info, name).

Search by URL (Поиск URL) – Пользователь может попробовать найти web-страницы, являющиеся подобными или принадлежащими к специфическому узлу.

Заключение

По итогам сделанной мной работы я могу заключить что; 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.

По моему мнению, самой лучшей иностранной поисковой системой является Google, так как для меня основное значение имеет точность и полнота предоставляемых данных. Но можно заключить также что, каждая поисковая система будь то Российская или зарубежная предоставляет различные возможности поиска, из различных баз данных, поэтому сказать точно какой именно лучше пользоваться было бы не правильно. Поэтому для удобства поиска и полноты информации следует пользоваться несколькими поисковиками вводя в них нужные запросы. По моему мнению, из многих Российских поисковиков выделяются Яндекс и Рамблер, для них характерно постоянное обновление баз данных что, обеспечивает именно

актуальность и точность предоставляемой информации.

Список литературы

1. Экслер А.Б. Самоучитель работы в Интернете – Москва.: NT Press, 2007г.
2. Кузьмин А.В. Золотарева Н.Н. Поиск в Интернете – Санкт – Петербург.: Издательство НиТ, 2006г.
3. Гусев В.С. Яндекс. Эффективный поиск – Москва, Санкт – Петербург, Киев.: Диалектика, 2007г.
4. Егоров А.Б. Поиск в Интернете – Санкт – Петербург.: НиТ, 2007г.
5. Гусев В.С. Поиск, Internet –Москва, Санкт – Петербург, Киев.: Диалектика, 2004г.
6. Гусев В.С. Google. Эффективный поиск – Москва, Санкт – Петербург, Киев.: Диалектика, 2007г.
7. www.citforum.ru – CIT forum, Поисковые системы в сети Интернет
8. www.ru.wikipedia.org – Википедия – свободная энциклопедия
9. www.clx.ru – Описание зарубежных поисковых систем
10. www.seop.ru – Search engine optimization project, рейтинг основных поисковых систем
11. www.gs.seo-auditor.com.ru – Рейтинг поисковых систем