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 мне этой темы, так как я часто пользуюсь всемирными сетями.

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. Что же такое поисковая система? Поисковая система – портал, осуществляющий поиск, сбор и сортировку информации в сети Интернет. Поисковые системы это инструмент, позволяющий пользователю глобальной сети в кратчайшие сроки найти интересующую его информацию.

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

Получая результат, пользователь оценивает работу системы, руководствуясь несколькими основными параметрами. Нашел ли он то, что искал? Если не нашел, то сколько раз ему пришлось перефразировать запрос, чтобы найти искомое? Насколько актуальную информацию он смог найти? Насколько быстро обрабатывала запрос поисковая машина? Насколько удобно были представлены результаты поиска? Был ли искомый результат первым или же сотым? Как много ненужного мусора было найдено наравне с полезной информацией? Найдется ли нужная информация, при обращении к поисковой системе, скажем, через неделю, или через месяц?

В данной курсовой работе мы подробно рассмотрим наиболее популярные поисковые системы.

Актуальность работы обусловлено колоссальным объемом информационных ресурсов в сети Интернет.

Цель курсовой работы - анализ наиболее популярных поисковых систем в сети Интернет.

Задачи: 1. Дать определение поисковой системы; 2. Изучить историю поисковых систем; 3. Понять принципы работы поисковых систем; 4. Анализировать популярные поисковые системы; 5. Разобраться в принципе поиска.

ГЛАВА 1. ПОИСКОВАЯ СИСТЕМА. ОСНОВНЫЕ СВЕДЕНИЯ

1.1 Понятие и история поисковых систем

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

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

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.[1]

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.

1.2 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.[2]

Это значит, что поисковая система должна думать точно также как думает пользователь, когда ищет ту или иную информацию. Обращаясь к поисковой системе, пользователь надеется максимально просто и быстро найти интересующую его информацию. После получения результата, он оценивает работу

системы, руководствуясь несколькими основными параметрами. Разработчики поисковых систем постоянно стараются совершенствовать алгоритмы и принципы поиска, пытаются всячески ускорить работу системы, добавляя новые функции и возможности, чтобы удовлетворить потребности пользователей.

1.3. Анализ и обзор поиска данных

Интернет («Яндекс» и «Google») - сервис статистики запросов на «Яндекс», включающих заданное слово или словосочетание, а также сервис «Google Тренды», анализирующий статистику поиска во всех доменах «Google» и позволяющий узнать, насколько популярны были те или иные поисковые запросы в определенный период времени. Сделанные в статье выводы и практические рекомендации могут использоваться как при проведении маркетинговых исследований, так и в целях обеспечения надлежащего уровня информационной безопасности общества (недопущения появления нелегального контента, экстремистских материалов и т.д.) в сети Интернет. Ключевые слова: контент-анализ, сеть Интернет, поисковые системы, маркетинговые исследования, информационная безопасность.

Анализ содержания текста. Предположим, что данный рисунок является аналогом некоторого текстового документа. Тогда в сценарии информационного поиска традиционные подходы видят данный рисунок как пять отдельных отрезков, расположенных в определенном порядке, а человек обычно понимает данный рисунок (его смысл) как печатную букву S или цифру 5, в зависимости от конкретного контекста. Анализ смысла текста. Рассмотрим другой пример. Допустим, есть некоторый текст «В финальном матче команда А победила команды Б со счетом 3:2». В сценарии информационного поиска, чтобы найти информацию о победителе данного соревнования с помощью традиционных поисковых систем, пользователь должен подбирать подходящие ключевые слова. В данном случае, если пользователь использует запрос «победитель соревнования», то система вернее всего вернет такое сообщение, как «по вашему запросу, ничего не найдено». Однако пользователь сможет найти нужную ему информацию косвенным путем, используя другой запрос, например «Финальный матч», для получения данного текста, а затем определит победителя соревнования самостоятельно. В отличие от традиционных подходов, в семантических поисковых системах пользователь может задавать вопросы на языке, близком к естественному. В рассматриваемой выше ситуации правильный запрос будет выглядеть примерно

так: «Кто победитель соревнования?», при этом система должна будет выполнить анализ смысла текста и смысл запроса для формирования соответствующих ответов. В результате сказанного выше можно дать следующее определение семантического поиска: семантический поиск – это метод информационного поиска, в котором релевантность документа запросу определяется семантически, а не синтаксически. В традиционных подходах поиска релевантность документов и запросов определяется синтаксически, путем вычисления встречаемости ключевых слов в документе, без учета их семантических особенностей. Семантическая релевантность оценивается по близости смыслов текстов, как это делает человек, т.е. семантические поисковые машины выполняют определение и описание смысла текста. Подходы семантического поиска используют именно такие технологии понимания текстов для улучшения качества поиска . Более подробно методы семантического поиска будут пояснены на примерах семантических поисковых систем, описанных ниже.

При сравнении подходов семантического поиска с традиционными подходами поиска по ключевым словам можно отметить, что теоретически они имеют ряд преимуществ над традиционными подходами в смысле повышения релевантности получаемых результатов. Это связано с тем, что релевантными результатами являются документы, удовлетворяющие информационные потребности пользователей и релевантность оценивается по смыслу текстов. Главным недостатком подходов семантического поиска в сравнении с традиционными подходами поиска является тот факт, что алгоритмы обработки смысла текстов зависят от особенностей конкретного анализируемого естественного языка, т.е. требуется создание специальных алгоритмов для разных естественных языков. При этом для каждого естественного языка должны учитываться его синтаксические и семантические особенности, отношения между словами и т.п. В связи с этим реализация подходов семантического поиска в многоязычных системах является очень сложной и трудоемкой работой.

Поиск изображеня происходит следующим образом.

Базы данных изображений могут быть очень большими и содержат сотни тысяч и даже миллионы изображений. В большинстве случаев эти базы проиндексированы только по ключевым словам. Эти ключевые слова вносит в базу оператор, который также распределяет все изображения по категориям. Но изображения могут быть найдены в базе и на основе собственного содержания. Под содержанием мы можем понимать цвета и их распределение, объекты на изображении и их пространственное положение и т.д. В настоящее время алгоритмы сегментации и

распознавания развиты недостаточно хорошо, тем не менее, сейчас уже существует несколько систем (в том числе коммерческих) для поиска изображений на основе их содержания.

К области поиска изображений по содержательным критериям сейчас возрастает интерес, он связан с ограниченностью методов, основанных исключительно на категоризации метаданных. Потенциальные области применения алгоритмов поиска по содержанию:

• поиск изображений в Интернете;

• определение объектов по фотографии;

• каталогизация изображений произведений искусства; • организация работы с архивами фотографических снимков;

• организация каталогов розничной продажи товаров;

• диагностика заболеваний (с помощью сравнения снимков);

• контроль распространения интеллектуальной собственности;

• контроль содержимого массивов изображений.

Человек может сравнивать изображения и выделять на них объекты визуально, на интуитивном уровне. Однако для машины изображение — набор данных, который требует сложной обработки. Есть определенные методы для сравнения изображений, основанные на сопоставлении знаний об изображениях в целом. В общем случае это выглядит следующим образом: для каждой точки изображения вычисляется значение определенной функции, на основании этих значений можно приписать изображению определенную характеристику, тогда задача сравнения изображений сводится к задаче сравнения таких характеристик. Однако эти методы приемлемо работают практически только в идеальных ситуациях. Этому препятствует ряд моментов:

• масштаб;

• расположение на сцене;

• фон и помехи;

• проекция, вращение и угол обзора.

Итак, задачу построения интересующей нас CBIR-системы можно сформулировать следующим образом. Требуется создать систему, позволяющую индексировать (организовывать) некое множество изображений таким образом, чтобы для заданного пользователем образца запрос к системе возвращал подмножество наиболее близких по содержанию изображений, то есть содержащих те же объекты, что и образец, но, возможно, отличающихся:

• масштабом;

• поворотом в плоскости изображения;

• расположением на сцене;

• наличием шумов, заслоняющих объектов;

• другими значениями яркости и контраста.

В большинстве существующих CBIR- систем содержание изображений представляется в виде набора низкоуровневых признаков, таких как цвет, текстура и форма. Недостаток такого подхода заключается в возможности потери большого количества нужной информации в случае, если объект изображен на одинаковых сценах, но в разных ракурсах. В последние годы во многие CBIR-системы начали добавлять модули, использующие ключевые точки и дескрипторы, для того, чтобы устранить этот недостаток.

SURF (Speed Up Robust Feature) — один из наиболее популярных методов выделения ключевых точек и их дескрипторов. Он широко применяется в области систем компьютерного зрения, так как обеспечивает высокую степень инвариантности повороту, масштабу и шумам, а также является одним из самых эффективных и быстрых современных алгоритмов этого класса. В совокупности с эффективным алгоритмом поиска ближайших соседей (Randomized KD-Tree) система представляет современное высокопроизводительное решение для большого круга задач, связанных с поиском изображений. Предлагаемая система изображена на рисунке. Наборы признаков заранее выделяются из изображений, находящихся в базе данных, и помещаются в виде многомерных векторов в базу признаков, при этом происходит индексирование. Для извлечения изображений из базы данных для образца, заданного пользователем, точно таким же способом выделяется набор признаков, который подается на вход алгоритма поиска ближайших точек (векторов) в базе признаков. Далее используется алгоритм отбора изображений, для которых суммарное расстояние от принадлежащих им точек до точек образца

является минимальным.

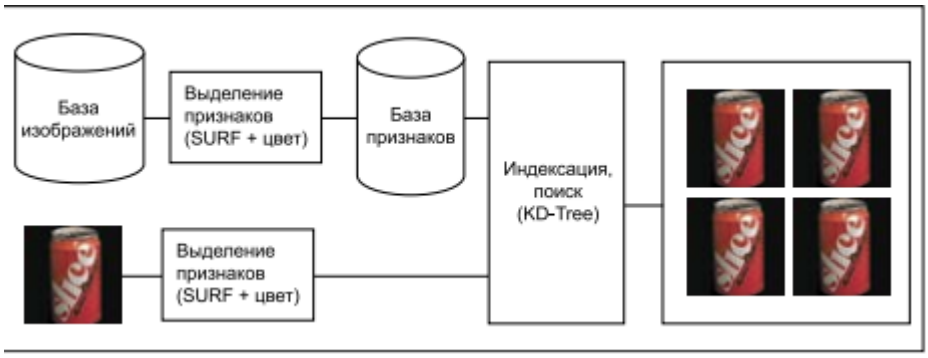

Рис. 1. Основные модули CBIR-системы

Каждую точку описывает вектор, составленный из SURF-дескриптора (64 числа с плавающей точкой одинарной точности) и вектора, определяющего среднее значение цвета в окрестности 5 на 5 соответствующей точки (6 чисел). Для поиска был выбран алгоритм Randomized KD-Tree — наиболее эффективный для конкретного случая. Так как метод SURF работает только в оттенках серого, часть информации об исходном изображении теряется. Для того чтобы исправить этот недостаток, необходимо добавить к каждой ключевой точке описание цветовых характеристик в некоторой ее окрестности. Выберем область 5 на 5 вокруг ключевой точки и вычислим средние значения для каждого RGB-канала среди 25 точек[3].

Для работы с базой изображений желательно иметь некоторый способ поиска изображений, который был бы удобнее и эффективнее, чем непосредственный просмотр всей базы. Большинство компаний выполняет всего два этапа обработки: отбор изображений, для включения в базу и классификация изображений посредством назначения им ключевых слов. Интернет-поисковики обычно получают ключевые слова автоматически из подписей к картинкам. С помощью обычных баз данных изображения можно находить на основе их текстовых атрибутов. При обычном поиске этими атрибутами могут быть категории, имена присутствующих на изображении людей, а также дата создания изображения. Для ускорения поиска содержимое базы может быть проиндексировано по всем этим полям. Тогда для поиска изображений можно будет воспользоваться языком SQL. Например, запрос:

SELECT * FROM IMAGEBD

WHERE CATEGORY=”МЭИ”

мог бы найти и вернуть все изображения из базы на которых изображён МЭИ. Но на самом деле всё не так просто. Такой тип поиска имеет ряд серьёзных ограничений. Назначение ключевых слов человеком является трудоёмкой задачей. Но, что гораздо хуже, эта задача допускает неоднозначное выполнение. Из-за этого некоторые из найденных изображений могут весьма и весьма сильно отличаться от ожиданий пользователя. На рисунке показана выдача google по запросу «МЭИ».

Рис. 2.Результаты поиска по запросу

Приняв, как факт, что использование ключевых слов не обеспечивает достаточной эффективности, мы рассмотрим ряд других методов поиска изображений[4].

Поиск по образцу. Вместо того, чтобы указывать ключевые слова, пользователь мог бы предъявить системе образец изображения, или нарисовать эскиз. Затем наша система поиска должна найти похожие изображения или изображения, содержащие требуемые объекты. Для простоты будем считать, что пользователь представляет системе грубый эскиз ожидаемого изображения и некоторый набор ограничений. Если пользователем предоставлен пустой эскиз, то система должна вернуть все изображения, удовлетворяющие ограничениям. Ограничения же логичней всего задавать в виде ключевых слов и различных логических условий, объединяющих их. В самом общем случае, запрос содержит какое-то изображение, которое сравнивается с изображениями из базы согласно применяемой мере расстояния. Если расстояние равно 0, то считается, что изображение точно соответствует запросу. Значения больше 0 соответствуют различной степени сходства рассматриваемого изображения с запросом. Поисковая система должна возвращать изображения, отсортированные по значению расстояния от эскиза[5].

На рисунке показаны поиска в системе QBIC с применением меры расстояния на основе цветового макета.

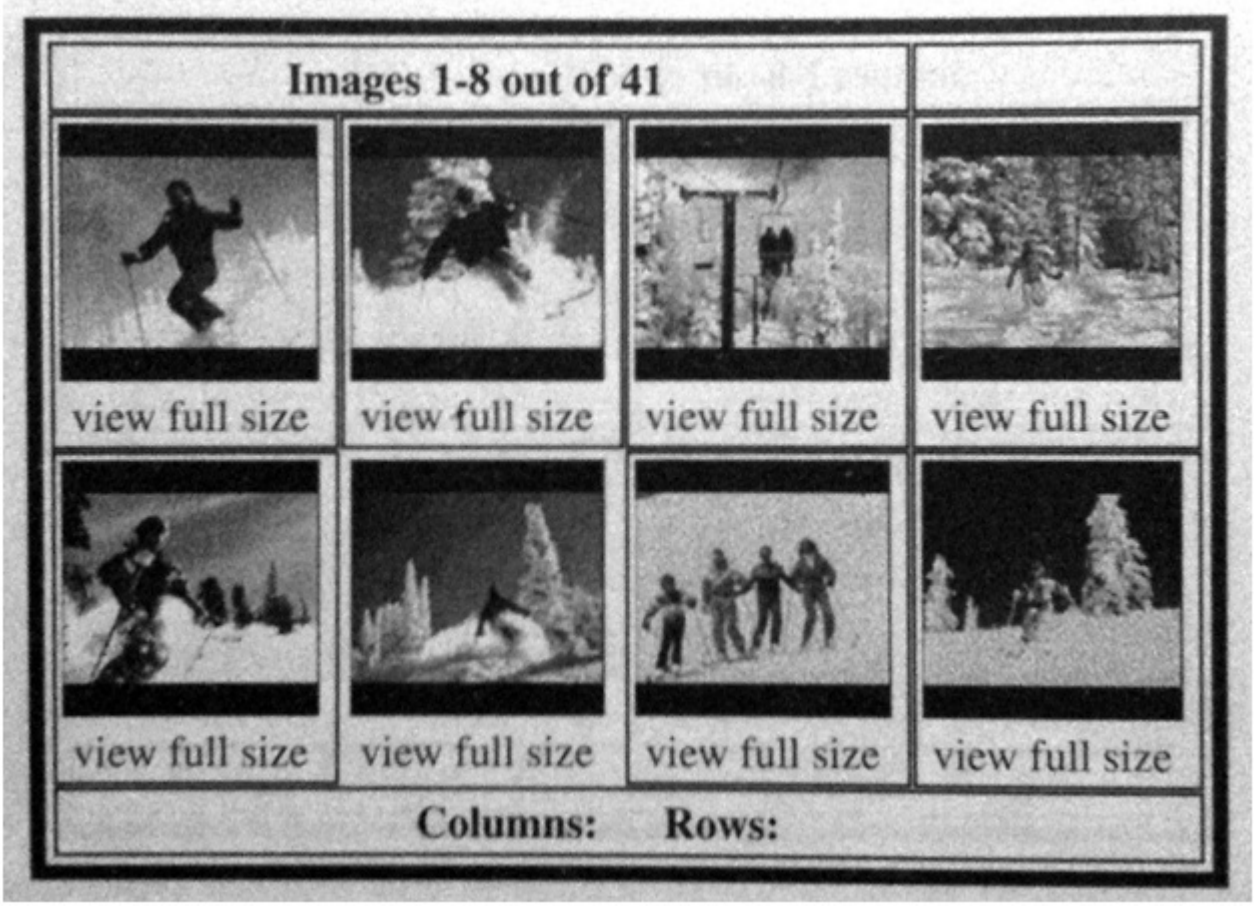

Рис. 3. Поиск в системе QBIC с применением меры расстояния на основе цветового макета[6]

Для определения сходства изображения из базы данных с изображением, указанным в запросе, обычно применяется некоторая мера расстояния или характеристики, с помощью которых можно получить численную оценку сходства изображений. Характеристики сходства изображений можно разделить на четыре основные группы:

1. Цветовое сходство

2. Текстурное сходство

3. Сходство формы

4. Сходство объектов и отношений между объектами[7]

Для простоты рассмотрим только методы цветового сходства. Характеристики цветового сходства часто выбираются очень простыми. Они позволяют сравнить цветовое содержание одного изображения с цветовым содержанием другого

изображения или с параметрами, заданными в запросе. Например, в системе QBIC пользователь может указать процентное соотношение цветов в искомых изображениях. На рисунке показан набор изображений, полученных в результате выполнения запроса с указанием 40% красного, 30% жёлтого и 10% чёрного цвета. Хотя найденные изображения содержат очень похожие цвета, но смысловое содержание этих изображений существенно отличается.

Похожий способ поиска основан на сопоставлении цветовых гистограмм. Меры расстояния на основе цветовой гистограммы должны предусматривать оценку сходства двух различных цветов. Система QBIC определяет расстояние следующим образом:

$$d_{hist}(I,Q) = \big(h(I) - h(Q)\big)^T A \big(h(I) - h(Q)\big)$$

где h(I), h(Q) – гистограммы изображений I,Q, A – матрица сходства. В матрице сходства элементы, значения которых близки к 1, соответствуют похожим цветам, близкие 0 соответствуют сильно различающимся цветам. Ещё одна возможная мера расстоянии основана на цветовом макете. При формировании запроса пользователю обычно предъявляется пустая сетка. Для каждой клетки пользователь может указать цвет из таблицы.

Характеристики сходства на основе цветового макета, в которых используется закрашенная сетка, требуют меры, которая учитывала бы содержание двух закрашенных сеток. Эта мера должна обеспечивать сравнение каждой клетки сетки, указанной в запросе, соответствующей клеткой сетки произвольного изображения из базы данных. Результаты сравнения всех клеток комбинируются для получения значения расстояния между изображениями[8]:

$$d_{color}(I,Q) = \sum_g d\big(C^I(g), C^Q(g)\big)$$

где C^I (g),C^Q (g) — это цвета клетки g в изображениях I,Q соответственно.

Поиск на основе текстурного сходства, а тем более на основе сходства формы гораздо сложнее, но, стоит сказать, что уже сделаны первые шаги в этом направлении. Например, в системе ART MUSEUM хранятся цветные изображения многих картин. Эти цветные изображения подвергаются обработке для получения промежуточного представления. Предварительная обработка состоит из трёх этапов[9]:

1. Уменьшение изображения до заданного размера и удаление шумов с помощью медианного фильтра.

2. Обнаружение границ. Во-первых, с помощью глобального порога, затем с помощью локального порога. В результате получается очищенное контурное изображение.
3. На очищенном контурном изображении удаляются избыточные контуры. Полученное изображение ещё раз очищается от шумов, и мы получаем требуемое абстрактное представление.

Когда пользователь представляет системе эскиз, над ним производятся такие же операции обработки, и мы получаем линейный эскиз. Алгоритм сопоставления имеет корелляционный характер: изображение делится на клетки, и для каждой клетки вычисляется корреляция с аналогичной клеткой изображения из базы данных. Для надёжности эта процедура выполняется несколько раз для разных значений сдвига линейного эскиза. В большинстве случаев этот метод позволяет успешно находить требуемые изображения.

Осталось только дождаться внедрения таких систем в привычные нам интернет-поисковики, и можно будет сказать, что проблема поиска картинок стала не такой уж и проблемой.

ГЛАВА 2. АНАЛИЗ ПОИСКОВЫХ СИСТЕМ В СЕТИ ИНТЕРНЕТ

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.[10]

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).

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.

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.[11]

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

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

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

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

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

Поисковая система большое значение придает показателю последнего изменения информации (Last-Modified). Если сервер не будет передавать эту информацию, то

процесс индексации данного ресурса будет происходить намного реже.

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

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

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

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

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

- Yandex/1.01.001 (compatible; Win16; I) – основной робот, занимающийся непосредственно индексацией сайтов;
- Yandex/1.01.001 (compatible; Win16; P) – робот-индексатор изображений;
- Yandex/1.01.001 (compatible; Win16; H) – робот, который выявляет «зеркала» индексируемых сайтов;
- Yandex/1.02.000 (compatible; Win16; F) – робот-индексатор пиктограмм ресурсов (favicons);
- Yandex/1.03.003 (compatible; Win16; D) – робот, который обращается к страницам, добавленным с помощью формы «Добавить URL»;
- Yandex/1.03.000 (compatible; Win16; M) – задействуется при переходе на страницу посредством ссылки «Найденные слова»;
- YaDirectBot/1.0 (compatible; Win16; I) – этот робот отвечает за индексацию страниц ресурсов, принимающих участие в рекламной сети Яндекс.

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

ресурса для поисковой системы.

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

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

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.[12]

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.

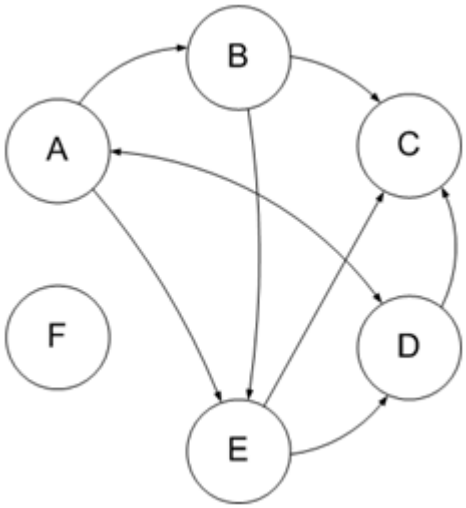

Рис.4. Ссылочный граф

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.

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.

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.

ЗАКЛЮЧЕНИЕ

Поисковые системы уже давно стали неотъемлемой частью Интернета. Поисковые системы сейчас - это огромные и сложные механизмы, представляющие собой не только инструмент поиска информации, но и заманчивые сферы для бизнеса.

Поисковые системы при использовании Интернет играют очень важную роль. В Интернете сосредоточено такое количество информации, что ее поиск уже превращается в отдельную задачу и отнимает очень много времени. Поисковые серверы выдают на запрос тысячи ссылок вместо нескольких страниц, где действительно имеется нужная информация. Пользователи всемирной сети Интернет, осознав преимущества, предоставляемые возможностью анализа пространственных данных, нуждаются в инструменте, позволяющем осуществлять быстрый и удобный поиск и доступ к цифровым снимкам местности и другой пространственной информации, сосредоточенной во многих правительственных, коммерческих и академических организациях.

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

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Аликберов А. П. «Несколько слов о том, как работают роботы поисковых машин». http://www.citforum.ru/internet/search/art_1. shtml
2. Баранов А. 3 правила успеха Интернет-маркетинга. РИОР, 2011. 232 с.
3. Гроховский Л., Хохловский О., Шестаков О., Рзаев Р. SEO для бизнеса. М: Топ Эксперт, 2015. 200 с
4. Гусев, В.С., «Яндекс. Эффективный поиск» - Москва, Санкт - Петербург, Киев.: Диалектика, 2010.
5. Гусев, В.С., «Поиск, Internet» - Москва, Санкт - Петербург, Киев.: Диалектика, 2010.
6. Гусев, В.С., «Google. Эффективный поиск» - Москва, Санкт - Петербург, Киев.: Диалектика, 2010г.
7. Егоров, А.Б., « Поиск в Интернете» - Санкт - Петербург.: НиТ, 2010.

8. Кузьмин А.В. Золотарева Н.Н. Поиск в Интернете - Санкт - Петербург.: Издательство НиТ, 2011.
9. Куприянова, Г.И., «Информационные ресурсы Internet» - М., 2012.
10. Неелова Н., Шпорт К., Моргачева А., Фролкина Е. Загребельный Г. SEMBOOK. Энциклопедия поискового продвижения. Питер, 2014. 520 с.
11. Федотова Л.Н. Анализ содержания – социологический метод изучения средств массовой коммуникации. - М.: Институт социологии РАН, 2001. - 202 с.
12. Экслер, А.Б., «Самоучитель работы в Интернете» - Москва.: NT Press, 2010.
13. Лебедев П. Проблемы и барьеры развития Рунета: экспертные мнения // Социальная реальность. - 2008. - № 7.
14. Доктрина информационной безопасности Российской Федерации. Утверждена Президентом РФ 9 сентября 2000 г., № Пр-1895 // Российская газета. – 2000. – 28 сентября. - № 187.
15. Храмцов П. А. «Поиск и навигация в Internet». http://www.osp.ru/cw/1996/20/31.htm

Интернет-ресурсы

1. www.clx.ru - Описание зарубежных поисковых систем
2. www.seop.ru - Search engine optimization project, рейтинг основных поисков
3. www.baidu.com - Поисковая система Baidu.
4. www.citforum.ru - CIT forum. Поисковые системы в сети Интернет
5. www.gpntb.ru - Перспективы развития поисковых систем
6. Отраслевой доклад «Интернет в России. Состояние, тенденции и перспективы развития. 2014» [Электронный ресурс] / Федеральное агентство по печати и массовым коммуникациям. - Режим доступа http://www.fapmc.ru/mobile/activities/reports/2014/internet-in-russia/main/custom/0/01/file.pdf (дата обращения: 04.06.2015).

1. ? Гусев, В.С., «Поиск, Internet» - Москва, Санкт - Петербург, Киев.: Диалектика, 2010. – С.34. ↑

2. ? Кузьмин А.В. Золотарева Н.Н. Поиск в Интернете - Санкт - Петербург.: Издательство НиТ, 2011. – С.29. ↑

3. ? Варламов А.Д., Шарапов Р.В. Поиск визуально подобных изображений на основе машинного обучения / А.Д. Варламов, Р.В.Шарапов // Электронные

библиотеки: перспективные методы и технологии, электронные коллекции: XIV Всероссийская научная конференция «RCDL’2012». Переславль-Залесский, 15-18 октября 2012 г.: труды конференции - Переславль-Залесский: Изд-во Университет города Переславля, 2012. - С. 152-159. ↑

4. ? Васильева. Н. Выбор шага квантования при построении цветовой гистограммы в задаче поиска изображений / Н. Васильева // Вестник Санкт-Петербургского Университета. - 2009. - № 2. - С. 155-164. ↑

5. ? Гонсалес Р., Вудс Р. Цифровая обработка изображений. Мир цифровой обработки.^ М.: Техносфера, 2005.— 1072 е. — (R. Gonzalez, R. Woods. Digital Image Processing). ↑

6. ? Васильева П., Марков И. СПбГУ на РОМИП' 2008: Синтез цветовых и текстурных признаков при поиске изображений по содержанию / П. Васильева, И. Марков // Труды Российского семинара по Оценке Методов Информационного Поиска РОМИП 2008. - С. 135-144. ↑

7. ? Stahl [Электронный ресурс]/ 2005-2016, – Режим доступа: http://www.stahl-online.de, свободный. ↑

8. ? Васильева Н. Методы поиска изображений по содержанию / Н. Васильева // Программирование. - 2009. - № 3. - С. 1-30. ↑

9. ? Нгуен Ба Нгок, А.Ф. Тузовский Обзор подходов семантического поиска.- Екатеренбург, 310 с. ↑

10. ? Экслер, А.Б., «Самоучитель работы в Интернете» - Москва.: NT Press, 2010. – С.116. ↑

11. ? Гусев, В.С., «Яндекс. Эффективный поиск» - Москва, Санкт - Петербург, Киев.: Диалектика, 2010. – С.95. ↑

12. ? Куприянова, Г.И., «Информационные ресурсы Internet» - М., 2012. – С.261. ↑