Содержание:

Введение

Всемирная сеть очень важна и полезна почти для всех. Любой пользователь Интернета может отыскать в нем много разной и интересной информации, а также использовать все широкие возможности сети. Для меня главными обстоятельствами в выборе темы «Анализ поисковых систем в сети Интернет», для моей курсовой работы, стали актуальность темы на сегодняшний день, а также достаточная открытость и известность этой темы.

Ресурсы Интернета уже давно не просто игрушка, превратившаяся в незаменимый инструмент для каждодневной работы людей различных профессий. Количество данных в сети стремительно растет, и пропорционально им растет и объем. Ученые утверждают, что объем информации, передаваемой по Интернету, увеличивается в два раза каждые шесть месяцев.

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.

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

Наиболее популярным и используемым способом поиска в Интернете является использование поисковых систем.

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 компьютерная система, предназначенная для поиска информации. Одно из наиболее известных применений поисковых систем — веб-сервисы для поиска текстовой или графической информации во Всемирной паутине. Существуют также системы, способные искать файлы на FTP-серверах, товары в интернет-магазинах, информацию в группах новостей Usenet.

Рассмотрим принцип работы поисковика, который довольно прост. Пользователю, пришедшему на сайт системы необходимо ввести в форму, располагающуюся на сайте ключевую фразу, по которой он ищет информацию, и послать запрос, нажав кнопку поиск. После чего он получит результат в виде списка текстовых ссылок на сайты соответствующие данному запросу. Это принцип работы поисковика со стороны пользователя. Ниже рассмотрим процесс работы (который не заметен пользователю) и внутреннее устройство.

1.2 Немного из истории

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

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

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

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

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

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

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

1.3 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 пользователь ту информацию, которую он ищет. Но, к сожалению нельзя научить пользователя производить «правильные» запросы к системе, т.е. запросы, которые соответствуют принципу работы поисковых систем. Вот почему разработчикам нужно создавать такие принципы работы и алгоритмы поисковых систем, которые бы позволяли пользователям находить искомую ими информацию.

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

1.4 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.

Поисковые машины классифицируются по области поиска информации:

1. Локальный поиск. Предназначен для осуществления поиска информации по какой-либо части всемирной сети, например, по одному или нескольким сайтам, либо по локальной сети. Примером служит поисковый скрипт на сайте или внутренние серверы крупных компаний. Локальные поисковые системы характеризуются наличием меньшего (по сравнению с глобальными системами) объема индексируемой текстовой информации, большей семантической ее однородностью, меньшим числом используемых национальных языков.

Поиск текстовых ресурсов в локальных поисковых системах характерен при разыскивании пользователем информации по описанию метадокумента (по автору, дате создания, названию, типу документа и т.п.). Ввиду возможности создания локальных поисковых систем с определенной направленностью (новостные, тематические и т.п.) локальные поисковые системы могут использовать специфические (не универсальные) критерии назначения релевантности (например, соответственно по дате создания документа, по его тематике и т.п.). При поиске текстовых ресурсов могут быть также использованы вероятностно-статистические методы, широко применяемые в глобальных поисковых системах.

Однако, учитывая прикладную направленность локальных поисковых систем на определенный круг пользователей, а также требования обеспечения полнотекстового поиска, наиболее актуальными являются алгоритмы точного поиска смысловой информации.

Успешный поиск точного ответа в локальных системах не может быть реализован без достаточного глубокого лексико-грамматического анализа текстовой базы и запросов пользователей, а также широкого привлечения эвристических методов

оценки их смыслового соответствия.

2. Глобальный поиск. Предназначены для поиска информации по всей сети Интернет либо по значительной её части. Владельцами таких поисковых машин являются поисковые системы Google, Яндекс и др. Поисковые машины осуществляют поиск информации различного типа, например текстов, видео, изображений, географических объектов, персональных данных и др. При этом файлы, с которыми может работать поисковая машина, могут быть как текстового формата (например .html, .htm, .txt, .doc, .rtf…), так и графического (.gif, .png, .svg…) или мультимедийного (видео и звук). Пока наиболее распространённым является именно поиск по текстовым документам. Глобальные поисковые системы характеризуются наличием большого объема разнородной текстовой информации, изложенной на различных национальных языках (более 30 основных национальных языков). Поэтому широко применяемые алгоритмы поиска для глобальных поисковых систем основаны на методах поиска по сигнатурам ключевых слов. Использование лингвистических методов в глобальных поисковых системах сводится в лучшем случае только к использованию морфологии для наиболее распространенных языков (часто только для английского языка). Использование морфологии позволяет расширить полноту поиска за счет отбора текстовых ресурсов, содержащие все возможные словоформы ключевых слов запроса пользователя. Для обеспечения более качественного отбора текстовых ресурсов, соответствующих запросу пользователя, в глобальных поисковых системах используются алгоритмы априорного назначения релевантности ресурсу (индекс цитирования, частота встречаемости ключевого слова на данном ресурсе и т.п.). Объектом поиска является ресурс текстовой информации – как правило, страница текста, имеющая уникальный URL . Полнотекстовый поиск глобальные поисковые системы обеспечивают только в пределах сравнительно небольшой части проиндексированного ресурса. Это обусловлено тем, что индексируется только ограниченное количество слов ресурса, отсчитываемое от начала документа.

В глобальных базах (вследствие их всеобъемлющего характера) с очень большой вероятностью может быть найден какой-нибудь подходящий информационный ресурс для практически большинства запросов пользователей даже без привлечения для этих целей лингвистического аппарата.

Число предложений в проиндексированных документах существенно превышает количество проиндексированных ресурсов, поэтому в глобальных поисковых системах используется именно поиск ресурсов. Полнотекстовый поиск смысловой

информации в предварительно отобранных ресурсах реализуется пользователем.

Поисковые машины по сети интернет осуществляют различный поиск информации. Например, музыка, картинки, личная информация, географическое положение и т.д. Поисковая машина может работать с файлами различных форматов (например .html,.htm,.txt,.doc,.rtf, …), мультимедийного (видео, звука и другой информации) или графического (.gif, .png, .svg,) типа. Но самым распространенным поиском является поиск текстовых документов (документы в формате doc, rtf, txt, web-страницы и др.). Но с технологической точки зрения поиск по звукам, видео, изображениям является более сложным, поэтому он не реализован массово. Например, такие системы как Яндекс.Картинки ищут картинки по альтернативным текстам, соответствующим этим изображениям, а не по самим изображениям. А в компании Google каталог поиска картинок составляется вручную, это тормозит обновление баз изображений, но значительно увеличивает релевантность запроса.

Модуль индексирования: Модуль индексирования состоит из трех вспомогательных программ (роботов):

Spider (паук) – программа, которая предназначена для скачивания веб-страниц. «Spider» полностью обеспечивает скачивание страницы, и все внутренние ссылки извлекает с этой страницы. С каждой страницы скачивается html-код. Роботы используют протоколы HTTP для скачивания страниц. «Spider» работает следующим образом. Робот передает на сервер запрос «get/path/document» и несколько других команд HTTP-запроса. В ответ роботу приходит текстовый поток, который содержит сам документ и служебную информацию.

Ссылки извлекаются из тэгов frame, base, area, frameset, и др. Многие роботы, наряду со ссылками, обрабатывают редиректы (перенаправления). Все страницы сохраняются в таких форматах как:

- дата, когда страница была скачана
- тело страницы (html-код)
- URL страницы
- http-заголовок ответа сервера

Crawler («путешествующий» паук) – эта программа, автоматически проходит по всем ссылкам, которые нашла на странице. Выделяет все ссылки, присутствующие на странице. Его задача – состоит в том, чтобы исходя из заранее заданного списка адресов или основываясь на ссылках, определить, куда дальше должен идти паук. Crawler, осуществляет поиск новых документов, еще неизвестных поисковой

системе, следуя по найденным ссылкам.

Indexer (робот - индексатор) - это программа, анализирующая веб-страницы, которые скачали пауки. Индексатор, применяя собственные лексические и морфологические алгоритмы, разбирает страницу на составные части и анализирует их. Разные элементы страницы подвергаются анализу, например, заголовки, текст, специальные служебные html-теги, ссылки структурные и стилевые особенности, и т.д.

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

База данных: Индекс поисковой системы или база данных - это информационный массив, в котором хранятся преобразованные параметры всех документов скачанных и обработанных модулем индексирования.

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

Работает поисковый сервер следующим образом:

- Запрос, который получен от пользователя подвергается морфологическому анализу. Генерируется информационное окружение каждого документа, содержащегося в базе (как раз оно и будет отображено в виде сниппета, т. е. текстовой информации соответствует запросу на странице выдачи результатов поиска).
- Все полученные данные передаются специальному модулю ранжирования в качестве входных параметров. После чего по всем документам происходит обработка данных, далее подсчитывается собственный рейтинг для каждого документа, который характеризует релевантность разных составляющих данного документа, хранящихся в индексе поисковой системы запроса, введенного пользователем.
- Этот рейтинг может быть составлен в зависимости от выбора пользователя дополнительными условиями (например, «расширенный поиск»).
- Далее генерируется сниппет, т. е., из таблицы документов извлекаются краткая аннотация, наиболее соответствующая запросу, заголовок и ссылка на сам документ для каждого найденного документа, и еще подсвечиваются все

найденные слова.
- Пользователю результаты поиска, которые мы получили, передаются в виде SERP (Search Engine Result Page) – страницы выдачи поисковых результатов.

Все эти компоненты работают во взаимодействии и тесно связаны друг с другом, именно они образовывают тот самый довольно сложный механизм работы поисковой системы, который требует огромных затрат ресурсов.

1.5 Поисковые системы в настоящее время

Во всем мире самые известные поисковые системы это: Google, Bing, Yahoo, Duckduckgo, Aol, Яндекс, Ask.com, Baidu, Mywebsearch.com, Quora, Ixquick, Waybackmachine.

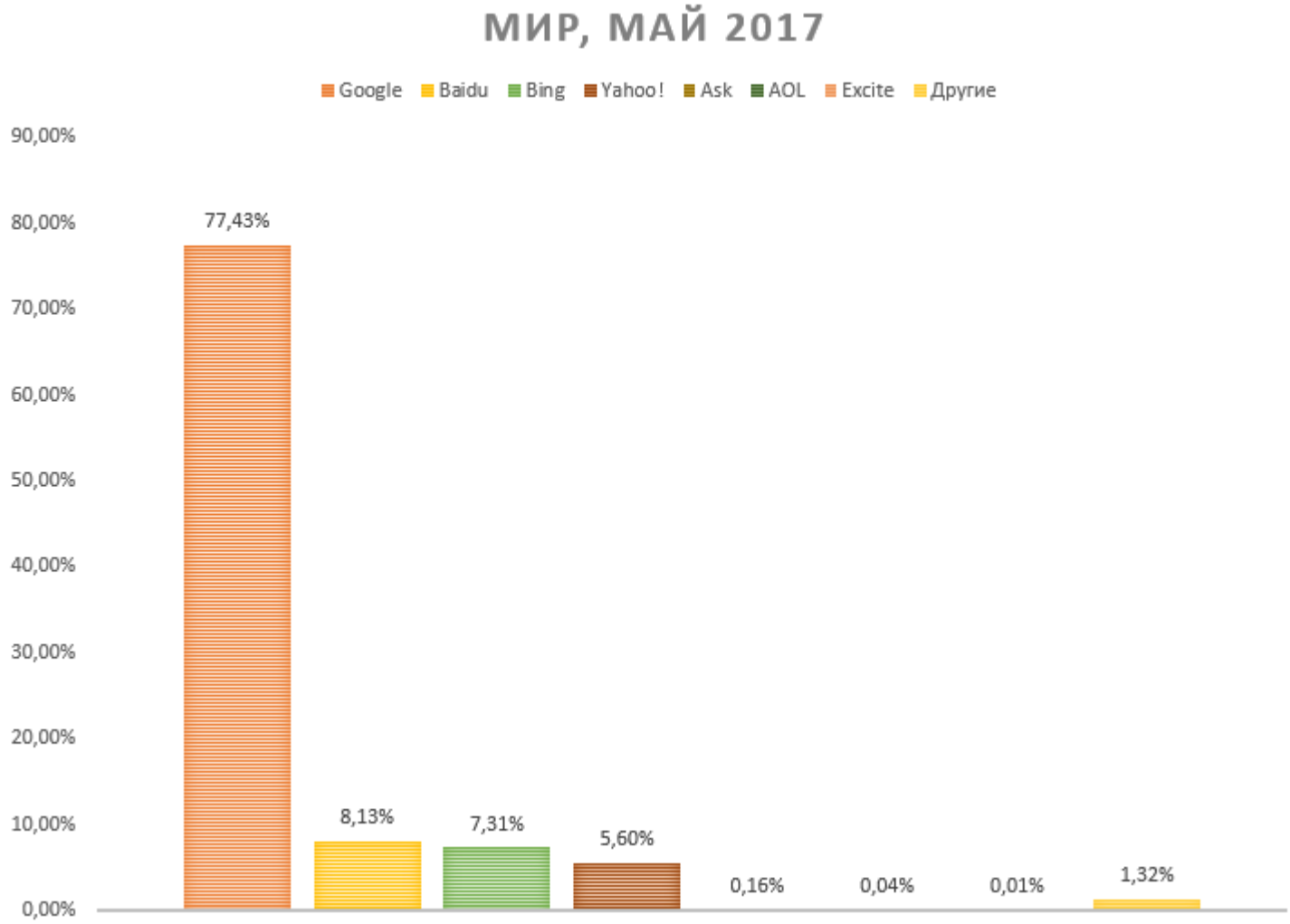
Русскоязычные — в основном все «русскоязычные» поисковые системы находят тексты и индексируют на нескольких языках — украинском, татарском, английском, белорусском и др. От «всеязычных» систем они отличаются тем, что практически всегда индексируют те ресурсы, которые расположены в доменных зонах, где на первом месте стоит русский язык и тем, что они своих роботов ограничивают русскоязычными сайтами другими способами. А всеязычные индексируют все документы подряд.

В России основной поисковой системой является «Яндекс», далее - Rambler, Aport, Mail.ru, Нигма.

По данным исследования который проводился в конце 2009 года доминирующие место в рейтинге стабильно занимает компания Google. В декабре на долю гиганта пришлось 41,3 миллиардов поисковых запросов, это – 62,4% рынка.

Второе место (с большим отрывом) у Yahoo! – 8,5 миллиардов запросов, 12,8% рынка и крупнейшего китайского поисковика Baidu.com – 3,4 млрд. запросов, 5,2% рынка. К слову, уверенные позиции последнего связаны с тем, что на территории Китая заблокированы и Google, и Yahoo!.

Диаграмма 1. Рейтинг мировых поисковых систем



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:

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

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 (около десятка биллионов 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 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.** [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:

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

А 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» 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.

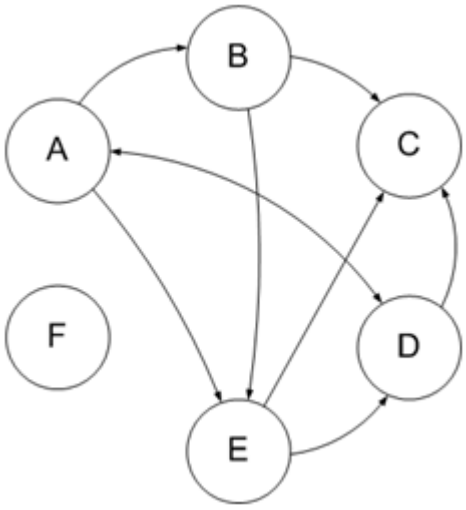


Рисунок 1. Ссылочный граф

На рисунке буквами 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.

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 обоих 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;

nv– количество страниц на рассматриваемом сайте v;

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

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

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

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

Изначально тематический индекс цитирования отражал 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 Yahooligans - Yahoo! для детей, система персональных каналов My Yahoo!, бесплатный E-mail сервис, система "Shop with Yahoo!" (покупайте с Yahoo!), совместный с MTV проект MTV unfURLed и многое другое.

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 дружественном 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 Yahooligans!

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 обеспечивается 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 будете 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 напечатали в блоке поиска. Это подобно вставке "AND" между словами или символом "+" перед словом.

Include this exact phrase (Включите эту точную фразу) – Эта опция позволяет исследовать результаты, 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

В наше время информация играет огромную роль во всех сферах жизнедеятельности. Людям, имеющим дело с большими объемами текстов - это и новости, и подшивки газет в электронном виде, и электронная почта, и Web-страницы, важно быстро находить в этом море информации действительно нужную. Без помощи поисковой системы это было бы нереально. Благодаря удобству в обращении и хорошим техническим характеристикам, различные поисковые системы могут помочь в этом и новичку, и опытному пользователю.

Поисковые системы и существующие к ним приложения, способны облегчить работу представителей многих профессий: Web-мастера, аналитика, руководителя, лингвиста. Информационный бум продолжается, происходит дальнейшее развитие электронно-компьютерных технологий, а следовательно и в будущем без поисковых систем обойтись будет крайне сложно.

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

Основные характеристики, которыми должны обладать поисковые системы:

Полнота – представляет собой отношение количества найденных по запросу документов к общему числу документов в сети Интернет, удовлетворяющих данному запросу.

Точность – определяется степенью соответствия найденных документов запросу пользователя.

Актуальность – характеризуется временем, проходящим с момента публикации документов в сети Интернет, до занесения их в индексную базу поисковой системы.

Скорость поиска – скорость поиска тесно связана с его устойчивостью к нагрузкам. Посетитель желает получить результаты как можно быстрее, а поисковая машина должна отрабатывать запрос максимально оперативно, чтобы не тормозить вычисление следующих запросов.

Наглядность – наглядность представления результатов является важным компонентом удобного поиска. По большинству запросов поисковая машина находит сотни, а то и тысячи документов. Вследствие нечеткости составления запросов или неточности поиска, даже первые страницы выдачи не всегда содержат только нужную информацию. Это означает, что пользователю зачастую приходится производить свой собственный поиск внутри найденного списка.

Как показывает статистика, пользователи русскоязычной части Интернета предпочитают несколько поисковых машин. Прежде всего, это мультиязычная платформа Google, являющаяся своеобразным эталонов универсального поискового механизма.

Лидер среди русскоязычных поисковых систем — Яндекс — индексирует документы форматов pdf, rtf, doc, txt, swf, rss и так далее. С помощью Яндекса можно искать информацию на русском, английском, украинском, белорусском, румынском, немецком и французском языках.
В настоящее время, практически каждая поисковая система имеет свои механизмы расчета рейтинга Интернет-страниц, и алгоритмы эти постоянно изменяются, совершенствуются. Однако в целом можно сказать, что наибольшее внимание современные поисковики уделяют внешним критериям оценки релевантности.

Список использованной литературы

1. http://uaweb.ua/publication/top_10_search_engine_2016.html

2.https://ru.wikipedia.org

3. https://www.seonews.ru/glossary/poiskovaya-mashina/

4. http://tuvanorchestra.ru/1-kurs/9-sostav-i-printcipy-raboty-poiskovykh-sistem/

5. http://www.asknet.ru/Technology/searchtask.htm

1. 6. www.citforum.ru

2. 7. Экслер А.Б. Самоучитель работы в Интернете - Москва.: NT Press, 2007г.
3. 8. Гусев В.С. Google. Эффективный поиск - Москва, Санкт - Петербург, Киев.: Диалектика, 2007г.
4. 9. Гусев В.С. Яндекс. Эффективный поиск - Москва, Санкт - Петербург, Киев.: Диалектика,2007г.

10. http://www.seonews.ru/news/.info_news/2385/

11. http://www.seo-gu.ru/im_stat.html

12. http://www.relevantno.ru/news/html/1138782965.html

13. http://www.vadimstepanov.ru/f_texts/column6.htm

14. http://book.itep.ru/4/45/retr4514.htm