Содержание:

ВВЕДЕНИЕ

Современному этапу развития человечества в XXI веке свойственен переходный период от индустриального общества к информационному. Одним из самых важных явлений данного перехода является появление и развитие глобальной информационной компьютерной среды. В связи с расширением информационных потоков появился вопрос о быстром и точном поиске нужной информации в компьютерной сети интернет.

В наши дни объем информации бесконечно растёт, а поэтому нет определенного предела совершенствованию информационно-поисковых систем.

Основной задачей разработчиков данных поисковых сервисов является улучшение качества поиска, а также эффективности использования вычислительных способностей и удобства в использовании системы. Для достижения данной цели постоянно необходимо менять и дорабатывать поисковые алгоритмы, а также создавать дополнительные сервисы и дорабатывать дизайн для привлечения и удобства «клиентов».

Однако для того, чтобы быть «онлайн» в данном направлении, при разработке необходимо закладывать современные технологии и большой запас отказоустойчивости, постоянно «заглядывать» в завтрашний день и примерять будущую нагрузку на сегодняшний поиск. Такой подход позволяет совершать важные и нужные шаги для повышения эффективности поиска в глобальной сети Интернет.

Целью работы является анализ поисковых систем в сети Интернет.

Для решения поставленной цели были поставлены следующие задачи:

- изучить теоретические основы поисковых систем в сети Интернет;

- провести анализ современных поисковых систем в Интернет;

- рассмотреть организацию поиска в поисковой системе Яндекс.

Объектом работы являются поисковые системы в сети Интернет.

Предметом работы является организация работы современных поисковых систем в сети Интернет.

При подготовке работы были использованы такие информационные источники как специализированная профессиональная литература, материалы из СМИ, данные интернет-ресурсов. Применены такие методы и приемы исследования как анализ, синтез, сравнение.

Глава 1. Теоретические основы поисковых систем в сети Интернет

1.1. Понятие поисковых систем в сети Интернет

Поисковая система – это программно-аппаратный комплекс, который предназначен для осуществления поиска в сети Интернет. Он помогает пользователям быстро найти необходимые сведения, реагируя на запрос пользователя выдачей списка ссылок на источники информации. Достаточно набрать в строке поиска интересующий вопрос или фразу, нажать на кнопку «Поиск» или «Search», и через несколько секунд поисковая система выдаст необходимую информацию[1].

Поисковые системы классифицируют по способу работы и по области использования. Каждая поисковая система имеет собственный алгоритм поиска, который определенным образом анализирует релевантность сайтов, чтобы выдать результат, наиболее соответствующий запросу пользователя.

Рассмотрим типы поисковых систем по способу работы.

Индексные поисковые системы собирают информацию в Интернете автоматически, с помощью специальных программ-роботов, посещающих веб-страницы. Они осуществляют всесторонний поиск по ключевым словам. Примерами таких поисковых систем являются Google, AltaVista, HotBot, Яндекс.

Индексная поисковая система состоит из трех основных компонентов:

1. Агент (паук или кроулер). Агент – это специальная программа, которая запускается на сервере поисковой системы с целью посещения веб-страниц. Когда агент находит новую страницу, удовлетворяющую алгоритму поисковой системы,

он индексирует ее, то есть добавляет в базу данных поисковой системы. Посещать страницы агенту помогает система гиперссылок, благодаря которой программа может бесконечно переходить с одной страницы на другую.

2. База данных поисковой системы. В ней хранятся все найденные и обработанные документы (индексы). Индекс позволяет быстро совершать поиск и обычно состоит из списка ключевых слов и информации о них (позиции в тексте, веса и др.). База регулярно обновляется, и именно из последнего ее обновления выдаются результаты для поставленного запроса. Частота обновления базы данных – критически важный параметр любой поисковой системы. Чем чаще происходит ее обновление, тем качественнее поисковая система.

3. Поисковый механизм. Поисковый механизм – интерфейс для взаимодействия пользователя и базы данных, то есть та самая программа, с которой мы непосредственно имеем дело[2].

Индексные поисковые системы работают по одному общему принципу. Сначала агент начинает сканирование сети с определенного адреса. На сервере создаются индексированные копии документов, своего рода вспомогательные файлы. Затем сохраненные документы просматриваются, определяются гиперссылки с этих страниц, по ним осуществляется переход на новые страницы. После сохранения копий найденных документов весь процесс повторяется. Все веб-страницы, проиндексированные поисковой системой, попадают в базу данных, что позволяет пользователю, формирующему запрос на поиск необходимой информации, мгновенно получить ссылки на нее.

Каталоговые системы поиска содержат тематически структурированный каталог серверов и чаще всего пополняются вручную модераторами. Эти системы устроены так же, как тематический каталог обычной библиотеки. Ссылки в них хранятся по теме категорий. Начав с основной страницы каталога, нужно выбрать ссылку, обозначающую главную категорию, а затем на последующих страницах указывать подкатегории до тех пор, пока не будут достигнуты ссылки на конкретные страницы. Каталог обычно имеет тематическую разбивку на подкаталоги, те в свою очередь могут подразделяться на более мелкие поддиректории и т. д. Ярким примером каталога является система Yahoo.

Индексные поисковые системы и поисковые каталоги отличаются так же, как содержание и алфавитный указатель в книге. Задача и содержания, и алфавитного указателя – помочь найти в книге нужный раздел. Содержание – это пример

каталогизации. Алфавитный указатель – пример индексации. Читатель находит в указателе нужный термин и получает номер страницы, на которой он встречается.

Метапоисковые системы – это системы, которые используют для поиска базы данных других поисковых систем. Они посылают запрос одновременно на несколько поисковых систем, каталогов и иногда в так называемую невидимую (скрытую) паутину – хранилище онлайн-информации, не считанной традиционными поисковыми системами. Собрав результаты, метапоисковая система удаляет дублированные ссылки и в соответствии со своим алгоритмом объединяет результаты в общем списке. Примером такой системы может служить российское решение Nigma, использующее для поиска Google, Yahoo, Апорт и Яндекс.

Специализированные поисковые системы, в отличие от поисковых систем общего назначения, которые ищут любую интересующую информацию, ищут информацию определенного вида, например, изображения, книги, организации, людей, то есть работают в какой-то конкретной области. Примерами таких систем могут служить moresoft.ru для поиска программ и файлов, bukinist.agava.ru для поиска книг и других электронных текстов, kinopoisk.ru для поиска информации о фильмах, Яндекс.Маркет для поиска описаний и цен товаров, представленных в Интернет-магазинах, drivers.ru для поиска драйверов, wink.com для поиска людей.

По области поиска поисковые системы можно разделить
на глобальные и локальные.

Глобальные предназначены для поиска информации по всей сети Интернет либо по значительной ее части, а локальные – по какой-либо части Сети, например, по одному или нескольким сайтам, либо по локальной сети. Часто локальные поисковые системы собирают информацию в пределах одного национального домена, как, например, yandex.ru.

Также существуют локальные поисковые машины, которые можно установить себе на компьютер, например Copernic Desktop Search для Microsoft Windows, Spotlight для Mac OS X, Tracker для Linux. Они значительно облегчают жизнь тех пользователей, которые хранят огромные архивы не рассортированных файлов.

Алгоритм создания эффективного запроса выглядит следующим образом:

- Формулировка задачи поиска. Для получения необходимой информации, в первую очередь, нужно понять, на какой именно вопрос пользователь ищет ответ.

- Ограничение области поиска. Выдача результатов может различаться в зависимости от региона, поэтому нужно добавить в запрос тот город, регион или страну, результаты по которым интересуют пользователя.

- Подбор ключевых слов, то есть слов и фраз, относящихся к теме поиска. Ключевые слова делят на высоко-, средне- и низкочастотные, это зависит от частоты запроса и определяется на основе статистики поисковой системы.

- Формулировка запроса. Важные слова необходимо поместить в начало запроса, для более эффективного поиска необходимо использовать язык запросов.

Использование поисковых систем может стать причиной проникновения на компьютер пользователя вредоносной программы. Выдавая результаты по запросам пользователей, поисковые системы могут выдавать адреса зараженных сайтов.

Также нужно иметь в виду, что поисковые системы выдают лишь ссылки на релевантные сайты, но не отвечают за достоверность информации, которая на этих сайтах содержится. Задача поисковых систем – максимально быстро и точно ответить на запрос, поэтому не стоит безоговорочно доверять всей информации, которая находится по выдаваемым ссылкам. Сайты, полученные при поиске, могут содержать некорректную или откровенно ложную информацию, которая может ввести в заблуждение пользователя, ведь далеко не все источники, скорее меньшая их часть, пишутся и проверяются действительно компетентными людьми. Например, информацию на таком популярном ресурсе как Википедия размещают все желающие, следствием чего является высокий процент ошибок в статьях. Рекомендуется крайне осторожно выбирать источники для школьных, студенческих и научных работ, да и вообще перепроверять любую информацию, особенно из совершенно незнакомой области.

Обозначим основные характеристики поиска:

- Полнота. Данная характеристика является одной из важнейших характеристик поиска, она представляет собой отношение величины найденных по запросу документов к общему их количеству в сети Интернет, относящихся к определенному запросу. К примеру, в Интернете есть 100 страниц, которые имеют словосочетание «как выбрать авто», а по такому же запросу было отобрано лишь 60 из общего числа, то в данном случае полнота поиска составляет 0,6. Понятно, что чем полнее сам поиск, тем больше вероятность, что пользователь найдёт именно тот документ, который ему нужен, конечно, если он вообще существует.

- Точность. Ещё одна важная функция поисковой системы – точность. Она определяет степень соответствия найденных страниц в Интернете запросу пользователя. Например, если по ключевой фразе «как выбрать автомобиль» найдется 100 документов, в половине из них содержится данное словосочетание, а в остальных просто есть в наличии такие слова (как грамотно выбрать автомагнитолу, и установить её в автомобиль»), то поисковая точность равна 50/100 = 0,5.

Чем поиск точнее, тем быстрее пользователь найдёт нужную ему информацию, тем меньше различных ненужных документов будет встречаться среди результатов, тем меньше найденных документов будут не соответствовать смыслу запроса.

- Актуальность. Это важная составляющая поиска, которую характеризует время, проходящее с момента публикации информации в сети Интернет до занесения её в индексную базу поисковой системы.

Например, на следующий день после появления информации о выходе нового iPad, много пользователей обратилась к поиску с соответствующими типами запросов. В большинстве случаев информация о данной новости уже доступна в поиске, хотя времени с момента её появления прошло еще мало. Это происходит благодаря наличию у крупных поисковых систем «быстрой базы», которая обновляется несколько раз в день.

Скорость поиска. Данная функция тесным образом связана с так называемой «устойчивостью к нагрузкам». Каждую секунду к поиску обращается большое число людей, подобная загруженность требует значительного уменьшения времени для обработки одного запроса. Тут интересы, как пользователя, так и поисковой системы полностью совпадают: пользователь хочет получить результат как можно скорее, а поисковая система должна отработать его запрос также максимально быстро, чтобы не тормозить обработку следующих запросов.

Наглядность. Наглядное представление результатов является важным элементом удобства поиска. По множеству запросов поисковая система находит тысячи, а в некоторых случаях и миллионы различных документов. Вследствие нечёткости составления ключевых фраз для поиска или его неточности, даже самые первые результаты запроса не всегда имеют только необходимые сведения.

Это значит, что человеку часто приходится осуществлять собственный поиск среди предоставленных результатов. Различные компоненты страниц выдачи поисковых систем помогают ориентироваться в результатах поиска[3].

1.2. Механизм поиска в поисковых системах

Поисковые системы можно сравнивать со справочной службой, агенты которой обходят организации, собирая информацию в базы данных. При обращении в службу выдается информация из данной базы. Данные в базе устаревают, поэтому агенты их периодически обновляют. Некоторые организации сами присылают данные о себе, и к ним агентам приезжать не нужно. Другими словами, справочная служба имеет две функции: создание и постоянное обновление данных в базе и поиск информации в базе по запросу клиентов.

Аналогично, поисковая машина состоит из двух частей: так называемого робота (или паука), который обходит серверы Интернета и формирует базу данных поискового механизма.

База робота, главным образом, формируется им самим (робот сам находит ссылки на новые ресурсы) и в гораздо меньшей степени - владельцами ресурсов, которые регистрируют свои сайты в поисковой машине. Помимо робота (червяка, паука, сетевого агента), формирующего базу данных, существует программа, определяющая рейтинг найденных ссылок.

Принцип работы поисковых машин сводится к тому, что они опрашивают свой внутренний каталог (базу данных) по ключевым словам, которые пользователь указывает в поле запроса, и выдают список ссылок, ранжированный по релевантности.

Следует заметить, что, отрабатывая конкретный запрос пользователя, поисковая система оперирует именно внутренними ресурсами (а не пускается в путешествие по Интернету, как часто думаю неопытные пользователи), а внутренние ресурсы, естественно, ограниченны. Несмотря на то, что база данных поисковой машины постоянно обновляется, поисковая машина не может проиндексировать все веб-документы: их количество крайне велико. Поэтому всегда существует вероятность, что необходимый ресурс просто неизвестен определенной поисковой системе.

Данную мысль наглядно иллюстрирует рисунок 1. Эллипс 1 ограничивает множество всех веб-документов, существующих на некоторый момент времени, эллипс 2 - все документы, проиндексированные определенной поисковой машиной, а эллипс 3 - искомые документы. Таким образом, найти с помощью определенной поисковой машины можно лишь ту часть искомых документов, которые ей

проиндексированы.

Рисунок 1. Схема, поясняющая возможности поиска

Проблема недостаточности полноты поиска состоит не только в ограниченности внутренних ресурсов поисковика, но и в том, что скорость робота ограничена, а количество новых веб-документов постоянно растёт. Рост внутренних ресурсов поисковой машины не может полностью решить проблему, поскольку скорость обхода ресурсов роботом конечна.

При этом считать, что поисковая машина содержит копию исходных ресурсов Интернета, было бы неправильно. Полная информация (исходные документы) хранится далеко не всегда, чаще хранится лишь её часть - так называемый индексированный список, или индекс, который гораздо компактнее текста документов и позволяет быстрее отвечать на поисковые запросы.

Для построения индекса исходные данные преобразуются так, чтобы объём базы был минимальным, а поиск осуществлялся очень быстро и давал максимум необходимой информации. Объясняя, что такое индексированный список, можно провести параллель с его бумажным аналогом - так называемым конкордансом, т. е. словарем, в котором в алфавитном порядке перечислены слова, употребляемые конкретным писателем, а также указаны ссылки на них и частота их употребления в его произведениях.

Очевидно, что словарь (конкорданс) гораздо компактнее исходных текстов произведений и найти в нём необходимое слово гораздо проще, чем перелистывать

книгу в надежде наткнуться на необходимое слово.

Схема построения индекса показана на рисунке 2. Сетевые агенты, или роботы-пауки, "ползают" по Интернету, анализируют содержимое веб-страниц и собирают информацию о том, что и на какой странице было обнаружено.

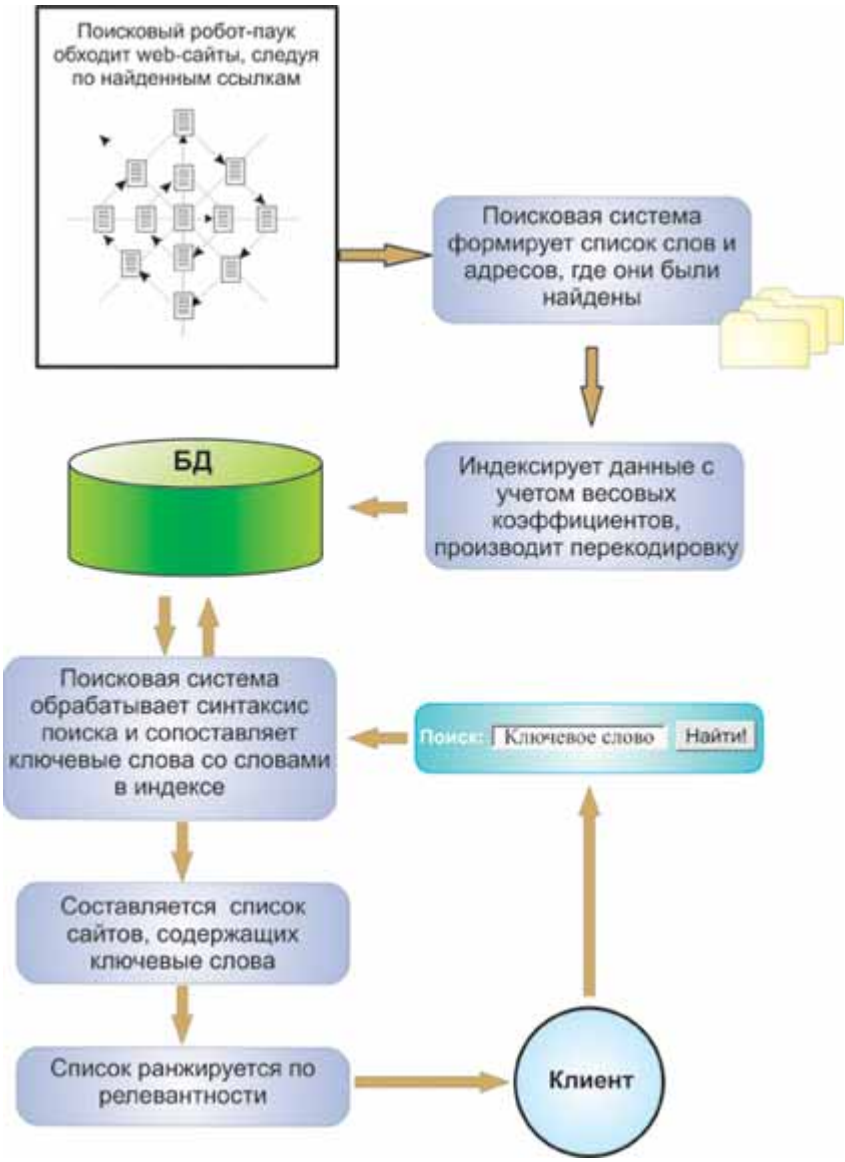

Рисунок 2. Схема построения индекса

При нахождении очередной HTML-страницы большая часть поисковых систем фиксируют слова, картинки, ссылки и другие элементы (по-разному в различных поисковых системах), содержащиеся на ней. Причём при отслеживании слов на странице фиксируется не только их наличие, но и местоположение, т.е. где данные слова находятся: в заголовке (title), подзаголовках (subtitles), в метатэгах (meta tags) или в других местах. При этом обычно фиксируются значимые слова, а союзы

и междометия типа "а", "но" и "или" игнорируются. Метатэги дают возможность владельцам страниц определить тематику и ключевые слова, по которым индексируется документ. Это может быть актуально в случае, если ключевые слова имеют несколько значений. Метатэги могут сориентировать поисковую систему при выборе из нескольких значений слова на единственно правильное. Однако метатэги работают надёжно только в том случае, когда заполняются честными владельцами сайтов. Недобросовестные владельцы веб-сайтов помещают в свои метатэги наиболее популярные в Интернете слова, не имеющие ничего общего с темой сайтов. В результате посетители попадают на незапрашиваемые сайты, повышая тем самым их рейтинг. Именно поэтому многие современные поисковые системы либо игнорируют метатэги, либо считают их дополнительными по отношению к тексту страницы. Каждый робот поддерживает свой список ресурсов, наказанных за недобросовестную рекламу[4].

Очевидно, что если пользователь ищет сайты по ключевому слову "собака", то поисковый механизм должен найти не просто все страницы, где упоминается слово "собака", а те, где это слово имеет отношение к теме сайта. Для того чтобы определить, в какой степени то или иное слово имеет отношение к профилю некоторой Web-страницы, необходимо оценить, насколько часто оно встречается на странице, есть ли по данному слову ссылки на другие страницы или нет. Короче говоря, необходимо ранжировать найденные на странице слова по степени важности. Словам присваиваются весовые коэффициенты в зависимости от того, сколько раз и где они встречаются (в заголовке страницы, в начале или в конце страницы, в ссылке, в метатэге и т.д.). Каждый поисковый механизм имеет свой алгоритм присваивания весовых коэффициентов - это одна из причин, по которой поисковые машины по одному и тому же ключевому слову выдают разные списки документов. Т. к. страницы постоянно обновляются, процесс индексирования должен выполняться постоянно. Роботы-пауки путешествуют по ссылкам и формируют файл, содержащий индекс, который может быть довольно большим. Для уменьшения его размеров прибегают к минимизации объёма информации и сжатию файлов. Имея несколько роботов, поисковая система может обрабатывать сотни страниц в секунду. Сегодня мощные поисковые машины хранят сотни миллионов страниц и получают десятки миллионов запросов каждый день.

При построении индекса решается также задача уменьшения числа дубликатов - задача нетривиальная, учитывая, что для корректного сравнения необходимо сначала определить кодировку документа. Ещё более сложной задачей является отделение очень похожих документов (их называют "почти дубликаты"), к примеру

таких, в которых отличается только заголовок, а текст дублируется. Подобных документов в Интернете очень много – к примеру, кто-то списал реферат и опубликовал его на сайте за своей подписью. Современные поисковые системы позволяют решать данные проблемы[5].

1.3. Оптимизация в поисковых системах

Одним из видом продвижения в сети Интернет является поисковая оптимизация или SEO-оптимизация – это совокупность мер по продвижению Интернет-ресурса в поисковых сетях[6].

Копирайтинг - это процесс написания уникальных статей, которые продвигают услугу, продукт, человека, мнение или идею. Контент сайтов, заголовки, слоганы, ключевые фразы, тексты рассылок — всё это должно быть интересным, уникальным, максимально соответствующим задачам сайтов.

SEO-копирайтинг — это создание уникального тематического контента с оптимизацией для поисковых систем. Такой контент:

- понятен каждому и прост;

- адаптирован для успешного продвижения сайтов;

- стимулирует покупательскую активность, увеличивая конверсию.

Профессиональная SEO-оптимизация – комплекс действий над сайтом, которые направлен на повышение позиций выдачи сайта в популярных поисковых системах (Google и Яндекс). Данная стадия раскрутки является обязательной для продвижения большинства коммерческих проектов в Сети. Неважно, какие услуги и товары предлагает сайт: главная цель – привлечь целевых клиентов на страницы сайта. Часто это делается при помощи платной рекламы (баннерной, контекстной, тизерной и т. д.), но также можно привлечь посетителей напрямую из поисковых систем. Оптимизацией занимается много компаний и студий интернет-маркетинга, но не все они дают стабильный результат. Качественное продвижение сетевых ресурсов – мероприятие поэтапное и длительное. Важно не просто повысить количество посетителей сайта, а привлечь потенциальных (целевых) клиентов.

SEO-продвижение обязательно включает техническую оптимизацию сайта, аудит и увеличение уровня релевантности — соответствия продвигаемых статей вашему

платному предложению на сайте. Современная поисковая оптимизация – реальный технологический инструмент, который повышает продажи гораздо эффективнее, чем традиционная реклама. В идеале начинать оптимизацию следует уже на этапе разработки сайта: следует сразу продумать и определить функциональное семантическое ядро – список ключевых слов (тем), которые наиболее точно определяют общее направление деятельности конкретного ресурса.

SEO-оптимизация под системы поиска строго ориентирована на цели бизнеса в XXI веке. Поисковые системы Google, Яндекс и другие давно стали неотъемлемой частью сетевого пространства. Если сайт не будет посещаемым, в том числе и SEO не оптимизированным, он будет бесполезен, так как о нем никто не узнает из-за отсутствия посетителей. Во всем мире бизнес постепенно перемещается в интернет-пространство. Все большее количество людей заказывает покупки через сайты коммерческих компаний и онлайн-магазины. Раскрутка фирмы в интернете – самый действенный на сегодня метод увеличения ее популярности[7].

Для эффективного SEO-продвижения важна совокупность многих факторов, среди которых выделяют:

- Плотность ключевых слов. Современные поисковые системы обладают отлаженными механизмами семантического анализа сайта, поэтому Интернет-ресурс должен обладать направленность ключевых слов должна соответствовать тематике сайта.

- Индекс цитируемости. Данный показатель характеризует авторитетность ресурса, показывающий, что на данный сайт ссылаются другие сайты. Наилучшим вариантом является, если на сайт ссылаются другие авторитетные ресурсы[8].

Для расчета данного индекса существуют различные алгоритмы ранжирования. Основными индексами цитируемости являются показатели от Google и Яндекс – PageRank и тематический индекс цитирования (ТИЦ).

- Пользовательские факторы. Это факторы, учитывающие поведение пользователя на ресурсе. К ним относятся посещаемость сайта, глубина просмотра сайта, различные переходы и действия, а также время, которое пользователь провел на сайте.

Таким образом, главной задачей SEO-продвижения является совокупность действий с контентом сайта для повышения позиций сайта в поисковых системах по сравнению с конкурентами.

По степени мероприятий по SEO-продвижению выделяют три вида SEO-оптимизации – белая, серая и черная.

Белая оптимизация является основным легальным компонентом всей структуры SEO-оптимизации. Совокупность мероприятий белой оптимизации направлены на создание качественного и уникального контента, на разработку юзабилити (usability), продвижение сайта с помощью социальных сетей и блогов и обмен ссылками.

К методам серой оптимизации относят не рекомендуемые мероприятия, которые однако формально не запрещаются. Примером серой оптимизации может служить перенасыщенность контента ключевыми словами.

К методам черной оптимизации относят запрещенные мероприятия по повышению сайта в поисковых системах, к которым относят спам и нацеленность контента не на пользователя, а на поискового робота.

Глава 2. Анализ современных поисковых систем в Интернет

2.1. Обзор рынка поисковых систем в России

На текущий момент на первом месте находится Google с долей 54,24%. Яндекс является второй по популярности поисковой системой в России с долей 42,27%. Остальные системы занимают незначительную долю, в частности, на третьем месте находится Mail.ru с долей 1,77% (рис. 3). Главные страницы поисковых систем Яндекс и Google представлены в Приложении.

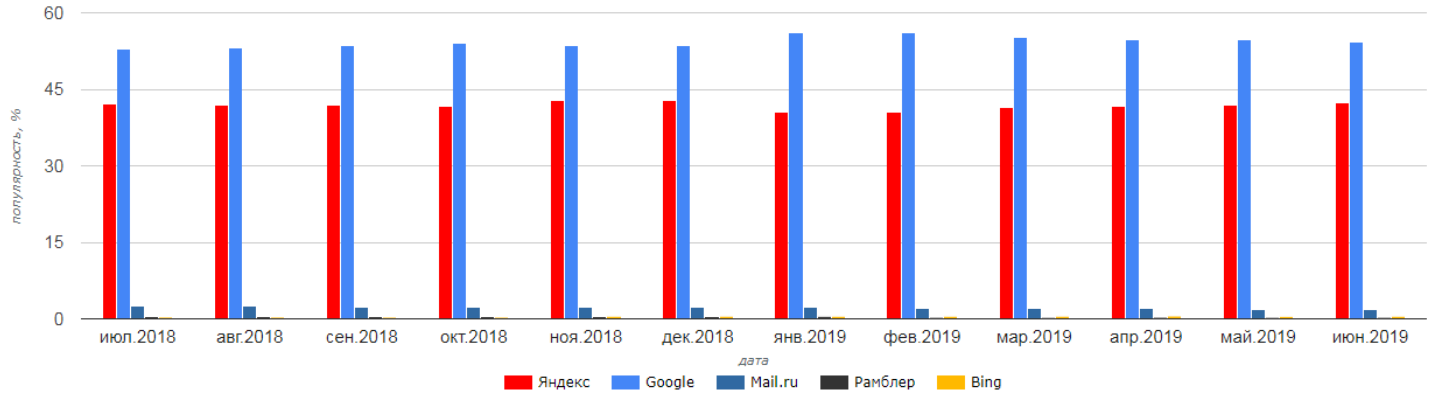

Рисунок 3. Самые популярные поисковые систем в России[9]

В табл. 1 представлена динамика изменения популярности поисковых систем.

Таблица 1

Динамика изменения популярности поисковых систем[10]

	янв. 2019	фев. 2019	мар. 2019	апр. 2019	май. 2019	июн. 2019
Яндекс	40.58%	40.61%	41.54%	41.78%	42.00%	42.27%
Google	56.09%	56.12%	55.13%	54.80%	54.79%	54.24%
Mail.ru	2.31%	2.16%	2.02%	2.02%	1.74%	1.77%
Рамблер	0.48%	0.24%	0.22%	0.26%	0.23%	0.24%
Bing	0.50%	0.50%	0.51%	0.57%	0.47%	0.50%
Yahoo!	0.17%	0.17%	0.17%	0.18%	0.18%	0.20%
Ask	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
Nigma	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
QIP	0.04%	0.03%	0.03%	0.04%	0.04%	0.04%

Рассмотрим самые популярные поисковые системы в России более подробно.

Google — самая популярная поисковая система в мире, которая занимает первое место в мировом рейтинге. Была создана ещё в 1998 году программистами Сергеем Брином и Ларри Пейджом. Обрабатывает свыше 41 млрд запросов в месяц, в индексе порядка 25 миллиардов веб-страниц, на сайт заходят более 200 миллионов

человек по всему миру и набирает более 72% запросов со всего мира. Поисковая система Гугл постоянно совершенствуется и улучшается. Позволяет пользователям искать информацию в мире, включая веб-страницы, изображения и видео. В 2017 году признан самым дорогим брендом в мире. Также корпорации Alphabet Inc. наряду с Google принадлежит браузер Хром и мобильная операционная система Андроид.

Яндекс — самая популярная поисковая система в Рунете. Была основана в 1997 году Аркадием Воложом и Ильёй Сегаловичем. Каждый день поиск Яндекса обрабатывает примерно на 280 миллионов запросов, а главной страницей Яндекса ежедневно пользуется 28 млн российских пользователей.

Mail.ru — это русскоязычный интернет-портал, принадлежащий крупнейшему IT-гиганту Рунета Mail.Ru Group и имеющий множество тематических проектов, в том числе ВКонтакте, Мой мир и Одноклассник». Ежемесячная аудитория портала составляет 54 миллиона человек и занимает 47-е место по популярности в мире, а в России — 5-е место. Создан в 1998 году авторами Евгением Голандом и Владимиром Шутовым. С начала 2006 года по 2009 год использовался поиск от «Яндекса».

Рамблер — популярный сервисный интернет-портал. Ежедневно на главную страницу заходит около полтора миллиона пользователей, а ежемесячно около шести миллионов. Был создан в 1996 году программистом Дмитрием Крюковым. До 23 июня 2011 года являлся одной из поисковых систем Рунета, но перестал существовать, перейдя на поисковый движок компании «Яндекс».

Bing — поисковая система, разработанная международной корпорацией Microsoft. Была основана в 2009 году. В настоящее время сайт Bing занимает 6-е место в списке самых популярных поисковых сайтов в мире. Посещаемость уже менее 200 миллионов в день. Наиболее активно внедряется в смартфоны на ОС Windows. Больше всего заходов из США (32%), Китая (16%) и Германии (4%).

Yahoo — это один из старейших и наиболее популярных в Интернете поисковиков. Создали его предприниматели Джерри Янг и Дэвид Фило в далёком 1995 году. Входит в первую двадцатку по посещаемости среди всех веб-сайтов в мировой сети. С 2009 года использует поиск Bing, а с 2017 года куплена компанией Verizon Communications. Крупный интернет-портал и поставщик услуг, предлагающий результаты поиска, настраиваемый контент. Наибольшая доля посетителей – граждане США.

Bing — поисковая система, разработанная международной корпорацией Microsoft. Была основана в 2009 году. В настоящее время сайт Bing занимает 6-е место в списке самых популярных поисковых сайтов в мире. Посещаемость уже менее 200 миллионов в день. Наиболее активно внедряется в смартфоны на ОС Windows. Больше всего заходов из США (32%), Китая (16%) и Германии (4%).

2.2. Сравнительный анализ поисковых систем

Теперь обратимся к положительным и отрицательным сторонам ранее рассмотренных наиболее популярных поисковых систем, тем самым продемонстрировав особенности, которыми должна обладать наиболее удобная система поиска.

Таблица 2

Сравнительный анализ поисковых систем

Поисковая система	**Преимущества**	**Недостатки**
1	2	3

Яндекс	1)Непрерывное развитие системы. 2)Качество выдачи растет, все больше удобных сервисов предлагает компания: каталог, карты, новости, прогноз погоды, почта. 3) глубокий морфологический анализ обрабатываемых терминов. 4) обладает хорошим механизмом распознавания одного документа в нескольких кодировках или на зеркальных серверах. 5) оригинально сконструированный механизм выдачи результатов. 6) огромная индексная база.	1) Разница в выдаче при наборе слова с большой (маленькой) буквы (иногда выдача меняется, иногда нет). 2) Частое выпадение секторов поисковой базы - когда исчезают части сайтов из выдачи и восстанавливаются через 2-5 дней. 3) Обновление индексов поисковой базы происходит недостаточно часто и регулярно.

Rambler	1) Система работает с большой скоростью поиска. 2) Обновление поискового индекса происходит несколько раз в день. 3) Поисковик всегда находит самые свежие документы и последние новости. 4) Обладает близким к оптимальному выводом результатов поиска. 5) производит ранжирование результатов в зависимости от частоты употребления и местоположения искомых терминов. 6) Один и тот же документ в различных кодировках показывается только один раз, а его конкретные адреса. суммируются в списке, идущим за резюме.	1) На величину индекса релевантности влияет время существования сайта в сети. Эта особенность позволяет пользователям находить ресурсы, которые давно существуют, успешно развиваются, а не сайты-однодневки. Но такой подход значительно затрудняет попадание в выдачу новых сайтов, информация на которых подчас оказывается актуальной и, возможно, более важной для пользователя. 2) невозможность осуществления поиска по целой фразе указывая в запросах предельное расстояние искомых терминов друг от друга.

Продолжение таблицы 2

1	2	3

Google	1) Очень мощная поисковая система, которая находится в постоянном развитии. 2) База индексов этой системы обновляется раз в два дня, качество выдачи очень высокое, найти необходимый документ или информацию довольно легко. 3) Система ориентирована в основном на ссылки, причем учитываются как входящие, так и исходящие ссылки с ресурса. 4) Способна выдавать результаты на запросы по семантике языка программирования (исходный код поиска).	1) Нередко встречаются ссылки на сайты с уже устаревшей информацией. 2) Случается, что ссылки, которые находятся в результатах поиска, ведут на сайт, находящийся в стадии разработки. 3) На запрос «фильм» и «фильмы» результаты поиска будут отличаться. 4) отсутствие возможности указать конкретную грамматическую форму слова, либо ударение также значительно усложняет процесс поиска информации.
Yahoo!	1) Содержит ссылки, которые наиболее полно отвечают указанной в запросе тематике. 2) Имеются интеллектуальные средства «отсечения» пустых, находящихся в разработке или чисто рекламных сайтов, далеких от искомой тематики. 3) всегда легко определить, в каком разделе находится нужная информация. 4) В случае если на Yahoo нет результатов, сразу выводятся результаты с AltaVista.	1) Возможна проблема с отсутствующими страницами, поскольку веб-мастера обычно забывают удалить свои сайты с поисковых систем, а на Yahoo нет механизма автоматического обновления. 2) Чисто русские ресурсы не добавляются, потому что их просто некому смотреть и оценивать содержимое. 2)Нет собственной поисковой машины. 3) Ищет слова, заданные в критерии поиска только в названии и описании страницы

Главный недостаток современных поисковых систем – это их централизация. А централизация означает, что вся информация хранится в одном месте, все работы и расчёты производятся в одном месте, все решения (результаты выдачи) принимаются в одном месте.

2.3. Организация поиска в поисковой системе Яндекс

Поисковая машина Яндекса отвечает на вопросы пользователей, находя нужные документы в интернете. А размеры современного интернета исчисляются в эксабайтах, то есть в миллиардах миллиардов байтов. Конечно же, Яндекс не обходит весь интернет каждый раз, когда ему задают вопрос. Поисковая система, так сказать, делает домашнее задание.

Поиск в интернете состоит из двух частей. Первая — поисковик обходит интернет, создавая его слепок на своих серверах. Вторая — пользователь задаёт запрос и получает ответ с серверов поисковика.

Яндекс ищет по поисковому индексу — базе данных, где для всех слов, которые есть на известных поиску сайтах, указано их местонахождение — адрес страницы и место на ней. Индекс можно сравнить с предметным указателем в книге или адресным справочником. В отличие от обычного предметного указателя, индекс содержит не только термины, а вообще все слова. А в отличие от адресного справочника, у каждого слова-адресата есть не одно, а очень много «мест прописки».

Подготовка данных, по которым ищет поисковая машина, называется индексированием. Специальная компьютерная система — поисковый робот — регулярно обходит интернет, выкачивает документы и обрабатывает их. Создается своего рода слепок интернета, который хранится на серверах поисковика и обновляется при каждом новом обходе.

У Яндекса два поисковых робота — основной и быстрый (он называется Orange). Основной робот индексирует интернет в целом, а Orange отвечает за то, чтобы в поиске можно было найти самые свежие документы, которые появились минуты или даже секунды назад. У каждого робота есть список адресов документов, которые нужно проиндексировать.

Когда при обходе робот видит на уже известных сайтах новые ссылки, он добавляет их в свой список, увеличивая количество индексируемых страниц. Впрочем, владелец сайта сам может помочь основному роботу Яндекса найти свой ресурс и подсказать, например, как часто обновляются его страницы — через сервис Яндекс.Вебмастер.

Сначала программа-планировщик выстраивает маршрут — очередность обхода документов. При этом планировщик учитывает важные для поисковой системы характеристики сайтов, такие как, например, цитируемость или частота обновления документов. После создания маршрута планировщик отдаёт его другой части поискового робота — «пауку». Паук регулярно обходит документы по заданному маршруту. Если сайт на месте, то есть работает и доступен, паук выкачивает запланированные в маршруте документы. Он определяет тип скачанного документа (html, pdf, swf и т.п.), кодировку и язык, а затем отправляет данные в хранилище.

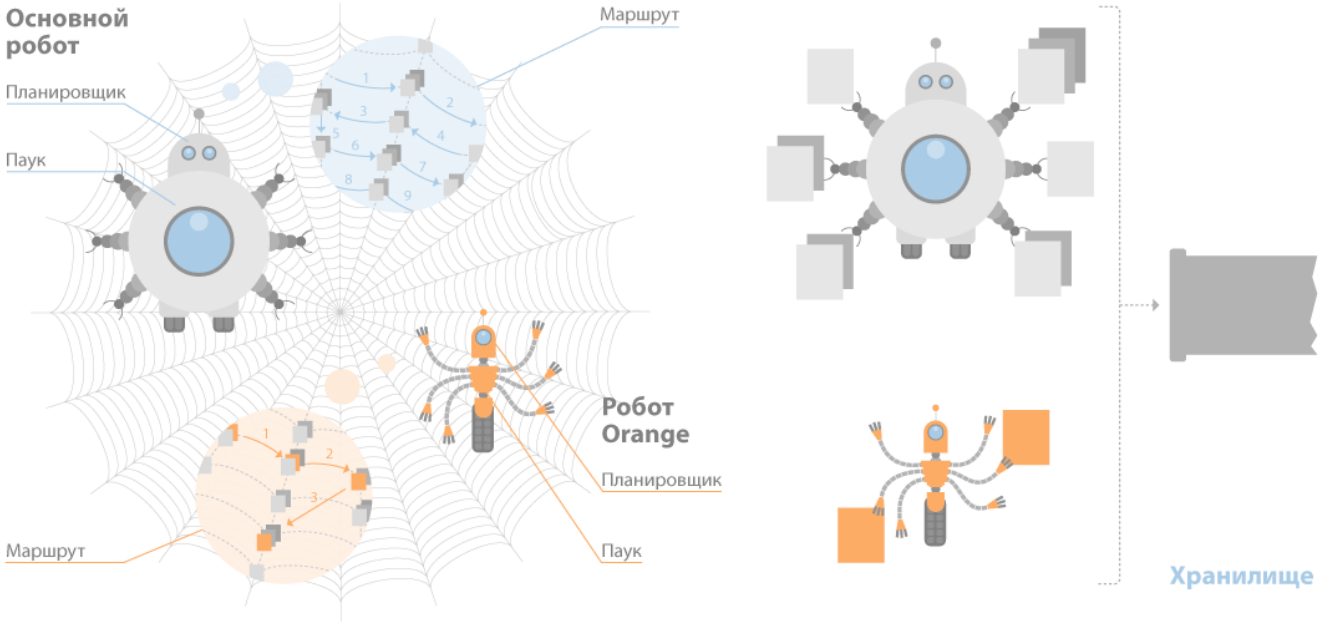

Рисунок 4. Роботы поисковой системы Яндекс[11]

Там программа разбирает документ по кирпичику: очищает от html-разметки, оставляя чистый текст, выделяет данные о местоположении каждого слова и добавляет их в индекс. Сам документ в исходном виде также остается в хранилище до следующего обхода. Благодаря этому пользователи могут найти в Яндексе и посмотреть документы, даже если сайт временно недоступен. Если сайт закрылся

или документ был удалён или обновлён, Яндекс удалит копию со своих серверов или заменит её на новую.

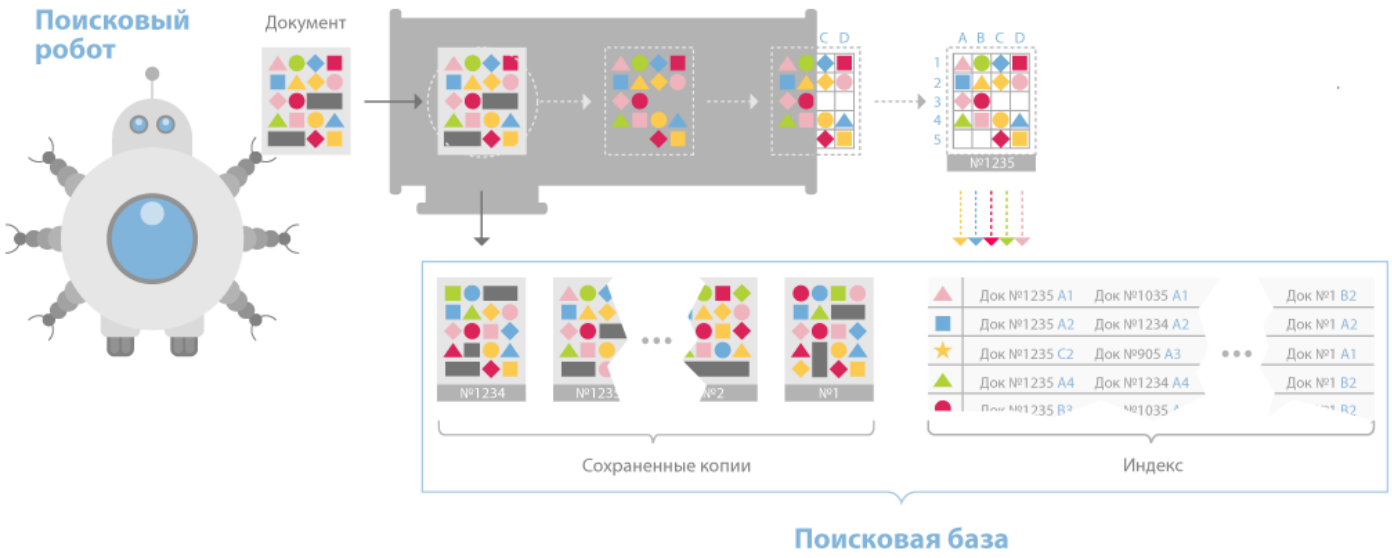

Рисунок 5. Поисковый робот системы Яндекс[12]

Поисковый индекс, данные о типе документов, кодировке, языке и сохраненные копии документов вместе составляют поисковую базу. Она обновляется постоянно, но, чтобы это обновление стало доступно пользователям, её нужно перенести на «базовый поиск». Базовый поиск — сервера, которые отвечают пользователям на запросы. Туда переносится не вся поисковая база, а только её полезная часть — без спама, дубликатов сайтов (зеркал) и других ненужных документов.

Обновление поисковой базы из хранилища основного робота попадает в поиск «пакетами» — раз в несколько дней. Этот процесс создаёт дополнительную нагрузку на сервера, поэтому производится ночью, когда к Яндексу обращаются на порядок меньше пользователей. Сначала новые части базы помещаются рядом с такими же частями из прошлого обхода. Затем они проверяются по целому ряду факторов, чтобы обновление не ухудшило качество поиска. Если проверка прошла успешно, новая часть базы заменяет собой старую.

Робот Orange предназначен для поиска в реальном времени. Его планировщик и паук настроены так, чтобы находить новые документы и выбирать из огромного их количества все, хоть сколько-нибудь интересные. Каждый такой документ Orange сразу обрабатывает и выкладывает на базовый поиск. Срочных документов не

очень много по сравнению с общим объемом интернета, поэтому обновление базы в реальном времени можно делать и при дневных нагрузках на сервера.

Страница результатов поиска — это ответ Яндекса на вопрос, который пользователь задал в поисковой строке. Она содержит не только ссылки на страницы, на которых нашлась нужная информация, но и дополнительные ответы, которые могут быть полезны пользователю — например, краткую справку об объекте, подходящий колдунщик или контекстные объявления Директа. Яндекс ведёт параллельный поиск по разным массивам информации, и на странице результатов поиска могут появляться картинки, видео и карты, музыкальный плеер, ссылки на товары на Маркете и другие данные. Перейти к ответам другого сервиса можно с помощью вертикального меню в левой части страницы.

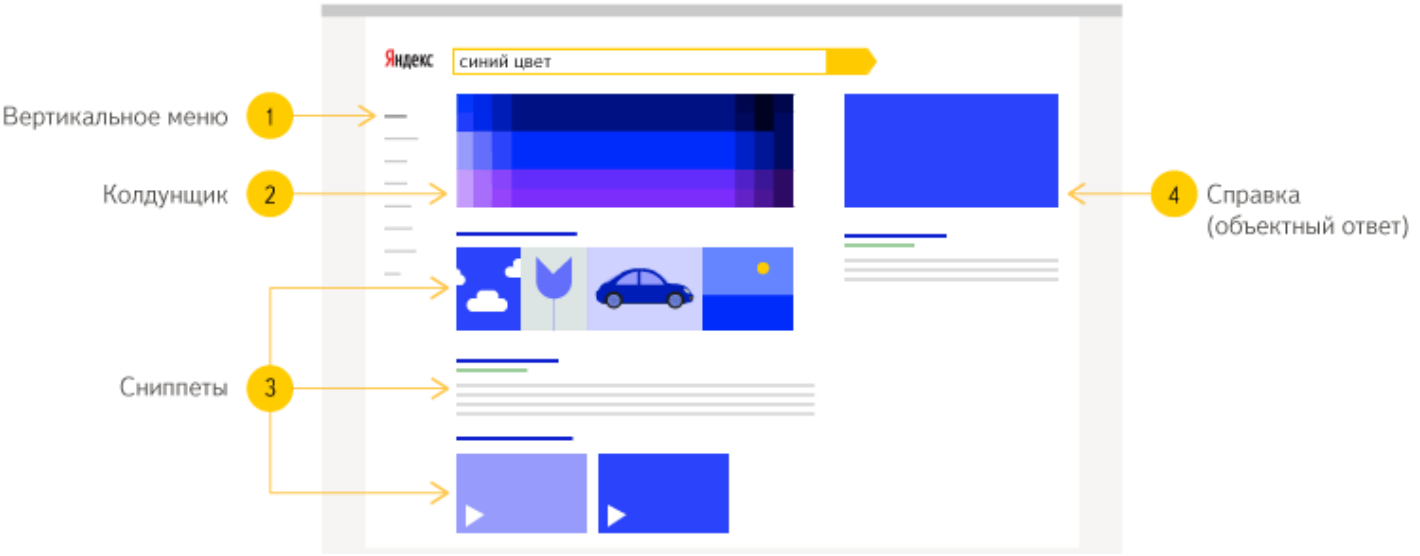

Рисунок 6. Страница результатов поиска Яндекс

Сниппет — это ссылка на найденную в интернете страницу, с заголовком, небольшим текстом, содержащим основную информацию о документе, и специальными элементами, которые могут меняться в зависимости от типа запроса.

Яндексу важно не просто показать релевантные ответы, но и описать их максимально информативно — так, чтобы пользователь мог понять, какой ответ подходит ему лучше всего сразу, не тратя время на переходы по нескольким ссылкам.

Для заголовка результата поиска Яндекс чаще всего использует заголовок самого документа. Если он слишком длинный, система выбирает фрагмент, который

больше всего подходит по смыслу к заданному запросу. Бывает, что у документа нет заголовка или заголовок не соответствует содержанию. Например, названия файлов в формате doc или pdf часто короткие и малоинформативные. В таких случаях Яндекс создаёт заголовок самостоятельно, основываясь на текстах ссылок на документ, заголовках в самом тексте документа и его содержании.

Для формирования описания страницы программа выбирает все фрагменты текста документа со словами из запроса. Каждый из таких фрагментов разбивается ещё на несколько частей — например, со словами из запроса в начале, в конце и в середине. Затем программа сравнивает их между собой и выбирает лучшие — они и попадают в сниппет.

При этом учитываются несколько десятков факторов. Одни из них повышают шансы попадания фрагмента в описание, другие — наоборот. Например, если слово содержится в длинном предложении, высока вероятность, что это часть повествования, а не навигационная ссылка. Значит, это хороший фрагмент для сниппета. Кроме того, Яндекс старается выбирать фрагменты из разных частей текста — так можно полнее описать содержание страницы. А вот фрагмент, схожий с заголовком текста страницы, вряд ли окажется в описании — чтобы не дублировать информацию.

Для каждого фактора компьютерная система рассчитывает коэффициент. С помощью машинного обучения она учится сама понимать значимость факторов, основываясь на данных от специалистов-асессоров. Они просматривают некоторые наборы сниппетов, вручную разделяют их на хорошие и плохие и сообщают эти оценки системе. Затем система уже без помощи людей строит формулу, по которой создает сниппеты.

При ответе на общие, неоднозначные запросы в сниппеты попадают уточняющие слова. Например, описания результатов поиска по запросу [буратино] будут встречаться слова «сказка», «мюзикл» и «огнемётная система». О том, какие слова помогают пользователю сориентироваться, Яндекс узнаёт, анализируя то, как люди переформулируют и уточняют свои запросы, и рассчитывая значимость этих уточнений.

Сниппет оформляется так, чтобы пользователю было легче его воспринимать. Заголовки выделены синим цветом — так традиционно выделяются ссылки в текстах веб-страниц. Узнать знакомый ресурс помогает небольшой фирменный значок сайта, слева от заголовка. Если заголовок или текст описания содержит

прописные буквы, Яндекс старается сделать их строчными — так проще читать.

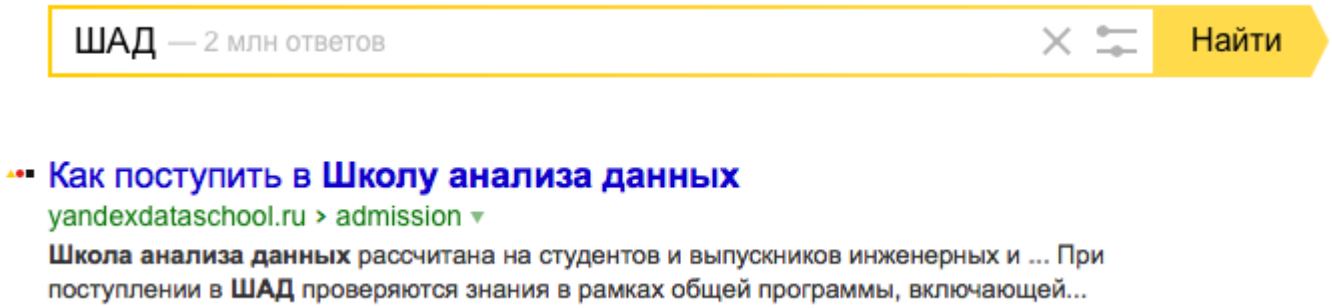

Рисунок 7. Оформление сниппетов

Чтобы было легче «зацепиться глазом», все слова из запроса в результатах поиска выделены жирным шрифтом. При этом Яндекс умеет сопоставлять аббревиатуры и их расшифровки, полные имена, сокращения и инициалы, числа и их текстовое написание. Например, по запросу [петр 1] Яндекс найдет документы, которые содержат и «Петр I», и «Петр первый», и выделит в сниппетах разные варианты написания имени.

Яндекс старается сделать так, чтобы пользователи могли быстро найти ответ — иногда даже сразу на странице результатов поиска. Для разных ответов нужна разная дополнительная информация. Например, если человек задаёт в запросе название организации, возможно, ему нужно узнать, где она находится или как с ней связаться. Чтобы не пришлось тратить время на поиски страницы с контактами на сайте организации, Яндекс добавляет в сниппет её телефон, физический адрес и кнопку, открывающую карту с нужным объектом.

Если Яндексу известна структура сайта, он показывает её пользователю. Под описанием появляются ссылки на его наиболее посещаемые страницы (например, «Контакты», «Галерея» или «Каталог товаров») — чтобы при желании пользователь мог перейти в нужный раздел, тратя меньше кликов и трафика. А адрес документа Яндекс преобразует в навигационную цепочку — названия разделов и подразделов сайта, из которых состоит путь до документа.

Для некоторых предметных областей Яндекс добавляет в ответ специальную информацию. Например, пользователь, который ищет какой-нибудь товар, увидит рейтинг магазина-продавца с Яндекс.Маркета, а ответ на запрос с моделью автомобиля будет содержать объявления о продаже подходящих машин. Благодаря таким сниппетам пользователь экономит время и трафик, а организация

получает посетителя сайта, заинтересованного именно в её услугах.

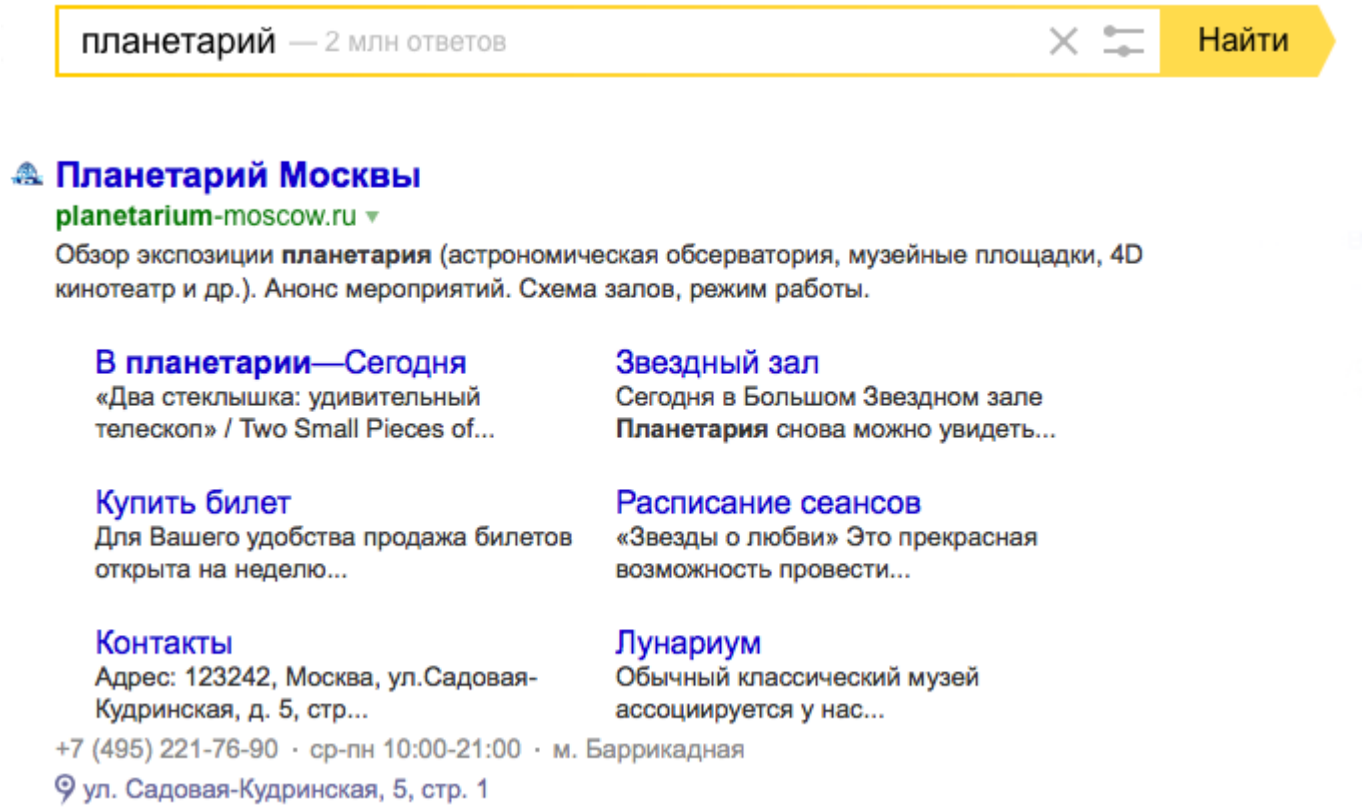

Рисунок 8. Результат поиска в Яндекс

Таким образом, каждый день пользователи задают Яндексу десятки миллионов запросов, и поисковая система должна не только точно отвечать, но и быстро обрабатывать весь этот поток.

ЗАКЛЮЧЕНИЕ

В век информационных технологий огромную роль играет интернет, а любое путешествие по просторам интернета невозможно без специальных поисковых систем, позволяющих комфортно просматривать любимые веб-страницы. Первоочередной задачей любой поисковой системы является доставление людям именно той информации, которую они ищут**.**

На сегодняшний день поисковые системы являются сложнейшими и громадными механизмами, представляющие собой не только инструмент для нахождения любой необходимой информации, но и довольно увлекательные сферы для бизнеса.

Работа с помощью поисковых систем позволяет многим пользователям глобальной сети осуществлять быстрый поиск нужной информации в кратчайшие сроки. В результате поисковые системы уже долгое время являются обязательной частью интернета и жизни общества.

Под поисковой системой понимается программное обеспечение, состоящее из базы данных документов, снабженной пользовательским интерфейсом, позволяющим пользователю получить упорядоченное подмножество этих документов как ответ на его поисковый запрос. Основная задача поисковой системы заключается в выборе наилучшего возможного подмножества в ответ на конкретный запрос, т.е. множества документов, которые наиболее соответствуют тому, что ищет пользователь (обычно в порядке убывания релевантности).

На текущий момент на первом месте в России находится Google с долей 54,24%. Яндекс является второй по популярности поисковой системой в России с долей 42,27%. Остальные системы занимают незначительную долю, в частности, на третьем месте находится Mail.ru с долей 1,77%.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Ашманов И. С. Продвижение сайта в поисковых системах / И. С. Ашманов. - М.: «Вильямс», 2016. - 304 с.
2. Байков В.Д. Интернет. Поиск информации. Продвижение сайтов / Д.В Байков. - СПб.: БХВ-Петербург, 2015. - 288 с.
3. Барсегян А. А. Технологии анализа данных. DataMining, VisualMining, TextMining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В., Степаненко, И. И. Холод. - СПб.: БХВ-Петербург, 2015. - 384 с.
4. Гайдамакин Н. А. Автоматизированные информационные системы, базы и банки данных / Н. А. Гайдамакин.- М. : «Гелиос», 2016.- 280 с.
5. Граппоне Д., Казн Г. Поисковая оптимизация сайтов: исчерпывающее руководство / Д. Граппоне, Г. Казн. - М.: Эксмо, 2015. - 528 с.
6. Дорофеев В. Яндекс Воложа: История создания компании мечты. - М.: Альпина Паблишер, 2017. – 275 с.
7. Евдокимов Н.В. Основы контентной оптимизации. Эффективная Интернет-коммерция и продвижение сайтов в Интернет. - М.: Вильямс, 2015. – 345 с.
8. Завьялов Д.В. О применении информационных технологий / Современные наукоемкие технологии. - 2018. - № 8-1. - С. 71-72

9. Информатика. Базовый курс: учебник / под ред. С. В. Симоновича. - СПб: Питер, 2016.- 110 с.
10. Кадеев Д. Н. Информационные технологии и электронные коммуникации / Д. Н. Кадеев. - М.: Электро, 2016. - 250 с.
11. Кириллов А. Поисковые системы: компоненты, логика и методы ранжирования // Бизнес-информатика. – 2018. – №4. – С. 51-59
12. Колисниченко Д. Н. Поисковые системы и продвижение сайтов в Интернете / Д. Н. Колисниченко. - М.: Диалектика, 2017. – 272 с.
13. Ланкастер Ф. У. Информационно-поисковые системы. Характеристики, испытание и оценка / Ф. У. Ланкастер. - М.: Наука, 2015. - 278 с.
14. Маннинг К. Введение в информационный поиск / К. Маннинг. – М.: Вильямс, 2015. - 200 с.
15. Попкова Е. Г., Ионов А. Ч., Токарева И. В. Эффективность рекламы в социальных сетях // Известия Волгоградского государственного технического университета. - 2017. - № 4 (131). – С. 41-48
16. Фомина Ю.А., Преображенский Ю.П. Принципы индексации информации в поисковых системах / Вестник Воронежского института высоких технологий. - 2017. - № 7. - С. 98-100
17. Юрасов А. В. Основы электронной коммерции. – М.: Горячая линия-Телеком, 2016. – 279 с.
18. Статистика поисковых систем LiveInternet [Электронный ресурс]. – Режим доступа: liveinternet.ru
19. Рейтинг поисковых систем SEO-AUDITOR [Электронный ресурс]. – Режим доступа: http://gs.seo-auditor.com.ru/sep
20. Поисковая система Яндекс [Электронный ресурс]. – Режим доступа: https://yandex.ru

ПРИЛОЖЕНИЯ

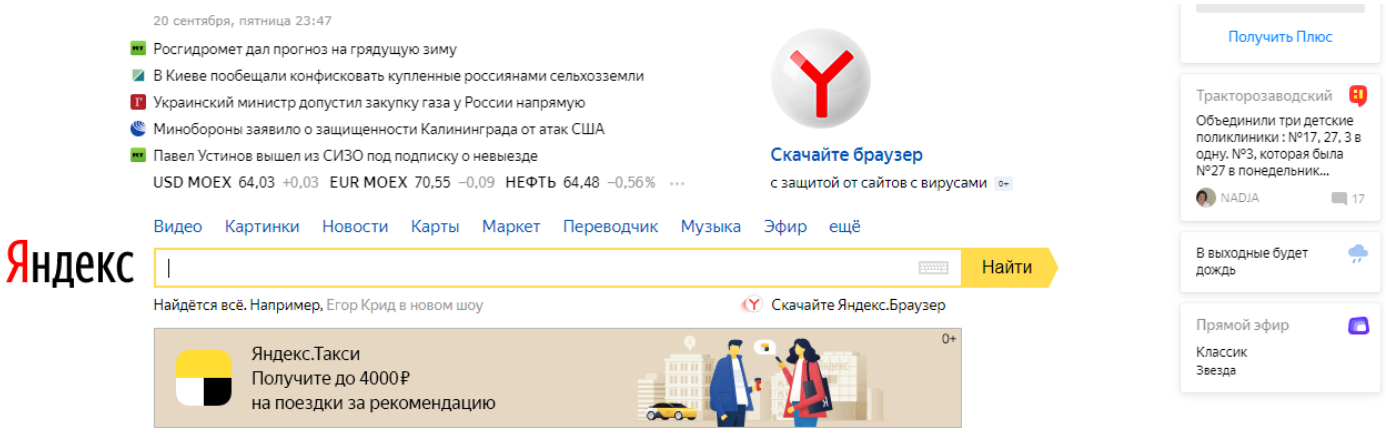

Рисунок 1. Главная страница поисковой системы Яндекс

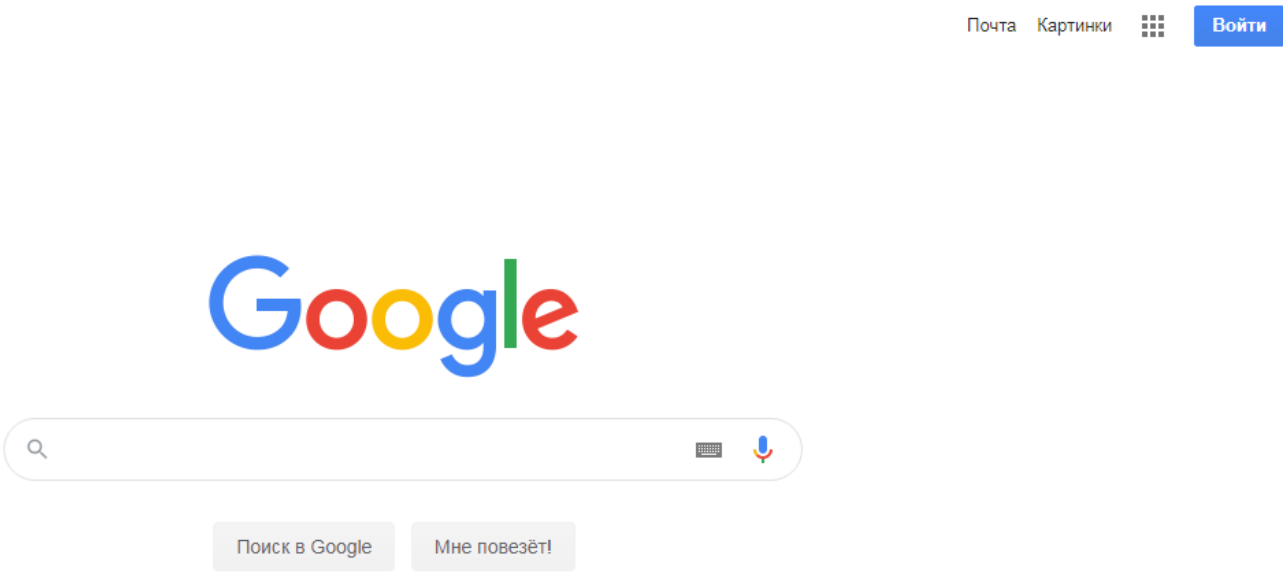

Рисунок 2. Главная страница поисковой системы Google

1. Байков В.Д. Интернет. Поиск информации. Продвижение сайтов / Д.В Байков. - СПб.: БХВ-Петербург, 2015. – С. 45 ↑

2. Гайдамакин Н. А. Автоматизированные информационные системы, базы и банки данных / Н. А. Гайдамакин.- М. : «Гелиос», 2016. - С. 75 ↑

3. Завьялов Д.В. О применении информационных технологий / Современные наукоемкие технологии. - 2018. - № 8-1. - С. 71 ↑

4. Кадеев Д. Н. Информационные технологии и электронные коммуникации / Д. Н. Кадеев. - М.: Электро, 2016. – С. 145 ↑

5. Кириллов А. Поисковые системы: компоненты, логика и методы ранжирования // Бизнес-информатика. – 2018. – №4. – С. 51 ↑

6. Евдокимов Н.В. Основы контентной оптимизации. Эффективная Интернет-коммерция и продвижение сайтов в Интернет. - М.: Вильямс, 2015. – С. 37 ↑

7. Юрасов А. В. Основы электронной коммерции. – М.: Горячая линия-Телеком, 2016. – С. 145 ↑

8. Попкова Е. Г., Ионов А. Ч., Токарева И. В. Эффективность рекламы в социальных сетях // Известия Волгоградского государственного технического университета. - 2017. - № 4 (131). – С. 45 ↑

9. Рейтинг поисковых систем SEO-AUDITOR [Электронный ресурс]. – Режим доступа: http://gs.seo-auditor.com.ru/sep ↑

10. Рейтинг поисковых систем SEO-AUDITOR [Электронный ресурс]. – Режим доступа: http://gs.seo-auditor.com.ru/sep ↑

11. Поисковая система Яндекс [Электронный ресурс]. – Режим доступа: https://yandex.ru ↑

12. Поисковая система Яндекс [Электронный ресурс]. – Режим доступа: https://yandex.ru ↑