МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

Некоммерческое акционерное общество

АЛМАТИНСКИЙ УНИВЕРСИТЕТ ЭНЕРГЕТИКИ И СВЯЗИ

ФАКУЛЬТЕТ «ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ»

Допущен к защите:
зав. кафедрой «Компьютерные технологии»
д. ф.-м. н., профессор_____Куралбаев З.К.
«___»_____20__г.

Магистерская диссертация

«Исследование методов и алгоритмов поиска в Web-поисковых системах»

специальность: 6М070400 – Вычислительная техника и программное обеспечение

Магистрантка_____Жданкова Е.С.
подпись (Ф.И.О.)

Руководитель диссертации_____Сербин В.В.
подпись (Ф.И.О.)

Алматы, 2014 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 4
1 Происхождение поисковых систем в сети интернет 9
1.1 История возникновения поисковых систем 9
1.2 Механизмы поиска по сайтам 10
1.3 Общие принципы работы поисковых систем 12
1.4 Основные поисковые системы и их популярность 15
1.5 Поисковый спам 17
1.5.1 Классификация спама 17
1.5.2 Проблемы поискового спама 20
1.5.3 Методы обнаружения спам контента 21
2 Алгоритмы поиска и фильтрации веб-поисковых систем 23
2.1 Эволюция поисковых систем и факторов ранжирования 23
2.2 Изучение методов и алгоритмов сортировки и ранжирования результатов поиска 25
2.3 Поисковая система Google 27
2.3.1 Возможности алгоритма поиска 27
2.3.2 Алгоритм индексации сайтов 29
2.3.3 Алгоритм ранжирования 30
2.3.3.1 Google PageRank 30
2.3.3.2 Алгоритм Google Колибри 31
2.3.4 Google и борьба с веб-поисковым спамом 34
2.3.4.1 Google Panda (Панда) 35
2.3.4.2 Google Penguin (Пингвин) 36
2.4 Поисковая система Яндекс 37
2.4.1 Поисковая выдача и ранжирование 39
2.4.2 Понятие ссылочного графа 39
2.4.3 Взвешенный индекс цитирования (вИЦ) 41
2.4.4 Алгоритмы фильтрации спам контента 44

2.5 Поисковая система Bing 44
3 Анализ и сравнение качества поиска существующих веб-поисковых систем 46
3.1 Эксперименты с поисковыми запросами 46
3.1.1 Анализ навигационного поиска 48
3.1.2 Анализ скорости индексации сайтов 49
3.1.3 Анализ полноты индекса 50
3.1.4 Анализ скорости выдачи результатов 51
3.1.5 Анализ раздражающих факторов 52
3.1.5.1 Анализ спама в поисковой выдаче 53
3.1.5.2 Анализ рекламной нагрузки в поисковой выдачи 54
3.1.6 Анализ вирусной угрозы в поисковой выдаче 55
3.1.7 Анализ качества снипетов 56
3.1.8 Анализ качества подсказок 57
3.1.9 Анализ устойчивости к опечаткам 57
3.1.10 Анализ запросов-синонимов 59
3.1.11 Анализ различия форм слов 60
3.1.12 Анализ SEO-прессинга 61
3.2 Результаты экспериментов 62
4 Описание работы программного продукта 63
4.1 Работа с поисковой системой 63
4.2 Система исправления и распознавания запросов 69
ЗАКЛЮЧЕНИЕ 75
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 77
ПРИЛОЖЕНИЕ А 79

АҢДАТПА

Диссертациялық жұмыс релевант құжаттың ізденісінің барлық нәзіктіктерін ара ауда интернет қарайды, от ізденістің жүйесінің алгоритмының эволюциясінің ерте кезеңдерінен, қайда ранжирования факторының шағын саны есептелді, дейін күрделі алгоритм пайдалан қазірде.

Айқында алға шық ізденістің жүйесі арқылы дүркін-дүркін эксперименттерді SEO облысында болды, ал да таныс- және талқыла ауда арада бүтіндікте ізденістің спамының және ақпараттың ізденісімен тоқулы мәселелер. Бас негіз зертте нәтижелерінің бол ұсын және әзірле мета-поисковая эксперименттің көсемдерін ша ақпараттың ізденісінің интернетке өрмекке біріктіру қойған жүйесінің.

АННОТАЦИЯ

Диссертационная работа рассматривает все тонкости поиска релевантных документов в сети интернет, от ранних стадий эволюций алгоритмов поисковых систем, где учитывалось малое количество факторов ранжирования, до сложных алгоритмов, используемых на сегодняшний день.

Была выявлена лидирующая поисковая система через многократные эксперименты в области SEO, а также изучить и обсудить проблемы, связанные с поисковым спамом и поиском информации в сети в целом. На основе результатов исследования была предложена и разработана мета-поисковая система, которая позволила объединить лидеров экспериментов по поиску информации в интернет паутине.

ABSTRACT

The thesis is considering all the intricacies of finding relevant documents on the Internet from the early stages of evolution of search engine algorithms, which take into account the small number of ranking factors to complex algorithms used today.

Was found leading search engine across multiple experiments in SEO, and to explore and discuss issues related to the search spam and searching for information on the network as a whole. Based on the results of the study was proposed and developed a meta- search engine, which brought together leaders of experiments to search information in the Internet web.

ВВЕДЕНИЕ

Современный этап развития цивилизации характеризуется переходом наиболее развитой части человечества от индустриального общества к информационному. Одним из наиболее ярких явлений этого процесса является возникновение и развития глобальной информационной компьютерной сети.

Проблема поиска и сбора информации - одна из важнейших проблем информационно поисковых систем. Конечно, нельзя сравнивать в этом отношении, скажем, средние века, когда поиск информации был проблемой потому, что этой информации было мало, и требовались усилия только для того, чтобы найти хоть что-то по более или менее значительному интересующему вопросу. Так, сначала появилась возможность пойти в библиотеку и, потратив там время на выбор нужной книги по каталогу, найти необходимую информацию. Но каталоги не решают полностью проблем поиска информации даже в рамках одной библиотеки, так как в каталожную запись входит относительно мало информации: заголовок, автор, место издания. Проблема поиска информации приобрела новый характер в 20-м столетии, с началом развития века информационных технологий. Теперь она заключается не в том, что информации мало, и поэтому ее трудно найти, а в том, что ее теперь наоборот становится все больше и больше, и от этого найти ответ на интересующий вопрос может оказаться тоже довольно сложной задачей. Проблема поиска информации значительно усложняется при использовании виртуальных источников. Здесь используется технология онлайновых каталогов, в результате применения которой пользователь имеет возможность выполнять поиск в каталогах сразу нескольких библиотек, чем, на самом деле, еще больше усложняет себе задачу, но, с другой стороны, увеличивает шансы решить ее.

На современном этапе все информационное пространство, в котором мы живем, все больше погружается в Internet. Internet становится основной формой существования информации, не отменив традиционных, такие как журналы, радио, телевидение, телефон, всевозможные справочные службы.

И основным инструментом доступа к информации в интернете являются поисковые системы, которые принимая запрос пользователя, возвращают упорядоченный список web-страниц, содержащих необходимую для пользователя информацию. Каждая поисковая система имеет свой собственный очень сложный и постоянно совершенствующийся алгоритм ранжирования, который является ее коммерческой тайной. Очевидно, что чем выше находится сайт в рейтинге поисковой системы по определенным запросам, тем больше посетителей будет у ресурса и, конечно, любой владелец сайта стремится, чтобы его ресурс занимал как можно более высокие позиции в выдаче поисковиков. Работа над содержимым сайта с целью повышения его рейтинга в результатах поискового запроса называется

его продвижением. Специалисты по продвижению сайтов в процессе работы сталкиваются с тремя основными проблемами.

Во-первых, современные алгоритмы ранжирования зависят не только от наполнения веб-станиц, но и от поискового запроса, поэтому зачастую методика продвижения, сработавшая с одним ресурсом, может не дать результата с другим. Во-вторых, так как процесс индексирования сайтов поисковой системой осуществляется с некоторой периодичностью, то до момента очередной индексации специалист не может видеть, как его усилия по продвижению повлияли на позицию сайта. В-третьих, среди огромного количества признаков, характеризующих пару запрос-документ, необходимо выбрать такие, изменение которых способно повлиять на позицию сайта максимальным образом.

Актуальность работы. На ранних стадиях эволюции в алгоритме web-поисковых системах учитывалось малое количество факторов, влияющих на ранжирование в выдаче результатов поиска, поэтому, зная базовые принципы работы поисковых систем, можно было достаточно легко манипулировать результатами, что и делали многие компании, которые занимались продвижением Интернет-ресурсов. Для того чтобы поддерживать качество поиска, самого важного аспекта, поисковые системы были вынуждены усложнять свои алгоритмы - количество учитываемых факторов возросли в сотни, и даже тысячи раз, стали все чаще появляться различные алгоритмы и фильтры.

При дальнейшей эволюции число таких факторов росло в геометрической прогрессии, постоянно увеличивался уровень конкуренции между поисковыми системами, выводя на поисковый рынок только тех игроков, которые могли бороться и предоставлять пользователям интернета релевантные результаты вне зависимости от умения seo-специалистов продвигать сайты.

Ресурсы Интернета давно перестали быть просто игрушкой, превратившись в незаменимый инструмент для повседневной работы людей многих профессий. Быстрый рост информации в сети сделали его океаном разнообразнейших данных, важность которых растет пропорционально их объему. По оценке специалистов объем информации, передаваемой по каналам всемирной паутины, удваивается каждые полгода. Ежедневно в сети появляются сотни тысяч новых документов, и естественно, что без систем поиска они в подавляющем своем большинстве остались бы не востребованными и не находимыми. Возникла необходимость создания таких средств, которые позволили бы легко ориентироваться в информационных ресурсах глобальных сетей, быстро и надежно находить нужные сведения. В интернете появились специальные поисковые средства.

Еще несколько лет назад бытовало такое мнение: в Интернете есть все, но найти там ничего невозможно. Однако с появлением и быстрым развитием поисковых каталогов, поисковых машин, и всевозможных поисковых

программ ситуация изменилась, и теперь в Сети срочно понадобившуюся информацию иногда можно найти быстрее, чем в книге, лежащей на столе.

Цель работы. Исследование существующих алгоритмов поиска в современных поисковых системах по показателям качества для выявления основных факторов, влияющих на ранжирование сайтов в результатах поиска.

В качестве исследуемых поисковых систем, для которых производилась оптимизация, были выбраны наиболее популярные и прогрессивные Яндекс (www.yandex.ru) и Google (www.google.ru).

Исследование остальных известных поисковых систем Рунет неактуально: Aport (www.aport.ru) – развитие прекращено в середине 2000-х годов, Rambler (www.rambler.ru) – развитие также остановлено, рассматривается использование поиска от Google. Портал Мэйл.ру (www.mail.ru) является в первую очередь развлекательным порталом и почтовым сервисом, поисковая система которого вторична.

Для достижения цели решены следующие задачи:

- проведен анализ современных поисковых систем, используемых в российском, казахстанском и мировом сегменте сети Интернет;
- разработана система факторов, участвующих в формулах ранжирования поисковых систем;
- проведен обзор алгоритмов поисковых систем.
- рассмотрены теоретические подходы к обоснованию проблемы поисковой оптимизации и ранжирования web-сайтов;
- уяснена сущность поисковой оптимизации и ранжирования как предметов исследования, охарактеризованы факторы ранжирования;
- исследована специфика работы современных поисковых машин.

Научная новизна полученных результатов заключается в разработке мульти поисковой системы сайтов, на основе исследования алгоритмов поиска, фильтрации, полноты индекса, эффективности ранжирования результатов, а также других показателей качества, влияющих на поиск информации в веб поисковых системах.

Практическая ценность работы заключается в возможности использования полученных научно-технических результатов при эксплуатации, исследовании, требующих относительного сравнения альтернативных алгоритмов и методов.

Достоверность и обоснованность результатов диссертации.

Положения и выводы, формулируемые в диссертации, доказываются с применением исследования алгоритмов поиска в веб-поисковых системах. Эффективность предложенных методов подтверждается посредством компьютерных экспериментов.

Методы исследований.

Для решения поставленных задач использованы методы теории множеств, теории систем массового обслуживания, сравнительного анализа, объектно-ориентированного анализа, рассчитана упрощенная формула для

определения релевантности страницы сайта, с помощью схем, графиков, и наблюдений.

Апробация работы. Диссертация явилась результатом многочисленных исследований, проведенных автором в области веб-поиска, алгоритмов, методов фильтрации и ранжирования интернет документов в глобальной сети интернет.

Выводы автора в области глобального поиска информации были полезны для разработки собственной мета-поисковой системы.

1 Происхождение поисковых систем в сети интернет

1.1 История возникновения поисковых систем

Одним из первых инструментов поиска в интернете (до Всемирной паутины) был Archie.

Первой поисковой системой для Всемирной паутины был «Wandex», уже не существующий индекс, который создавал «World Wide Web Wanderer» — бот, разработанный Мэтью Грэем (англ. Matthew Gray) из Массачусетского технологического института в 1993. Также в 1993 году появилась поисковая система «Aliweb», работающая до сих пор. Первой полнотекстовой (т. н. «crawler-based», то есть индексирующей ресурсы при помощи робота) поисковой системой стала «WebCrawler», запущенная в 1994. В отличие от своих предшественников, она позволяла пользователям искать по любым ключевым словам на любой веб-странице — с тех пор это стало стандартом во всех основных поисковых системах. Кроме того, это был первый поисковик, о котором было известно в широких кругах. В 1994 был запущен «Lycos», разработанный в университете Карнеги Мелона.

Одним из первых способов организации доступа к информационным ресурсам сети стало создание каталогов сайтов, в которых ссылки на ресурсы группировались согласно тематике. Первым таким проектом стал сайт Yahoo, открывшийся в апреле 1994 года. После того, как число сайтов в каталоге Yahoo значительно увеличилось, была добавлена возможность поиска информации по каталогу. Это, конечно же, не было поисковой системой в полном смысле, так как область поиска была ограничена только ресурсами, присутствующими в каталоге, а не всеми ресурсами сети Интернет.

Каталоги ссылок широко использовались ранее, но практически утратили свою популярность в настоящее время. Причина этого очень проста – даже современные каталоги, содержащие огромное количество ресурсов, представляют информацию лишь об очень малой части сети Интернет.

Самый большой каталог сети DMOZ (или Open Directory Project) содержит информацию о 5 миллионах ресурсов, в то время как база поисковой системы Google состоит из более чем 8 миллиардов документов.
Первой полноценной поисковой системой стал проект WebCrawler появившийся в 1994 году.

В 1995 году появились поисковые системы Lycos и AltaVista. Последняя долгие годы была лидером в области поиска информации в Интернет.

В 1997 году Сергей Брин и Лари Пейдж создали Google самую популярную на сегодняшний момент поисковую систему в мире.

23 сентября 1997 года была официально анонсирована поисковая система Yandex, самая популярная в русскоязычной части Интернет [1, с.4].

В настоящее время существует 3 основных международных поисковых системы – Google, Yandex, Yahoo и Bing (MSN Search), имеющих собственные базы и алгоритмы поиска. Большинство остальных поисковых

систем (коих можно насчитать очень много) использует в том или ином виде результаты 3 перечисленных. Например, поиск AOL (search.aol.com) использует базу Google, а AltaVista, Lycos и AllTheWeb – базу Yahoo, который в последнее время пытается также использовать выдачу Bing.

1.2 Механизмы поиска по сайтам

Интернет состоит из миллионов сайтов и содержит экзабайты информации. Чтобы люди могли узнать о существовании этой информации и воспользоваться ей, существуют web-поисковые системы. Они реализуют право человека на доступ к информации — любой информации, которая нужна в данный момент. Поисковая система — это техническое средство, с помощью которого пользователь интернета может найти данные, уже размещенные в сети.

Обычно запросу соответствуют тысячи, а иногда и миллионы веб-страниц. Алгоритмы – это автоматические процессы и формулы, которые преобразуют запросы пользователя в ответы и адреса интернет ресурсов где находится необходимая информация.

Средства поиска и структурирования, иногда называемые поисковыми механизмами, используются для того, чтобы помочь людям найти информацию, в которой они нуждаются. Средства поиска типа агентов, пауков, кроулеров и роботов используются для сбора информации о документах, находящихся в Сети Интернет. Это специальные программы, которые занимаются поиском страниц в Сети, извлекают гипертекстовые ссылки на этих страницах и автоматически индексируют информацию, которую они находят для построения базы данных. Каждый поисковый механизм имеет собственный набор правил, определяющих, как собирать документы. Некоторые следуют за каждой ссылкой на каждой найденной странице и затем, в свою очередь, исследуют каждую ссылку на каждой из новых страниц, и так далее. Некоторые игнорируют ссылки, которые ведут к графическим и звуковым файлам, файлам мультипликации; другие игнорируют ссылки к ресурсам типа баз данных WAIS; другие проинструктированы, что нужно просматривать прежде всего наиболее популярные страницы.

Агенты - самые "интеллектуальные" из поисковых средств. Они могут делать больше, чем просто искать: они могут выполнять даже транзакции от Вашего имени. Уже сейчас они могут искать сайты специфической тематики и возвращать списки сайтов, отсортированных по их посещаемости. Агенты могут обрабатывать содержание документов, находить и индексировать другие виды ресурсов, не только страницы. Они могут также быть запрограммированы для извлечения информации из уже существующих баз данных. Независимо от информации, которую агенты индексируют, они передают ее обратно базе данных поискового механизма.

Общий поиск информации в Сети осуществляют программы, известные как пауки. Пауки сообщают о содержании найденного документа, индексируют его и извлекают итоговую информацию. Также они просматривают заголовки, некоторые ссылки и посылают проиндексированную информацию базе данных поискового механизма.

Кроулеры просматривают заголовки и возвращают только первую ссылку.

Роботы могут быть запрограммированы так, чтобы переходить по различным ссылкам различной глубины вложенности, выполнять индексацию и даже проверять ссылки в документе. Из-за их природы они могут застревать в циклах, поэтому, проходя по ссылкам, им нужны значительные ресурсы Сети. Однако, имеются методы, предназначенные для того, чтобы запретить роботам поиск по сайтам, владельцы которых не желают, чтобы они были проиндексированы.

Агенты извлекают и индексируют различные виды информации. Некоторые, например, индексируют каждое отдельное слово во встречающемся документе, в то время как другие индексируют только наиболее важных 100 слов в каждом, индексируют размер документа и число слов в нем, название, заголовки и подзаголовки и так далее. Вид построенного индекса определяет, какой поиск может быть сделан поисковым механизмом и как полученная информация будет интерпретирована [2, с.15].

Агенты могут также перемещаться по Интернет и находить информацию, после чего помещать ее в базу данных поискового механизма. Администраторы поисковых систем могут определить, какие сайты или типы сайтов агенты должны посетить и проиндексировать. Проиндексированная информация отсылается базе данных поискового механизма так же, как было описано выше.

Люди могут помещать информацию прямо в индекс, заполняя особую форму для того раздела, в который они хотели бы поместить свою информацию. Эти данные передаются базе данных.

Когда кто-либо хочет найти информацию, доступную в Интернет, он посещает страницу поисковой системы и заполняет форму, детализирующую информацию, которая ему необходима. Здесь могут использоваться ключевые слова, даты и другие критерии. Критерии в форме поиска должны соответствовать критериям, используемым агентами при индексации информации, которую они нашли при перемещении по Сети.

База данных отыскивает предмет запроса, основанный на информации, указанной в заполненной форме, и выводит соответствующие документы, подготовленные базой данных. Чтобы определить порядок, в котором список документов будет показан, база данных применяет алгоритм ранжирования.

В идеальном случае, документы, наиболее релевантные пользовательскому запросу будут помещены первыми в списке. Различные

поисковые системы используют различные алгоритмы ранжирования, однако основные принципы определения релевантности следующие:

– количество слов запроса в текстовом содержимом документа (т.е. в html-коде);

– тэги, в которых эти слова располагаются;

– местоположение искомых слов в документе;

– удельный вес слов, относительно которых определяется релевантность, в общем количестве слов документа.

База данных выводит ранжированный подобным образом список документов с HTML и возвращает его человеку, сделавшему запрос. Различные поисковые механизмы также выбирают различные способы показа полученного списка - некоторые показывают только ссылки; другие выводят ссылки с первыми несколькими предложениями, содержащимися в документе или заголовок документа вместе с ссылкой [2, с.19].

На ранних стадиях эволюции в алгоритме web-поисковых системах учитывалось малое количество факторов, влияющих на ранжирование в выдаче результатов поиска, поэтому, зная базовые принципы работы поисковых систем, можно было достаточно легко манипулировать результатами, что и делали многие компании, которые занимались продвижением интернет-ресурсов. Для того чтобы поддерживать качество поиска, самого важного аспекта, поисковые системы были вынуждены усложнять свои алгоритмы - количество учитываемых факторов возросли в сотни, и даже тысячи раз, стали все чаще появляться различные алгоритмы и фильтры.

При дальнейшей эволюции число таких факторов росло в геометрической прогрессии, постоянно увеличивался уровень конкуренции между поисковыми системами, выводя на поисковый рынок только тех игроков, которые могли бороться и предоставлять пользователям интернета релевантные результаты вне зависимости от умения seo-специалистов продвигать сайты.

1.3 Общие принципы работы поисковых систем

Чтобы рассмотреть в сети процесс поиска информации, необходимо остановится на следующих этапах:

– собрать информацию в сети Интернет с различных сайтов;
– исследование по запросу;
– ранжирование результатов;
– индексация сайтов.

Несмотря на то, что каждая поисковая система имеет свои особенности в алгоритмах поиска и ранжирования результатов, принципы работы всех поисковых систем общие.

Рассмотрим каждый из этапов отдельно.

Сбор данных осуществляется специальной программой, которая есть у каждой поисковой системы – поисковый робот. Он предназначен для перебора страниц Интернета с целью занесения информации о них в базу данных поисковой системы. Как только вы запустили свой сайт и дали роботу какой-нибудь поисковой системы понять, что появился новый ресурс (с помощью внешних ссылок на ваш сайт, добавления его сервисы добавления addurl (англ., добавить сайт) или используя другие методы), робот приходит к вам, начинает ходить по страницам и собирать с них данные (это могут быть текстовый контент, картинки, видео и др. файлы).

Этот процесс называется сбором данных (англ. crawling) и он может происходить не только при запуске сайта. Робот составляет для сайта расписание, когда он должен на него зайти в следующий раз, проверить старую информацию и добавить новые страницы, если таковые имеются.

Важно, чтобы общение вашего сайта с роботом было приятным для обеих сторон. В ваших интересах, чтобы робот долго не задерживался на сайте, чтобы не грузить лишний раз сервер, и в то же время необходимо, чтобы он правильно собрал все данные со всех нужных страниц. В интересах робота также сделать сбор быстрым, чтобы приступить к обработке следующего сайта в таблице расписаний. Для этого вам необходимо убедиться, что сайт доступен, что нет проблем с навигацией по сайту (flash и javascript меню роботы ещё плохо распознают), что отсутствуют «битые» страницы (выдющие 404 ошибку), не заставлять робота ходить по страницам, которые доступны только зарегистрированным пользователям и так далее. Также следует помнить, что для web-пауков существует ограничение на глубину проникновения (уровень вложенности страниц) и максимальный размер сканируемого текста (обычно 256кб).

Управлять доступом на разные ресурсы для поискового робота можно с помощью файла robots.txt, в котором содержатся правила и ограничения доступа к содержимому сайта на http-сервере. Карта сайта sitemap.xml (XML-файл с информацией для поисковых систем о страницах веб-сайта, которые подлежат индексации) также может помочь роботу, если по каким-либо причинам навигация по сайту ему затруднена.

Робот может ходить по вашему сайту долгое время, однако это не значит, что он сразу появится в поисковой выдаче. Страницам сайта необходимо пройти такой этап, как **индексация** – составление для каждой страницы обратного (инвертированного) файла индекса. Индекс – все страницы, которые известны поисковой системе, то есть которые прошли индексацию. Индекс служит для того, чтобы быстро по нему производить поиск и состоит обычно из списка слов из текста и информации о них (позиции в тексте, вес и др.). Инвертированным файлом индекса называется такой индекс поисковой системы, в котором перечислены слова коллекции документов, а для каждого слова перечислены все места, в которых оно встретилось.

После того, как прошла индексация сайта или отдельных страниц, они появляются в основной выдаче поисковой системы и их можно найти по ключевым словам, присутствующим в тексте. Процесс индексации обычно происходит довольно быстро после того, как робот скачает информацию с вашего сайта.

При поиске, первым делом, анализируется запрос, введенный пользователем (происходит препроцессинг запроса), в результате которого вычисляются веса для каждого из слов. Весом слова называется отношение частоты использования этого слова к общему количеству слов, выраженное в процентах. [3, с.25]

Далее, поиск производится по инвертированным индексам, находятся все документы в коллекции (базе данных поисковой системы), которые наиболее подходят под данный запрос. Другими словами, вычисляется схожесть документа запросу примерно по следующей формуле:

$$similarity(Q,D)=SUM(w(qk)\cdot w(dk)), \quad (1)$$

где similarity(Q,D) – схожесть запроса Q документу D;

w(qk) – вес k-го слова в запросе;

w(dk) – вес k-го слова в документе.

Документы, наиболее схожие с запросом, попадают в результаты поиска.

После того, как наиболее схожие документы были отобраны из основной коллекции, они должны ранжироваться, чтобы в верхних результатах отражались наиболее полезные для пользователя ресурсы. Для этого используется специальная формула ранжирования, которая для разных поисковых систем имеет разный вид, однако для всех из них основными факторами ранжирования являются:

- вес страницы (вИЦ, PageRank);
- авторитетность домена;
- релевантность текста запросу;
- релевантность текстов внешних ссылок запросу, а также множество других факторов ранжирования.

Поскольку точная формула не раскрывается поисковыми системами, существует упрощенная формула:

$$Ra(x)=(m\cdot Ta(x)+p\cdot La(x))\cdot F(PRa), \quad (2)$$

где Ra(x) – итоговое соответствие документа a запросу x;

Ta(x) – релевантность текста (кода) документа a запросу x;

La(x) – релевантность текста ссылок с других документов на документ a запросу x;

PRa – показатель авторитетности страницы a, константа относительно x;

F(PRa) – монотонно неубывающая функция, причем F(0)=1, можно допустить, что F(PRa)=1+q·PRa;

m,p,q – некие коэффициенты.

При ранжировании документов используются, как внутренние факторы, так и внешние. А также можно их разделить на зависимые от запроса факторы (релевантность текста документа или ссылок) и независимые от запроса. Конечно же, эта формула даёт очень общее представление об алгоритмах ранжирования документов в результатах поиска [4, с.3].

1.4 Основные поисковые системы и их популярность

В мире всего три национальных поисковика, которые выстояли перед натиском американского поискового гиганта - Google: в Чехии самой популярной поисковой системой пока остаётся Seznam (seznam.cz), в Китае с огромным преимуществом в поисковом секторе лидирует Baidu.com, а в России самый популярный поисковик - это Яндекс (yandex.ru). **По данным LiveInternet.ru** (www.liveinternet.ru/stat/kz/), за 2013 года (в среднем за 3 месяца) поисковые предпочтения казахстанских пользователей выглядят так (в соответствии с рисунком 1):

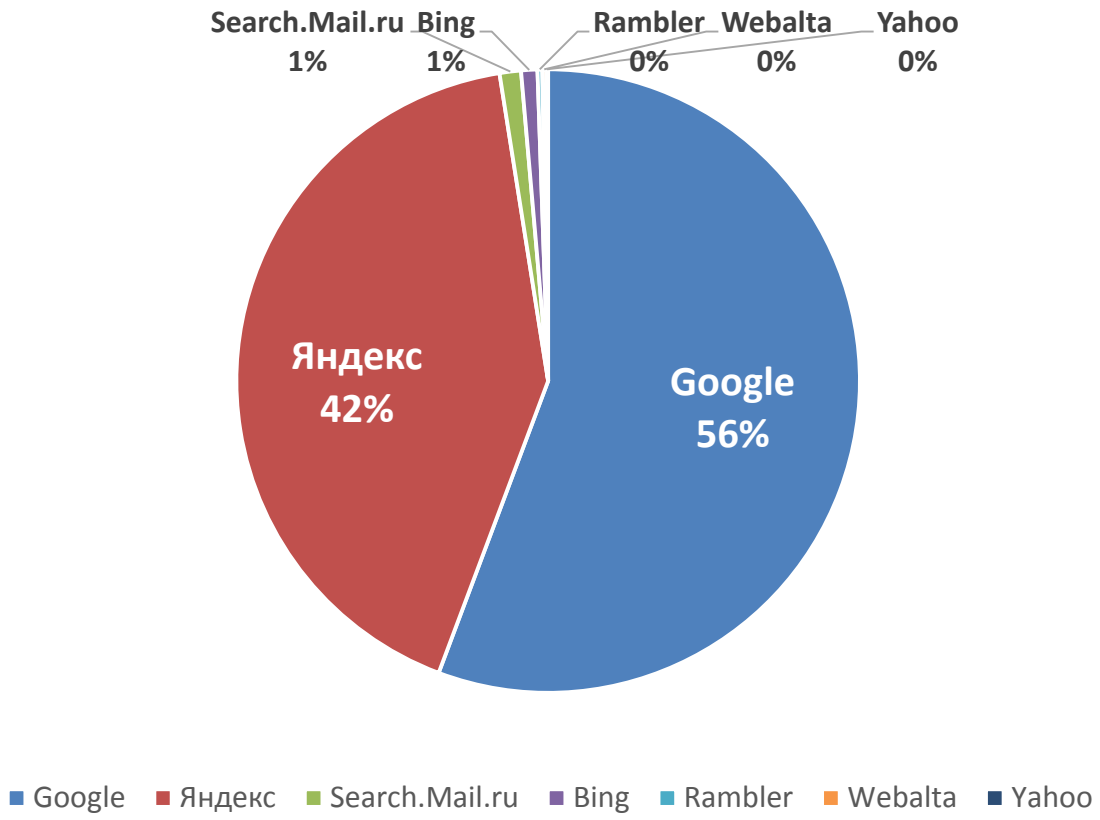

Рисунок 1- Рейтинг поисковых систем казахстанских пользователей

Рамблер с 2011 года перестал быть самостоятельным поисковиком и стал использовать поисковые технологии Яндекса. Mail.ru до недавнего времени использовал технологии Гугла, но 1 июля 2013 было официально объявлено, что Mail.ru отныне будет обрабатывать все запросы с помощью собственной поисковой системы. Однако стоит отметить, что на сегодняшний день выдача поисковика Mail.ru соответствует выдаче Google без фильтров, тем самым можно предположить, что до сих пор своя выдача не используется. Таким образом, на российском поисковом рынке как в целом и в казахстанском существует 2 серьёзных игрока: Яндекс, который вместе с Рамблером ищут информацию 50% пользователей рунета и казнета, а также Гугл с Mail.ru.

Для российских пользователей в соответствии с рисунком 2:

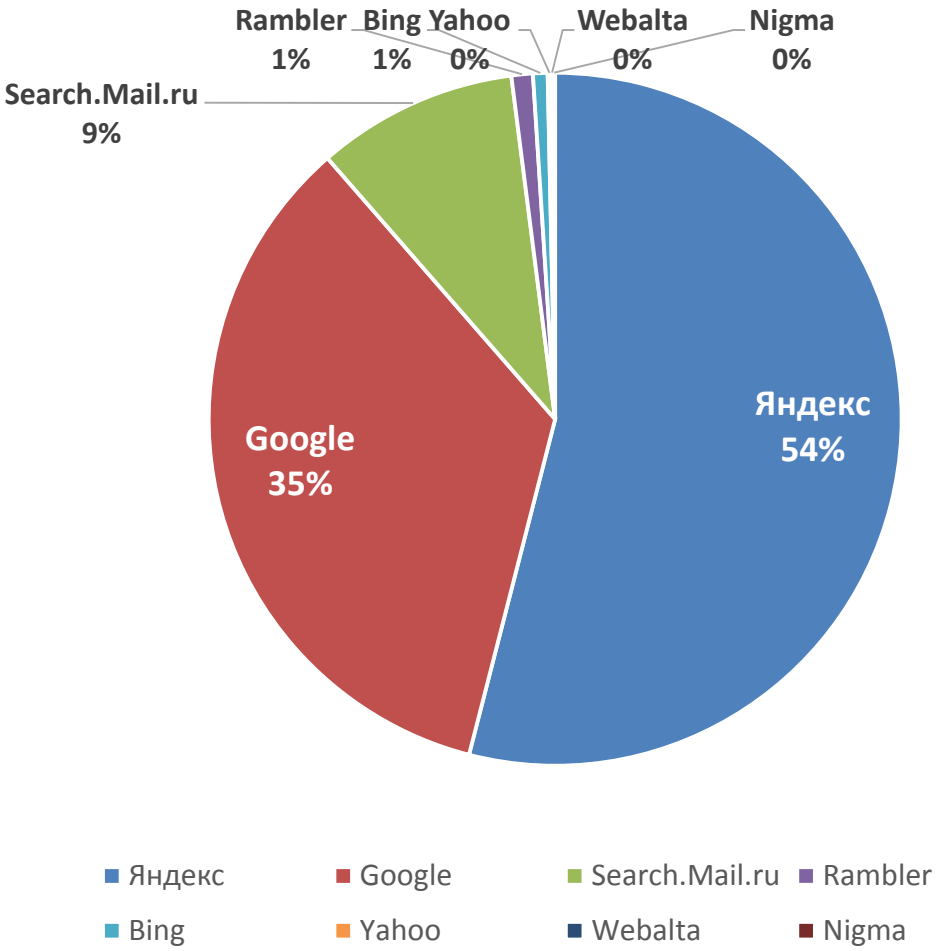

Рисунок 2- Рейтинг поисковых систем для российских пользователей

Русскоязычные в мире (согласно крупнейшему статистическому сайту Alexa.com) в соответствии с рисунком 3:

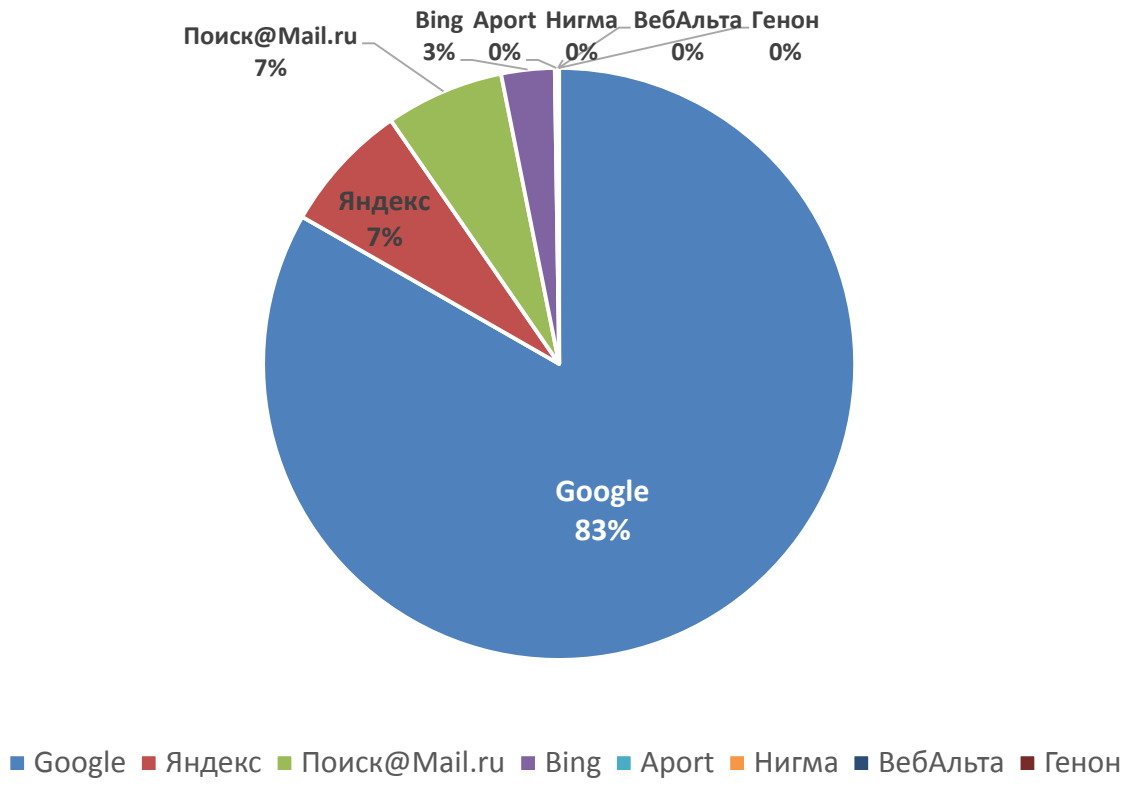

Рисунок 3- Рейтинг поисковых систем

1.5 Поисковый спам

Поисковый спам (**спамдексинг** или **веб-спам**) — сайты и страницы в Интернете, созданные с целью манипуляции результатами поиска в поисковых машинах — в конечном счёте, для обмана пользователя. Это наиболее распространенный инструмент, так называемой, “черной” оптимизации. Сегодня поисковый спам применяется в том или ином виде практически повсеместно, поскольку, несмотря на запрет его использования для поискового продвижения, при правильном подходе он может достаточно существенно, кратковременно посодействовать раскрутке web-сайта и росту посещаемости [5, с.6].

Более того, очень часто начинающие web-мастера используют те или иные методы поискового спама, сами о том не подозревая. Например, покупка ссылок – это также один из методов поискового спама, однако далеко не всегда она приводит к наложению на ресурс фильтров поисковиков.

1.5.1 Классификация спама

Контекстный (или текстовый) спам. Суть данного метода поискового спама заключается в перенасыщении web-документов сайта ключевыми

словами. Наиболее распространенный способ применения контекстного спама – это перенасыщение основного текста страниц ресурса ключевой фразой или словом. Однако существуют и более изощренные способы его применения, среди которых использование для переоптимизированного текста шрифта, цвет которого совпадает с цветом его фона, использование для ключевых слов шрифта размером в 1 пиксель, перенасыщение ключами тегов web-документов и так далее рассмотрим в соответствии с рисунком 4.

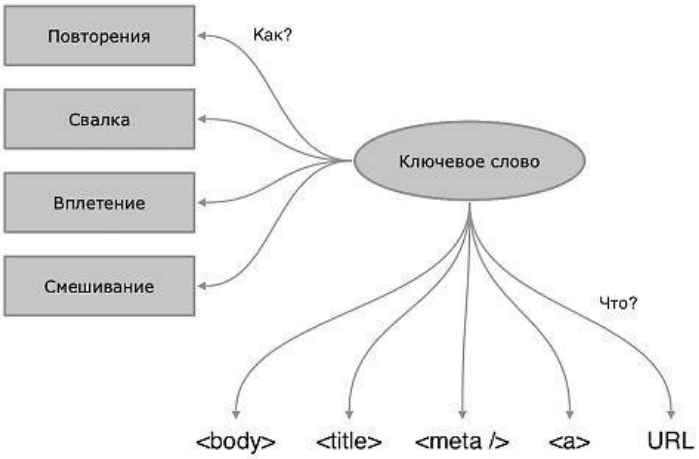

Рисунок 4 – Контекстный спам

Текстовый спам это, прежде всего, повторения ключевого слова (переоптимизация) в различных информационных составляющих сайта (теле документа, заголовке, мета-тегах, ссылках и ЧПУ). Но также сюда относятся:

- **Свалка**. Когда страница завалена множеством разных ключевых слов, чтобы повысить вероятность захода на неё хотя бы по ряду низкочастотных запросов.
- **Вплетение**. Когда берутся цитаты с других источников (например, новостных) и в них вплетается нужная ключевая фраза (или несколько).
- **Смешивание**. Когда контент агрегируется из различных источников в надежде синтезировать что-то уникальное.

Сегодня применять контекстный спам нецелесообразно, поскольку поисковые алгоритмы достаточно эффективно выявляют его использование, за что сайты попадают под фильтры или же пессимизируются поисковыми алгоритмами.

Ссылочный спам. Это наиболее распространенный метод реализации поискового спама. Его суть заключается в неестественных манипуляциях с внешней и внутренней ссылочной массой интернет-ресурса рассмотрим в соответствии с рисунком 5.

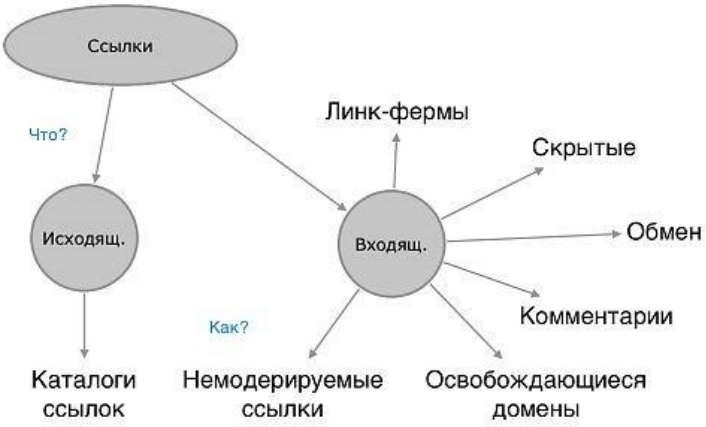

Рисунок 5 – Ссылочный спам

Здесь предлагается анализировать не только входящие ссылки, но и исходящие, где к спаму относят каталоги ссылок (часто нетематические). Что касается входящих, то это:

- **Линк-фермы**. Специально созданные сайты для размещения на них ссылок. Часто это сеть сателлитов, сплогов или специальные ссылочные системы.
- **Скрытые ссылки**. Ссылки на сайте-доноре либо скрыты в коде, либо имеют цвет фона, чтобы их не видели посетители (или сам владелец сайта, если ресурс был взломан).
- **Обмен ссылками**. Участие в системах по обмену ссылками. Также сюда относится обмен со своими же ресурсами (сателлитами).
- **Ссылки из комментариев**. Уже в 2005 году понимали малозначимость ссылок из комментариев. Особенно, если они имели такой вид: “У вас хороший сайт. Загляните также на мой про <a>казино”.
- **Ссылки с/на освобождающиеся домены**. Чаще всего освобождающиеся домены с имеющейся ссылочной массой покупают именно с целью манипулирования выдачей.
- **Немодерируемые ссылки**. Все типы ссылок, которые не проходят модерацию. Сюда относятся и гостевые книги, доски объявлений, а также ряд других сайтов (каталогов).

Ссылочный спам считается самым безопасным, так как поисковая система не всегда имеет возможность точно определить, каким именно образом появилась та или иная ссылка сайта [6, с.578].

Однако чрезмерное использование ссылочного спама может привести к серьезным последствиям, среди которых и наложение фильтра АГС Яндекса. Поэтому, насыщая ссылочную массу покупными ссылками, следует помнить о том, что они должны обладать естественными признаками: стоять на тематических площадках, иметь различные анкоры и так далее.

Поведенческий спам. Данный метод поискового спама является относительно молодым. Сегодня поисковые системы лишь разрабатывают методы борьбы с поведенческим спамом, что и дает возможность многим оптимизаторам достаточно эффективно использовать его.

Накрутка поведенческих факторов web-сайта, как правило, осуществляется путем покупки активностей на биржах, поэтому его применение требует определенных финансовых затрат.

Помимо ссылочного, контентного и поведенческого спама в классификации выделены другие общеизвестные типы, такие как клоакинг, редирект (дорвеи) и скрытый контент.

1.5.2 Проблемы поискового спама

Для описания проблемы определения поискового спама ниже приведена небольшая выдержка из документа Spam: It's Not for Inboxes Anymore (научный труд ребят из стэндфордского университета по классификации веб-спама, на который, впоследствии, ссылались авторы других трудов) [7, с.265].

В ряде случаев веб-спам очевиден. Например, при публикации этой статьи (в 2005 г), топовым результатом поиска Google по запросу "Kaiser pharmacy online" была страница, внешне аналогичная выдаче Google, за исключением заголовка **Gogle**, и все ссылки вели на ряд сайтов, продающих дешевые лекарства, отпускаемые по рецепту. Естественно, эти грубые результаты намеренно подрывают репутацию Google и, несомненно, должны быть помечены как "спам".

Однако, рассмотрим другой случай. Сайт "World News Network" ранжируется третьим в Google по запросам "world news" и "news network", наряду с такими компаниями, как **CNN**,**BBC** и **FOX**. Пользователю очевидно, что **WNN** не является источником новостей; это компания, владеющая тысячами сайтов, каждый из которых по отдельности является онлайн-газетой определенной тематики в определенном географическом регионе (например, Бизнес в Каире или Азиатское Побережье). Дальнейший просмотр сайтов покажет, что многие статьи нетематичны. Например, определенные статьи, принадлежащие разделу "Бизнес в Каире", мало относятся к бизнесу в Каире, и больше раскрывают основные события, такие как война в Ираке. Эти WNN-ресурсы, плотно ссылаются друг на друга, при этом, нет других неаффилированных сайтов, ссылающихся на них. Несмотря на то, что статьи берутся из достойных источников, вроде **Associated Press**,**CNN** и **New York Times**, ссылки на действительный источник ставятся редко. Таким образом, здесь имеется серая область. Агрегация новостей может быть полезной, но дает ли WNN что-либо истинно ценное или существует лишь для получения прибыли с рекламы на его страницах?

Размытость картины также зависит от субъективного взгляда пользователя. При поиске "miserable failure" в Google и Yahoo первым результатом была

официальная биография президента **Джоржа Буша** (**Мишель Мур** и **Хиллари Клинтон** также были в десятке). Спамные техники, конечно же, влияли на рейтинг, но в данном случае вебспам неочевиден, так как одобряете ли вы результат или нет, зависит от личных предпочтений. Таким образом, даже если ценность заключается в довольно странном результате, она всё же есть [8, с.823].

Таким образом, определение веб-спама не так просто, как может показаться. С одной стороны, веб-спам — это явное влияние на ожидание людей. С другой стороны, это серьезная проблема. Ежедневно производится множество попыток повлиять на людей – рекомендации, телевидение и газеты, политическая пропаганда и евангелизм. Одному пользователю что-то может казаться некорректным, обманчивым, неубедительным и раздражающим, другие могут в этом находить соответствие своим взглядам. Автоматически генерируемый мусор, никому не несущий ценности, легко классифицировать как спам. Однако, что на счет большой серой области, где контент несет для кого-то хотя бы небольшую ценность? И кому решать, где кончается серая область и начинается сплошной бесполезняк?”

Главная проблема, порождаемая поисковым спамом, заключается в том, что он генерирует множество мусорного контента, затрудняя эффективную работу поисковых серверов, искажает объективное ранжирование интернет-ресурсов и релевантность поисковых результатов. В итоге это во многом обесценивает Интернет как источник получения объективной информации.

Если ранние поисковые машины могли доверять ключевым словам, указаниям на частоту обновления, то, в связи с активным использованием этих методов для «обмана» поисковых систем, более поздние версии поисковиков были вынуждены практически полностью игнорировать эти указания, критически относясь к каждой из страниц сайта, что сделало затруднительным поиск «добропорядочных» страниц с редким содержимым и прописанными ключевыми словами. Например, страница с текстом средневековой песни и ключевыми «Средние века, поэзия, Восточная Европа», не обладающая большим количеством ссылок с других сайтов, не содержащая в тексте слова «Средние века, поэзия», вряд ли будет найдена по этим ключевым словам.

1.5.3 Методы обнаружения спам контента

Один из методов определения контентного поискового спама это статистический анализ различных его информационных составляющих. В одном из документов специалисты из Microsoft (Detecting Spam Web Pages throught Content Analysis) показывают явные зависимости спамности контента от количества используемых слов, средней их длины, сжимаемости текста, используемых водянистых слов и других параметров [9, с.6].

Например, ниже показано распределение страниц в зависимости от количества используемых слов, а также красной линией вероятность спама в

зависимости от того же числа слов на странице, показан в соответствии с рисунком 6.

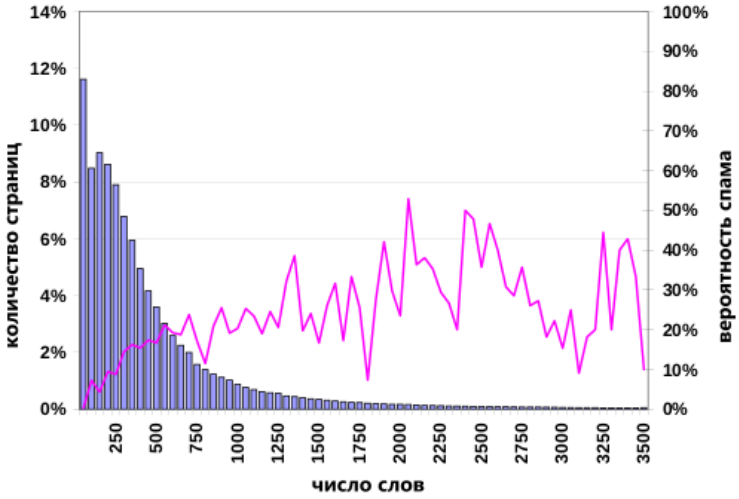

Рисунок 6 – Распределение страниц в зависимости от количества используемых слов

Здесь четко видно, что страницы с количеством слов больше 500-600 явно выделяются из общей массы и вероятность спама в них больше, чем для остальных. Та же зависимость, только по заголовку TITLE показана в соответствии с рисунком 7.

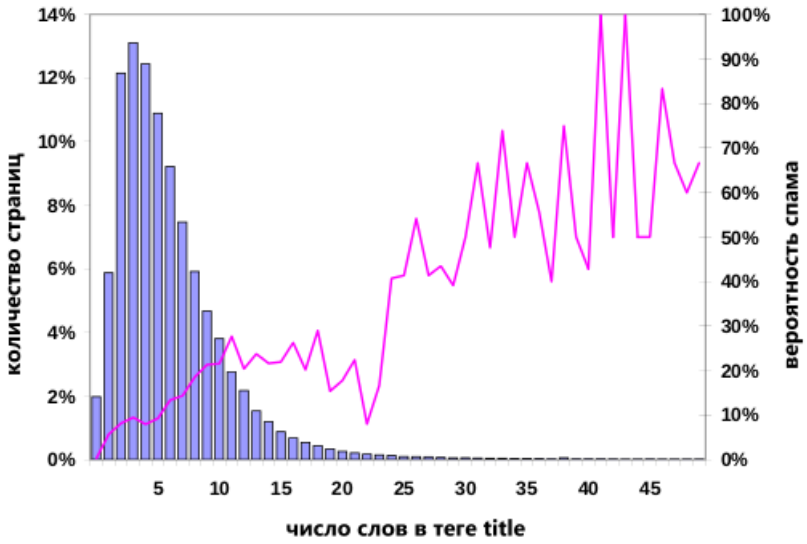

Рисунок 7 – Зависимость по заголовку Title

Заголовки длиннее 10-15 слов встречаются в сети достаточно редко и чем длиннее заголовок, тем больше вероятность попадания документа в спам-зону. Ещё интересна зависимость спамности контента от его способности к сжатию (архивации) показана в соответствии с рисунком 8.

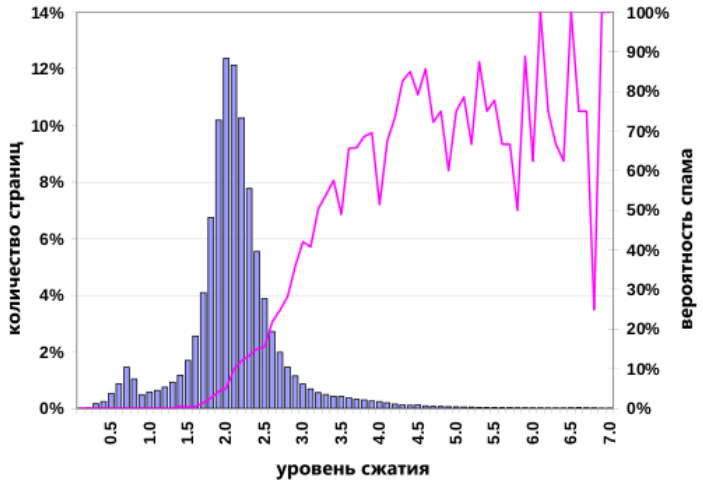

Рисунок 8 – Зависимость спамности контента

Спамный контент часто хорошо сжимается. Из-за большого разнообразия ссылочного поискового спама также существует большое разнообразие методов его идентификации. Например, Anti-Trust Rank (по аналогии с TrustRank) по начальной выборке заведомо спамных сайтов находит аналогичные ресурсы. SpamRank меряет количество незаслуженного PageRank с помощью анализа обратных ссылок. Другие алгоритмы позволяют определять линк-фармы, сетки сайтов, обменивающихся ссылками, и другие спам-техники.

2 Алгоритмы поиска и фильтрации веб-поисковых систем

2.1 Эволюция поисковых систем и факторов ранжирования

Поисковые системы — это очень мощные аппаратно-программные комплексы. Основной целью, которую преследуют их разработчики, является индексирование страниц и документов в глобальной сети интернет для дальнейшего нахождения и выдачи согласно запросам пользователей. Для того, чтобы предоставленная информация была актуальной и качественной, разработчики неустанно совершенствуют формулы и методы ранжирования.

Ранжирование – это процесс, при котором поисковые системы сортируют сайты на странице результатов поиска, присваивая одним сайтам лучшие места, другим – худшие. «Место» сайта зависит от разных факторов.

Обычно поисковики держат в тайне свои критерии ранжирования, однако наиболее общеизвестными характеристиками являются следующие (показан в соответствии с рисунком 9):

- количество слов/словосочетаний, представленных в запросе пользователя, в найденной веб-странице;
- расположение слов/словосочетаний запроса в найденном документе (заголовки, текст, выделенный текст и т.д.);

– прямое соответствие формам слов в запросе (падеж, число, склонение, часть речи и т.д.);
– расстояние между словами, указанными в запросе, и словами в найденной веб-странице;
– ссылочный вес веб-страницы, т.е. насколько часто на данную страницу ссылаются другие веб-ресурсы по указанному запросу.

Рисунок 9 – Схема ранжирования

Индексация сайта – это обход роботами поисковых систем сайтов, и внесение их в базу поисковых машин. После чего в базы поисковых систем заноситься информация о сайте и о страницах сайта (ссылки, изображения, тексты, видео и прочее…) [10, с.1].

Индексация – это процесс, при котором информация из Интернета загружается и анализируется поисковой машиной. В этом процессе каждому слову ставятся в соответствие его «координаты», т.е. номер документа, в котором имеется данное слово, а зачастую и позиция слова в документе. База данных поисковой машины, где хранится найденная информация, называется индексом. Главная задача всего этого – обеспечение высокой скорости поиска и предоставление актуальной информации пользователю представлен в соответствии с рисунком 10.

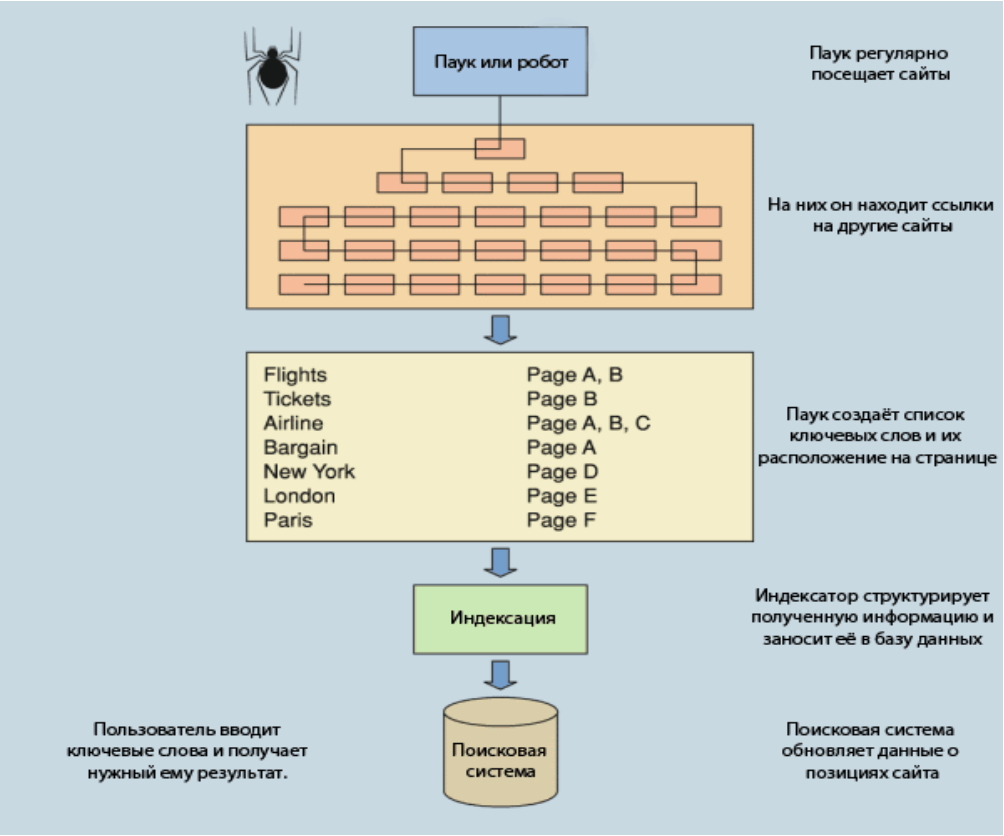

Рисунок 10 - Индексация страниц

2.2 Изучение методов и алгоритмов сортировки и ранжирования результатов поиска

Задача поисковика - при выдаче **результатов поиска** обеспечить максимальное совпадение слов в поисковом запросе со словами, найденными на той или иной веб-странице или в тексте ссылок, ведущих на неё.

Однако перед тем, как быть показанными в результатах поиска, сайты могут быть переставлены в зависимости от того или иного фактора. Вот некоторые из них:

1. Фильтрация дублирующегося (или близкого к этому) контента

Поисковики не любят, чтобы страницы с одинаковым контентом переполняли результаты поиска, поэтому похожие страницы могут быть отфильтрованы.

2. Удаление из результатов поиска нескольких релевантных страниц с одного и того же сайта

Не так уж редко случается так, что релевантными поисковому запросу оказываются сразу несколько страниц одного сайта. В таком случае поисковик может показать ссылку на другую страницу с того же сайта сразу после ссылки на самую релевантную страницу либо просто выдаст

дополнительно ссылку "Ещё результаты с этого сайта". Бывает, что дополнительные результаты вообще не отображаются.

3. Персональные интересы пользователя

Поисковик может перестраивать результаты поиска в зависимости от прошлых результатов и другой активности в интернете того или иного пользователя. Этот вид изменения ранжирования используется в так называемом "персонализированном поиске".

4. Взаимосвязанность сайтов

Поисковик может перестраивать сайты в зависимости от количества и качества ссылок между ними.

5. Сортировка в зависимости от дислокации пользователя

Вполне возможно, что пользователь предпочтёт увидеть в результатах поиска те сайты, которые расположены в его регионе проживания. Иногда пользователи сами включают опцию поиска сайтов только из определённой страны, иногда поисковик пытается самостоятельно определить дислокацию пользователя на основе его IP.

6. Язык запроса

Опции ограничения по языкам могут быть установлены пользователем в браузере либо через настройки поисковика; иногда поисковик сам выбирает язык результатов поиска в зависимости от поискового запроса, настроек либо характеристик результатов поиска.

7. Демографические характеристики

При использовании этого метода для ранжирования результатов поисковик может учитывать такие параметры, как дислокация, демографическая информация, информация о группах, к которым относится пользователь.

8. Хронологические данные

Сайты могут быть рассортированы в зависимости от времени создания и изменения страниц, возраста ссылок на них и многих других временнЫх параметров.

9. Тематическая схожесть

Учитываются заголовки, стоп-слова и другие связанные с текстом параметры.

10. Заинтересованность пользователя в коммерческой информации

К примеру, Yahoo! Mindset позволяет пользователю самостоятельно отсортировать результаты поиска в зависимости от того, хочет он видеть информацию коммерческого или более информативного, некоммерческого характера.

11. Наличие у сайта версии для мобильных устройств

Поиск от Microsoft может отфильтровать те сайты, которые не могут быть отображены на мобильных устройствах (например, КПК).

12. Доступность людям с ограниченными возможностями

Google не так давно запустил специализированный поиск, который ранжирует сайты в зависимости от их доступности людям с ограниченными возможностями.

13. Отношение пользователей

Google умеет сортировать страницы в зависимости от того, нравятся они пользователям или нет.

14. Дополнительные параметры поиска и схожесть текста

Google умеет ранжировать результаты поиска по новостям, основываясь на обработке и автоматическом добавлении дополнительных параметров к поисковому запросу, а также на схожести документов.

15. Поведение пользователя

16. Одобрение сообществом

Метод основывается на мнении социальных групп, пользователей социальных сетей и т.п.

17. Избыток информации

Отфильтровывается избыточная информация, засоряющая результаты поиска.

18. Новостные сюжеты

Методика от IBM позволяет преобразовывать результаты поиска в сюжеты новостей, которые будут раскрываться на сайтах, выдаваемых в результатах поиска.

19. Оценка актуальности на основе сообщений в блогах, новостях и т.п.

Результат поиска ставится в зависимость от того, куда именно ссылаются блоги и форумы в данный период времени. Приоритет получает сайт, имеющий наиболее актуальную на данный момент информацию.

20. Актуальность (временной фактор) и использование

Этот метод близок к сочетанию двух предыдущих. Он предполагает группировку понятий с учётом того, как они развивались во времени и как люди влияли на эти изменения.

2.3 Поисковая система Google

Процесс поиска информации и создания страницы результатов включает множество компонентов. Google постоянно совершенствует поисковые технологии, чтобы повысить точность поиска. Google использует уникальные инновационные решения, такие как Сеть Знаний и Живой поиск Google. Есть и другие важные системы, которые постоянно отлаживаются и оптимизируются. Ниже приведен список проектов, которые позволяют получить представление о многообразии процессов, используемых в поиске Google [11, с.12].

2.3.1 Возможности алгоритма поиска

Поддерживает поиск в документах форматов PDF, RTF, PostScript, Microsoft, Word, Microsoft Excel, Microsoft PowerPoint и других.

Для улучшения качества алгоритмов поиска в Google существуют целые проекты, направленные на максимальное получение качественных результатов за меньший промежуток времени. Это такие проекты как:

Ответы

Немедленный показ ответов и информации по запросам о погоде, результатах соревнований и отдельных фактах.

Автозаполнение

Предсказание поискового запроса, в том числе распознавание терминов с несколькими значениями.

Книги

Поиск в миллионах книг различных библиотек и издательств всего мира, предварительный просмотр и доступ к текстам.

Актуальность

Последние новости и информация. Поиск с указанием определенных дат.

Живой поиск Google

Показ результатов поиска во время ввода запроса.

Картинки

Уменьшенные изображения найденных картинок, по которым можно сразу выбрать нужную страницу.

Индексация

Использование систем сбора и хранения документов в Интернете.

Сеть знаний

Результаты из базы данных реально существующих людей, мест и вещей с учетом связей между ними.

Для мобильных устройств

Функции поиска, разработанные специально для мобильных устройств: планшетных ПК и смартфонов.

Новости

Результаты из периодических изданий в Интернете и блогов со всего мира.

Понимание запросов

Понимание смысла запросов пользователей.

Уточнения

Расширенный поиск, похожие запросы и другие инструменты, позволяющие уточнять поисковые запросы.

Безопасный поиск

Удаление из результатов поиска веб-страниц, изображений и видео, предназначенных только для взрослых.

Методы поиска

Новые способы поиска, например поиск по картинке и голосовой поиск.

Качество сайта и страницы

Использование набора критериев для определения надежности, достоверности и авторитетности источника. (К таким критериям можно

отнести PageRank. Это один из первых алгоритмов Google, который оценивает ссылки между страницами, чтобы определить их релевантность.)

Описания страниц

Предварительный просмотр информации, например заголовок страницы и краткое описание для каждого результата.

Правописание

Исправление возможных орфографических ошибок и предложение альтернативных вариантов.

Синонимы

Распознавание слов с похожими значениями.

Перевод и интернационализация

Поиск с учетом страны и языка пользователя.

Универсальный поиск

Различные материалы (картинки, новости, карты и видео), а также персональные результаты на одной общей странице результатов поиска.

Контекст пользователя

Более точные результаты с учетом географического положения, Истории веб-поиска и других критериев.

Видео

Уменьшенные изображения видеороликов в результатах поиска, которые позволяют сразу выбрать нужное видео [12, с.62].

2.3.2 Алгоритм индексации сайтов

Поисковый робот Google имеет User Agent — Googlebot (Поисковый робот), который является основным роботом, сканирующим содержание страницы для поискового индекса. Помимо него существуют ещё несколько специализированных роботов:

1. Googlebot-Mobile — робот, индексирующий сайты для мобильных устройств;
2. Google Search Appliance (Google) gsa-crawler — поисковый робот нового аппаратно-программного комплекса Search Appliance;
3. Googlebot-Image — робот, сканирующий страницы для индекса картинок;
4. Mediapartners-Google — робот, сканирующий контент страницы для определения содержания AdSense;
5. Adsbot-Google — робот, сканирующий контент для оценки качества целевых страниц AdWords.

Существует теория эффекта песочницы, которая утверждает, что сайты, которые имеют новые доменные имена, частые смены владельцев или нэймсерверов, помещаются в «песочницу» (зону ожидания) и пребывают там, пока механизм Google не сочтёт сайт готовым из неё выйти.

Также существует обратная теория, называющаяся «бонус новичка», в которой при первоначальной индексации сайта, в силу некоторых аспектов (например, нескольких ссылок с авторитетных ресурсов), сайту сразу присваивается высокий PageRank и хорошие места в поисковой выдаче. По истечении некоторого времени, после снятия этого эффекта сайту присваиваются его реальные показатели.

2.3.3 Алгоритм ранжирования

Современные алгоритмы Google используют более 200 различных сигналов или "ключей", чтобы понять, что именно ищет пользователь. Учитывается такие параметры, как наличие слов на страницах сайтов, актуальность информации, местоположение пользователя и показатель PageRank интернет ресурсов и многое другое [13, с.2].

PageRank - один из алгоритмов ссылочного ранжирования. Алгоритм применяется к коллекции документов, связанных гиперссылками (таких, как веб-страницы из всемирной паутины), и назначает каждому из них некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов. Это числовая величина, характеризующая «важность» веб-страницы. Чем больше ссылок на страницу, тем она «важнее». Кроме того, «вес» страницы A определяется весом ссылки, передаваемой страницей B. Таким образом, PageRank — это метод вычисления веса страницы путём подсчёта важности ссылок на неё.

PageRank является одним из вспомогательных факторов при ранжировании сайтов в результатах поиска, Google использует показатель PageRank найденных по запросу страниц, чтобы определить порядок выдачи этих страниц посетителю в результатах поиска.

2.3.3.1 Google PageRank

Согласно документу (на англ. языке), подробно описывающий модель PageRank, введеной Сергеем Брином и Ларри Пейджем основная формула, описывающая PR выглядит следующим образом:

$$PR(A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right) \quad (3)$$

где 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 в Google знают только

там, остальные принимают его равным 0.85 (видимо из экспериментальных данных);

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

Ti — i-ая ссылающаяся страница;

C — количество внешних ссылок на странице-доноре [14, с.35].

2.3.3.2 Алгоритм Google Колибри

Google использует синонимы, чтобы представить пользователям больше ответов на поисковый запрос, расширить его и позволить подобрать наиболее подходящий вариант.

Кстати, из этого же поста становится понятным, почему использование лишь словаря синонимов или высокорейтинговых запросов не гарантирует релевантную поисковую выдачу (прочитайте опыт Ванессы, когда она вводила в качестве поисковой фразы «взять домашнее животное с питомника» и ей Google предлагал варианты только с кошками, но не было ни одного с собаками).

Амит Сингал также в одной из своих старых публикаций говорит о том, что полагаться на одни только синонимы – далеко не идеальное решение, т.к. два слова могут как являться синонимами, так и нет, в зависимости от контекста, в котором они используются.

Например, оба слова «coche» и «automóvil» с испанского могут переводиться как «машина». А вот слово «carro» в латиноамериканском варианте испанского языка означает исключительно «машина», в испанском же переводится как «вагон».

Таким образом, для подбора более подходящих ответов необходимо использовать семантический принцип анализа слов (т.е. по смыслу), который позволит Google лучше и быстрее понимать поисковые фразы.

Именно для этих целей и был создан алгоритм Колибри. Похоже, Google не заменил предыдущие алгоритмы, такие как PageRank и многие другие, которые он использовал для сортировки страниц. Во время анонса нового алгоритма представители компании сообщили, что Google начал использовать Hummingbird несколько недель назад и что он охватывает 90% всех поисковых запросов, представлен в соответствии с рисунком 11.

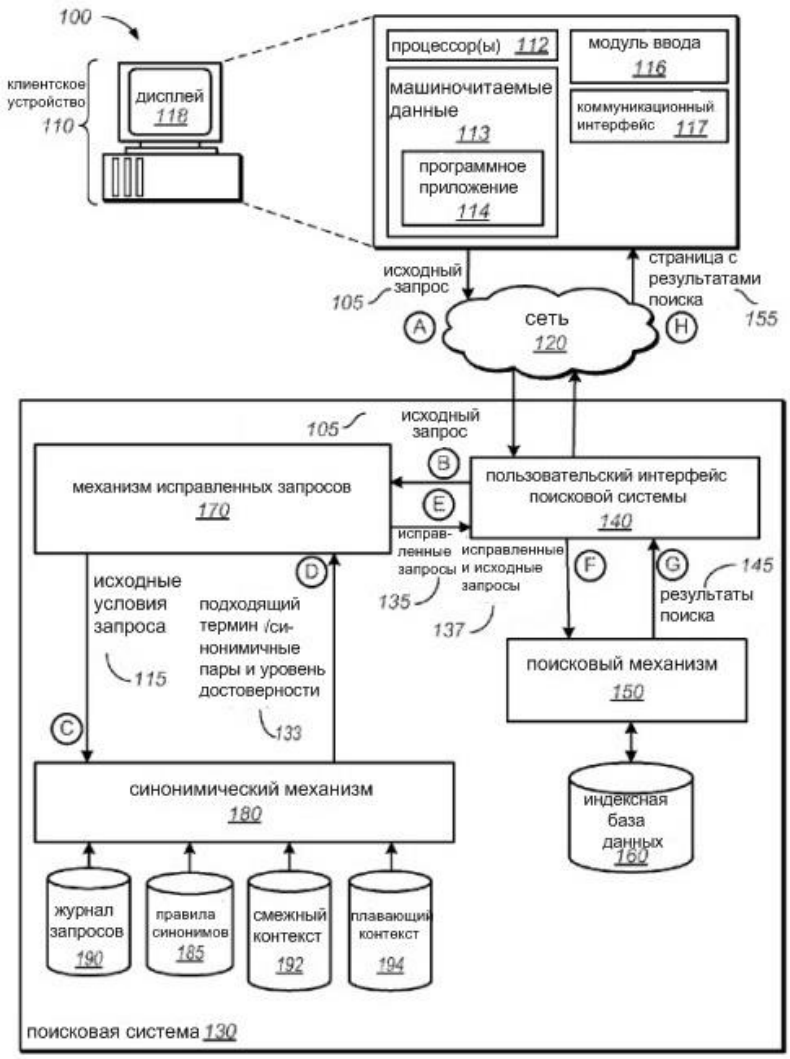

Рисунок 11 - Схема работы нового алгоритма

Слова нельзя рассматривать как отдельные «элементы», все они обозначают какие-то понятия, поэтому Google начал объединять слова в концепции. У любого объекта может быть связь с другими объектами, которая может измениться в зависимости от контекста. В этом смысле слова подобны людям, городам, книгам и другим именам собственным – с чем обычно работает Граф Знаний.

Механизмы, которые использует Google для определения области поиска, особенно важны в процессе устранения возможных неправильных толкований различных слов. Так уточняется поиск, что, соответственно, повышает точность поисковой выдачи.

Данная техника не особо отличается от семантического поиска, используемого в Графе Знаний. Объединение всех этих элементов уже сейчас позволяет Google (теоретически):

- лучше понимать цель поискового запроса;
- расширять область поиска, которая может соответствовать запросу;

- упрощать выдачу информации, т.е. если запрос А, запрос В и запрос С по существу обозначают одно и то же, Google не предлагает три варианта поисковой выдачи, а лишь один;
- предлагать более точную поисковую выдачу благодаря семантической обработке поисковой фразы (основываясь как на прямых, так и косвенных элементах персонализации). Google теперь предлагает результаты, которые имеют более высокую вероятность точно удовлетворить потребности пользователей;
- как следствие, Google может предоставить и лучшую поисковую выдачу для контекстной рекламы, т.к. в 99% случаях до запуска алгоритма Колибри контекстная реклама не отображалась в поисковой выдаче при многословных запросах.

Данный алгоритм направлен на расширение запроса или расширение подхода, который поможет лучше понять длинные естественные языковые запросы, которые используются в речи вместо коротких ключевых слов запроса, которые могут быть набраны в поисковой строке.

Например, запрос, в котором алгоритм может проявить свой потенциал, выглядит так: «Самое лучшее заведение, в котором можно попробовать пиццу по-чикагски». Google может использовать синоним и правила эквивалентного запроса в сочетании с анализом других слов, не отбрасывая их в самом запросе, чтобы понять контекст термина запроса и возможную замену для него, чтобы перефразировать или поменять термины поиска и показать лучшие результаты.

Google может проанализировать запрос «Самое лучшее заведение, в котором можно попробовать пиццу по-чикагски» и понять, что пользователь ищет результаты, в которых предпочтительнее использовать слово “ресторан” вместо “заведение”.

Слово “ресторан” вместо “заведение” может рассматриваться как возможный синоним или эквивалент на основе правил замены, сфокусированных на сочетаемых условиях, которые могут появиться в результатах запроса при поиске данных условий или на сочетаемых терминах при сеансах запроса.

Анализ Google различных объектов поиска, таких как связь, между запросами, может быть в некоторых случаях более подходящим для пользователя. В данном случае результаты поиска основываются на таком факторе, как длительность нахождения на странице, когда пользователь выбирает ее из списка результатов поиска представлен в соответствии с рисунком 12.

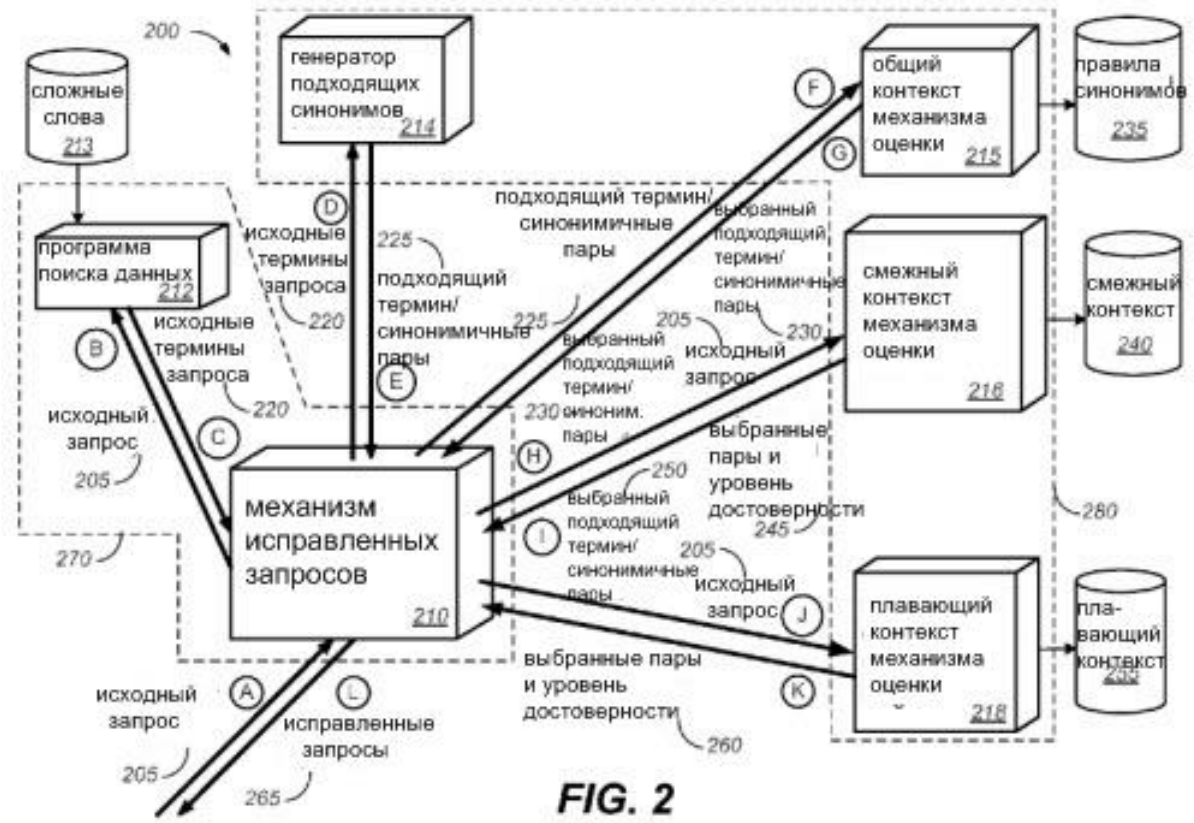

FIG. 2

Рисунок 12 - Механизм анализа поисковых запросов

2.3.4 Google и борьба с веб-поисковым спамом

При ранжировании сайтов поисковыми системами в формуле учитывается не только релевантность сайта, его авторитетность и все остальные показатели качества, но также и коэффициент спамности. Причем, этот коэффициент не бинарный (спам или не спам), а непрерывный, как многие другие ранки.

Для работы с поиском в Google даже существует разные отделы, один посвящен качеству поиска (Search Quality Team) и его задача как раз обеспечивать качество выдачи (работают с формулой ранжирования). Глава этого отдела **Амит Сингал** (Amit Singhal). Другой отдел (Webspam Team) борется с поисковым спамом во главе с **Мэттом Каттсом** (Matt Catts). Соответственно, алгоритм “Панда” (в 2011 г) это разработка отдела качества, в то время как алгоритм “Пингвин” (в 2012-ом) это порождение отдела по борьбе с веб-спамом [15, с.958].

Каждый день в сети появляются миллионы страниц со спамом. Чтобы они не мешали пользователям, Google использует эффективные методы автоматической и ручной проверки сайтов. Сайты, содержащие спам, стремятся пробиться в верхние строки результатов поиска с помощью таких мошеннических приемов, как многократно повторяющиеся ключевые слова, покупка ссылок, передающих PageRank, или использование невидимого

текста. Это снижает эффективность поиска, поскольку действительно полезные сайты оттесняются в конец списка результатов, и найти нужные сведения становится труднее.

Сайты, размещающие веб-спам, бывают очень разными. Некоторые содержат бессмысленный, автоматически сгенерированный текст, который никто не может понять; другие сайты используют более тонкие приемы [16, с. 96].

К счастью, алгоритмы Google позволяют идентифицировать подавляющее большинство разновидностей спама и автоматически понижать рейтинг сайтов, на которых они встречаются. Оставшихся нарушителей сотрудники Google отсеивают вручную. Для удаления спам страниц Google использует различные алгоритмы, рассмотрим основные из них:

2.3.4.1 Google Panda (Панда)

Ещё зимой 2011 года поисковая система Google ввела в действие новый алгоритм ранжирования, под безобидным с одной стороны названием Panda. Разработчиками нового алгоритма стали сотрудники поискового гиганта Амит Сингал и Мэтт Катс. В первое время алгоритм обкатывался на американской выдаче, но вскоре ему стало тесно на просторах США и Канады, и он перебрался на весь остальной мир, а в СНГ получил более понятное прозвище фильтр Google Panda. Новый алгоритм радикально отличается от прежних разработок своей ориентированностью на человеческий фактор. Однако алгоритм учитывает при ранжировании и обычные критерии. В данном случае было бы правильнее оперировать SEO-терминологией, в которой человеческий фактор называется поведенческим.

Поисковый алгоритм Панда как никакой другой до него придирчиво учитывает поведение пользователя на сайте. Учитывается, например, не только время, в течение которого пользователь оставался на сайте, но и то, какие действия он там производил, к примеру, совершал ли переходы по ссылкам, так или иначе относящимся к предмету, который содержался в ключевом запросе. Очевидно, что для этого разработчикам необходимо было колоссальное усложнение алгоритма, поэтому, к слову сказать, алгоритм так долго запускался для всего интернета – создатели до бесконечности шлифовали поисковые инструменты [17, с.153].

Данный алгоритм ранжирования отбрасывает вниз по выдаче запросы, а то и целые сайты, которые, по мнению разработчиков, не отвечают требованиям качественного продвижения. Так на попадание под фильтр алгоритма могут смело рассчитывать те сайта, где есть:

- Низкий процент уникального контента;
- Большое количество нерелевантных заголовку страниц;
- Переоптимизация по ключевым словам;
- Высокий процентный показатель отказов от просмотра, как на отдельных страницах, так и на всём сайте в целом;
- Низкая средняя величина времени пользователя на проекте;

- Нерелевантная реклама;
- Негативное процентное соотношение уникального и неуникального текста;
- Однообразно форматированный и структурированный контент;
- Узкий анкор-лист внутренних ссылок.

Фильтр непредсказуем, Google Панда постоянно обучается и как он поведет себя в будущем предсказать сложно. Единственное можно добавить по переоптимизированному контенту. Плотность ключевых слов страниц сайта не должна составлять больше 5 - 6%.

Особенности фильтра Panda заключаются в том, что из относительно небольшого количества нарушений на страницах сайта алгоритм может ударить по всему домену сайта, понизив в выдаче и качественные и некачественные материалы, а в исключительных случаях не исключается возможность и полного бана домена сайта.

2.3.4.2 Google Penguin (Пингвин)

Впервые имя этого алгоритма прозвучало в апреле 2012 года, а в начале мая пришел и первый апдейт, и, как итог: позиции большинства сайтов упали. Алгоритм выполняет те же задачи, что и другие алгоритмы и фильтры. Задача же проста - очистка поисковой системы от некачественных ресурсов. Получить наказание от Google Penguin можно за следующее:.

- Активные ссылочные связи с другими сайтами (партнерами) для передачи друг-другу ссылочного веса и манипулирования поисковыми алгоритмами;
- Продажа ссылок с сайта (нетематические исходящие ссылки);
- Использование в большом количестве шаблонных страниц (для привлечения поискового трафика и заработка на рекламе);
- Чрезмерная насыщенность ключевыми словами для внутренних и внешних ссылок;
- Использование материала с других интернет ресурсов;
- Спам в каталогах и блогах некачественными статьями;
- Переоптимизация страниц на сайте (большое количество ключевых слов при создании текста и мега-тегов).

Стоит отметить что этот фильтр уже четыре раза обновляли. Полная хронология апдейтов выглядит так:

- Впервые алгоритм Penguin вышел 24 апреля 2012 и затронул примерно 3% поисковых запросов;
- Первый апдейт Penguin до версии 1.1 состоялся 26 мая 2012 и затронул менее 0,1% поисковых запросов;
- Второе обновление фильтра произошло 5 октября 2012 и затронуло примерно 0,3% поисковых запросов;
- Penguin 2.0 официально представлен 22 мая 2013. По официальному заявлению обновлённый фильтр затронул до 2,3% поисковых запросов.

В общем, новый алгоритм теперь умеет глубже сканировать и намного лучше разбираться в качестве интернет-ресурса. Необходимо учесть и еще один момент - это реклама на ресурсе. Перенасыщенность рекламой главной страницы, ведет к потере позиции в выдаче, Пингвин в этом преуспел [18, с.157].

2.4 Поисковая система Яндекс

Поисковый запрос от пользователя, после анализа загруженности поисковой системы, отправляется на наименее загруженный сервер. Для обеспечения такой возможности сервера Яндекса объединены в кластеры и даже кластеры кластеров. Затем пользовательский запрос обрабатывается программой под названием «Метапоиск». Метапоиск осуществляет анализ запроса в реальном времени: определяет географическое положение пользователя, проводит лингвистический анализ и т. д. Также, программа определяет, относится ли запрос к категории наиболее популярных или недавно заданных. Выдача на такие запросы некоторое время хранится в памяти (кэше) метапоиска, и в случае совпадения показываются заранее сохранённые результаты. Если запрос является редким и совпадений в кэше не найдено, система перенаправляет его на программу «Базового поиска».

Тот анализирует индекс системы, также разбитый по разным дублирующимся серверам (это укоряет процедуру). Затем полученная информация снова попадает на метапоиск, данные ранжируются и показываются пользователю в готовом виде [19, с.20].

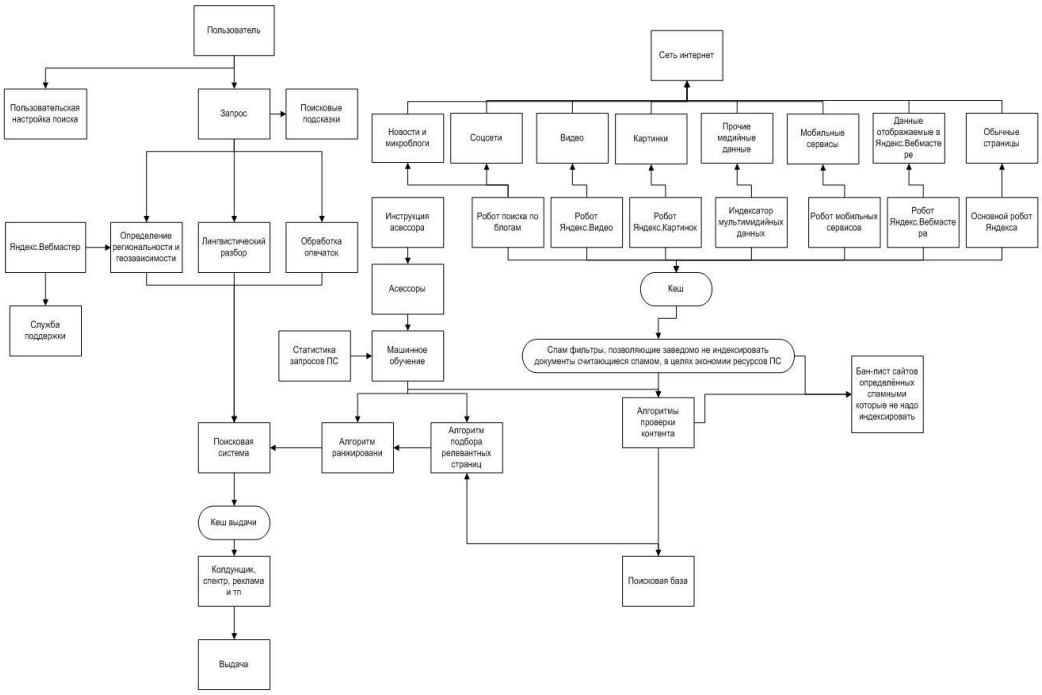

Рисунок 13 - Упрощенная схема работы поисковой системы, системы кеширования, индексирования и ранжирования результатов

Индексирование.

В целом «Яндекс» индексирует следующие типы файлов: html, pdf, rtf, doc, xls, ppt, docx, odt, odp, ods, odg, xlsx, pptx.

Поисковая система способна также индексировать текст внутри объектов Shockwave Flash (если текст не помещен на само изображение), если эти элементы передаются отдельной страницей, имеющей MIME-тип application/x-shockwave-flash, и файлы с расширением .swf.

В Яндексе работают 2 сканирующих робота — «основной» и «быстрый». Первый отвечает за интернет в целом, второй индексирует сайты с часто меняющейся и обновляемой информацией (новостные сайты и информационные агентства). В 2010 году «быстрый робот» получил новую технологию под названием «Orange», разработанную совместно калифорнийским и московским подразделениями «Яндекса».

В логах сервера роботы Яндекса представляются следующим образом:

- Mozilla/5.0 (compatible; YandexBot/3.0) — основной индексирующий робот;
- Mozilla/5.0 (compatible; YandexBot/3.0; MirrorDetector) — робот, определяющий зеркала сайтов. Если найдутся несколько сайтов с одинаковым содержимым, в результатах поиска будут показан только один;
- Mozilla/5.0 (compatible; YandexImages/3.0) — индексатор «Яндекс.Картинок»;
- Mozilla/5.0 (compatible; YandexVideo/3.0) — индексатор «Яндекс.Видео»;
- Mozilla/5.0 (compatible; YandexMedia/3.0) — робот, индексирующий мультимедийные данные;
- Mozilla/5.0 (compatible; YandexBlogs/0.99; robot) — робот поиска по блогам, индексирующий комментарии постов;
- Mozilla/5.0 (compatible; YandexAddurl/2.0) — робот, обращающийся к странице при добавлении её через форму «Добавить URL»;
- Mozilla/5.0 (compatible; YandexFavicons/1.0) — робот, индексирующий иконки сайтов (favicons);
- Mozilla/5.0 (compatible; YandexDirect/3.0) — робот, индексирующий страницы сайтов, участвующих в Рекламной сети Яндекса (РСЯ);
- Mozilla/5.0 (compatible; YandexDirect/2.0; Dyatel) — «простукивалка» «Яндекс.Директа»;
- Mozilla/5.0 (compatible; YandexMetrika/2.0) — робот «Яндекс.Метрики»;
- Mozilla/5.0 (compatible; YandexCatalog/3.0; Dyatel) — «простукивалка» «Яндекс.Каталога»;
- Mozilla/5.0 (compatible; YandexNews/3.0) — индексатор «Яндекс.Новостей»;

– Mozilla/5.0 (compatible; YandexAntivirus/2.0) — антивирусный робот «Яндекса»;

С 2009 года Яндекс поддерживает технологию Sitemaps.

2.4.1 Поисковая выдача и ранжирование

Поисковая выдача для каждого пользователя формируется индивидуально на основе его местоположения, языковых запросов, интересов и предпочтений по результатам предыдущих и текущей поисковой сессии. Тем не менее, ключевым фактором при ранжировании поисковых результатов является их релевантность, соответствие поисковому запросу. Релевантность определяется на основе формулы ранжирования, которая постоянно обновляется на основе алгоритмов машинного обучения.

Долгое время ключевым фактором ранжирования у Яндекса было количество сторонних ссылок на конкретный сайт. Каждой странице в Интернете присваивался своеобразный индекс цитирования, аналогичный индексу для авторов научных статей: чем больше ссылок, тем лучше. Подобный механизм работал как в тИЦ Яндекса, так и в ПэйджРэнк Гугла. Тем не менее, в настоящий момент в целях предотвращения «накрутки» Яндекс применяет многофакторный анализ, из восьмисот факторов ранжирования которого на ссылочные приходится лишь 70. Сегодня гораздо большую роль играет содержание сайта и наличие-отсутствие там ключевых слов, удобство чтения текста, название домена и его история, наличие мультимедийного контента. [20, с.164]

Поиск производится на русском, английском, французском, немецком, украинском, белорусском, татарском, казахском языках.

Результаты поиска можно отсортировать по релевантности и по дате (кнопки снизу поисковой выдачи).

2.4.2 Понятие 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 в соответствии с рисунком 14.

Из ссылочного графа (др. название веб-граф) можно определять различные параметры сайтов, такие как: индекс цитируемости, авторитетность ресурса, вероятность нахождения пользователя на том или ином сайте, и другие. Для хранения веб-графа в машинном виде, используют другое представление данных, а именно матрицы смежности и инцидентности и, возможно, матрицу достижимости.

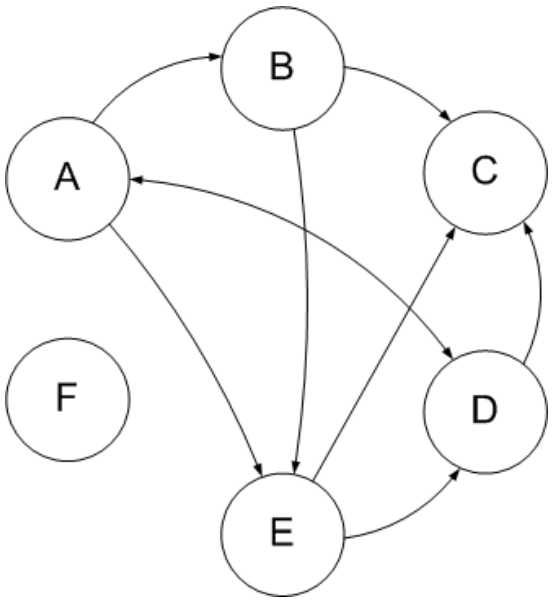

Рисунок 14 – Ссылочный граф.

где A, B, …, F — определенные сайты в индексе поисковой системы; стрелки изображают направление связей — односторонние либо двусторонние.

Каждая строка в матрице инцидентности соответствует определенной вершине графа, а столбцы соответствуют его связям. В ячейку на пересечении i-ой строки с j-м столбцом матрицы записывается 1 в случае если связь j «выходит» из вершины i, −1 если связь «входит» в вершину, любое число отличное от 0, 1, -1 если связь является петлей, и 0 во всех остальных случаях. Такой способ представления связей между ресурсами является самым емким и неудобным для хранения, но облегчает нахождение циклов в графе (сателлиты и сайты, участвующие в кольцевом обмене, легко обнаруживаются) [21, с.100].

Для приведенного выше примера, матрица инцидентности будет иметь вид:

	(AB)	**(BC)**	**(BE)**	**(DC)**	**(AD)**	**(DA)**	**(EC)**	**(ED)**	**(AE)**
A	1	0	0	0	1	-1	0	0	1
B	-1	1	1	0	0	0	0	0	0
C	0	-1	0	-1	0	0	-1	0	0
D	0	0	0	1	-1	1	0	-1	0
E	0	0	-1	0	0	0	1	1	-1
F	0	0	0	0	0	0	0	0	0

В матрице смежности столбцы и строки соответствуют вершинам графа. В каждой ячейке этой матрицы записывается число, определяющее наличие связи от вершины-строки к вершине-столбцу или наоборот. Для нашего случая матрица смежности будет иметь вид:

	A	**B**	**C**	**D**	**E**	**F**
A	0	1	0	1	1	0
B	0	0	1	0	1	0
C	0	0	0	0	0	0
D	1	0	1	0	0	0
E	0	0	1	1	0	0
F	0	0	0	0	0	0

В случае взвешенного графа каждому ребру присваивается определенный вес w и, соответственно, в матрице смежности вместо единиц будут присутствовать веса связей.

Что касается поисковых систем, то при составлении ссылочных графов и расчете зависимых показателей они не учитывают ряд немодерируемых ресурсов, таких как линкопомойки, гостевые книги, сетевые конференции, каталоги и другие сайты, где кто угодно может добавлять ссылки без контроля со стороны владельца ресурса и, таким образом, влиять на итоговый ссылочный граф, который позволяет находить ряд параметров сайтов, являющимися факторами ранжирования в поисковой системе.

2.4.3 Взвешенный индекс цитирования (вИЦ)

Согласно теории поиска, пользователи информационно-поисковых систем определяют ценность документа путем так называемых информационных ключей, таких как: заголовок статьи, автор, название сборника, индекс цитируемости, дата публикации работы, и далее переходят к более значимым документам. Чем больше ключей они учтут при поиске, тем лучше они могут вычислить ценность того или иного документа. Информационные ключи, основанные на лексическом сходстве документов запросам, весьма ограничены. В качестве параметра значимости статьи обычно рассматривается её индекс цитируемости. Однако, он также имеет ряд ограничений. Этот фактор не отражает структуру ссылок в каждой дисциплине (тематике), а также слабозначимые работы и труды с большой значимостью могут иметь одинаковый индекс цитируемости.

Опираясь на изложенную модель, люди, занимающиеся ранжированием результатов поиска в электронных библиотеках, предложили не только подсчитывать количество цитат, но также ввести понятие веса ссылки, обосновывая это тем, что качество статьи определяется не только количеством, но и качеством ссылающихся источников. То есть, значимость источника ссылки является таким же важным информационным ключом для пользователей, определяющих ценность документа.

Раннее, для определения значимости документа использовали импакт-фактор. Однако этот фактор имеет некоторые недостатки. Главные из них в том, что импакт-фактор является не нормализованной величиной, а также рассматривает ссылки за короткий период времени (2-3 года) и не учитывает цитирования с конференций, которые часто играют важную роль в исследованиях, благодаря их популярности.

Поэтому и был введен термин **фактор популярности** (англ. *Popularity Factor*), который для определенного сборника вычислялся по следующей формуле:

$$PF(v,t) = \frac{n_v}{N} * \sum_{i \in P} \frac{PF(i,t) \times w(i)}{N(i)}. \quad (4)$$

где, PF(v,t) — фактор популярности сборника публикаций в рассматриваемый год t;

P — множество сборников, которые ссылаются на сборник v в этомгоду;

nv — количество статей, опубликованных в сборнике v в этом году;

N — общее число статей, опубликованных в этом году (при этом, nv/N — вероятность того, что пользователь читает сборник v);

w(i) — частота цитируемости сборником i сборника v;

N(i) — общее число ссылок в i-ом сборнике.

При этом PF(v,t) является нормализованной величиной, где сумма квадратов PF для всех сборников равна единице.

Все вышесказанное касалось электронных библиотек, однако, описанная модель также применяется поисковыми системами для интернет-документов, используя популярность как один из факторов для ранжирования документов. В разных поисковиках этот фактор называется по разному: PageRank в Google, вИЦ в Яндексе, ИЦ в Апорте (из-за разных названий и возникала путаница среди вебмастеров).

Как описывал о работе ПС Яндекс Илья Сегалович:

Введение ссылочного поиска и статической ссылочной популярности (мы называем этот фактор ВИЦ — взвешенный индекс цитирования — аналог известного PageRank) помогает поисковым системам справляться с примитивным текстовым спамом, который полностью разрушает традицонные статистические алгоритмы информационного поиска, полученные в свое время для контролируемых коллекций.

Александр Садовский по этому поводу дал следующий комментарий:

Аналогичную возможность для русскоязычного интернета предоставляет Yandex.Bar. Яндекс.Бар показывает “взвешенный индекс цитирования” гораздо точнее, чем это делает Google — данные

отображаются с точностью до сотых. В то же время, показываемое число касается только главной страницы сайта, и не зависит от веса его конкретной страницы, которая в данный момент загружена в окне браузера.

Еще один источник информации про вес PageRank для русскоязычных сайтов — каталог поисковой системы Апорт. Если сайт присутствует в каталоге, то рядом с его описанием указывается ИЦ главной страницы сайта. Отсюда можно сделать вывод, что вИЦ в яндексе это тот же PageRank в Google с точностью до деталей реализации. Также обратите внимание, что формула (1) совпадает с формулой вычисления PageRank, где w(i)=1 (количество учитываемых одинаковых ссылок) и не учитывается переход пользователя не по ссылке (параметр 1-d отсутствует).

$$PR(A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right) \quad (5)$$

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

Тематический индекс цитирования (тИЦ) — заключающаяся в определении авторитетности интернет-ресурсов с учётом качественной характеристики — ссылок на них с других сайтов. тИЦ рассчитывается по специально разработанному алгоритму, в котором особое значение придаётся тематической близости ресурса и ссылающихся на него сайтов. Данный показатель в первую очередь используется для определения порядка расположения ресурсов в рубриках каталога «Яндекса». Все ссылающиеся сайты обязательно должны быть проиндексированы Яндексом. При этом на соответствующих страницах каталога указываются лишь округлённые значения, которые помогают приблизительно ориентироваться в авторитетности ресурсов раздела.

тИЦ определяется суммарным весом ссылающихся сайтов. Не могут влиять на тИЦ сайты, где любой человек может поставить свою ссылку без ведома администратора ресурса. тИЦ имеет систему апдейтов (пересчётов показателей), и обычно его обновление происходит 1 раз в 2—3 месяца.

Тематический индекс цитирования (тИЦ) определяет «авторитетность» интернет-ресурсов с учетом качественной характеристики ссылок на них с других сайтов. Эту качественную характеристику мы называем «весом» ссылки. Рассчитывается она по специально разработанному алгоритму.

Большую роль играет **тематическая близость** ресурса и ссылающихся на него сайтов. Само по себе количество ссылок на ресурс также влияет на значение его тИЦ, но тИЦ определяется не **количеством** ссылок, а суммой их **весов**.

Поэтому Яндекс постоянно совершенствует алгоритмы ранжирования и делает их устойчивыми к внешнему влиянию (например, к попыткам некоторых вебмастеров обмануть поисковую систему). Также как и в Google,

Яндекс имеет алгоритмы и фильтры, которые убирают спам из результатов поиска, оставляя только релевантные интернет-ресурсы [22, с.259].

2.4.4 Алгоритмы фильтрации спам контента

Основным алгоритмом является фильтр АГС — призванный минимизировать фактор постороннего влияния на результат поисковой выдачи. С помощью данного алгоритма «Яндекс» заносит сайты в черный список. Также как и в Google со временем фильтр усовершенствовался, получая названия АГС-17, АГС-30 и АГС-40. Официально о внедрении алгоритма АГС-17 компания Яндекс объявила в конце 2009 года – именно тогда значимая часть некачественных интернет-ресурсов выпала из индексирования. И уже через полгода появилась версия АГС-30, дополнившая и усовершенствовавшая фильтр АГС-17.

Основные причины АГС

Факторов, благодаря которым могут быть применены санкции в Яндексе (АГС), - более 100. Причем, они могут в большей или меньшей степени присутствовать на интернет-ресурсе – это не обязательное условие, чтобы сайт попал под АГС. Принцип работы фильтра базируется на системе накопления – за каждое несоответствие стандартам поисковой системы фильтр АГС Яндекса начисляет баллы. А при достижении критического значения, сайт частично выпадает из индекса поисковой системы.

Качество поиска — это самый важный аспект для любой поисковой системы. Если она будет плохо искать, люди просто перестанут ей пользоваться.

2.5 Поисковая система Bing

Поисковая система Bing — одна из крупнейших в мире. В мировом поиске занимает 3 место; в России уступает Google, Yandex, Mail.ru, Rambler.

Имеет существенные возможности по сравнению с другими поисковиками: результаты поиска можно просмотреть на одной странице, также есть возможность динамического корректирование объёма информации, отображаемой для каждого результата поиска. Поисковик использует как длительную, так и временную персонализацию. Bing запоминает, что пользователь искал один раз, и что интересует его постоянно, а также анализирует периодичность запросов. [23, с.101]

В одном из своих патентов компания Microsoft приводит схему, как формируется поисковая выдача с учетом модуля классификатора веб-спама (для общего представления) показанного в соответствии с рисунком 15.

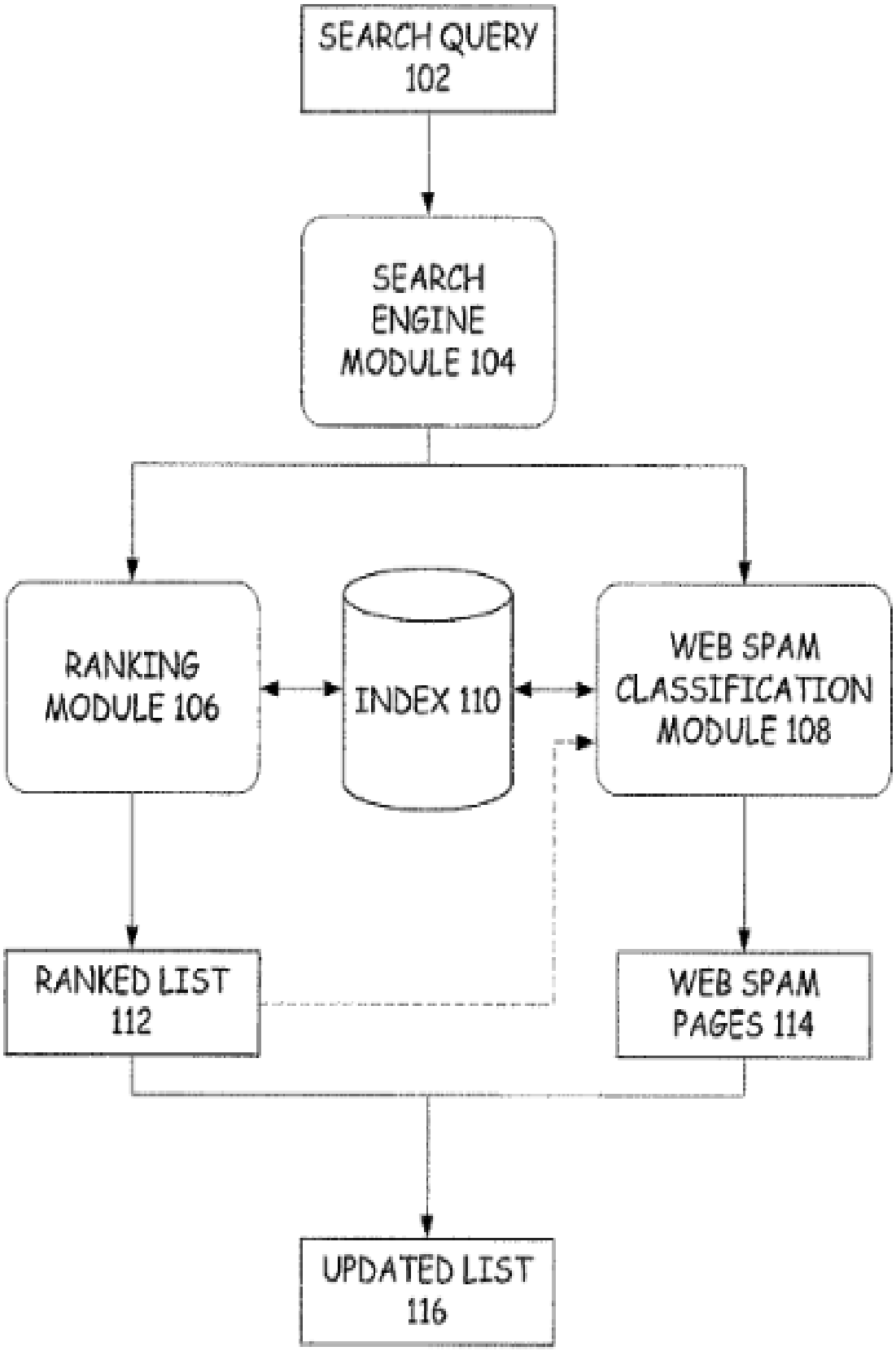

Рисунок 15 - Поисковая выдача с учетом модуля классификатора веб-спама

3 Анализ и сравнение качества поиска существующих веб-поисковых систем

3.1 Эксперименты с поисковыми запросами

Для того чтобы понять какой из поисковых систем следует отдавать предпочтении при поиске, был проведен ряд экспериментов позволяющих определить качество ранжирования сайтов в результатах поиска в таких крупных поисковых систем как Google, Bing, Яндекс, Mail.ru, Rambler, Yahoo.

Методика проведения эксперементов

Для того чтобы определить различные параметры поисковых систем было выбрано около 100 доменов в зоне .KZ из каждой категории сайта рейтинга http://zero.kz/. Исходя из типа анализа к каждому домену были подобраны ключевые запросы. Затем с помощью софта для каждого домена и поисковой системы был проведен анализ, благодаря результатам которого были построены графики.

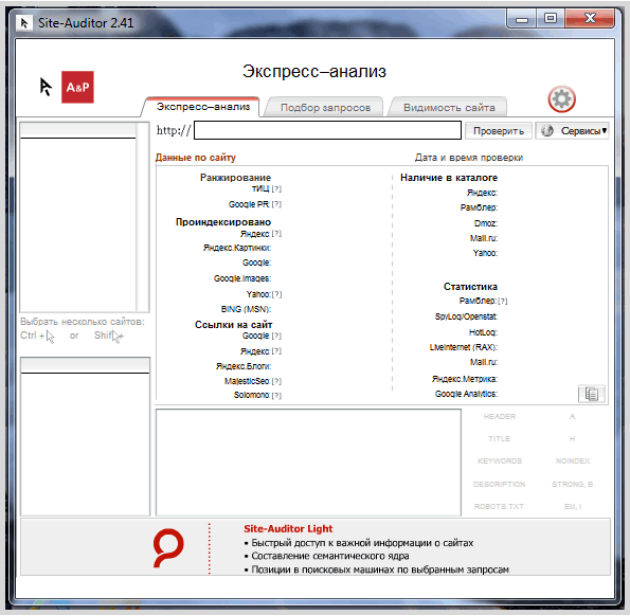

Рисунок 15 – Программа Site-Auditor 2.41

Программы для проведения анализа

Site Auditor

Подбор ключевых слов и видимость сайтов по ним было проведено с помощью программы SiteAuditor.

SiteAuditor создала, постоянно обновляет компания «Ашманов и партнеры». Она является совершенно бесплатной, оказывая огромную помощь в оптимизации и продвижении проекта.

Эта программа позволяет проверить позиции ресурса, а также следить за различными seo-показателями, например тИЦ и ПР сайта, Alexa Rank.

Следовательно, вы получаете очень удобный инструмент, который поможет оптимизировать страницы для завоевания более высоких позиций поисковой выдачи.

При помощи утилиты Site-Auditor можно быстро собрать данные, необходимые для оценки видимости сайта в поисковых системах. Одним кликом можно получить все необходимые данные из основных поисковых сервисов, используемых в российском сегменте Интернета, а именно: *Яндекс*, *Рамблер*, *Апорт*, а также в наиболее популярных международных сервисах *Google* и *Yahoo*.

Утилита позволяет получить позиции сайта в выдаче Яндекс, Рамблер и Google по выбранному списку запросов (семантическому ядру). Все собираемые данные сохраняются на вашем компьютере и позволяют вам отслеживать историю изменений.

App.KeyGrabber

Утилита для подбора ключевых запросов через Яндекс wordstat. Позволяет после авторизации в Яндекс с помощью API подбирать ключевые слова.

AllSubmitter

В подборе ключевых слов и анализа видимости сайтов также помогала программа AllSubmitter представлена в соответствии с рисунком 16.

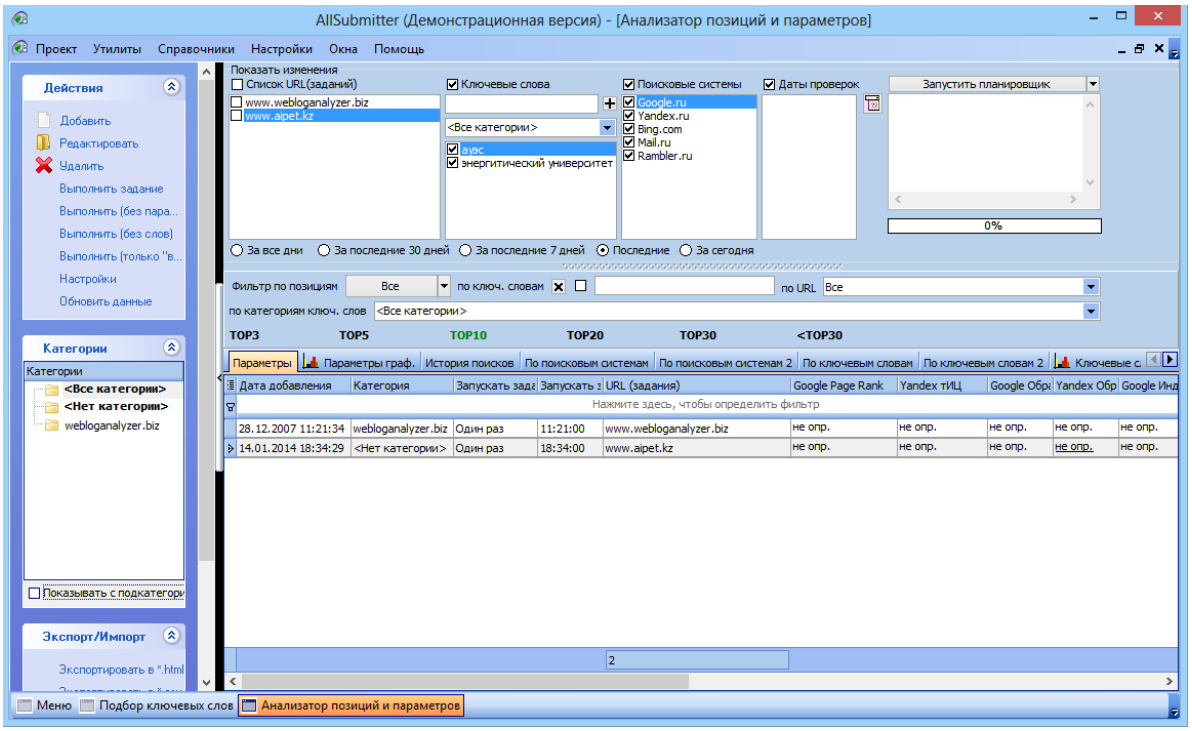

Рисунок 16 – Программа AllSubmitter

AllSubmitter содержит множество параметров для анализа ключевых слов. Которые помогают оценить сложность продвижения. Все проверки, очистки происходят, одним кликом мышки.

3.1.1 Анализ навигационного поиска

Навигационными называются запросы, направленные в первую очередь на поиск конкретного сайта, файла или страницы. Таковы запросы, содержащие названия компаний и организаций (например, [казахмыс] или [сайт акимата алматы]), печатных изданий и сайтов (например, [tengrinews] или [алматыжарнама]), указывающие на конкретную страницу или раздел сайта (например, [салон красоты на абая 187]). Навигационными часто оказываются и запросы, связанные с именами или псевдонимами людей, имеющих персональные сайты или ведущих блоги.

Разумеется, навигационные запросы могут иметь и другие смыслы. Задавая запрос [мдв] или [Роза Рымбаева], пользователь может искать и независимую информацию о структуре или человеке — и всё-таки их официальные сайты, безусловно, должны присутствовать в выдаче поисковой машины, занимая там высокие позиции.

Навигационным называется запрос, с помощью которого пользователь ищет определенный сайт. Таковы, например, запросы "казкомерцбанк", "караван", "тудей кз", "газета кз" и т.п.

Лучшим результатом в ответ на навигационный запрос является искомый сайт (маркер) на первой позиции поисковой выдачи.

Для оценки качества навигационного поиска производился опрос поисковиков по 100 запросам, случайно выбираемым из общего массива навигационных запросов (более 100). Каждому запросу был приписан один или несколько сайтов-маркеров. (Несколько сайтов-маркеров возникает обычно в тех случаях, когда сайт доступен по нескольким веб-адресам: может, например, открываться несколько зеркал одного и того же сайта, либо один из сайтов может осуществлять редирект на другой.)

В Топ-10 поисковой выдачи искались вхождения сайтов-маркеров. Если для запроса задано несколько маркеров — учитывалось вхождение любого из них. Далее рассчитывалась доля запросов, для которых на первой странице результатов поиска найден сайт-маркер. Эта доля (в процентах) и служит сводным показателем качества навигационного поиска.

Согласно полученным данным был построен график, чем выше показатель, тем лучше, показан в соответствии с рисунком 17.

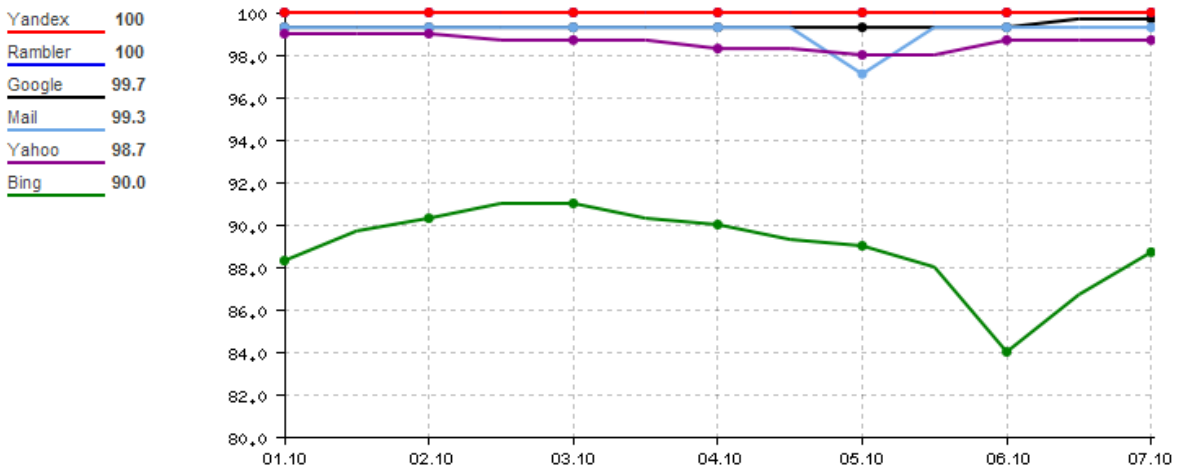

Рисунок 17 - Показатель качества навигационного поиска

Лучшим результатом в ответ на навигационный запрос является искомый сайт (маркер) на первой позиции поисковой выдачи. Из графика можно сделать вывод что максимально качественный навигационный поиск у Яндекса, Рамблера (видно что Рамблер использует выдачу Яндекса т.к. результаты идентичные) и Google с Mail.ru.

3.1.2 Анализ скорости индексации сайтов

Насколько быстро новые страницы сайтов попадают в поисковую выдачу?

Было отобрано около 100 сайтов, которые сами сообщают поисковым системам о том, что они обновились через формальную «карту сайта», sitemap.xml. Все эти сайты хорошо известны поисковикам: убедились в том, что их случайные страницы во всех или в подавляющем большинстве поисковых систем попадают в «десятку» по запросам, совпадающим с их заголовками (тег <title>). А вот насколько быстро начинают находиться по соответствующим им запросам новые страницы?

В течении 7 дней была выкачена с отобранных сайтов карта сайта, сравнивали ее с сохраненной версией и смотрели, какие новые страницы появились (с каждого сайта берется не более 3 страниц в день). То же самое, естественно, могут делать и поисковые системы — для этого файлы sitemap.xml и создаются.

Из адреса и заголовка найденной таким образом новой страницы сайта формируется маркер, который будет активен 30 дней. В каждые 5 из этих дней мы проверяли, находится ли страница в поисковых машинах, если взять в качестве запроса ее заголовок: а) при поиске в кавычках с ограничением по данному сайту (так можно оценить скорость индексации); б) при обычном поиске по Интернету (это позволяет оценить скорость попадания новых страниц в поисковую выдачу).

Основная оценка, которую выдает анализ — средняя «видимость» активных маркеров за месяц представлен в соответствии с рисунком 18.

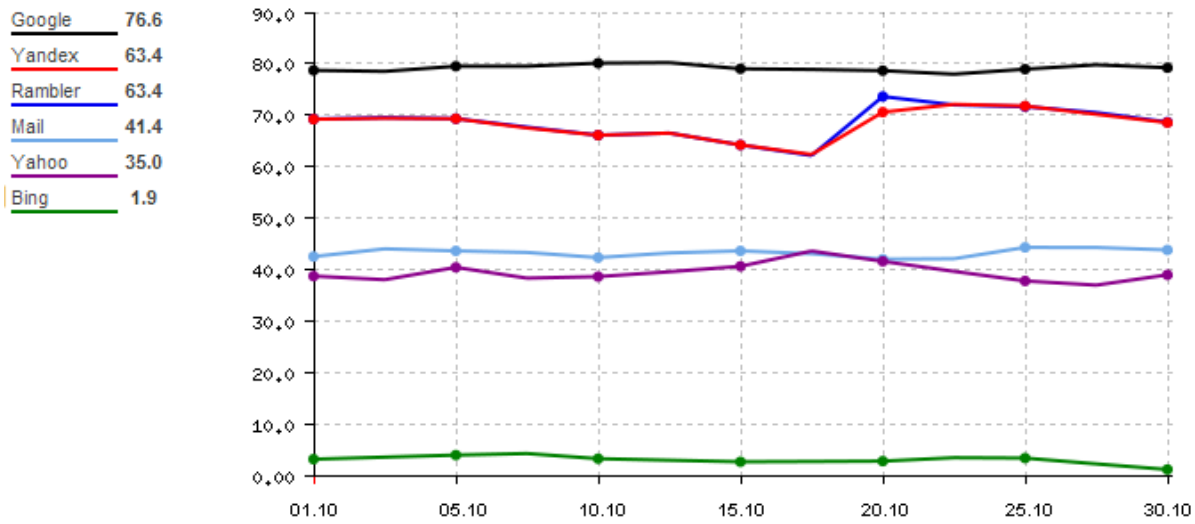

Рисунок 18 – Анализ скорости индексации сайтов

Из графика можно понять что самые шустрые роботы индексаторы у Google. Самое медленные, у Binga.

3.1.3 Анализ полноты индекса

Для качества поиска довольно важно, насколько полно индексируется Интернет поисковой машиной. Если поисковая база поисковика маленькая, то часть веб-страниц могут оказаться вообще недоступными через данный поисковик.

Конечно, по большинству частотных и среднечастотных запросов любая поисковая машина гарантированно выдаст вам десятки или сотни тысяч результатов. При этом, однако, реальный размер базы у разных поисковиков может отличаться в несколько раз.

Как понять истинный размер базы? Это непросто и вот почему:

а) Нельзя доверять декларации поисковиков "всего найдено N документов", поскольку у каждого поисковика количество найденного подсчитывается по-своему. Кто-то считает только документы, где есть все заданные в запросе слова, а кто-то дописывает в "хвост" вообще все документы. где встретилось хотя бы одно слово из запроса. Поисковики любят мериться мускулами и не чужды маркетинговых трюков.

б) Более того, в действительности ни один поисковик никогда не покажет вам все заявленные 10 000 результатов. Листая результаты поиска, Вы сможете увидеть десятки, в лучшем случае сотни страниц результатов поиска, после чего вас так или иначе отключат.

Таким образом, можно быть уверенным, что мы знаем истинное количество страниц в Интернете с данным словом, только когда результатов – максимум сотня-другая.

в) доверять декларации поисковика о том, что им заиндексировано М миллиардов документов – тоже нельзя, потому что, опять-таки способ подсчёта неизвестен (считаются ли дубли, например, и какие именно), а также потому, что может оказаться, что половина Рунета всё-таки оказалась вне этих условных М миллиардов страниц и не попала в индекс вообще.

Поэтому был выбран наиболее простой и надёжный способ анализа полноты поисковой базы по редким запросам.

Редкими здесь называем запросы, которые встречаются в Интернете всего несколько десятков раз. Впрочем, обычно такие запросы – редкие и в смысле частоты поиска их в поисковиках, поскольку если запрос – частотный, то под него моментально подстраиваются спамеры, производя множество фальшивых страниц с этим словом или словосочетанием.

Был построен набор редких запросов, по каждому из которых поисковики выдают не более нескольких десятков результатов. Проверив "окраины Интернета" на предмет того, видит ли их каждая из поисковых машин. Для этого раз в сутки проверялись все поисковики на предмет того, сколько страниц найдёно каждым поисковиком по каждому запросу.

После чего был вычислен показатель полноты базы по следующей формуле: Проводится поиск по запросу во всех сравниваемых поисковых машинах. Для машины, в которой найдено максимальное число страниц, полнота считается равной 1, для остальных – определяется пропорционально. Затем полнота усредняется по множеству специфических запросов. Это и есть значение полноты индекса представлено в соответствии с рисунком 19.

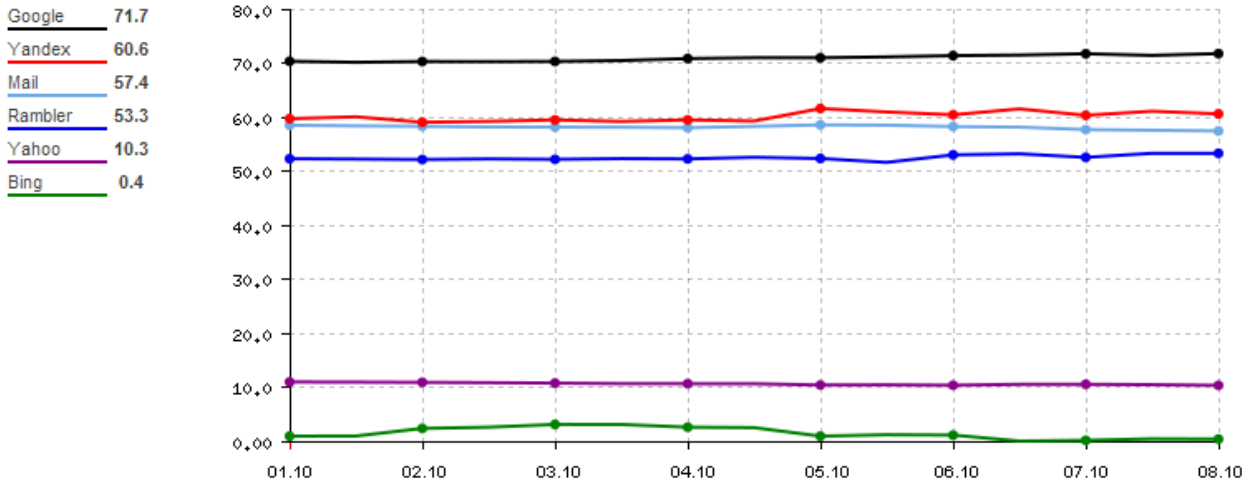

Рисунок 19 - Анализ полноты индекса

3.1.4 Анализ скорости выдачи результатов

Поиск — сервис, которым пользуются миллионы людей. И у каждого из них от запроса до получения страницы с результатами поиска должен проходить минимум времени — одна-две секунды, а лучше — доли секунды. Чуть медленнее — и капризный пользователь обидится, а если ситуация будет повторяться, предпочтет другой поисковик, благо есть из чего выбирать.

Этот анализ показал, насколько быстро результаты поиска «прилетают» к пользователям.

Разумеется, скорость поиска зависит и от «погоды» в Сети: на результаты по любому поисковику в любом городе в любой конкретный день может влиять загруженность канала, качество работы провайдера и многие другие факторы. Однако если медленные ответы у какого-то из поисковиков случаются чаще, чем у других, и повторяются изо дня в день — это уже повод задуматься.

Основная метрика анализа — среднее время, затрачиваемое поисковиками на выдачу результатов пользователям, показан в соответствии с рисунком 20. Чем ниже, тем лучше.

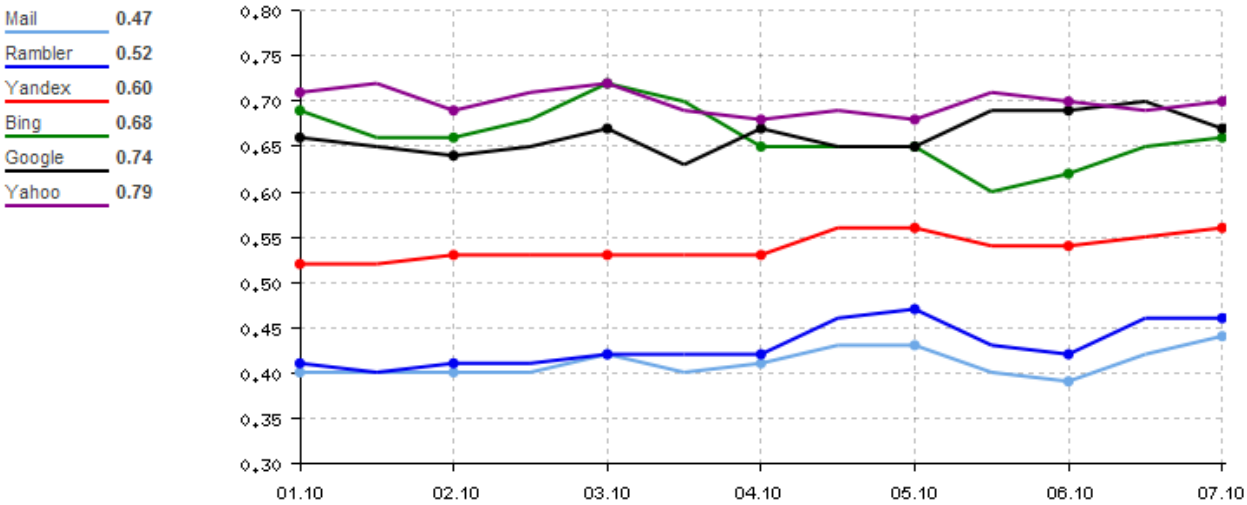

Рисунок 20 – Анализ скорости выдачи результатов

3.1.5 Анализ раздражающих факторов

Как бы хорошо ни работал поисковик, есть детали, которые могут сильно испортить пользователю впечатление от качественных результатов и заметно повлиять на его лояльность сервису. К таким, безусловно, относятся опасности заражения вирусами, раздражающие рекламные баннеры, поисковый спам и т.д.

Конечно, количество рекламы или опасных скриптов на сайтах не зависит от поисковиков. Однако концентрация нежелательного содержимого в поисковой выдаче – полностью в их компетенции. Соответственно, в их

интересах ранжировать сайты с большим количеством раздражающих факторов заметно ниже качественных и безопасных.

3.1.5.1 Анализ спама в поисковой выдаче

Поисковым спамом могут быть названы текст, ссылка, технология, программный код и иные элементы веба, созданные не для повышения удобства пользователей, целью которых является быстрый и легкий поиск полной, профессионально оформленной, достоверной информации, определяемой поисковым запросом. Но для которых точно известно, что причиной их создания является желание веб-мастера улучшить позиции продвигаемого сайта в результатах поиска.

Были рассмотрены следующие категории поискового спама: дорвеи, увод пользователя к другим страницам, спамерские каталоги, спамерский заимствованный контент, сайт, маскирующийся под корпоративный (псевдофирма), каталоги, доски объявлений, домены на продажу, вторичный, заимствованный контент, любые партнерки, сайт линкоподдержки, заспамленный форум.

В качестве эксперимента было просмотрено Top10 выдачи по выборке поисковых запросов, отмечая сайты, которые, с точки зрения обычного пользователя, содержат элементы поискового спама. Сводные данные были сведены в график, который показан в соответствии с рисунком 21.

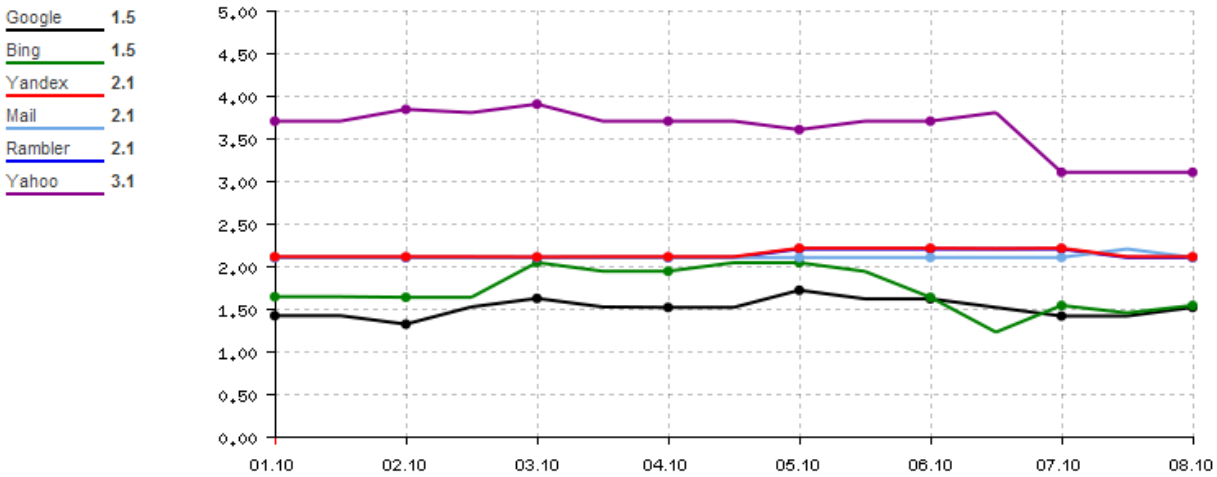

Рисунок 21 – Анализ спама в поисковой выдаче

Содержание спама в поисковых системах за пол месяца (ниже, лучше).

Как видно из графика можно сделать вывод что поисковые системы Google и Bing содержат минимальное количество поискового спама. Поисковая система Mail.ru несмотря на использование выдачи Google, как говорилось выше не использует их фильтры, по этой причине спама в выдаче

поисковика намного больше. Яндекс и Рамблер имеют одни и те же данные показывая еще раз что Рамблер полностью использует выдачу Яндекса.

3.1.5.2 Анализ рекламной нагрузки в поисковой выдачи

Сайты с одинаковым или похожим содержанием могут сильно различаться по количеству рекламы. В таком случае для пользователя предпочтительнее тот поисковик, который поднимает в поисковой выдаче сайты с меньшим количеством рекламы и избегает рекомендовать пользователям сайты с наиболее навязчивой рекламой.

В основе анализа лежит специально разработана технология распознавания рекламы, основанная на оценке скриптов, iframe'ов и других элементов кода страницы, которые могут вызывать рекламу. Для каждой страницы из результатов поиска вычисляется ее рекламная нагрузка, учитывающая «веса» всех найденных на странице рекламных элементов. Показатель каждого поисковика в данном анализе — это средняя рекламная нагрузка всех найденных страниц с учетом позиционных коэффициентов.

Рекламная нагрузка каждой страницы складывается из «весов» всех рекламных элементов на этой странице. Блоки контекстной рекламы и небольшие немигающие баннеры получают вес 1. Тизеры и крупные, мигающие баннеры получают вес 3. Некоторые рекламные элементы заставляют пользователя обязательно кликнуть по ним — хотя бы для того, чтобы их закрыть. Сюда относятся так называемые кликандеры (дополнительные окна, которые открываются при клике в любом месте на странице) и баннеры, закрывающие контент основной страницы и прокручивающиеся вместе с ней. Такие рекламные элементы получают вес 9. Наконец, если пользователю приходится совершить более одного клика, рекламный элемент получает 18 баллов.

Для каждой позиции в результатах поиска рекламная нагрузка сайта умножается на специальный коэффициент, который отражает вероятность клика на данной позиции — от 1,5 у первой до 0,4 у десятой. Позиционные коэффициенты подобраны так, что в среднем они дают единицу — таким образом, если «зарекламленные» сайты распределены по результатам поиска равномерно, то средняя рекламная нагрузка результатов поиска на большом массиве запросов от использования коэффициентов не изменится.

Для анализа подобраны запросы на тему музыки, видео и программного обеспечения, поскольку именно по таким запросам обвешенные рекламой сайты находятся особенно часто, и это нередко создает для пользователей реальную проблему.

Таким образом, если поисковая машина получила, скажем, рекламную нагрузку 12, это означает, что на странице, которую пользователь откроет из результатов поиска по «мультимедийным» запросам, в среднем окажется один кликандер или скролируемый баннер и один тизер или крупный нескроллируемый баннер. При этом необходимо учитывать, что у конкретного пользователя могут выводиться не все рекламные элементы,

давшие вес странице по результатам анализа ее кода: часть рекламы может отсекаться браузерам, какие-то рекламные модули выводятся только однократно, и т. п.
Все собранные данные были собраны в таблицу, по которым и был построен график, в соответствии с рисунком 22. Чем ниже, тем лучше.

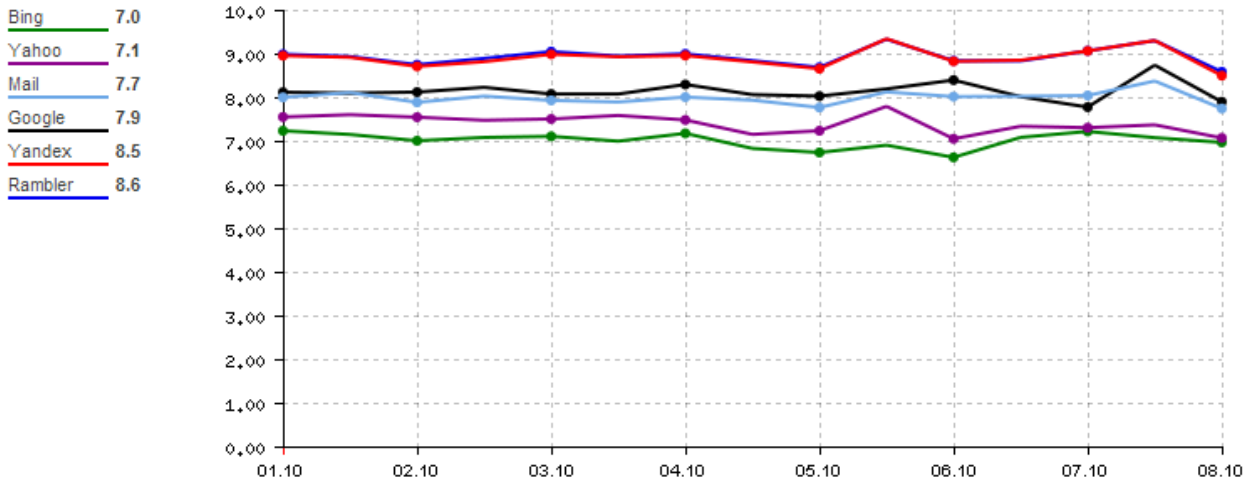

Рисунок 22 – Рекламная нагрузка на сайты

Из графика видно что наличие рекламы меньше всего в не особо популярных поисковых системах. Также можно сделать вывод что выдача Google без фильтров в Mail.ru дает пользователю меньшее число рекламы.

3.1.6 Анализ вирусной угрозы в поисковой выдаче

Поисковики стараются по возможности оградить пользователя от сайтов, на которых он может «подхватить» вирус или стать жертвой мошенников. Насколько эффективна эта защита?

Данный анализ показал, какой процент из найденных сайтов содержит хотя бы одну возможную угрозу. При оценке угроз на сайте учитывались вирусные скрипты и фреймы. Кроме того, учитывались широко распространенная реклама, при клике на которую пользователю предлагается загрузить файл, который впоследствии используется для мошенничества.

Количество и виды угроз на сайте (будь то 1 или 5) показывает, но не влияет на общую оценку сайта. Вероятность того, что некий фрагмент кода содержит вирус, оценивается индивидуально на основе данных антивирусных программ, отзывов вебмастеров и других источников. В этом анализе учитывались все виды угроз.

Некоторые поисковики умеют выводить в результатах поиска предупреждения о том, что найденный сайт может представлять угрозу.
Однако в силу специфики запросов, используемых в данном анализе, такое предупреждение может быть недостаточным. Данный анализ использует запросы на тему скачивания, прослушивания и просмотра музыки и видео -

те же, что и анализе рекламной нагрузки. По таким запросам существуют тысячи сайтов, распространяющие примерно одинаковый контент. В этой ситуации представляется правильным не просто предупреждать пользователя о возможной угрозе но и понижать зараженный сайт в выдаче, чтобы на его месте выдавался незараженный, показано в соответствии с рисунком 23. Чем ниже, тем лучше.

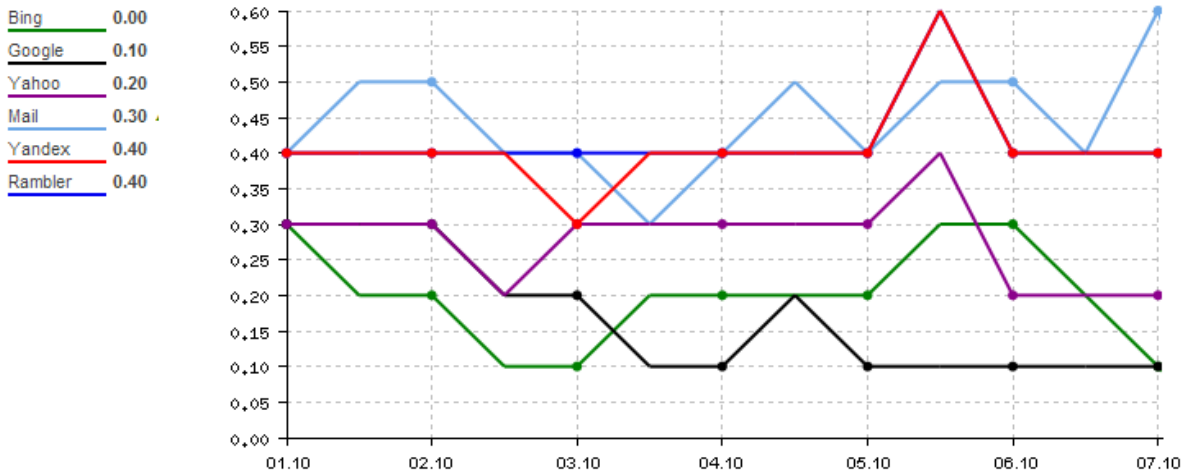

Рисунок 23 - Анализ вирусной угрозы в поисковой выдаче

3.1.7 Анализ качества снипетов

Качество снипетов является важной составляющей общего качества поиска: именно по снипетам пользователь определяет, есть ли смысл открывать ту или иную страницу выдачи поисковика, и ситуации, когда снипет плохо отражает содержание страницы, вызывают ощутимое раздражение. Напротив, хорошие снипеты помогают пользователю быстрее найти то, что нужно, т.е. решить задачу, поставленную поисковику.

Помимо упомянутого соответствия снипета содержимому страницы, являющемуся основным критерием оценки, на оценку так же влияет качество текста снипета (фразы, оборванные на середине или просто отдельно стоящие слова снижают оценку) и его состав (наличие в снипете html-кода, малопонятных списков из меню страницы и т.п. не добавляют снипету наглядности). Снипет, состоящий только из заголовка, также оценивается невысоко.

Каждый снипет получает оценку по пятибалльной шкале от 0,2 до 1, эти оценки усредняются по всей выдаче, после чего полученные оценки усредняются по всем запросам, полученное число и представлено в сводной таблице, в соответствии с рисунком 24.

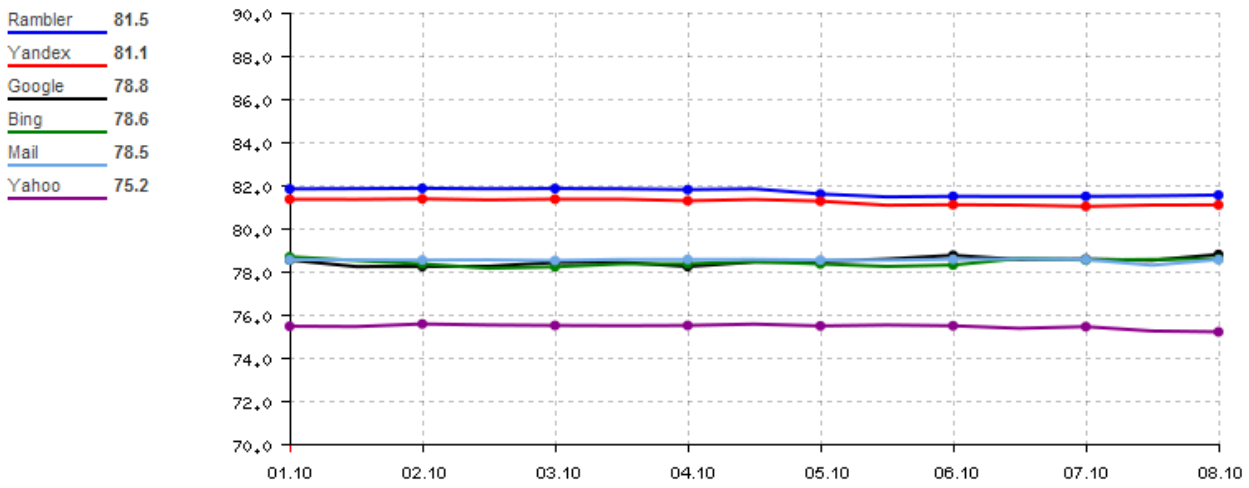

Рисунок 24 – Анализ качества снипетов

3.1.8 Анализ качества подсказок

Большинство поисковых машин, заподозрив при вводе поискового запроса опечатку, пытаются предложить пользователю правильное (по мнению поисковика) написание запроса. Качество подсказок является важным дополнением к общему качеству поиска.

В данном анализе искалась правильная подсказка в поисковой выдаче по запросу с заведомой опечаткой и оценивалось количество случаев, когда подсказка содержит "правильную" форму запроса.

Чем больше правильных подсказок было дано, тем выше показатель данного поисковика. Именно в этом порядке сверху вниз отсортированы поисковики таблице данного анализа в соответствии с рисунком 25.

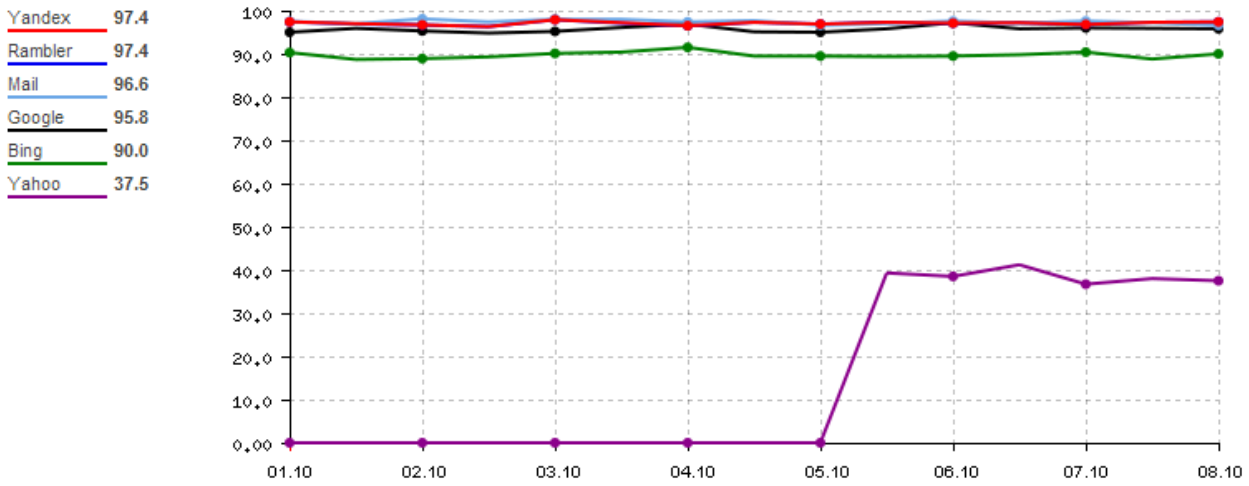

Рисунок 25 – Анализ качества подсказок

3.1.9 Анализ устойчивости к опечаткам

Человек – не робот и может ошибиться. В том числе и при вводе поискового запроса в поисковую строку поисковой машины. Он может просто допустить опечатку, введя соседний символ (и тогда вместо "запрос" получится "звпрос"), он может по ошибке ввести символ дважды или пропустить вовсе (и тогда вместо "запрос" получится "зпрос" или "заппрос"), наконец, он может не знать правильного написания слова и ввести его "как слышится" (и тогда вместо "яндекс" получится "яндыкс").

Поисковик в этом случае может придерживаться одной из трех стратегий:

1) никак не обрабатывать эти случаи и искать строго то что ввел пользователь;

2) понять что имела место опечатка, тем не менее найти то что введено и рядом предложить правильную по мнению поисковой машины форму – "возможно, вы имели ввиду [правильное написание запроса]";

3) понять что имела место опечатка, не искать ошибочное написание а сразу искать правильную форму.

В зависимости от выбранной стратегии, пользователь либо не увидит, что он ошибся при написании, либо увидит это и сделает лишний клик (если захочет), либо даже не догадается о собственной опечатке.

В данном анализе сравнивалась поисковая выдача по "правильному запросу" и по нескольким формам возможных его опечаток. Оценивалась степень близости поисковой выдачи по запросу с опечаткой к выдаче по "правильному" запросу

Кроме сознательного исправления опечаток, совпадения могут возникать в четырех случаях:

1) случайно;

3) страница содержит как правильную форму, так и форму с опечаткой;

4) ложное срабатывание морфологии поисковика (например, приведение неизвестного поисковику слова "гриби", которое есть опечатка слова "грибы" к слову "гриб");

5) продвижение одних и тех же сайтов как по запросам в правильном ниписании, так и по запросам с опечатками.

Все эти случаи в рамках данного анализа дают шум – случайное совпадение результатов.

Степень близости оценивалась аналогично тому как это делась в анализаторе апдейтов, только сравнивались другие запросы.

Чем большее совпадение результатов было зафиксировано, тем выше результат поисковика по данному анализу. Именно в этом порядке сверху вниз отсортированы поисковики в таблице данного эксперимента, в соответствии с рисунком 26.

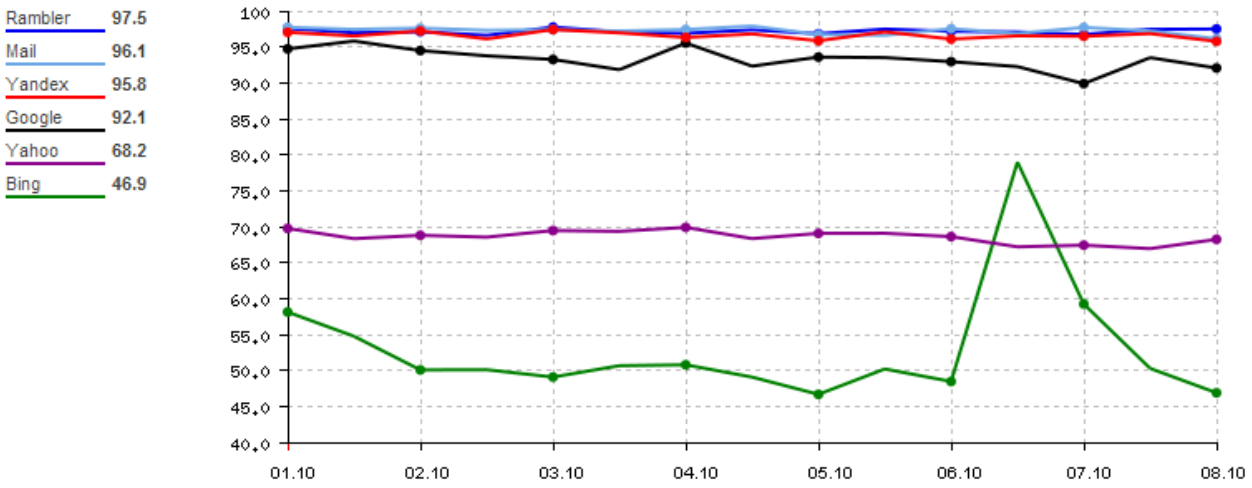

Рисунок 26 – Анализ устойчивости к опечаткам

3.1.10 Анализ запросов-синонимов

Один и тот же вопрос можно задать десятком способов. Например, для пользователей запросы "как узнать адрес по номеру телефона", "поиск адреса по номеру телефона" и "найти адрес по номеру телефона" имеют одинаковый смысл, это запросы-синонимы.

Запросы-синонимы появляются по разным причинам:

- при использовании общепринятых сокращений – "погода в Санкт-Петербурге" и "погода в Спб";
- использовании транслитераций – "характеристики toyota camry" и "характеристики тойота камри";
- использовании разных падежей – "сценарий встречи Нового Года" и "сценарий встреча Новый Год";
- использовании перестановок слов – "ремонт АКПП", "АКПП ремонт";
- использовании жаргонизмов – "скачать мультфильмы", "скачать мультики";
- использовании слов, не несущих смысловой нагрузки – "быстрый интернет", "супер быстрый интернет";
- использовании синонимов слов, входящих в запрос – "грустные стихи", "печальные стихи".

Поисковые системы в ответ на запрос должны подобрать наилучший набор ссылок. И, конечно, чаще всего в результатах поиска люди ожидают увидеть информацию, соответствующую смыслу, а не букве запроса. Из этого следует, что выдача в ответ на синонимичные запросы должна быть одинаковой, несмотря на то, что формулировки отличаются друг от друга.

Анализ поиска по запросам-синонимам проверил, насколько совпадает выдача по разным запросам, имеющим одинаковое значение. Образно можно сказать, что анализ в какой-то мере показал способность поисковика "понимать" смысл запроса. В таблице поисковые системы отсортированы по убыванию этого показателя, в соответствии с рисунком 27.

Все примеры, используемые в данном эксперименте, реальные, они получены с помощью сервиса статистики поисковых запросов "Рамблера" (http://adstat.rambler.ru/wrds/).

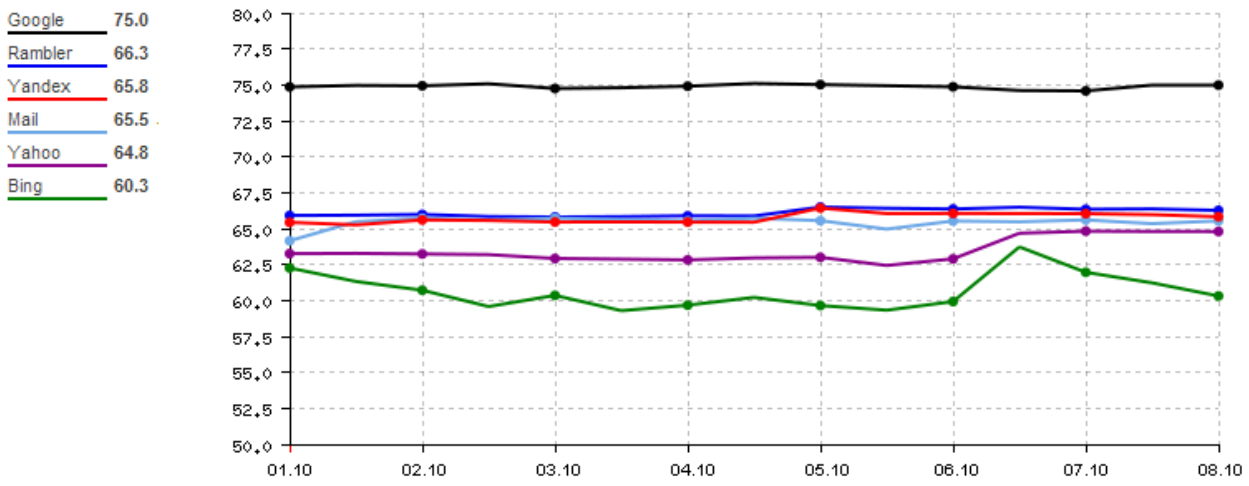

Рисунок 27 – Анализ запросов-синонимов

3.1.11 Анализ различия форм слов

Поисковики должны уметь находить слова во всех формах. Но иногда формы двух слов очень похожи или просто совпадают — и в таких случаях важно ограничить расширение запроса, не найти лишнего. Человека, задающего запрос про двигатели на паромах, вряд ли заинтересует мойка двигателя паром, а спрашивающего про компанию Роскофе — отели в городе Роскоф.

Название «Грамматика» — достаточно условное: речь идет о способности поисковых машин различать похожие грамматические формы разных слов.

В этом анализе были специально подобраны трудные запросы (некоторые из них — достаточно редкие), для которых существует заметная вероятность неверной интерпретации. Каждому запросу соответствовал набор маркеров, иногда довольно обширный, присутствие которых в снипете определяет, как именно, «в плюс» или «в минус» засчитывается поисковику соответствующий результат. Иногда это отдельные словоформы — например, слово «ливень» будет однозначным минусом для запроса про Авраама Ливни, а слово «куру» — плюсом для запроса про болезнь куру. Некоторые маркеры, однако, устроены сложнее и используют не только форму слова, но и контекст, так ответ «к походу готов» очевидно, не будет засчитан для запроса [походы готов].

Результатом поисковика в этом анализе является общее количество найденных им правильных форм — то есть, несмотря на принадлежность к группе «Ошибок», эксперимент определяет процент результатов, в которых ошибок нет, в соответствии с рисунком 28.

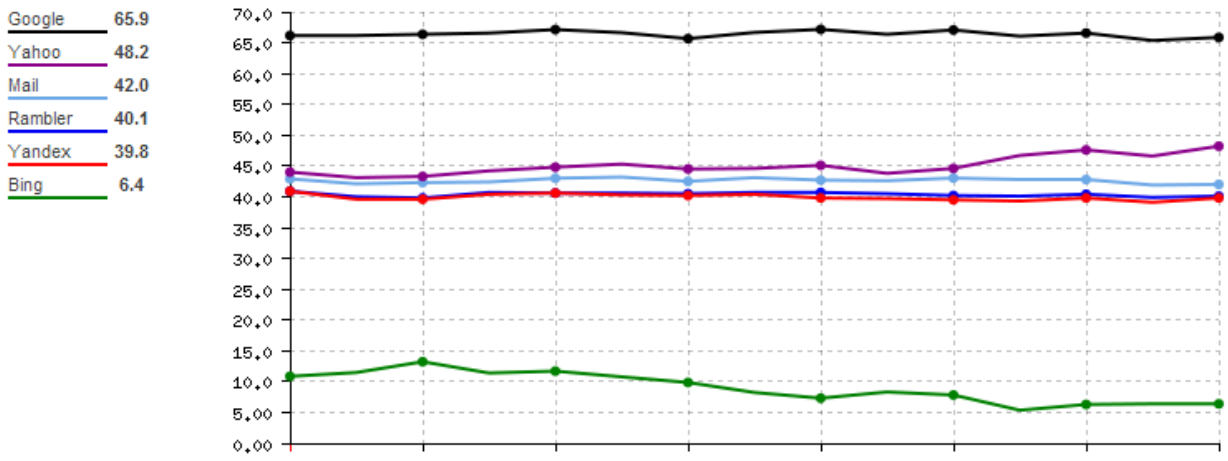

Рисунок 28 - Анализ различия форм слов

3.1.12 Анализ SEO-прессинга

Многие запросы невозможно понять однозначно. Например, дизайн, автомобили, спорт и др. Такие запросы называются информационными. Лучшим ответом на них является подборка, в которой пользователям предлагаются ссылки на ресурсы по различным смысловым направлениям запроса. Так, в выдаче в ответ на запрос "дизайн" должны быть ссылки на сайты о веб-дизайне, ландшафтном дизайне, дизайне интерьера и др.

Сформировать качественную политематическую подборку ссылок непросто. Особенно в условиях, когда оптимизаторы рассматривают популярные информационные запросы как цели для продвижения сайтов своих клиентов. В результате такого SEO-"прессинга" наверх пробиваются только ресурсы, продвижение которых наиболее окупается, и выдача становится однообразной, состоящей из ссылок на сайты с однотипными коммерческими предложениями.

Для этого анализа был поиск в титульных фразах и снипетах Top10 выдачи по анализируемым запросам однотипные строки. Суммарный показатель представляет собой процент найденных однотипных строк от общего количества сайтов в Top10 выдачи по анализируемым запросам. Чем выше этот показатель – тем более велико SEO-давление на данный поисковик.

В качестве признака монотематичности использовалась вхождение в заголовок или цитату страницы характерных слов или фраз. Сводным показателем является доля результатов поиска, в которых "фразы-маркеры" встречаются.

Лучшим считается поисковик с наименьшим сводным показателем по данному анализу. Именно в таком порядке поисковики отсортированы в таблице данного эксперимента, в соответствии с рисунком 29.

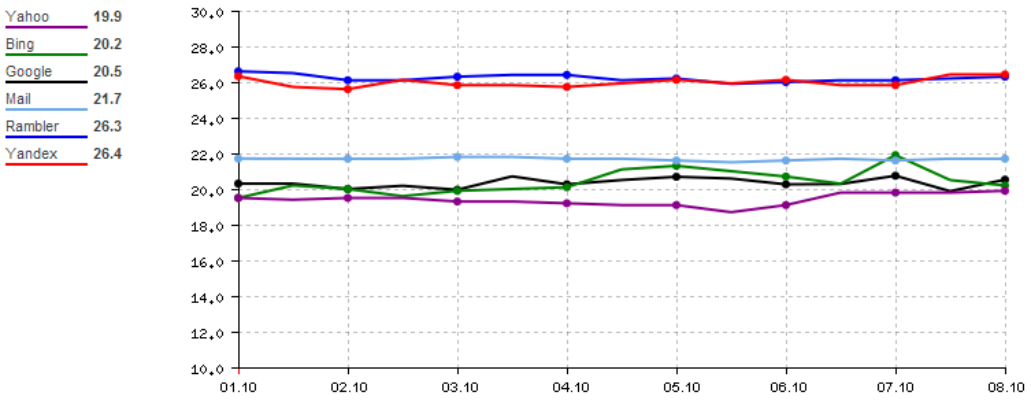

Рисунок 29 - Анализ SEO-прессинга

3.2 Результаты экспериментов

Интегральный показатель качества поиска призван помочь оценить общее качество работы каждого поисковика.

Для расчёта использовались результаты всех экспериментов, оценивающих качество отдельных аспектов поиска (не учитывались анализы: их результаты - не оценка, а информация).

При вычислении значения этого показателя результаты поисковиков в каждом из анализе были переведены в стобалльную шкалу нормированием на лучший в анализаторе результат. Это было сделано для того, чтобы учесть заметные различия значений показателей в каждом анализе. Так например, в анализе навигационного поиска все значения лежат в диапазоне 85-100, а в анализаторе спама — в интервале от 1 до 10.

Полученные значения умножены на коэффициенты, отражающие представления о значимости соответствующего вида или функциональности поиска в интернет-поиске в целом. После этого все значения суммируются и делятся на сумму коэффициентов, чтобы получить число от 1 до 100 — значение интегрального показателя качества поиска.

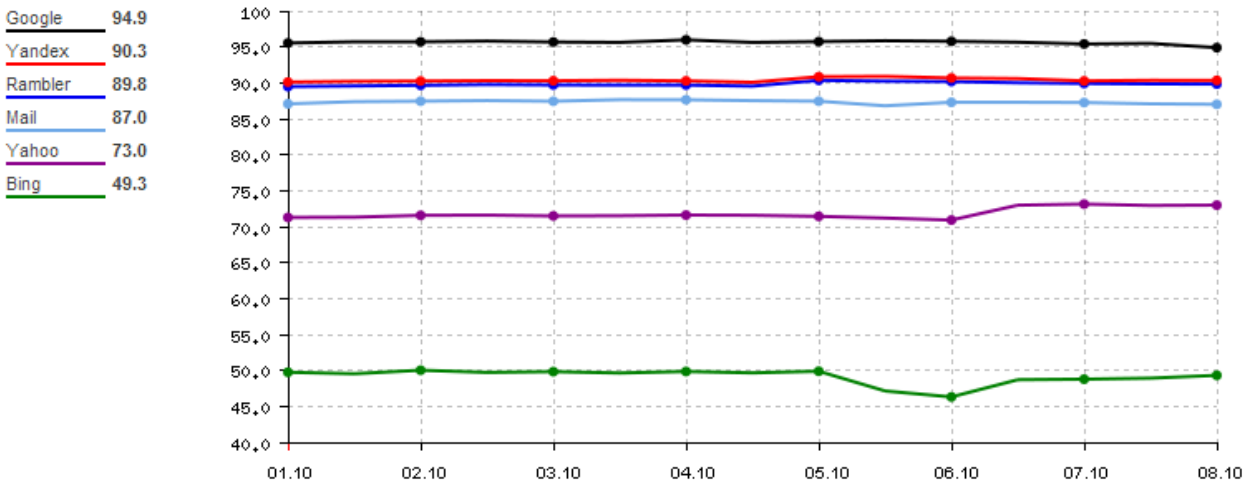

Рисунок 30 - Интегральный показатель качества поиска

4 Описание работы программного продукта

4.1 Работа с поисковой системой

Начало работы начинается с набора адреса поисковой системы www.ym.kz. Главная страница состоит из центрального блока, на котором расположилась поисковая форма с выбором типов поиска: Веб, Картинки, Видео, Новости, Блоги. При наборе поискового запроса появляются подсказки, при нажатии на которые сразу идет поиск. Есть кнопки добавления сайта в социальные сервисы и закладки. Далее после набора и нажатия кнопки “Найти” пользователь переходит на страницу результатов поиска. Также главную страницу поисковой системы “Умный поиск” можно сделать стартовой в браузере и добавить сайт в избранное.

Вкладки на главной страницы позволяют пользователю без перезагрузки страницы изменить тип поиска и искать нужный им вид данных. Вкладки находятся сразу над полем поискового запроса. Ссылка с Казахстанским флагом позволяет перевести интерфейс поисковой системы на казахский язык.

В соответствии с рисунком 31, представлено основное окно интерфейса поисковой системы.

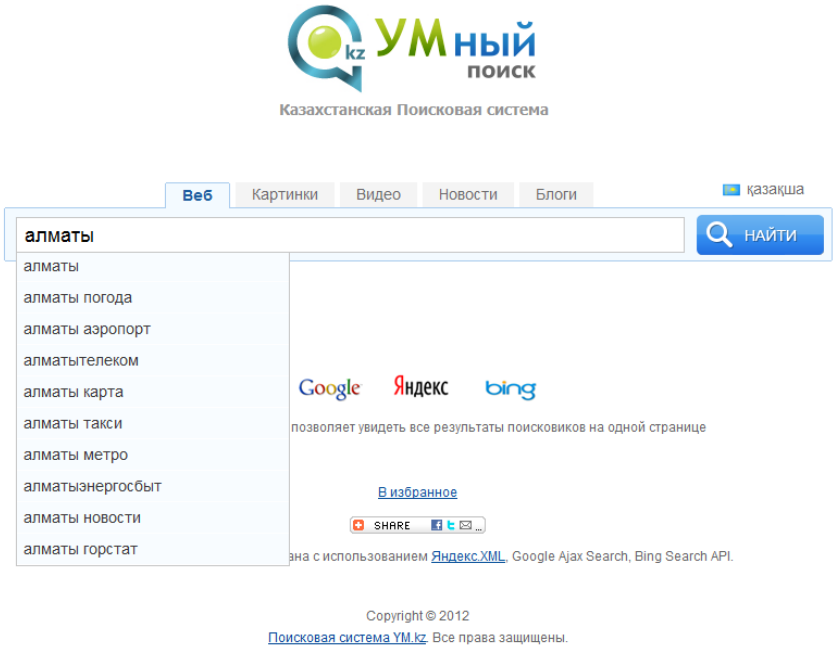

Рисунок 31 – Основное окно интерфейса

Отображение результатов происходит в реальном времени, с подзагрузкой данных, используя Ajax. На рисунке 32 предоставлена страница с результатами поиска по запросу “Алматы”.

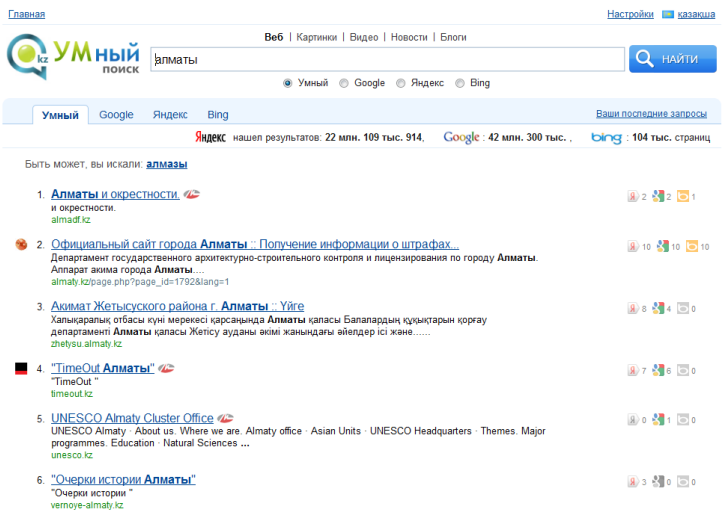

Рисунок 32 – Страница с результатами поиска

При нажатии на кнопку “Еще результаты” в реальном времени на страницу через Ajax подгружаются дополнительные результаты поиска. Вместе с этим страница автоматически прокручивается ниже, позволяя пользователю увидеть новые результаты. На рисунке 33 видна часть подгруженных данных при загрузке дополнительных результатов поиска.

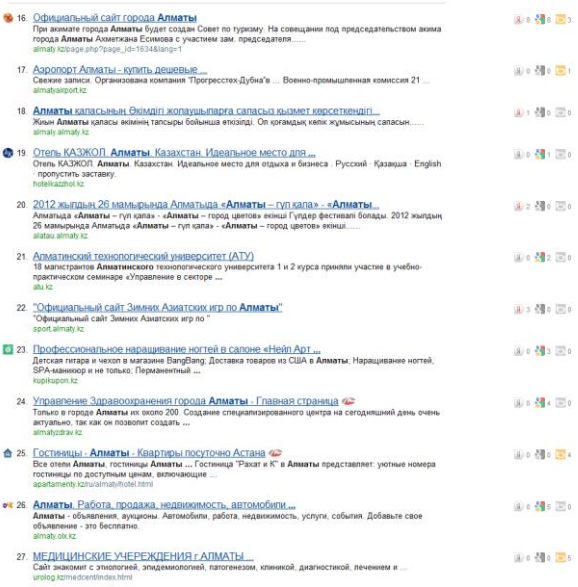

Рисунок 33 – Загрузка дополнительных результатов поиска

Если пользователь хочет изменить тип поиска, нужно выбрать его выше поискового поля. Можно изменить тип поиска, например на “Картинки”. Пример поиска по запросу “Алматы” в разделе картинки предоставлен в соответствии с рисунком 34.

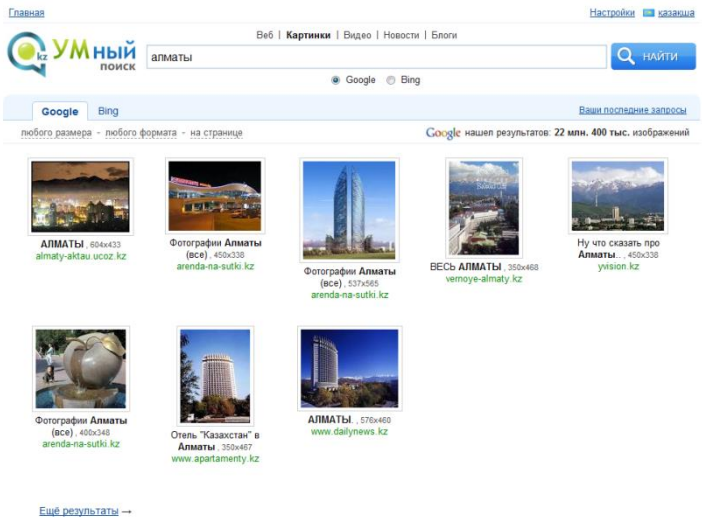

Рисунок 34 – Страница результатов поиска
по картинкам

Если же пользователь выберет тип поиска “Видео”, то будет произведен поиск по видеороликам. При этом структура отображения результатов меняется в зависимости от типа поиска. На рисунке 35 видна страница поиска по видеороликам.

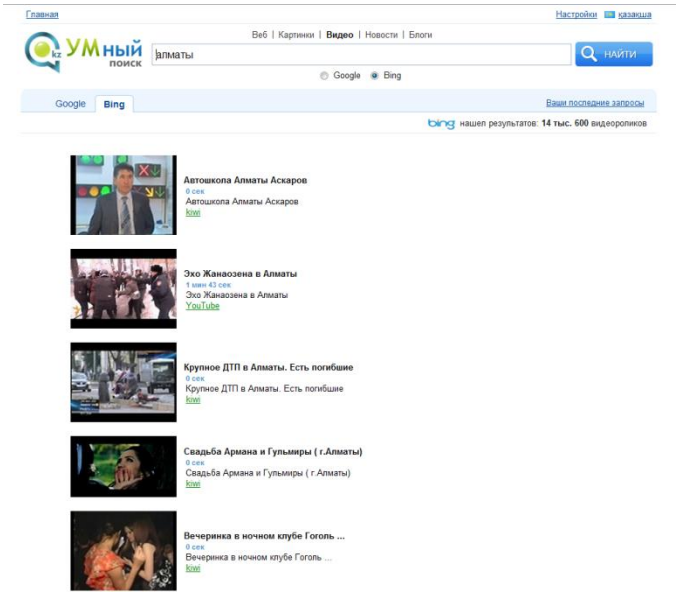

Рисунок 35 – Страница результатов поиска
по видеороликам

При выборе типа поиска “Новости” пользователь получит самые свежие результаты Казахстанских новостных агентств. Например по запросу “Алматы” мы видем данные таких источников как “ Today.kz”, “ Tengri

News”, “ КазИнформ”, “ Zakon.kz”, “ GAZETA.KZ” и других. На рисунке 36 предоставлена страница результатов поиска по новостям.

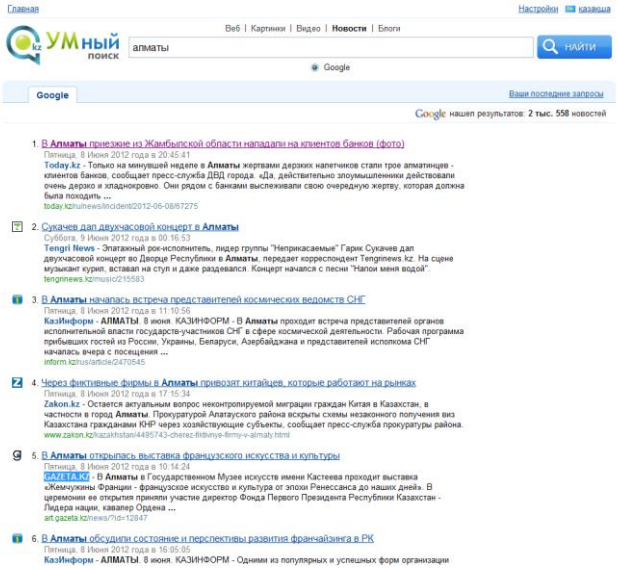

Рисунок 36 – Страница результатов поиска
по новостям

Если же пользователь выберет тип поиска “Блоги” то соответственно получит результаты поиска по Казахстанским блогам. На рисунке 37 показана страница результатов поиска по блогам.

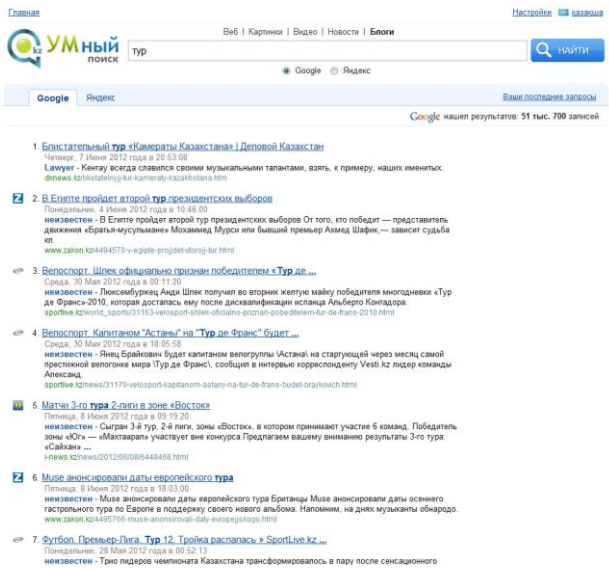

Рисунок 37 – Страница результатов поиска
по блогам

Рассмотрим блок слева, который открывается при нажатии на ссылку “Ваши последние запросы”. Данный блок содержит историю поиска пользователя и состоит из ссылок:

– Закрыть, ссылка при нажатии на которую скроется данный блок.

– Очистить, ссылка которая очистит и скроет блок последних запросов.

На рисунке 38 показан скриншот блока “Ваши последние запросы”.

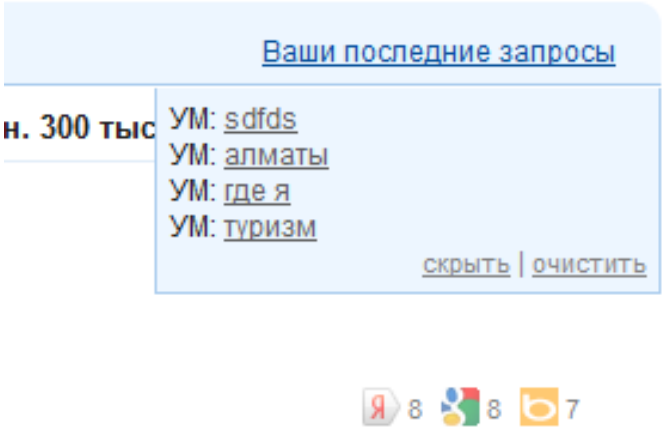

Рисунок 38 – Блок “Ваши последние запросы”

Если нажать ссылку вверху страницы “Настройки”, то откроется маленькое окошко по настройкам персонализации выдачи. А именно включение и отключение подсказок к набору запросов и автоматическому скроллингу страницы. На рисунке показано 39 окно настроек.

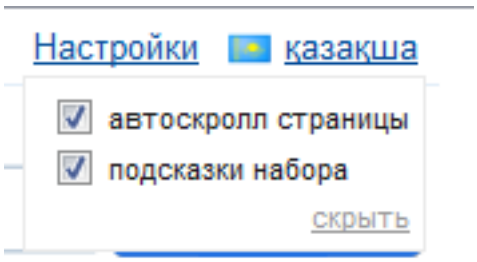

Рисунок 39 – Окно настроек

Измененные параметры сохраняются в cookie браузера пользователя, благодаря чему при последующей работе с поисковой системой “Умный поиск” не придется снова заходить и изменять настройки в окне настроек.

Единственный минус такого решения в том, если пользователь будет пользоваться разными браузерами.

Если нажать на черный текст результата когда текст отличен от казахского и русского, появляется ссылка “Перевести результат на русский”, при нажатии на которую без перезагрузки страницы идет перевод только данного результата поиска. На рисунке 40 показан мгновенный перевод результатов поиска.

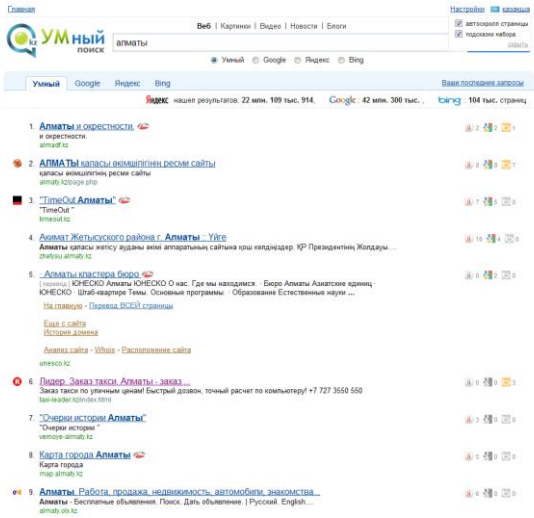

Рисунок 40 – Мгновенный перевод результата поиска

Если необходим поиск только по определенному домену, необходимо нажать на зеленую часть домена в адресе страницы в выдаче. После это без перезагрузки страницы произойдет загрузка и мгновенное отображение результатов поиска только этому домену. На рисунке 41 показан поиск по нужному домену.

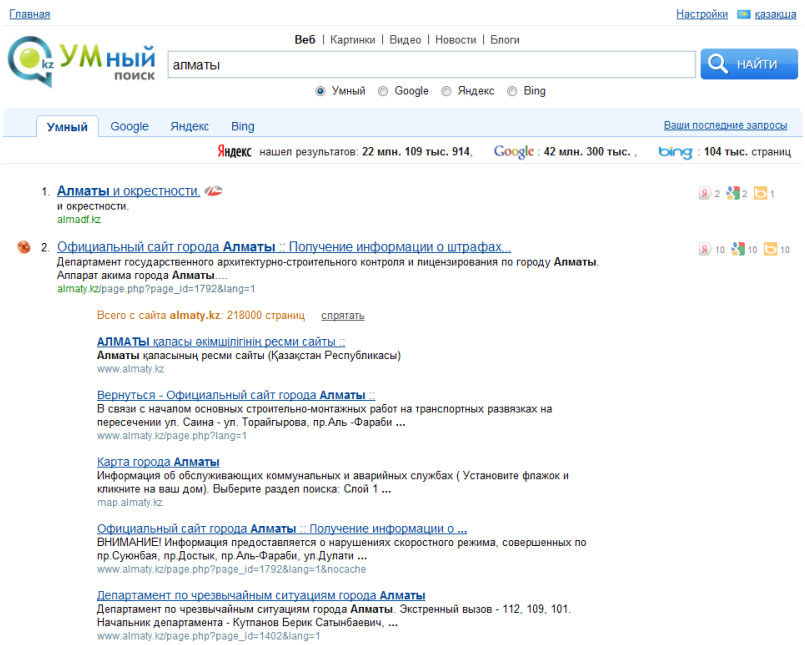

Рисунок 41 – Мгновенный поиск по нужному домену

Если сайт имеет иконку Мегалайна, это означает что данный интернет ресурс расположен на территории Казахстана и трафик для пользователей интернета Megaline бесплатен. На рисунке 42 видно иконку принадлежности сайта к Казахстанской сети интернет.

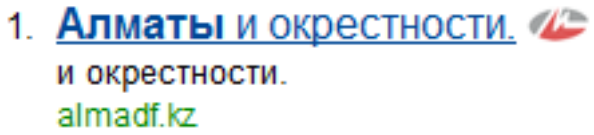

Рисунок 42 – Иконка принадлежности сайта.
к Казахстанской сети интернета

Сбоку от каждого сайта, расположены данные позиции сайта в поисковых машинах. На рисунке 43 показаны позиции сайта в других поисковых машинах.

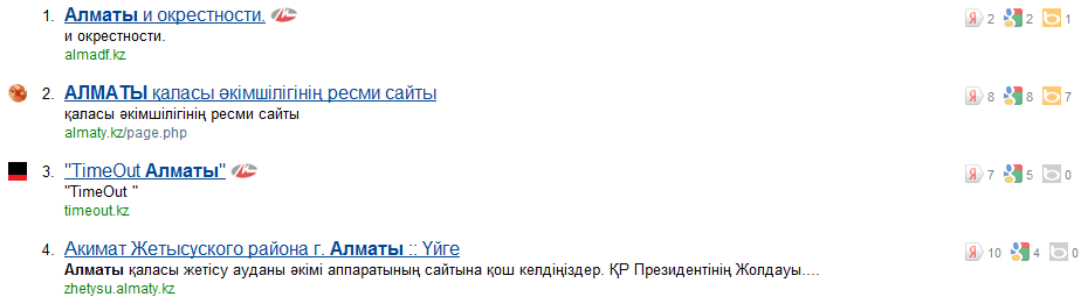

Рисунок 43 – Позиция сайта в других поисковых машинах

4.2 Система исправления и распознавания запросов

Все вводимые пользователем поисковые запросы изначально попадают в систему исправления и распознавания запросов. Если система обнаружила, что поисковых запрос, имеет ошибку, будет предложен исправленный вариант. Исправление запросов показано на рисунке 44.

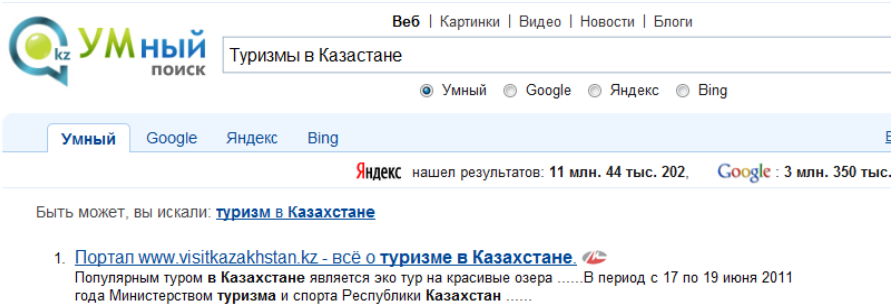

Рисунок 44 – Исправление запросов

Если пользователь наберет “переведи <какие то слова>” или “перевод <какие то слова>” то произойдет автоматическое распознавание текста и перевод на язык. Русский на английский и наоборот. Работа автоматического перевода запроса приведена на рисунке 45.

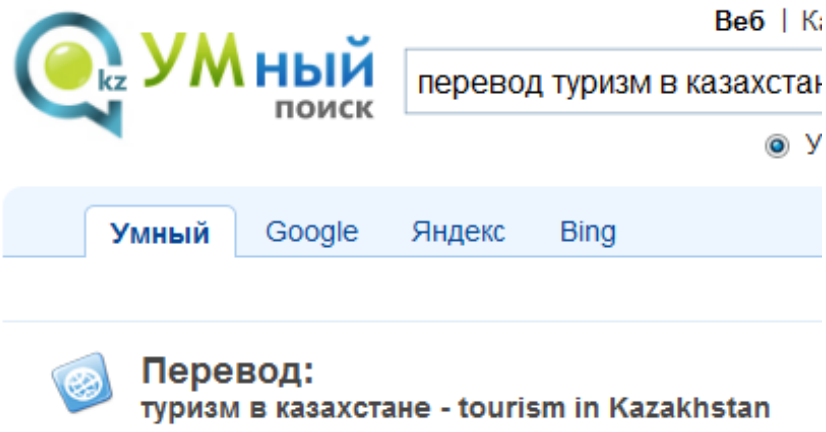

Рисунок 45 – Автоматический перевод результатов поиска

Если пользователь набрал в поисковой строке “погода в <любой город>” или “прогноз погоды в <любой город>” выдается прогноз погоды в набранном городе. Если же город не указан, с помощью Geoip идет определение города пользователя и отображение погоды в нем. На рисунке 46 показано отображение погоды по запросу.

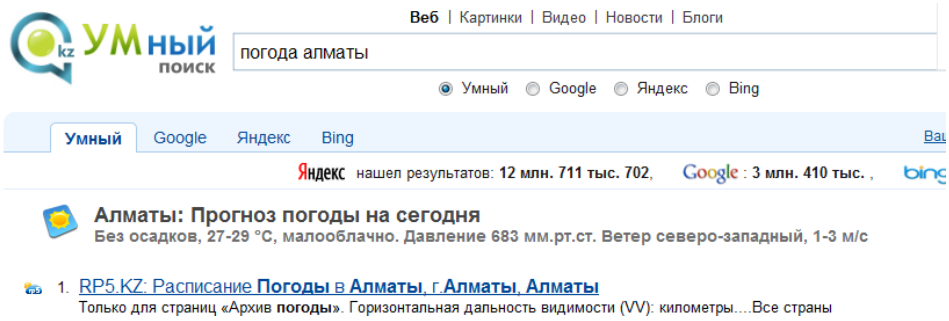

Рисунок 46 – Отображение погоды по запросу

По умолчанию прогноз погоды отображается на сегодня, если время в данном городе меньше 12 часов дня. Если же время больше 12, отображается погода на завтра. Если уточнить запрос и набрать “погода на завтра” то соответственно будет отображена погода на завтра. Так как город не указан, в этом запросе он также будет автоматически определен по IP адресу пользователя.

При наборе слов “время <населенный пункт>”, или “который час в <населенный пункт>” выдается время в набранном населенном пункте. Если же пункт не указан, с помощью Geoip идет определение города пользователя и отображение точного времени в определенном городе. На рисунке 47 показано определение точного времени с определением населенного пункта по IP.

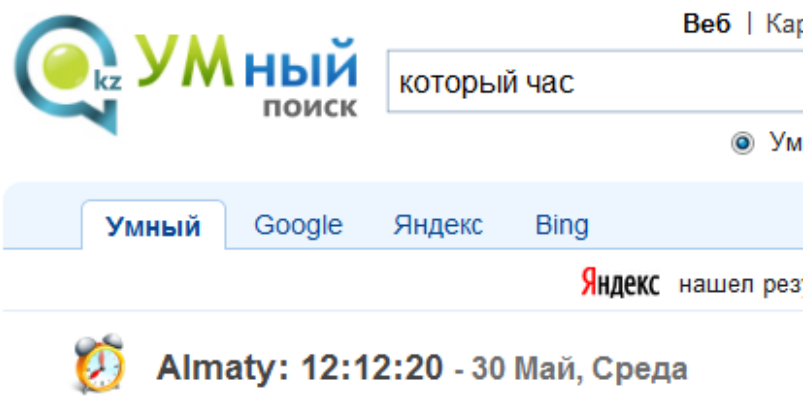

Рисунок 47 – Определение точного времени

Если в запросе встречается слово какой либо знак зодиака, то поиск автоматически делает запрос и выводит астрологический прогноз для этого знака зодиака. На рисунке 48 предоставлена работа отображение астрологического гороскопа в результатах поиска.

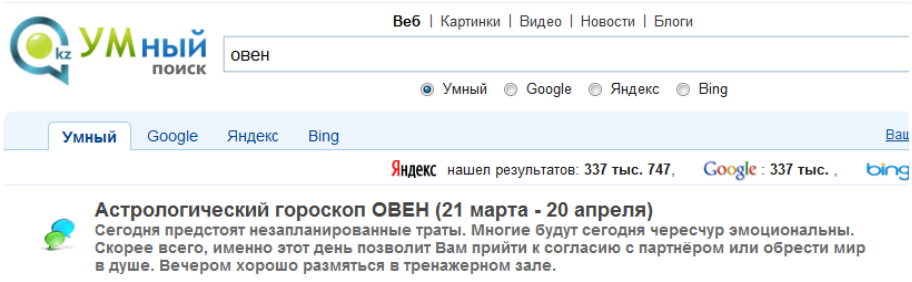

Рисунок 48 – Астрологический гороскоп
в результатах поиска

При этом запросе пользователя через Ajax автоматически идет запрос к XML API сервера hyrax.ru и получение XML данных астрологического гороскопа на сегодня. Затем XML данные разгруппировываются и отображаются пользователю в соответствующем виде и подходящем дизайне уже в виде отформатированного HTML.

Если набрать в поисковой строке “закат” или “восход” то поисковая машина также сразу покажет время закат и восхода солнца для населенного пункта пользователя который определяется через GeoIP.

На рисунке 49 показано определение восхода и захода солнца.

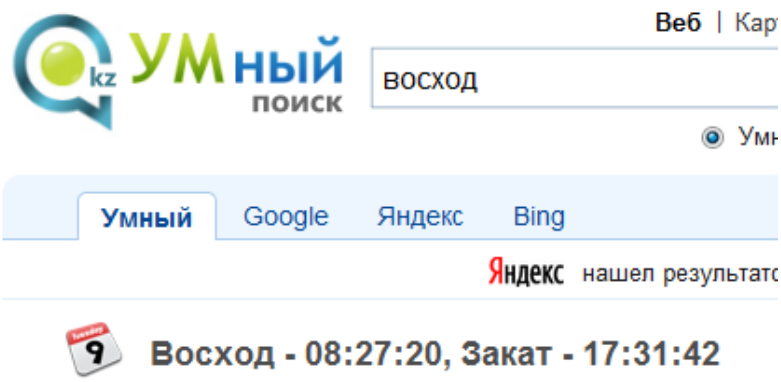

Рисунок 49 – Определение восхода и захода солнца

При запросе “маршрут <номер>”, “автобус <номер>” поисковая система отображает сразу маршрут выбранного автобуса, трамвая либо троллейбуса. На рисунке 50 показан вывод схемы маршрута городского транспорта.

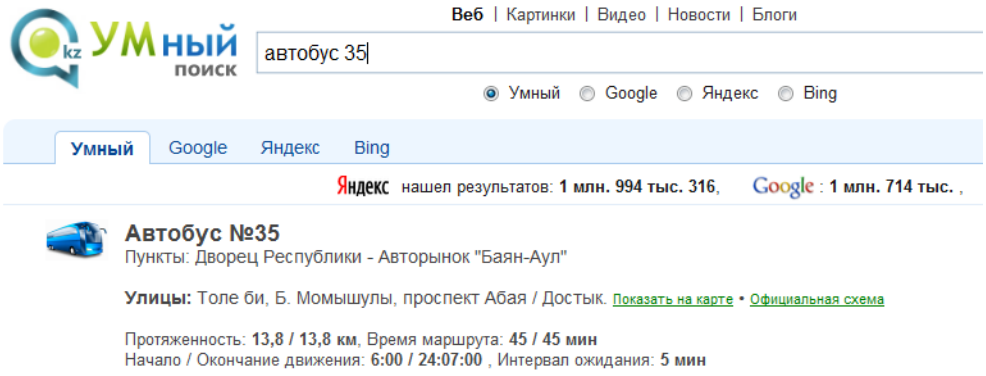

Рисунок 50 – Схема маршрутов
общественного транспорта

При этом у пользователя есть возможность посмотреть официальную схему проезда этого общественного транспорта.

Если же пользователь нажмет на ссылку “Показать на карте” то автоматически появится карта Яндекс с отображением выделенных улиц маршрута данного транспорта. Также отображаются данные о начале и окончание движении автобуса, троллейбуса или трамвая.

Если же пользователь наберет “мой ip” или “ip” то поисковая система сразу же покажет эти данные. На рисунке 51 показано автоматическое определение IP адреса пользователя при запросе.

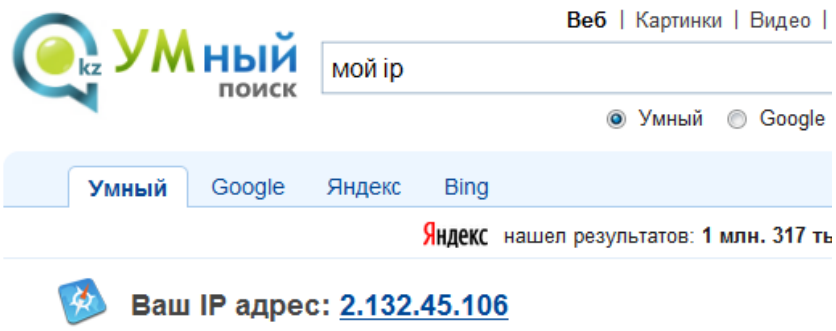

Рисунок 51 – Автоматическое определение
IP адреса пользователя

Если пользователь наберет какой то математический пример, например “sin(3)*3+2” поисковая система сразу покажет результат. Для примера Мы ввели запрос “sin(3)*3+2” что показано на рисунке 52.

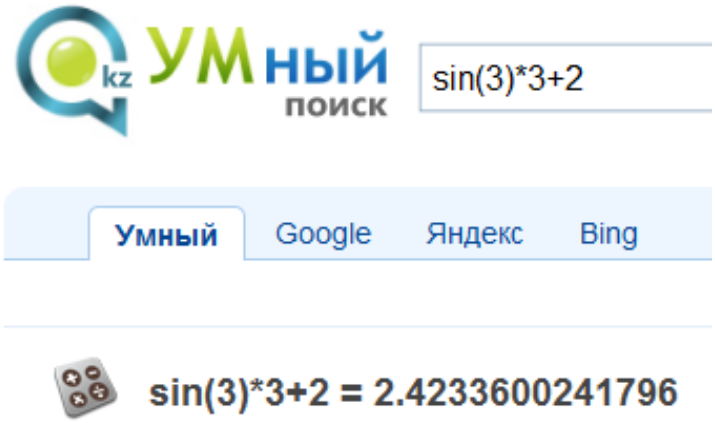

Рисунок 52 – Пример математического калькулятора

Математический калькулятор поддерживает функции: abs, acos, acosh, acot, acoth, asin, asinh, atan, atanh, combin , combina, cos, cosh, cot, coth, degrees, even, exp, fact, factdouble, int, iseven, isodd, lcm, lcm_add, ln, log, pi, power, round, sign, sin, sinh, sqrt, sqrtpi, tan, tanh.

Длина и количество знаков в математическом запросе не ограничивается. Запрос на вычисление идет посредствам Ajax.

При этом поисковые результаты на другие сайты не отображаются, что позволяет использовать математический калькулятор очень быстро.

При наборе в поисковой строке слов вида “40 тенге в доллара” и других, поисковая система моментально переведет 40 тенге в доллары. Поддерживаются различные написания и более 30 видов валют. На рисунке 53 показана работа автоматического конвертера валют при наборе в поисковой строке.

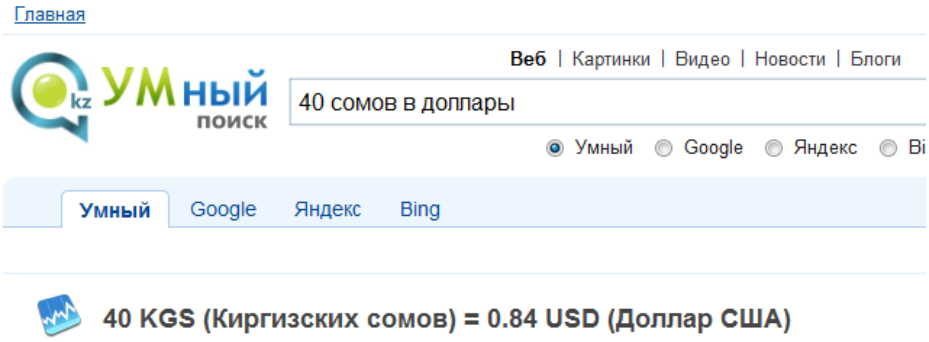

Рисунок 53 – Автоматический конвертер валюты
в поисковой строке

Дизайн самого поискового результата также продуман до мелочей. Около каждого результата отображается иконка сайта. В тексте и заголовке идет выделение найденных ключевых поисковых фраз – полужирным шрифтом. Домен в адресе результата выделен зеленым цветом и является ссылкой для автоматического поиска по этому домену. При нажатии на текст результата, открываются дополнительные навигационные ссылки к этому ресурсу.

Дизайн поискового результата представлен на рисунке 54.

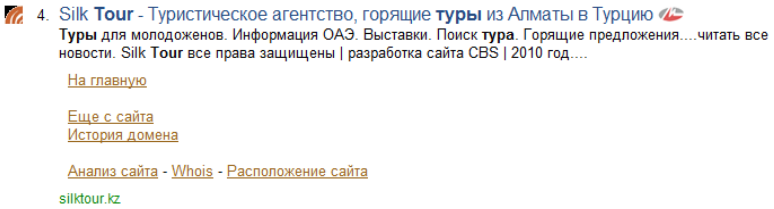

Рисунок 54 – Дизайн поискового результата

ЗАКЛЮЧЕНИЕ

В данном диссертационной работе проведен анализ и исследование поисковых систем мира. Результаты анализа поисковой отрасли, во-первых, позволили определить основные факторы и принципы ранжирования сайтов и отдельных веб-документов в поисковой выдаче.

Во-вторых, после экспериментов и анализа основных и второстепенных аспектов качества поиска существующих популярных систем можно уверенно утверждать, что лучшей поисковой системой во всемирной паутине на сегодня является Google.

В процессе выполнения работы были сделаны следующие выводы:

- проведен обзор и анализ алгоритмов поисковых систем, используемых в российском, казахстанском и мировом сегменте сети Интернет;
- выявлены показатели, влияющие на качество алгоритмов поисковых систем;
- проведены эксперименты для получения значений по каждому показателю качества каждого алгоритма поисковой системы;
- на основе проведенных экспериментов сделан анализ алгоритмов поисковых систем;
- разработка программного модуля на основе собственного алгоритма поисковой системы.

Однако, заключение пишется в конце и предполагает конечность. Но рост информации бесконечен, а потому нет предела совершенствованию поисковых машин. Важнейшей задачей разработчиков поисковых систем является улучшение качества поиска, движение в сторону большей эффективности и удобства в использовании системы. С этой целью постоянно меняются поисковые алгоритмы, фильтры, создаются дополнительные сервисы, дорабатывается и оптимизируется дизайн. Также совершенствуется отрасль СЕО-оптимизации, где специалисты в данной области идут на различные ухищрения для того чтобы добиться высоких результатов в поисковой выдаче.

На основе полученных данных разработана поисковая система сайтов YM.kz с учетом всех полученных данных, которая обеспечит эффективный и релевантный поиск нужного интернет ресурса казахстанскому пользователю. Ее преимущества и новизна заключаются в уникальной системе ранжирования и смешивания результатов, уникальным дизайном и выдачей результатов поиска, большом функционале и только региональном локальном поиске по Казахстану.

Для того, чтобы выжить в мире динамичного Интернета, при разработке необходимо закладывать большой запас устойчивости, постоянно заглядывать в завтрашний день и примерять будущую нагрузку на сегодняшний поиск. Все, что сегодня программируется в YM.kz, рассчитано "на вырост". Такой подход позволяет заниматься не только постоянной борьбой и приспособлением поисковой машины к растущим объемам

информации, но и реализовывать что-то новое, действительно важное и нужное для повышения эффективности поиска в сети Интернет.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Э.А. Якубайтис «Информатика-электроника-сети». М., «Финансы и статистика», 2004. – 200 с.
2. Плаксина А. А. Интернет в образовательном процессе учебного заведения / А. А. Плаксина, Э. А. Тихонов. - Дистанционное и виртуальное обучение. - 2010 № 7. - С. 15-24.
3. Сергей Людкевич, Евгений Есипов, "Промо.Текарт", 2003. - С. 13-29
4. Андрей Гулин, Михаил Маслов, Илья Сегалович, “Алгоритм текстового ранжирования Яндекса на РОМИП-2006”. - С. 1-6.
5. D. Fetterly, M. Manasse, and M. Najork, “Spam, Damn Spam, and Statistics,” Proc. Int’l Workshop on the Web and Databases (WebDB), ACM Press, 2004, pp. 1-6.
6. Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, “Combating Web Spam with TrustRank,” Proc. Int’l Conf. Very Large Databases (VLDB), Morgan Kauf- mann, , 2004, pp. 576-584.
7. A. Benczúr et al., “SpamRank—Fully Automatic Link Spam Detection,” Proc. Int’l Workshop Adversarial Information Retrieval on the Web (AIRWeb), 2005. pp. 258-584.
8. B. Wu and B. Davison, “Identifying Link Farm Spam Pages,” Proc. Int’l Conf. World Wide Web, ACM Press, 2005, pp. 820-829.
9. Alexandros N. UCLA Computer Science Dept. , Mark M. Microsoft Research, “Detecting Spam Web Pages through Content Analysis”, 2012, pp. 1-6.
10. Индексация сайта в поисковых системах, http://seo.in.ua/blog/indeksaciya-sajta-v-poiskovyx-sistemax, 2006. - С. 1-6.
11. Осипов Г.С., и др. Семантический поиск в среде Интернет. // Препринт. -Переславль-Залесский: ИПС РАН, 2003. – 350 с.
12. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011. – 600 с.
13. Людкевич С. Ранжирование документов в поисковых машинах. http://www.promo-techart.ru/analvsis/range.htm, 2006. - С. 1-3.
14. 3. S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine // Computer Networks and ISDN Systems. 1998. 30. № 1. P. 107–117.
15. G. Mishne, D. Carmel, and R. Lempel, “Blocking Blog Spam with Language Model Disagreement,” Proc. Int’l Workshop Adversarial Information Retrieval on the Web (AIRWeb), 2005. pp. 955-969.
16. H. Zhang et al., “Making Eigenvector-Based Repu- tation Systems Robust to Collusion,” Proc. Int’l Workshop Algorithms and Models for the Web-Graph (WAW), Springer-Verlag, 2004, pp. 92-104.

17. Гусев В.С. Google: эффективный поиск. Краткое руководство. М.: «Вильяме», 2006. - 354с.
18. Крохина О. И., Полосина М. Н., Рубель А. В., О. И. Сахно, Селин Е. В., М. С.Ханина. Первая книга SEO-копирайтера. Как написать текст для поисковых машин и пользователей. — М.: «Инфра-Инженерия», 2012. — 216 с.
19. Троян Г.М. Поиск в русскоязычной части Интернет: поисковая система Yandex // Радиолюбитель. Ваш компьютер. – № 1-3, 2000. — 252 с.
20. Гулин А., Карпович П., Расковалов Д., Сегалович И. Оптимизация алгоритмов ранжирования методами машинного обучения // Тр. Росс. сем. по оценке методов информационного поиска. СПб.: НУ ЦСИ, 2009. 163–168.
21. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of Web resources. Proceedings of the 26th International Conference on Very Large Databases (VLDB’00), 2000. pp. 92-104.
22. Ашманов И. С., Иванов А. А. Продвижение сайта в поисковых системах. — М.: «Вильямс», 2007. — 304 с.
23. Колисниченко Д. Н. Поисковые системы и продвижение сайтов в Интернете. — М.: Диалектика, 2007. — 272 с.

ПРИЛОЖЕНИЕ А

Исходный файл скрипта система алгоритма смешивания и ранжирования результатов поиска (ym.php)

```
<?php
import_request_variables("gP", "");
if(!isset($start)){
    $start=0;
}else{
    $start=$start/15;
}
 // Подключаем БД
require_once("db.php");
function highlight_words($node)
{
  $stripped = preg_replace('/<\/?(title|passage)[^>]*>/', '', $node->asXML());
  return str_replace('</hlword>', '</b>', preg_replace('/<hlword[^>]*>/', '<b>', $stripped));

}
// Функция многопоточного получения данных
function multi_read($connomains, $max_threads, $option){
$mh = curl_multi_init();
foreach ($connomains as $i => $url) {
$conn[$i] = curl_init($url);
curl_setopt($conn[$i], CURLOPT_RETURNTRANSFER, 1);
if($option[$i]!=""){
    curl_setopt($conn[$i], CURLOPT_POSTFIELDS,$option[$i]);
}
curl_multi_add_handle ($mh,$conn[$i]);
}
do {
$mrc = curl_multi_exec($mh, $active);
} while ($mrc == CURLM_CALL_MULTI_PERFORM);
while ($active and $mrc == CURLM_OK) {
if (curl_multi_select($mh) != -1) {
do {
$mrc = curl_multi_exec($mh, $active);
} while ($mrc == CURLM_CALL_MULTI_PERFORM);
}
}
if ($mrc != CURLM_OK) {
print "Curl multi read error $mrc\n";
```

```
}
```



```
foreach ($connomains as $i => $url) {
if (($err = curl_error($conn[$i])) == '') {
$res[$i]=curl_multi_getcontent($conn[$i]);
} else {
print "Curl error on handle $i: $err\n";
}
curl_multi_remove_handle($mh,$conn[$i]);
curl_close($conn[$i]);
}
curl_multi_close($mh);
return($res);
}
function get_res($query,$start,$context){
    $options_array = array( <<<DOC
<?xml version='1.0' encoding='utf-8'?>
<request>
   <query>$query domain:kz</query>
   <groupings>
      <groupby attr="" mode="flat" groups-on-page="10" docs-in-
group="1" />
   </groupings>

   <page>$start</page>
</request>
DOC
,"","""
                    );
        $search_array = array(

        "http://xmlsearch.yandex.ru/xmlsearch?user=dekodalabs&key=03.21655
416:9794869862fc97e50b2a5bb34d610998&lr=159",

        "http://api.bing.net/xml.aspx?Appid=5C08C95FC0715E88226405E208
BF29D4AAE15FCD&sources=web&query=".urlencode($query."
site:kz")."&web.count=10&Version=2.0&web.offset=".($start*10)."",

        "http://ajax.googleapis.com/ajax/services/search/web?v=1.0&filter=1&k
ey=ABQIAAAA9cuS2ToykTODWTgzr5qAuRQxMDBcEALRwwtQ3-
ky4PqXSeOE-RSNJ5UfBIDDwMbwsqYZGrr2BtwDtA&q=".urlencode($query."
site:kz")."&rsz=large&start=".($start*8)
                                        );
```

```
$data = multi_read($search_array, count($search_array), $options_array);
    $domains=array();
    $result=array();
    $pos=array();
      // Паринг Yandex XML
    $xmldoc = new SimpleXMLElement($data[0]);
       $error = $xmldoc->response->error;
      $found_yandex = $xmldoc->response->found;
      $found = $xmldoc->xpath("response/results/grouping/group/doc");
      $i=1;
      foreach ($found as $item) {
      $text="";
        if ($item->passages) {
foreach ($item->passages->passage as $passage) {
                                    if(mb_strlen($text,'UTF-8')<255){
                                    $text.="" . highlight_words($passage) . "...";
                                    }
                }
              }else{
                                $text=$item->title;
                          }
                $text=mb_substr($text, 0, 250,'UTF-8');
                $dom=$item->domain;
                $yandex_d="".str_Replace("www.","",$dom[0][0]);
                $result[$yandex_d]=array($item->url,    highlight_words($item-
>title), $text);
                $pos[$yandex_d]=array($i,0,0);
                array_push($domains,$yandex_d);
                $i++;
                }
    //print_r($result);
    //print_R($pos);
    // Парсинг Google
      $json = json_decode($data[2]);
      $found_google=$json->responseData->cursor->resultCount;
      $i=1;
      foreach($json->responseData->results as $item) {
                                    $google_d=str_Replace("www.","",$item-
>visibleUrl);

                if(in_array($google_d, $domains)){
```

```
                        $temp2=$pos[$google_d];
```

Продолжение приложения А

```
        $pos[$google_d]=array($temp2[0],$i,0);
                }else{
                        $result[$google_d]=array($item->url,  $item->title,$item-
>content);
                        $pos[$google_d]=array(0,$i,0);
                        array_push($domains,$google_d);
                }

                $i++;
            }
            //print_r($google);
        //print_r($google_d);
        //Парсинг выдачи Bing
        $sxe = new SimpleXMLElement($data[1]);
            $resultsearch = $sxe->children('web', TRUE);
            $i=1;
            $found_bing=$resultsearch->Web->Total;
            if ($found_bing > 0) {

                foreach($resultsearch->Web->Results->WebResult as $item) {
                        $bing_d=str_Replace("www.","",parse_url($item->Url,
PHP_URL_HOST));
                $bing[$i]=array();
                array_push($domains,$bing_d[$i]);
                if(in_array($bing_d, $domains)){
                        $temp2=$pos[$bing_d];
                        $pos[$bing_d]=array($temp2[0],$temp2[0],$i);
                }else{
                        $result[$bing_d]=array($item->Url,  $item->Title,  $item-
>Description);
                        $pos[$bing_d]=array(0,0,$i);
                        array_push($domains,$bing_d);
                }
                                $i++;
                }
            }
         //print_r($bing);
        //print_r($bing_d);
    //print_r($result);
    //$domains=array_unique($domains);
        //echo "<br>".$domains[0].."";
```

// АЛГОРИТМ СМЕШИВАНИЯ и разделение результатов на 3 группы

Продолжение приложения А

```
$suma1=array();$suma2=array();$suma3=array();
    $gr1=array();$gr2=array();$gr3=array();
    foreach ($pos as $k => $v){
    $u=0;
      if($v[0]!=0){$u++;}
      if($v[1]!=0){$u++;}
      if($v[2]!=0){$u++;}
            $o=$v[0]+$v[1]+$v[2];
      if($u==3){
            $gr1[$k]=$result[$k];
            $suma1[$k]=$o;
      }elseif($u==2){
            $gr2[$k]=$result[$k];
            $suma2[$k]=$o;
      }else{
            $gr3[$k]=$result[$k];
            $suma3[$k]=$o;
      }
    }
```

// СОРТИРОВКА результатов согласно сумме вхождения в поисковые системы в каждой группе

```
    array_multisort($suma1, $gr1);
    array_multisort($suma2, $gr2);
    array_multisort($suma3, $gr3);
echo "Callback('$context',{\"results\":[";
    $i=1;
    $results="";
    foreach ($gr1 as $k => $v){
      $l=$pos[$k];
    $results.=$v[0]."^|".$v[1]."^|".$v[2]."^|".$k."^|".$l[0]."^|".$l[1]."^|".$l[2]
."^|".($l[0]+$l[1]+$l[2])."*|*";
      $i++;
      echo
'{"Y":'".$l[0].'","G":'".$l[1].'","B":'".$l[2].'","unescapedUrl":'".$v[0].'","url":'".mb_
htmlentities($v[0]).'","visibleUrl":'".$k.'","cacheUrl":"","title":'".mb_htmlentities($
v[1]).'","titleNoFormatting":"","content":'".mb_htmlentities($v[2]).'"},';

}
    foreach ($gr2 as $k => $v){
      if($i<16){
            $l=$pos[$k];
```

```
            $results.=$v[0]."^|".$v[1]."^|".$v[2]."^|".$k."^|".$l[0]."^|".$l[1]."^|".$l[2]
."^|".($l[0]+$l[1]+$l[2])."*|*";
                        $i++;
                        echo
'{"Y":"".$l[0]."","G":"".$l[1]."","B":"".$l[2]."","unescapedUrl":"".$v[0]."","url":"".mb_
htmlentities($v[0])."","visibleUrl":"".$k."","cacheUrl":"","title":"".mb_htmlentities($
v[1])."","titleNoFormatting":"","content":"".mb_htmlentities($v[2])."}';
                        if($i<16){echo ",";}
                    }
            }
            foreach ($gr3 as $k => $v){
                if($i<16){
                        $l=$pos[$k];
        $results.=$v[0]."^|".$v[1]."^|".$v[2]."^|".$k."^|".$l[0]."^|".$l[1]."^|".$l[2]."^|
".($l[0]+$l[1]+$l[2])."*|*";
                        $i++;
                        echo
'{"Y":"".$l[0]."","G":"".$l[1]."","B":"".$l[2]."","unescapedUrl":"".$v[0]."","url":"".mb_
htmlentities($v[0])."","visibleUrl":"".$k."","cacheUrl":"","title":"".mb_htmlentities($
v[1])."","titleNoFormatting":"","content":"".mb_htmlentities($v[2])."}';
                        if($i<16){echo ",";}
                    }
            }
            $found_google=str_Replace(",","",$found_google);
                $total=(ceil(($found_yandex+$found_google+$found_bing)/3));
        echo
"],\"cursor\":{\"estimatedResultCount\":\"".$total."\",\"Y\":\"$found_yandex\",\"G\
":\"$found_google\",\"B\":\"$found_bing\",\"currentPageIndex\":0,\"moreResultsU
rl\":\"\"}}, 200, null, 200)";
        //print_r($gr1);
            //print_r($gr2);
            //print_r($gr3);
            return
array($results,$total,$found_yandex,$found_google,$found_bing);
        }
        function mb_htmlentities($str, $encoding = 'utf-8') {

            $str=str_replace('<',"\u003c",$str);
            $str=str_replace('>',"\u003e",$str);
            $str=str_replace("","&quot;",$str);
            $str=str_replace("\","&#39;",$str);
```

```
        return $str;
```

Продолжение приложения А

```
    }
    $query=$q;
    // Делаем запрос в бд, Проверяем существование
    $test = mysql_query("SELECT id,query,date,pages,total,yandex,google,bing
FROM queries WHERE query = '$query'") or die(mysql_error());
    $test_res = mysql_fetch_row($test);
    if($test_res[0]!=""){
        //echo "Уже есть!";
        if(((strtotime(date('Y-m-d')) - strtotime($test_res[2]))/3600/24)>3 ){ //
проверяем что результат не старше 3 дней
                $get_result=get_res($query,$start,$context);
                mysql_query("UPDATE queries SET date='".date('Y-m-d').'",
pages=0 WHERE id=".$test_res[0]."'); // Обновляем дату
            mysql_query("DELETE FROM result WHERE id=".$test_res[0]."');
                mysql_query("INSERT INTO result VALUES (".$test_res[0].",
'0','".mysql_escape_string($get_result[0]).'")') or die(mysql_error());
                        //echo "Обновлено.";
        }else{
            if($start!=0 && $test_res[3]<$start){
                $get_result=get_res($query,$start,$context);
                mysql_query("UPDATE queries SET pages='".$start."'
WHERE id=".$test_res[0]."'); // Обновляем дату
                mysql_query("INSERT INTO result VALUES (".$test_res[0].",
'".$start."','".mysql_escape_string($get_result[0]).'")') or die(mysql_error());
                //echo "Увеличиваем страницу";
            }else{
                $pager = mysql_query("SELECT page,data FROM result
WHERE id = '".$test_res[0]."' AND page='".$start."') or die(mysql_error());
                $page_res = mysql_fetch_row($pager);
                        //echo $page_res[1]."<br>";
                $datatop=explode("*|*",$page_res[1]);
                echo "Callback('$context',{\"results\":[";
                $i=1;
                    foreach($datatop as $it) {
                        $item=explode("^|",$it);
                        if($item[0]!=""){
                            echo
'{"Y":"'.$item[4].'","G":"'.$item[5].'","B":"'.$item[6].'","unescapedUrl":"'.$item[0].
'","url":"'.mb_htmlentities($item[0]).'","visibleUrl":"'.$item[3].'","cacheUrl":"","titl
```

```
e":'".mb_htmlentities($item[1]).'","titleNoFormatting":"","content":'".mb_htmlentit
ies($item[2]).'"}';
                                        $i++;   if($i!=count($datatop)){echo ",";}
```

Окончание приложения А

```
            }
                                        }

                    echo
"],\"cursor\":{\"estimatedResultCount\":\"".$test_res[4]."\",\"Y\":\"".$test_res[5]."\
",\"G\":\"".$test_res[6]."\",\"B\":\"".$test_res[7]."\",\"currentPageIndex\":0,\"more
ResultsUrl\":\"\"}}, 200, null, 200)";

                //      echo "Выводим страницу.";
                }
                    }
        }else{

            $get_result=get_res($query,$start,$context);

            mysql_query("INSERT INTO queries VALUES (NULL,
'".mysql_escape_string($query).'","'.date('Y-m-d').'",
'0','".$get_result[1].'","'.$get_result[2].'","'.$get_result[3].'","'.$get_result[4].'")") or
die(mysql_error());
            $id=mysql_insert_id();
            mysql_query("INSERT INTO result VALUES (".$id.",
'0','".mysql_escape_string($get_result[0]).'")") or die(mysql_error());
            //echo "Добавлен новый.";
         }
        ?>
```