Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет»
(национальный исследовательский университет)
Высшая школа экономики и управления
Кафедра «Информационные технологии в экономике»

ПРОЕКТ ПРОВЕРЕН
Рецензент, редактор
Интернет-издания
ГПЧО «Обл-ТВ»
_____(П.С. Агасиев)
«____» _____ 2018 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.т.н., с.н.с,
_____(Б.М. Суховилов)
«____» _____ 2018 г.

Повышение эффективности методики поисковой оптимизации веб-сайта

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–38.04.05.2018.115.ПЗ ВКР

Руководитель работы, ст. преподаватель
_____(В.В. Костерин)
«____» _____2018 г.

Автор работы,
студент группы ЭУ-312
_____(Е.А. Гусева)
«____» _____2018 г.

Нормоконтролер, к.т.н., доцент
_____(О.С. Буслаева)
«____» _____2018 г.

Челябинск 2018

АННОТАЦИЯ

Гусева Е.А. Повышение эффективности методики поисковой оптимизации веб-сайта, Челябинск: ЮУрГУ, ЭУ-312, 2018. – 90 стр., 21 ил., 4 табл., библиографический список – 51 наим.

Исследование посвящено повышению эффективности методики поисковой оптимизации веб-сайтов.

Рассмотрены основные понятия, характеристики и устройство поисковой системы.

Получена система критериев, участвующих в формулах ранжирования поисковых систем, на основе метода экспертных оценок.

Разработан метод поисковой оптимизации на основе факторов, участвующих в формулах ранжирования поисковых систем, позволяющий повышать посещаемость веб-сайтов.

С помощью метода поисковой оптимизации была повышена посещаемость сайта 1obl.ru.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Информационные технологии стали неотъемлемой частью жизни каждого современного человека. В соответствии с данными на 5 октября 2017 года, россияне от 18 лет, использующие Интернет не реже 1 раза в месяц, составляют 70% (или 81,8 млн. человек), а ежедневно пользующиеся Интернетом 60% (или 70,4 млн. человек). Поэтому тяжело представить фирму, завод, лабораторию, образовательное учреждение и даже квартиру, которые не были бы подключены к сети Интернет. С одной стороны Интернет открывает доступ к огромному объему информации, с другой, предоставляет возможность разместить собственные информационные ресурсы.

В связи с этим возникла проблема поиска информации. Сегодня в большинстве случаев ее решают поисковые системы [21], которые определяют список веб-страниц, соответствующих (релевантных) запросу пользователя. Таких страниц могут быть миллионы, и задача поисковой системы – расположить найденные веб-страницы в порядке убывания релевантности.

Рост объемов информации, индексируемый поисковыми системами, порождает постоянное развитие алгоритмов поисковой оптимизации. За последние несколько лет реализовано множество нововведений. Например, представители Яндекса предоставили алгоритм [28], при котором на первую позицию должны попадать компании и фирмы, предлагающие действительно высокое качество (с положительной репутацией как в сети, так и в оффлайне).

В связи с этим появляется вопрос анализа влияния индексации сайта в поисковой системе на посещаемость. В исследовании в качестве примера взят сайт челябинского областного телевидения – «Первый областной» (1obl.ru).

Цель работы – повышение эффективности поисковой оптимизации по сравнению традиционными методами.

Задачи работы:

– проанализировать основные современные поисковые системы, используемые в российском сегменте сети Интернет;

– изучить алгоритмы поисковой оптимизации;

– разработать методологию поисковой оптимизации веб-сайта;

– применить разработанную методику на практике (увеличить посещаемость сайта 1obl.ru);

– рассмотреть другие способы повышения посещаемости.

Объектом исследования является сайт «Первый областной».

Предметом исследования являются алгоритмы поисковой оптимизации наиболее популярных и прогрессивных поисковых систем (Яндекс и Google).

Практическая значимость результатов работы заключается в повышении посещаемости сайта «Первый областной» (1obl.ru) при применении разработанного метода поисковой оптимизации.

ГЛАВА 1. ОБЗОР И АНАЛИЗ ПОИСКОВЫХ СИСТЕМ СЕТИ ИНТЕРНЕТ

1.1 Понятие, характеристики и устройство поисковой системы

Поисковая система [35] – это набор технических и программных средств, предназначенный для поиска информации в Интернете. Такая система реагирует на запрос пользователя, который задается в виде текстовой фразы (поискового запроса) и выдает упорядоченный по релевантности список ссылок на веб-сайты. К задачам поисковых систем относится предоставление пользователю тематической информации в соответствии с его запросом.

Полнота – одна из основных характеристик поисковых систем, представляющая собой отношение найденных поисковой системой веб-страниц к общему числу страниц в Интернет, соответствующих запросу. Допустим, фраза «происшествия в Челябинске» встречается 650 000 раз на различных страницах сети Интернет, а при запросе этой фразы в поисковой системе найдено 519 000 документов – тогда полнота поиска составляет примерно 0.8. Чем больше полнота, тем больше шансов у пользователя найти искомую информацию [27].

Точность – еще одна из важнейших характеристик поисковой системы, которая определяет, насколько релевантны найденные страницы запросу пользователя [46], а так же насколько быстро осуществляется поиск. Например, по запросу «происшествия в Челябинске» найдено 650 000 результатов, 300 из которых содержат фразу «происшествия в Челябинске», а остальные 350 содержат фразу «происшествия в Челябинске и Челябинской области», то точность поиска составляет чуть меньше 0,5 (отношение релевантных страниц к общему числу страниц, выданных поисковой системой). Максимальная точность равна 1, если количество не релевантных страниц равно нулю. Точность поиска равна 0, если по поисковому запросу не найдено ни одной релевантной страницы.

Актуальность – еще одна важная составляющая, которая характеризуется временем между публикацией документа в сети Интернет и попаданием его в базу поисковой системы. Поисковые системы помимо основной базы документов имеют так называемую «быструю» базу, которая обновляется постоянно. Напри-

мер, новость о каком-либо происшествии разместили на множестве веб-страниц, и через несколько минут пользователи ввели поисковые запросы по этому происшествию. Т.к. документы, содержащие новость, были сохранены в «быстрой» базе, поисковая система смогла предоставить пользователю релевантную информацию. Размер «быстрой» базы на порядки меньше, и данные, содержащиеся в ней, периодически (1-2 раза в неделю) переносятся в основную базу.

Скорость поиска связана с устойчивостью поисковой системы к нагрузкам. Каждую неделю только жители Челябинской области задают поиску Яндекса примерно 25 млн. запросов. А за день пользователи всей России совершают 18,5 млн. поисковых сессий. Такая загруженность требует снижения обработки отдельного запроса. Пользователь хочет получить ответ как можно быстрее, а поисковая система – дать ему ответ, чтобы затем обрабатывать следующие запросы.

Наглядность представления результатов поиска – важный компонент удобного поиска. Как правило, поисковые системы предлагают возможность сохранения индивидуальных настроек для пользователя. За последнее время результаты поиска дополнились рекламой, новостями, различными сервисами, что не может не отвлекать пользователя (например, сервис Яндекс.Директ).

Устройство поисковой системы

Каждая крупная поисковая система имеет свою собственную архитектуру, но для всех них можно выделить общие компоненты.

Паук (поисковый робот) – программа, предназначенная для поиска новых документов в Интернете. Паук передает запрос на сервер, где расположен веб-сайт для получения информации, а в ответ получает непосредственно саму страницу и служебную информацию. Из документа извлекаются все гиперссылки, по которым отправляются аналогичные запросы. Таким образом, переходя по гиперссылкам, паук собирает информацию обо всех документах, расположенных в сети. Существуют также альтернативные способы «приглашения» паука на веб-сайт – каждая поисковая система имеет форму для добавления нового документа или целого веб-сайта. Помимо задачи перехода по гиперссылкам для нахождения всех

документов сети Интернет, в обязанности паука входит составление расписания обхода найденных ранее документов на предмет изменений. Для каждого веб-сайта расписание составляется индивидуально, и, в общих чертах, скорейшее возвращения паука тем вероятнее, чем чаще на веб-сайте появляются новые, уникальные документы, а также чем чаще цитируют данный сайт в сети Интернет (появляются гиперссылки на его документы).

Скачанные пауком страницы переходят к роботу-индексатору. Он обрабатывает их и делает пригодными для поиска. Перед отправкой запроса на сервер для получения документа робот-индексатор запрашивает содержимое файла robots.txt, если таковой существует в корневой директории сайта. Robots.txt – файл ограничения доступа роботам к содержимому веб-сайта. В случае если документ разрешен для скачивания, робот-индексатор составляет обратный (инвертированный) файл и сохраняет его в базе данных [20]. Инвертированный файл в самом простом случае представляет собой структуру, состоящую из двух частей:

– списка, содержащего все слова, которые были найдены во всех документах;

– указатели на все документы, а точнее – места в этих документах, в которых содержится каждое слово.

По этой структуре в дальнейшем и происходит поиск при запросе пользователя к поисковой системе, а сама структура называется ее индексом. Аналогичной структурой обладает «быстрая» база, документы для которой индексирует так называемый «быстроробот» (в Яндекс такой робот имеет название Orange). В такую базу, как правило, попадают страницы новостных сайтов, блогов, а также документы многих ежедневно пополняемых веб-сайтов. Стоит отметить, что документы сохраненные роботом-индексатором в основной базе, не сразу участвуют в результатах поиска, в отличие от документов в быстрой базе, которые могут появляться в результатах поиска через минуты или даже секунды после индексации «быстророботом». Обновление поисковой базы происходит 1-2 раза в неделю.

Обход документов веб-сайта пауком не гарантирует, что робот-индексатор сохранит документ в поисковом индексе. Если веб-сайт содержит множество неуни-

кальной информации, содержит вирусы, всплывающие рекламные окна или использует в своих документах различные виды спама для обмана поисковой системы, такие документы могут никогда не попасть в поисковый индекс.

Модуль поиска отвечает за анализ запроса пользователя, за поиск по инвертированной базе [35], ранжирование и представление документов пользователю. При поиске первым делом анализируется запрос, введенный пользователем. Часто пользователи вводят «длинные» запросы, состоящие из 3 и более слов, и возникает проблема – точного совпадения с запросом в индексе поисковой системы нет. В этом случае на помощь приходит переформулировка запроса. В Яндекс за это отвечает так называемый «колдунщик» [40].

Переформулировка – это процесс, в результате которого различные запросы преобразуются и поиск в индексе осуществляется по новому, уточненному запросу. В результате вместо «ничего не найдено» пользователь получает в достаточной степени релевантный ответ. До конца 2007 года в Яндекс можно было увидеть переформулированный запрос. К примеру, в то время отбрасывались частицы, предлоги, изменялись словоформы, редко употребляемым словам в многословных запросах отдавалось меньшее предпочтение, а затем осуществлялся поиск. В настоящее же время «колдунщик», являясь специальным модулем Яндекса, параллельно с поиском в интернете осуществляет поиск по собственным сервисам. Таким образом «колдунщику» удается выдавать ответ на запрос пользователя поверх всех результатов поиска («как по волшебству», отсюда и название – «колдунщик»). Например, это могут быть точное время, погода, цвет и даже математическая формула (рисунок 1).

После переформулировки запроса осуществляется поиск по индексной базе, и находятся все документы, удовлетворяющие уже новому запросу. После того, как наиболее схожие документы были отобраны, их необходимо упорядочить по релевантности (выполнить ранжирование) [15]. За этот процесс отвечает формула ранжирования, которую обычно и называют алгоритмом поисковой системы. Формула ранжирования содержит множество факторов, которые влияют на реле-

вантность документа запросу; для разных поисковых систем эти множества различны.

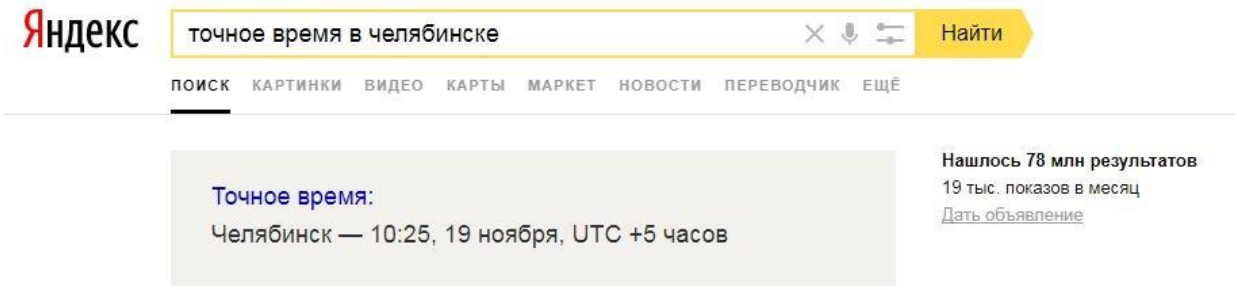

Рисунок 1 – Ответ на запрос поверх всех результатов поиска

1.2 Понятие и существовавшие ранее формулы ранжирования

Формулы ранжирования учитывают множество факторов при определении релевантности документа запросу пользователя. В первых версиях Яндекс и Google факторы ранжирования можно было разделить на две группы [3]:

– внутренние;

– внешние.

К внутренним факторам относились свойства самого документа – наличие в нем слов запроса, их точное вхождение в ключевые HTML-теги документа (*<title>, <hl>*), плотность в документе (отношение вхождений слов запроса в документ к общему числу слов в документе, выраженное в процентах), и т.д. Для каждого запроса вычисляется значение Score документа – показатель релевантности документа запросу, на основании которого и производится ранжирование [24]. Вопрос был в том, какие слагаемые должны быть в формуле расчета Score (формула 1). В результате экспериментов были отобраны слагаемые отвечающие за встречаемость слов из запроса в документе (W_{single}), за встречаемость пар слов из запроса з документе (W_{pair}) и за встречаемость текста запроса целиком (W_{phrase}). Помимо этого есть два слагаемых, дающих преимущество из-за наличия всех слов запроса в документе ($W_{allwords}$) и за наличие многих слов запроса в одном предложении ($W_{halfphrase}$), а также слагаемое за похожесть документа (W_{prf}); k_1, k_2, k_3 – коэффициенты. Итоговая формула расчета текстовой релевантности (1):

$$Score = W_{single} + W_{pair} + k_1 * W_{allwords} + k_2 * W_{halfphrase} + W_{prf}. \quad (1)$$

Для улучшения результатов поиска также использовался подход «Pseudo-relevance feedback». Суть подхода заключается в том, что поиск проводится в два этапа. На первом этапе используется простой метод, описанный выше. После этого документы, найденные на первых позициях, объявляются релевантными, и ищутся «похожие». Можно использовать любую меру похожести, но в данном случае используется 2 разные меры, которые можно реализовать с достаточной для реальных применений производительностью. Рассмотрим слагаемые более подробно.

Встречаемость слов в документе

Существует множество вариаций стандартного слагаемого отвечающего за встречаемость слова в тексте, известного как подход TF-IDF (TF – долгосрочная частота, IDF – частота обратного документа) [38]. В данном случае используется модификация bm25 (формула 2):

$$Score(D,Q) = \sum_{i=1}^{n} IDF(q_1) * \frac{f(q_1,D)*(k_1+1)}{f(q_1,D)+k_1*\left(1-b+b*\frac{|D|}{avgdl}\right)}, \quad (2)$$

где IDF – релевантность каждого из слов;

$|D|$ – число документов в коллекции;

$f(q_1, D)$ – частота слова;

k_1 – коэффициент.

Помимо учета количества слов в документе можно учитывать html-форматирование и позицию слова в документе. В Яндекс это учитывается в виде отдельного слагаемого [41]. Учитывается наличие слова в первом предложении, во втором предложении, внутри выделяющих html-тегов (формула 3):

$$W_{single} = \log(p) * (TF_1 + 0{,}2 * TF_2), \quad (3)$$

где $TF_1 = \frac{TF}{TF+k_1+k_2*DocLengt}$ при $k_1 = 1,\ k_2 = \frac{1}{350}$;

$TF_2 = \frac{Hdr}{1+Hdr}$, где Hdr – сумма весов слова за форматирование.

Учет пар слов

Помимо вхождения в текст одиночных слов запроса в описываемом алгоритме учитывается вхождение пар слов запроса в документ. После экспериментов с различными способами учета в Яндекс выделили четыре варианта учета пар и определили соответствующие веса. За разные случаи дается разный вес. Пара учитывается, когда слова запроса встречаются в тексте подряд (+1), через слово (+0.5) или в обратном порядке (+0.5). Плюс еще специальный случай, когда слова, идущие в запросе через одно, в тексте встречаются подряд (+0.1) [39]. Для хорошего ранжирования всегда достаточно вхождения лишь пары слов из запроса. Вес за пары вычисляется по формуле 4:

$$W_{pair} = 0{,}3 * (log(p_1) + \log(p_2)) * \frac{TF}{1+TF}, \quad (4)$$

где p_1 и p_2 – p для первого и второго слова пары из слагаемого W_{single}; TF – количество вхождений пары в текст с учетом весов вхождений.

Учет длины документа и форматирования для пар слов не дают выигрыша, поэтому они не учитываются.

Учет всех слов запроса в документе, учет фраз

Важным фактором, помимо перечисленных, является наличие в документе всех слов запроса [18]. За наличие всех слов запроса добавляется дополнительный «бонус» $W_{allwords}$, пропорциональный сумме idf слов запроса (формула 5):

$$W_{allwords} = 0{,}2 * \sum \log(p_i). \quad (5)$$

Если в документе присутствуют не все слова, то за каждое отсутствующее слово $W_{allwords}$ домножается на коэффициент 0.03 (формула 6):

$$W_{allwords} = 0{,}2 * \sum \log(p_i) * 0{,}03^{N_{miss}}, \quad (6)$$

где N_{miss} – количество отсутствующих в документе слов запроса.

Если в документе отсутствует много слов запроса, то возникает проблема. С

точки зрения эффективности поиска выгодно удалять документы, в которых нет хотя бы половины (считая по сумме idf) слов запроса. С точки зрения полноты поиска необходимо вернуть как можно больше возможно релевантных документов. Следовательно, удалять документы нельзя. Помимо наличия слов запроса в документе можно учесть наличие в документе текста запроса целиком. Это слагаемое называется W_{phase}, TF в данном случае – количество вхождений запроса в текст документа (формула 7):

$$W_{phase} = 0{,}1 * \sum \log(p_i) * \frac{TF}{1+TF}. \quad (7)$$

Плюс к этому еще небольшой «бонус» $W_{halfphase}$ дается за наличие в тексте предложений, содержащих значительное количество слов запроса [28]. «Значительное» в данном случае означает, что сумма idf слов запроса в предложении больше половины суммы idf всех слов запроса. TF здесь – количество таких предложений в тексте (формула 8):

$$W_{halfphase} = 0{,}02 * \sum \log(p_i) * \frac{TF}{1+TF}. \quad (8)$$

Pseudo-relevance feedback

Для дальнейшего улучшения результатов поиска используется метод PRF (pseudo-relevance feedback) [47].

Метод relevance feedback без приставки «псевдо-» заключается в том, что поиск ведется в две стадии. На первой находятся какие-то документы, которые предъявляются пользователю. Пользователь помечает документы, релевантные запросу. Второй этап поиска использует эти оценки.

Традиционный метод использования relevance feedback – добавить в запрос новые слова, встречающиеся в документах, помеченных как релевантные. Такой способ сложен в практическом использовании, так как требует повторного исполнения всего запроса, причем более сложного, чем изначальный. В Яндекс используется другой способ – документам, которые похожи на помеченные экспертом, дастся бонус. Мера похожести может быть любой. Используется мера похожести,

основанная на тегах, которые присваивались каждому документу.

Метод relevance feedback [39] можно применять и без участия пользователя, если предположить, что система достаточно хороша и на первых позициях находит релевантные документы. Тогда первые N документов просто объявляются релевантными, и повышается ранг документов, похожих на них. Допускается, что степень релевантности зависит от позиции документа в выдаче [44].

Для расчета похожести нужны признаки документов, по которым будет определяться похожесть. Используется два набора тегов.

Первый набор признаков – автоматическая классификация документов по темам Яндекс.Каталога. Для классификации использовался алгоритм Байеса в интерпретации Пола Грэма.

Теорема Байеса – это метод подсчёта обоснованности верований (гипотез, заявлений, предложений) на основе имеющихся доказательств (наблюдений, данных, информации). Изначальная вера плюс новые свидетельства равно новая, улучшенная вера. Вероятность того, что убеждение истинно с учётом новых свидетельств равна вероятности того, что убеждение было истинно без этих свидетельств, помноженной на вероятность того, что свидетельства истинны в случае истинности убеждений, и делённой на вероятность того, что свидетельства истинны вне зависимости от истинности убеждений [45].

Таким образом, каждому документу автомат приписывает одну тему. Точность алгоритма – 63%, полнота – 46%, F1 – 54% (величины – микроусредненные, измерены по рубрикам 2-го уровня Яндекс.Каталога). В результате PRF дополнительный бонус получают документы той же темы, что и первые документы выдачи [44].

Второй набор признаков использует слова, встречающиеся в документе [38]. Идея метода заключается в том, что некоторые группы слов часто встречаются вместе. Найдя такие группы, можно назначить им признаки. После этого каждому документу можно назначить признак, если в нем встречается много слов из группы этого признака. Для построения таких групп использовался принцип минимальной длины описания (Minimal Description Length, MDL).

Рассматривалась матрица, по строкам которой расположены документы, по столбцам – слова. В пересечение записывалась 1, если в документе встречается это слово, и 0 – в обратном случае. Дальше нужно построить максимально компактное описание этой матрицы с помощью групп слов. Использовалось описание в следующем виде. Для каждой группы слов имеется список слов, для, каждого документа – список слов этого документа и список «поправочных» слов. Объединение слов документа дает нам множество «предсказанных» слов для документа. «Поправочные» слова – это слова, которые есть в документе, но их нет в «предсказанных». Или, наоборот, слова, которые есть в «предсказанных», но отсутствуют в документе. Подбиралось оптимальное с точки зрения количества информации описание исходной матрицы. В результате получились группы слов и списки документов, в которых используются эти группы слов. Они и были использованы в качестве второго набора признаков. В результате PRF со вторым набором признаков бонус получают документы, использующие сходную с лидерами лексику [45].

Для определения схожести двух документов по признакам используется взвешенное по idf тега скалярное произведение (формула 9):

$$tadW_i = \sum R_{pos[k]} * tagged_{k.i}\ ,$$

$$Similar_k = \frac{\sum tagW_i * tagged_{k.i} * (idf_i)^2}{\sqrt{[\sum (tagW_i{}^2) * (idf_i)^2] * [\sum tagged_{k.i} * (idf_i)^2]}}, \quad (9)$$

$$idf_i = \frac{D_i}{D},$$

где $R_{pos[k]}$ – релевантность k-го документа;

Pos[k] – позиция k-го документа в выдаче первого прохода;

$tagged_{k.i} = 1$, если у k документа есть тег i, в противном случае $tagged_{k.i} = 0$;

D_i – количество документов с i-м тегом;

D – количество документов в коллекции [7].

В качестве бонуса для k-го документа добавляется слагаемое (формула 10):

$$R_{pos[k]} = Similar_k * k_2 * \sum \log(p_i). \quad (10)$$

Все эти выкладки описывают подход к определению релевантности документов на основе анализа их текста. Однако, как показало время, анализа одной текстовой составляющей для качественного поиска недостаточно. С ростом популярности поисковые системы превратились в неплохой источник трафика для веб-сайтов, и веб-мастера для его привлечения наполняли страницы словами, которые используют посетители поисковых систем в запросах. Результаты поиска ухудшались, необходим был технологический прорыв, который уменьшил бы влияние на результаты поиска со стороны веб-мастеров. В 1996 году Сергей Брин и Лэрри Пэйдж, основатели поисковой системы Google [10], придумали модель ссылочного ранжирования, которая использовалась для оценки релевантности документа и при этом не зависела от наличия в нем слов запроса или их плотности. Основная идея состояла в том, что, посчитав количество гиперссылок на веб-сайт, можно определить степень его популярности. По аналогии с эксклюзивными статьями на новостных ресурсах, когда автор в статье ссылается на другую ранее опубликованную новость, а количество таких ссылок на конкретный ресурс служит мерилом авторитетности и оперативности издания. Далее Пейдж развил свою идею: ссылки бывают разные, некоторые из них обладают большей значимостью («весом»), другие – меньшим. Чтобы определить, какая ссылка наиболее значима, необходимо проанализировать путь попадания на них. На те сайты, на которые ведет большее количество ссылок, являются более важными и наоборот. Программу определения значимости Пэйдж назвал Page Rank (PR) [27] – от англ. «страница» и «ранжировать». Для каждой страницы, проиндексированной Google, рассчитывается PR, который зависит, от числа ссылок в Интернете на эту страницу (естественно, проиндексированных Google) и важности ссылающихся страниц. Следует отметить, что не все ссылки могут учитываться – отфильтровываются ссылки с сайтов, специально предназначенных для «накрутки» PR (к примеру, неструктурированные сайты, содержащие большое количество ссылок на ресурсы

разнообразной тематики) [13]. Более того, некоторые ссылки могут давать отрицательный вклад при расчете PR (формула 11):

$$PR(A) = \{1 - d\} + d * \left(\left(\frac{PR(T_1)}{C(T_1)} \right) + \cdots + \left(\frac{PR(T_n)}{C(T_n)} \right) \right), \quad (11)$$

где d – вероятность, с которой пользователь перейдет по одной из ссылок в документе, а не закроет браузер (обычно ее принимают равной 0.85, что означает, что документ может передать 85% своего веса);

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

T_i – i-ый ссылающийся документ;

$PR(T_i)$ – PR i-ого документа;

C – количество ссылок в документе [37].

Поскольку ссылающихся документов может быть много, и общее их количество, проиндексированное поисковой системой Google, измеряется миллионами и постоянно растет, то представлять PR в абсолютных значениях было бы неудобно. Чтобы уложить все PR страниц в шкалу от 0 до 10, ввели логарифмическую шкалу (формула 12):

$$TLPR = log_{base}(PR) * a, \quad (12)$$

где PR – Page Rank проиндексированной Google страницы;

$TLPR$ – так называемый тулбарный PR;

$base$ – основание логарифма, которое зависит от количества документов, проиндексированных поисковой системой;

a – коэффициент приведения, который удовлетворяет неравенству $0 < a \leq 1$;

Пересчет PR – длительный процесс, требующий больших машинных ресурсов. Он происходит раз в 2–4 месяца и длится около недели.

Таким образом, к внешним факторам относились ссылки и все, что с ними связано: их количество, вес, анкер (от англ. anchor – «якорь», текст, при нажатии на который происходит переход в другой документ). При прочих равных анкор ссылки имеет решающее значение. Считается, что если на документ стоит ссылка

с анкором «происшествие», то, с некоторой долей вероятности, он содержит информацию об автомобильной аварии. Документ, в котором стоит ссылка, называется донором; документ, на который ведет ссылка, – акцептором. В случае если акцептор не содержит слов запроса пользователя в поисковой системе, он все равно может показываться в результатах, т.к. анкоры указывающих на этот документ ссылок содержат слова запроса. В таком случае рядом со ссылкой стоит подпись «найден по ссылке» (Яндекс) или «слова присутствуют только в ссылках на эту страницу» (Google) [17].

Со временем веб-мастера стали проставлять множество ссылок для манипулирования результатами поиска, появились биржи по покупке и продаже ссылок, которые существуют до сих пор. Ссылочное ранжирование усложнялось, модифицировалось, но до сих пор остается одним из главных факторов ранжирования в поисковых системах.

В последние 2–3 года добавились запросные факторы – геозависимость, т.е. для хорошего ответа, поисковой системе необходимо учитывать регион, из которого был задан запрос.

1.3 Поисковые системы Рунет

В 1996 году в Рунет появились поисковые системы Апорт (www.aport.ru) и Рамблер (www.rambler.ru). 23 сентября 1997 года был открыт Яндекс (www.yandex.ru), осенью 1997 года в США для студентов и преподавателей стала доступной поисковая система Google (www.google.com). В связи с бурным ростом Рунет и объемов информации, индексируемой поисковыми системами, необходимо обладать мощными дата-центрами для соответствия современным реалиям. Одним из критериев качества поиска и, соответственно, положительного имиджа, перспективности поисковой системы является частота обновления поискового индекса, которое также требует значительных мощностей. За прошедшие десятилетие Апорт превратился из полноценной поисковой системы в прайс-агрегатор (ресурсы, специализирующийся на сборе данных о наличии продукции в зарегистрированных в системе ресурса интернет-магазинах). Рамблер же стал просто ка-

талогом сайтов, то есть редко изменяемым списком документов, (обновления поисковой базы происходят крайне редко 1 раз в 2 месяца). Яндекс прошел путь от единственного сервера, установленного пол столом одного из разработчиков Дмитрия Тейблюма, до разветвленной независимой сети дата-центров, которая включает в себя тысячи серверов [39]. Поисковая система Google также прошла путь от одного сервера к огромному распределенному дата-центру, включающему на сегодняшний день свыше 100 тысяч серверов. Google – самая популярная поисковая система в мире, которая осуществляет поиск и по Рунет. На сегодняшний день в России наиболее востребованы, актуальны и перспективны лишь две поисковые системы – Яндекс и Google.

1.3.1 Яндекс

Поисковая система Яндекс индексирует и осуществляет поиск по следующим форматам документов: HTML, PDF, RTF, DOC, XLS. Стоит также отметить параллельный поиск Яндекс, который заключается в одновременном поиске по основной базе и по другим сервисам [46]. В них входят новости, картинки, видео, блоги, карты и маркет (платные рекламные объявления). Результаты параллельного поиска могут располагаться над результатами основного поиска, справа от них и даже внутри. Поисковая система Яндекс имеет не один, а целую группу индексирующих роботов (таблица 1). Распознать их можно через лог-файлы веб-сервера по полю User-agent, IP-адреса роботов постоянно меняются, и осуществлять идентификацию по ним нецелесообразно.

Таблица 1 – Группа индексирующих роботов

User-agent	Название
Mozilla (compatible; YandexBot)	основной индексирующий робот
Mozilla (compatible; YandexBot; MirrorDctcctor)	робот, определяющий зеркала сайтов
Mozilla (compatible; YandexImages)	индексатор Яндекс.Картинок

Окончание таблицы 1

Mozilla (compatible; Yandex Video)	индексатор Яндекс.Видео
Mozilla (compatible; YandexMedia)	робот, индексирующий мультимедийные данные
Mozilla (compatible; YandexBIogs; robot)	робот поиска по блогам, индексирующий комментарии постов
Mozilla (compatible; YandexAddurl)	робот, обращающийся к странице при добавлении ее через форму
Mozilla (compatible; YandexFavicons)	робот, индексирующий пиктограммы сайтов (favicons)
Mozilla (compatible; YandexDirect)	робот, индексирующий страницы сайтов, участвующих в рекламной сети Яндекса
Mozilla (compatible; YandexDirect)	«простукивалка» Яндекс.Директа
Mozilla (compatible; YandexMetrika)	робот Яндекс.Метрики
Mozilla (compatible; YandexCataIog; Dyatel)	«простукивалка» Яндекс.Каталога
Mozilla (compatible; YandexNews)	индексатор Яндекс.Новостей
Mozilla (compatible; YandexNews)	индексатор Яндекс.Новостей

Основной индексирующий робот – индексирует основной объем текстовой информации, размещенной в сети. Индексирует HTML, а также другие типы документов, содержащих текстовые данные.

Робот, определяющий зеркала сайтов – так называемый «зеркальщик», определяет зеркала веб-сайтов, в том числе и как отображать веб-сайт, с «www» или без (например, https://www.1obl.ru или https://1obl.tv). Апдейт зеркальщика – учет изменений, найденных роботом, происходит довольно редко, 1 раз в 1–2 месяца.

Индексатор Яндекс.Картинок – отвечает за индексацию картинок в Интернет. Индексирует все популярные форматы картинок. Апдейт происходит в среднем раз в неделю, иногда чаще.

Индексатор Яндекс.Видео – отвечает за поиск видео. Ранжирование осуществляет за счет анализа текста, окружающего файл с видео на странице, а также популярности ролика в блогах и т.д.

Робот, индексирующий мультимедийные данные – индексирует документы в формате Adobe Flash.

Робот поиска по благим, индексирующий комментарии постов – специальный робот, индексирующий посты в блогах. Как правило, сами записи в блогах после опубликования практически никогда не изменяются, в отличие от списка комментариев, который постоянно растет. Видимо, для того чтобы не нагружать основного индексирующего робота, и был создан рассматриваемый. Блоги, как правило, имеют ограниченный список «движков» – платформ, на которых они построены, и с определением, является ли конкретный сайт блогом, проблемы не возникает.

Робот, обращающийся к странице при добавлении ее через форму «Добавить URL» (http://webmaster.yandex.ru/addurl.xml) – при добавлении нового веб-сайта или документа через форму на странице Яндекс, происходит обращение данного робота. Посещение основного индексирующего робота может занять от нескольких дней до нескольких месяцев.

Робот, индексирующий пиктограммы сайтов (favicons) – робот, индексирующий пиктограммы веб-сайтов, которые затем отображаются рядом со ссылкой в результатах поиска.

Робот, индексирующий страницы сайтов, участвующих в Рекламной сети Яндекса – робот, индексирующий веб-сайты на которых показываются рекламные объявления Яндекс (Яндекс.Директ).

«Простукивапка» Яндекс.Директа – робот, проверяющий работоспособность веб-сайтов, размещающих на своих страницах рекламные объявления Яндекс.Директ, а также веб-сайты, рекламирующиеся в нем.

Робот Яндекс.Метрики – робот, проверяющий работоспособность страниц, на которых установлен код Яндекс.Метрики (позволяет анализировать поведение посетителя на веб-сайте).

«Простукивалка» Яндекс.Каталога – робот, проверяющий на работоспособность веб-сайты, размещенные в Яндекс.Каталогс.

Индексатор Яндекс.Новостей – специальный робот, индексирующий часто обновляемые новостные ресурсы, которые участвуют в проекте Яндекс.Новости (https://news.yandex.ru/).

Робот мобильных сервисов – информации об этом роботе на официальном блоге Яндекс нет, но, судя по названию, этот робот индексирует wap-сайты.

Помимо перечисленных, в 2005 году Яндекс запустил «быстрый робот» (далее – быстроробот), который работает одновременно с основным индексирующим и предназначен для оперативного обнаружения и индексации актуальных страниц (рисунок 2). По словам Яндекс быстрый робот использует некую информацию о востребованных пользователями документах и на основании этого находит новые и измененные страницы, делая их доступными в результатах поиска в течение короткого времени. Это время измеряется в минутах, а страницы, обнаруженные быстророботом, можно определить в результатах поиска по пометке, когда документ был проиндексирован [40].

Для того чтобы быстроробот посещал веб-сайт, достаточно добавлять по 1 новому, уникальному документу ежедневно.

Результаты работы любого из вышеперечисленных роботов, за исключением быстроробота и Индексатора Яндекс.Новостей, можно увидеть лишь после обновления поисковой базы (так называемого «апдейта»). Как правило, апдейты в Яндекс происходят 1-2 раза в неделю. На сегодняшний день они делятся на два типа: текстовые и ссылочные. В первом случае в основную базу, по которой происходит поиск, добавляются новые страницы. При этом, естественно, данные из базы быстроробота удаляются и переходят в основную базу. Так как в основной базе обновляются измененные и появляются новые документы, изменяется и чис-

ло ссылок на веб-сайты. Эти ссылки на данном этапе не учитываются, т.е. не дают вклад в ссылочное ранжирование. В ссылочный апдейт происходит учет найденных ранее ссылок, без добавления в основную базу новых документов.

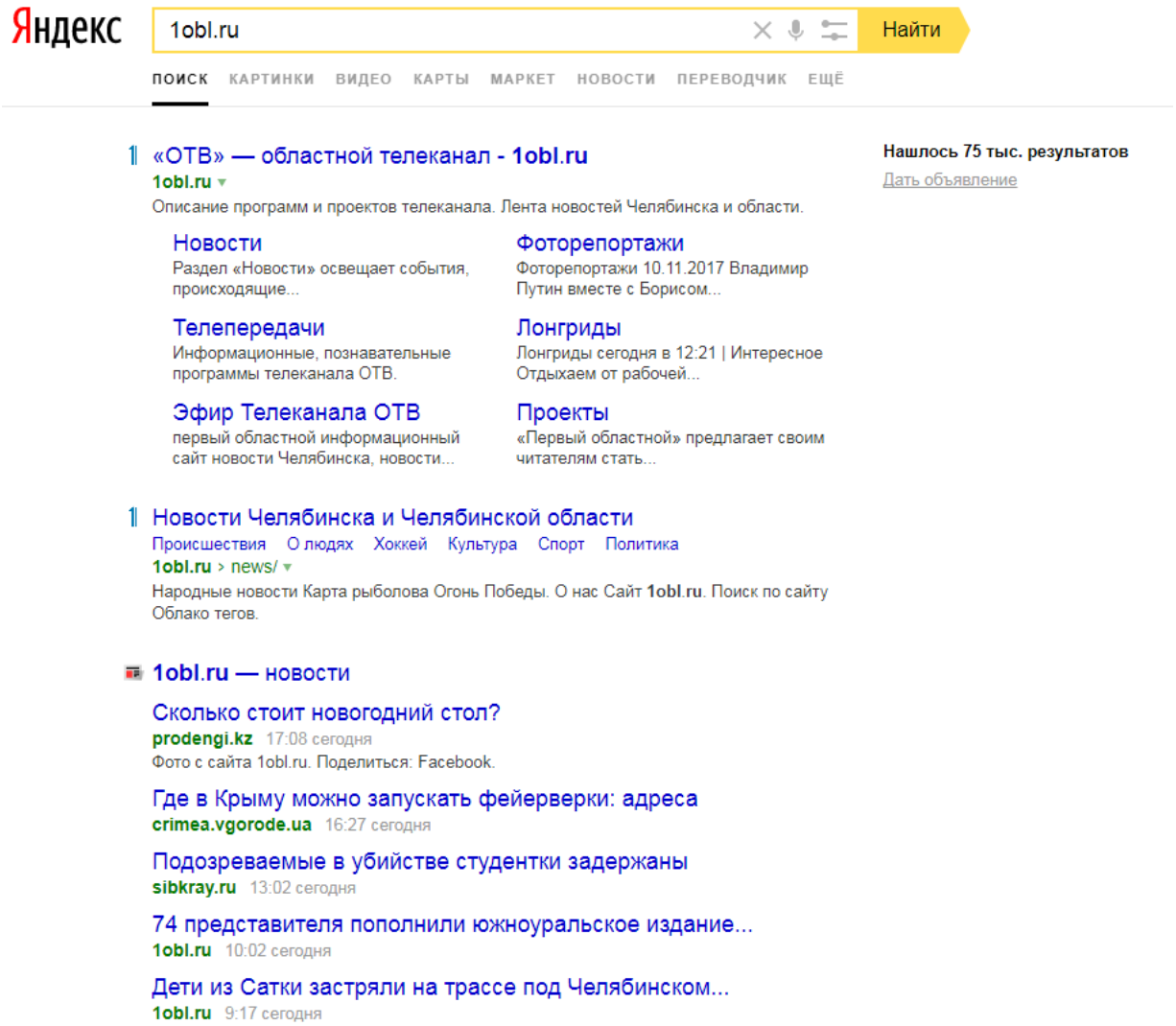

Рисунок 2 – Страницы, найденные быстроботом Яндекс

Язык поисковых запросов

По умолчанию Яндекс ищет все формы слова, указанного в запросе. Например, при запросе «рассказал» поиск будет производиться по глагольным формам: «рассказать», «расскажу», «рассказывать» (но не по однокоренным словам типа «рассказ», «рассказчик»).

Для использования расширенных возможностей поиска Яндекс позволяет вводить запросы, используя собственный язык (таблица 2).

Таблица 2 – Примеры собственного языка запросов

Пример	Значение
«Смотрите на 1obl.tv противостояние»	Слова идут подряд в точной форме
«Трактора» и * «Динамо»	Пропущено слово в цитате (в данном случае «минского»)
Матч & эфир	Слова в пределах одного предложе-
Статистика && «Трактор»	Слова в пределах одного документа
«Трактора» | «Динамо» | онлайн	Поиск любого из слов
«Трактора» << победа	Неранжирующее «и»: выражение после оператора не влияет па позицию документа в выдаче
Сегодня, челябинский «Трактор»/ 2 обыграл	Расстояние в пределах двух слов в любую сторону (то есть между заданными словами может встречаться одно слово)
«Трактор» на своем льду &&/ 3 одержал победу	Расстояние в 3 предложения в любую сторону
«Трактор» на своем льду ~~ поражение	Исключение слова «пойму» из поиска
Смотрите /+2 трансляцию	Расстояние в пределах двух слов в прямом порядке
«Трактор» ~ проиграл	Поиск предложения, где слово «Трактор» встречается без слова «проиграл»
Хоккейный клуб /(-1 +2) юбилей	Расстояние от одного слова в обратном порядке до двух слов в прямом порядке
!«Трактор» !матч !20 !ноября	Слова в точной форме с заданным регистром

Окончание таблицы 2

статистика && (+последних | !игр)	Скобки формируют группы в сложных запросах
!! хоккей	Словарная форма слова
title:(в стране)	Поиск по заголовкам документов
url:.www.1obl.ru	Поиск по URL
1obl tv	Поиск с учетом фрагмента URL
host: 1obl.tv	Поиск по хосту

Среди всех вышеперечисленных конструкций полезным будет поиск по URL – он позволяет определить количество документов, проиндексированных Яндекс для каждого домена. Также полезен поиск точной фразы – он позволяет найти дубликаты текстов в Интернет. Поиск точной словарной формы в сочетании с поиском точной фразы дают возможность правильно воспользоваться статистикой поисковых запросов Яндекс.

Статистика поисковых запросов

Для оценки частотности того или иного запроса в Яндекс необходимо воспользоваться статистикой поисковых запросов (http://wordstat.yandex.ru).

Сервис позволяет проанализировать частотность запросов с учетом регионов и сезонности (необходимо выбрать регион из списка).

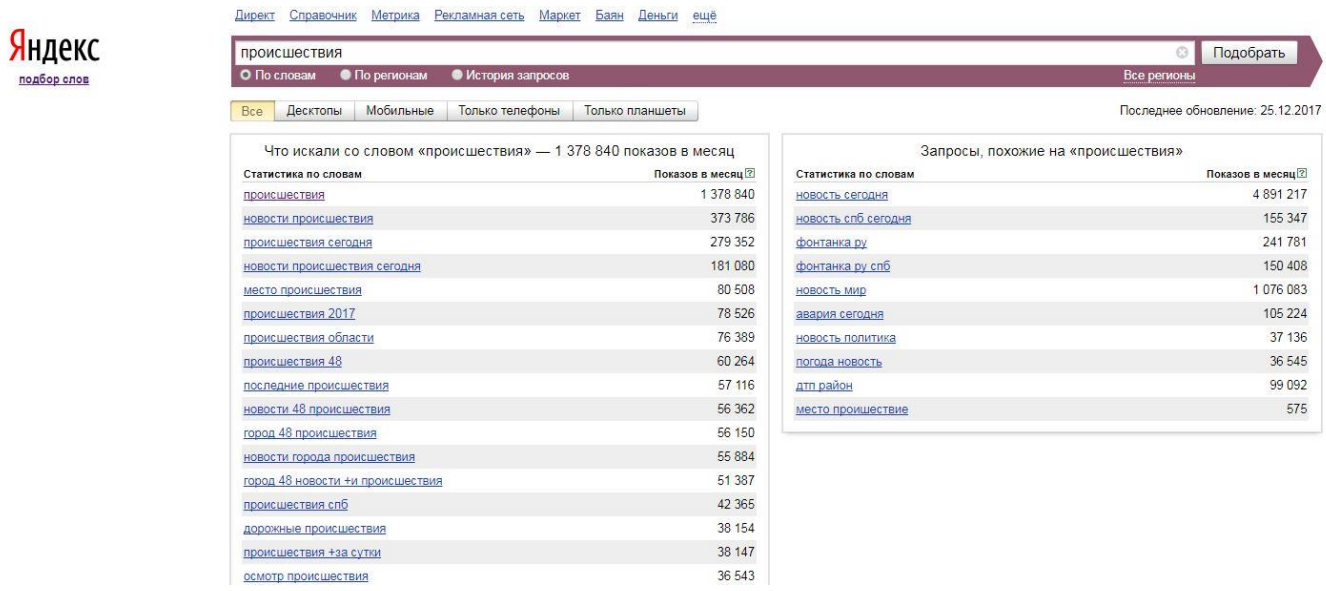

Рисунок 3 – Статистика запросов Яндекс

Статистика разбита на две колонки (рисунок 3). В первой показана статистика по запрашиваемым словам, в данном примере «происшествие» – 1 378 840 показов в месяц. Эта цифра показывает суммарное количество запросов, учитывая как двухсловный запрос «новости происшествия», так и его расширения в виде трех-, четырехсловных и более длинных запросов («новости города происшествия»). Исходя из рисунка 3 число 373 786 – статистика запрашиваемости двухсловного запроса «новости происшествия». Но это не совсем так. Для точного определения, сколько же в месяц в Яндекс спрашивают «новости происшествия», необходимо использовать язык поисковых запросов Яндекс. Вводим «новости! Происшествия!» – поиск точной фразы с учетом словоформ (рисунок 4).

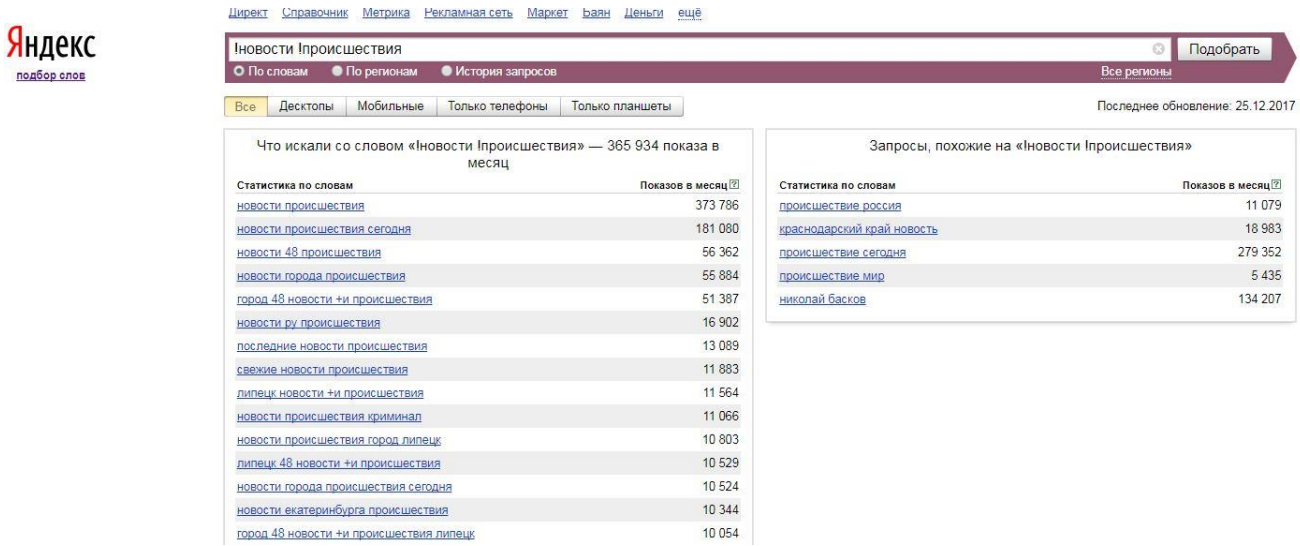

Рисунок 4 – статистика запросов Яндекс

Количество запросов – 365 934. Именно столько ищут фразу «новости происшествия». 90% веб-мастеров, использующих статистику запросов Яндекс, совершают подобную ошибку.

В правой колонке (см. рисунок 4) показаны запросы, которые вводили пользователи после того, как спросили у Яндекс «новости происшествия». Скорее всего пользователи не нашли ответ на свой вопрос и пытались его переформулировать. Таким образом, в правой колонке можно получить смежные, тематические запросы.

Метод шинглов

В случае полных совпадений никакой сложности в определении схожести документов нет. Поисковая машина отбрасывает HTML-теги, графические элементы, сквозное меню и т.д., оставляя лишь текст и сравнивая его с уже проиндексированными документами. Однако, если текст документа немного изменен либо «разбавлен» другими словами, такой подход уже не дает результата. Для определения так называемых «нечетких» дублей используют метод шинглов [14].

Этапы метода шинглов:

1. Канонизация текста;
2. Разбиение на шинглы;
3. Вычисление хэшей шинглов с помощью 84х статических функций;
4. Случайная выборка 84 значений контрольных сумм;
5. Сравнение, определение результата.

Канонизация текста

Для последующего анализа необходимо привести текст к нормальной форме. Из текста удаляются предлоги, союзы, HTML теги, знаки препинания, – то есть части речи, а также другие элементы, не несущие смысловой нагрузки.

Существительные приводятся к именительному падежу, единственному числу, либо от них остаются только корни. Классический метод шинглов не использовал канонизацию, но существующие реалии заставили поисковые системы его усовершенствовать. После того, как Яндекс перестал отображать в результатах поиска дубликаты, веб-мастера стали использовать различные генераторы текстов для придания «уникальности». В скопированный текст добавлялись предлоги, частицы, прилагательные и наречия заменялись синонимами. В Интернете появилось множество программ, называемых «синонимайзерами», которые выполняли все вышеперечисленные действия автоматически. Это и заставило разработчиков алгоритмов поисковой системы Яндекс усовершенствовать классический метод шинглов.

На выходе получается текст, очищенный от «мусора» и готовый для сравнения (рисунок 5).

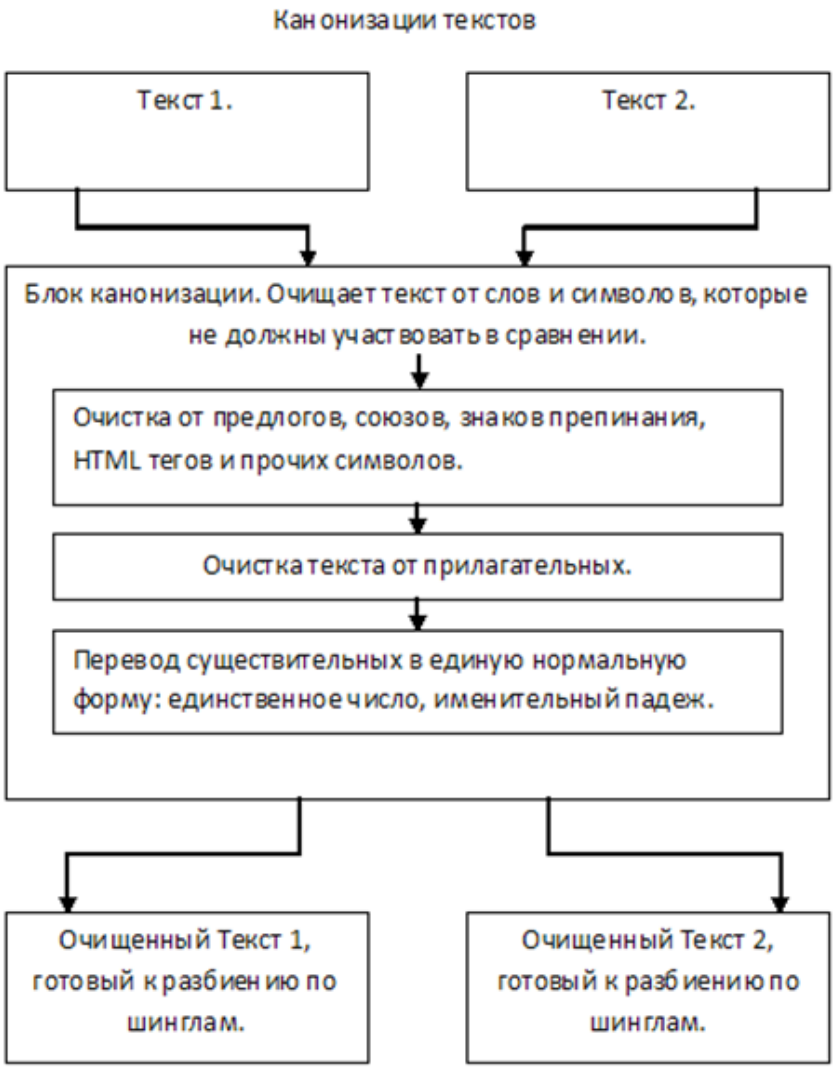

Рисунок 5 – Канонизация текстов

Разбиение на шинглы

Шинглы (англ) – «чешуйки», выделенные из статьи подпоследовательности слов. Длина шингла – это количество слов, которые берутся для сравнения. Рассмотрим пример, в котором длина шингла равна 10 (в Яндекс используется длина шингла равная 5-6 словам). Выборка происходит «внахлест».

При разбиении текста на шинглы таким образом (рисунок 6), получается набор шинглов в количестве, равном количеству слов минус длина шингла плюс один (формула 13):

$$N_{shi} = N_{num\_words} - L_{shi} + 1, \quad (13)$$

где N_{num_words} – количество слов, L_{shi} – длина шингла.

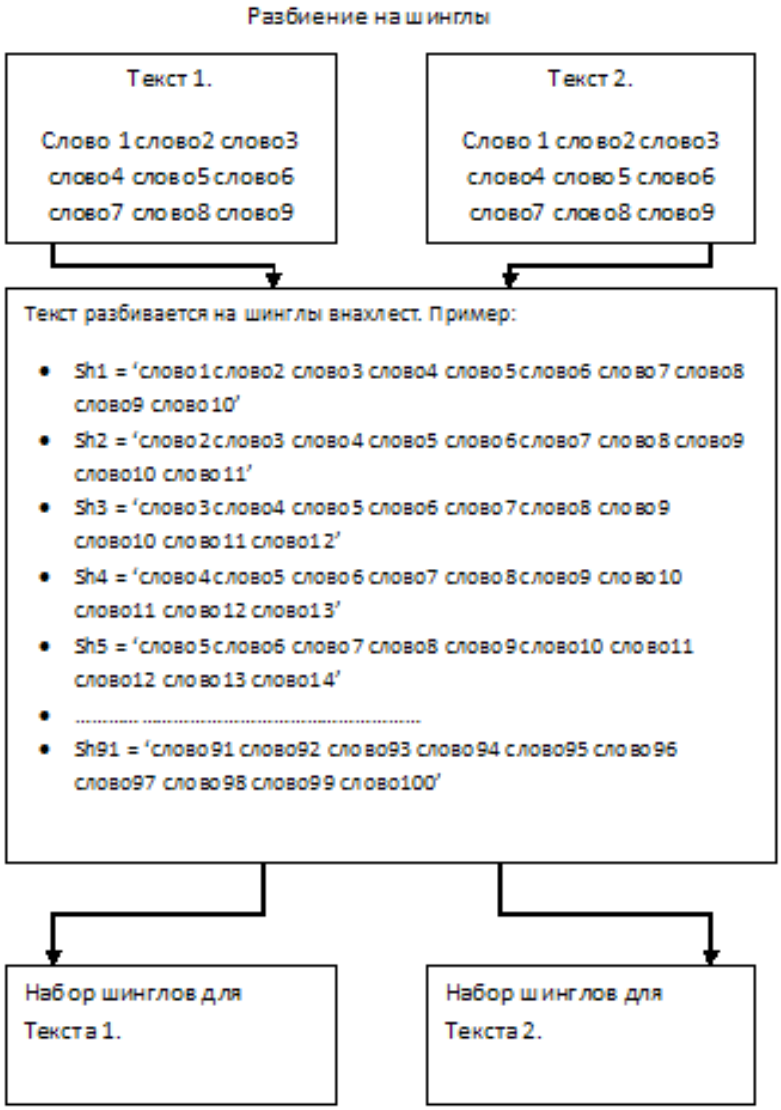

Рисунок 6 – Разбиение на шинглы

Вычисление хэшей шинглов с помощью 84-х статических функций

Принцип сравнения текстов по алгоритму шинглов заключается в сравнении случайной выборки контрольных сумм шинглов двух текстов между собой [37].

Проблема алгоритма заключается в количестве сравнений, ведь это напрямую отражается на производительности. Увеличение количества шинглов для сравнения характеризуется экспоненциальным ростом операции, что критически отразится на производительности [31].

Предлагается представить текст в виде набора контрольных сумм, рассчитанных через 84 уникальные между собой статические хэш функции. Для каждого шингла рассчитывается 84 значения контрольной суммы через разные функции (например, SEA1, MD5, CRC32 и т.д., всего 84 функции). Таким образом, каждый из текстов будет представлен в виде двумерного массива из 84-х строк, где каждая строка характеризует соответствующую из 84-х функций контрольных сумм.

Из полученных наборов будут случайным образом отобраны 84 значения для каждого из текстов и сравнены между собой в соответствии функции контрольной суммы, через которую каждый из них был рассчитан. Таким образом, для сравнения будет необходимо выполнить всего 84 операции (рисунок 7):

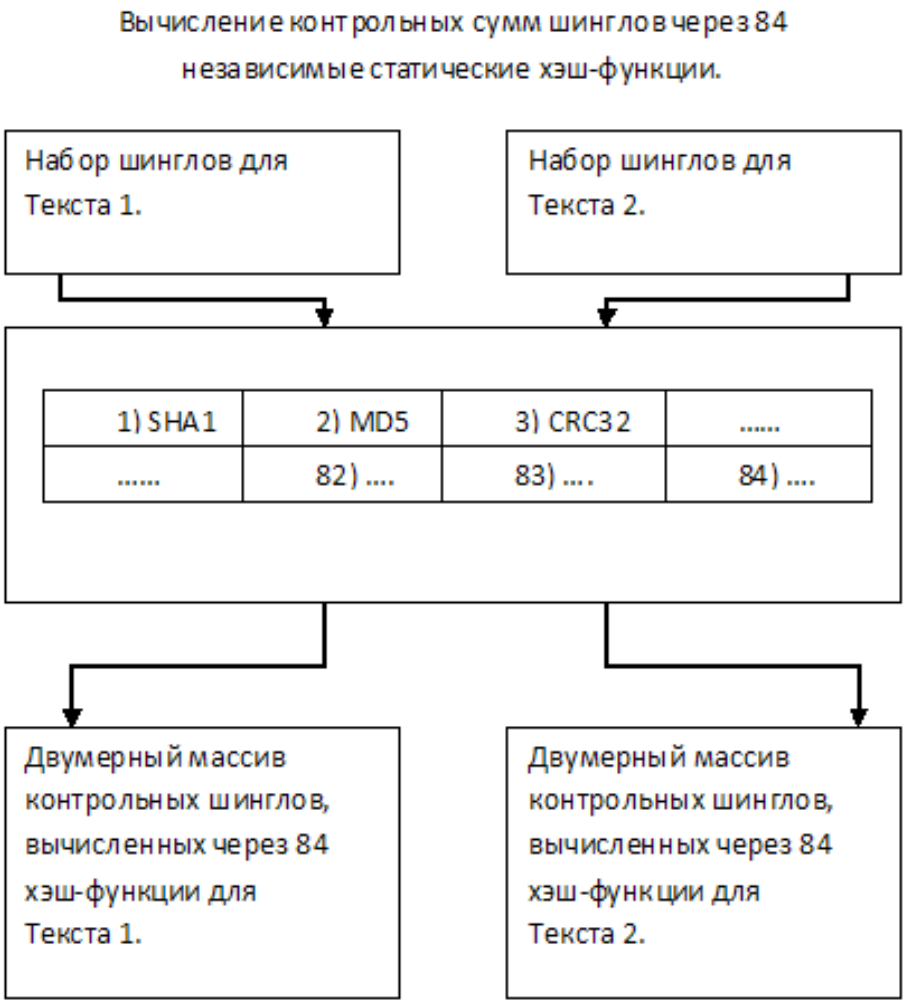

Рисунок 7 – Вычисление контрольных сумм

Случайная выборка 84-х значений контрольных сумм

Сравнивать элементы каждого из 84-х массивов между собой – ресурсоемко. Для увеличения производительности выполняется случайная выборка контрольных сумм для каждой из 84-х строк двумерного массива, для обоих текстов. К примеру, можно выбрать минимальное значение из каждой строки.

На выходе получается набор из минимальных значений контрольных сумм шинглов для каждой из хэш функций (рисунок 8):

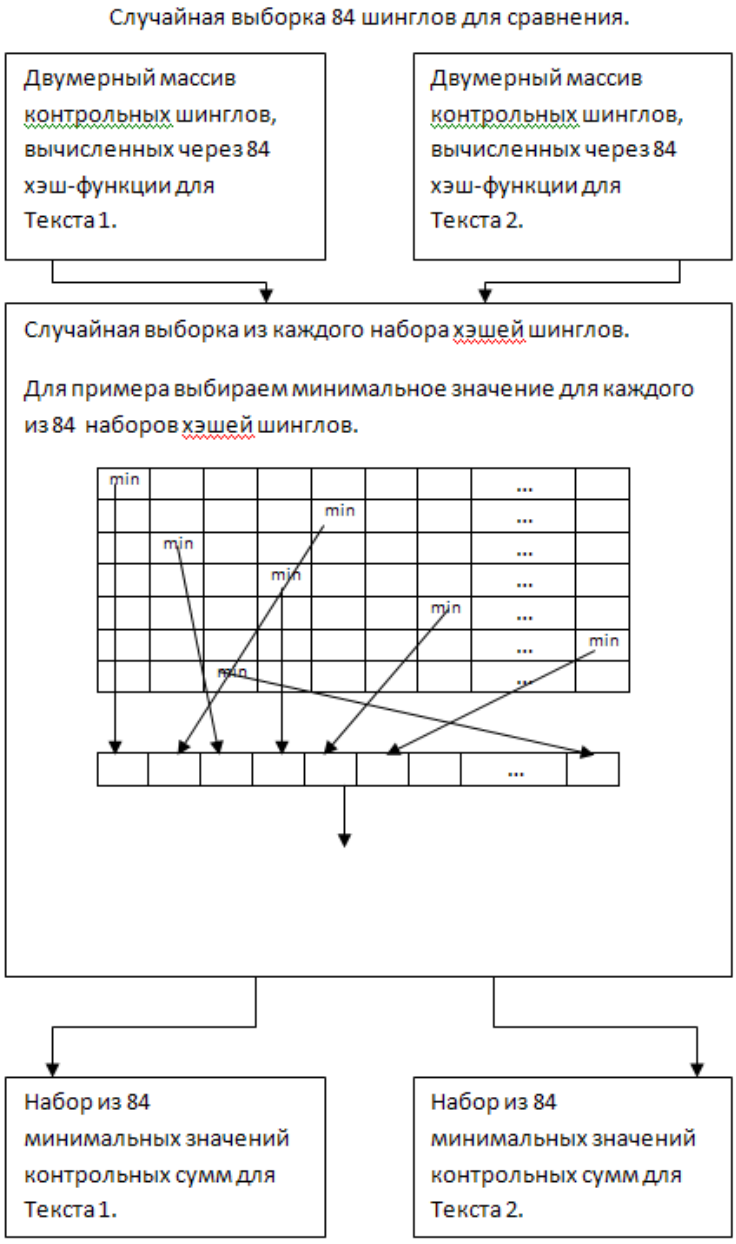

Рисунок 8 – Случайная выборка шинглов для сравнения

Сравнение, определение результата

При сравнении между собой 84-х элементов первого массива с соответствующими 84-ю элементами второго массива рассчитывается отношение одинаковых значений (рисунок 9):

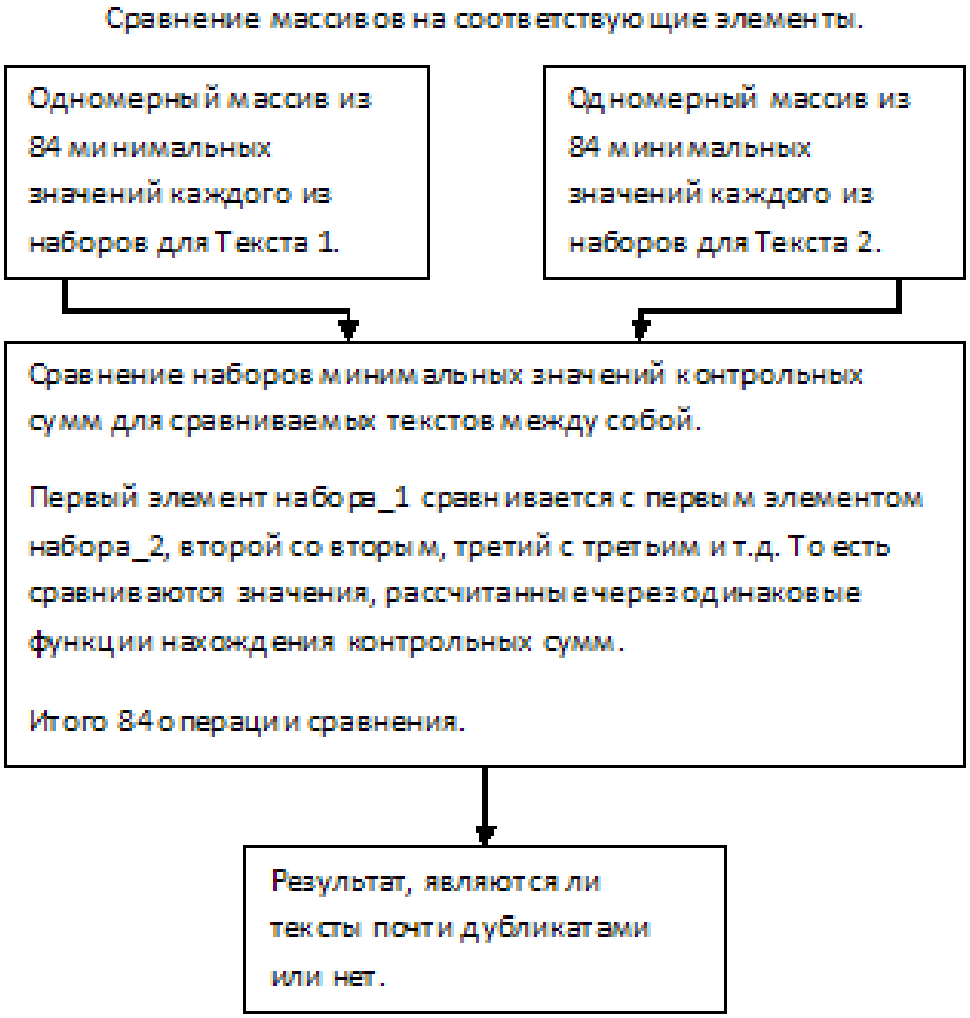

Рисунок 9 – Сравнение массивов и получение результатов

Поиск с учетом региона

От 15 до 30% всех запросов, которые пользователи задают Яндекс, – геозависимые [45]. Это запросы, по которым, пользователи рассчитывают получить локальную (региональную) информацию, к примеру, об происшествиях в своем городе. На такие запросы Яндекс ответит по-разному, для каждого города/региона предложив пользователю различные результаты поиска.

Многие запросы, наоборот, не зависят от региона, в котором находится пользователь, например, при поиске книг, фильмов или рецептов. Такие запросы называются геонезависимыми, и результаты поиска в данном случае не зависят от региона пользователя [41].

Существуют также запросы, которые в разных регионах имеют разное значение. К примеру, по запросу «орбита» москвичи чаще всего ищут кинотеатр, в Ростове-на-Дону – автосалон, израильтяне – интернет-портал.

Умение различать запросы пользователя повышает качество поиска и позволяет дать пользователю точный ответ без повторных уточняющих запросов.

Геозависимые запросы определяются статически, т.е. в запросе необязательно должно быть указано название города. К примеру, запрос «купить кондиционер» подразумевает, что пользователь ищет услугу в пределах своего города/населенного пункта. Местонахождение пользователя определяется на основе IP-адреса [10]. Регион поиска по умолчанию привязан к IP-адресу, возможны также ручные настройки (https://yandex.ru/tune/geo).

Для геозависимых запросов Яндекс показывает различные результаты поиска для разных регионов. При этом в результатах поиска могут участвовать и авторитетные общероссийские веб-сайты, но приоритет все же отдается региональным сайтам.

Принадлежность веб-сайта к тому или иному региону определяется по многим признакам, в их числе – его IP-адрес, контактная информация, указанная на веб-сайте, регион, которому посвящена информация на нем [17].

Для веб-сайтов, зарегистрированных в Яндекс.Вебмастер (http://webmaster.yandex.ru) можно указать регион самостоятельно. Также регион могут присвоить модераторы при добавлении веб-сайта в каталог Яндекс (http://yaca.yandex.ru)

Веб-сайты, на страницах которых указано множество адресов из различных регионов (например, почта России), могут признаваться общероссийскими, т.е. получить регион «Россия» и участвовать при поиске из любого города. Список

регионов/городов, для которых осуществляется поиск – (http://search.yaca.yandex.ru/geo.c2n).

История алгоритмов

Первая значительная смена формулы ранжирования с момента запуска поисковой системы Яндекс произошла 20 декабря 2007 года. Именно тогда Яндекс стал использовать переформулирование запросов и практически на любой запрос стал выдавать сайты, подходящие под ключевые слова. До этого момента Яндекс соответствовал надписи на собственном логотипе при отсутствии найденных результатов: «Найдется все... со временем!», т.к. на некоторые запросы ответов не находилось.

Практически сразу вслед за этой сменой формулы раткирования последовали еще две: алгоритм «8 sp 1» начал использоваться 17 января и 5 февраля 2008 года.

Примерно в это время началось массовое использование поискового спама – попытки искусственно повышать позиции со стороны веб-мастеров. Все последующие версии алгоритмов помимо новшеств для пользователей поисковой системы включали также новые способы борьбы с поисковым спамом.

Последующие версии алгоритмов Яндекс называл именами городов. Формула ранжирования от 16 мая 2008 года была названа «Магадан». Помимо текста на странице Яндекс стал учитывать переведенные и транслитерованные URL-адреса, повысилась точность распознавания фамилий и географических названий.

Вскоре специалисты Яндекс внесли корректировки в алгоритм, и 2 июля 2008 года был запущен «Магадан 2.0». Важнейшим нововведением алгоритма было определение первоисточника.

9 июля 2008 года был проведен запуск бета-версии алгоритма «Находка». Доработывали этот алгоритм долго, и итоговая версия «Находки» появилась лишь 11 сентября 2008 года. В ней был реализован инновационный подход к машинному обучению, изменен учет «стоп-слов», расширен тезаурус.

Следующим алгоритмом в Яндекс стал «Арзамас» (Анадырь), запущенный в работу 10 апреля 2009 года. С этим алгоритмом Яндекс научился лучше понимать

русский язык. Теперь при запросе «реконструкция Маяковской» показывается также страница со словами «станция Маяковская открылась после реконструкции», и т.д. Также именно в этом алгоритме появилась геозависимая выдача.

Алгоритм дорабатывался не один раз. Проводились изменения формулы ранжирования, добавления новых регионов. В итоге к последней версии Яндекс предоставлял локальные результаты поиска для 19 регионов.

17 ноября 2009 года был запущен алгоритм «Снежинок». Количество учитываемых при ранжировании параметров увеличилось в разы, благодаря чему повысилось качество поиска Яндекс. В этой версии впервые использовался новый алгоритм машинного обучения «Матрикснет» [38].

22 декабря 2009 года Яндекс расширил количество городов, для которых производится локальное ранжирование. Обновленная версия получила название «Конаково» (неофициальное название обновленного «Снежинска»).

10 марта 2010 года Яндекс анонсировал новую версию – Снежинок 1.1, работа которого началась 17 марта. В ней была улучшена общая формула ранжирования для геонезависимы запросов.

Вслед за «Снежинск 1.1» 13 сентября 2010 появился «Обнинск» – новая формула ранжирования для геонезависимых запросов. По словам разработчиков Яндекс, формула ранжирования выросла почти в 2,5 раза (до 280 Мб).

Помимо постоянного развития алгоритмов ранжирования Яндекс предлагает пользователям уникальные сервисы. Например, Яндекс первым предложил мониторинг автомобильных пробок в крупнейших городах. К тому же Яндекс на сегодняшний день – это наш, «русский» поисковик, единственный из большой тройки (Яндекс, Рамблер, Апорт), который успешно развивается. Доверие пользователей к этой поисковой системе огромно, она по праву является самой популярной в России и будет оставаться такой еще долгие годы. Даже если развитие интернет – проекта приостанавливается, он будет еще популярен как минимум 3-5 лет. Поэтому исследование формулы ранжирования Яндекс – важная задача, которой уделена большая часть данной работы.

1.3.2 Google

Поисковая система Google (http://www.google.ru) индексирует и осуществляет поиск по следующим форматам документов: HTML, PDF, RTF, DOC, XLS). Google в отличие от Яндекс использует для индексирования всего лишь одного робота, User-agent которого – «Googlebot». Он также осуществляет поиск по картинкам и медиафайлам. Робот переходит по ссылкам, находя в документах конструкции SRC и HREF формата HTML. Самостоятельно сообщить Google о новом сайте можно, используя форму на странице http://www.google.m/addurl.html. Google не имеет быстроробота, но скорость индексации Googlebot очень высока, и новые документы с качественных веб-сайтов попадают в поисковую базу в течение нескольких часов [37].

Google позволяет переводить запрос с русского языка на множество других и осуществлять поиск документам на различных языках. Веб-сайт Google доступен в доменных зонах различных стран (.ru, .ua, .com и т.д.), алгоритмы которых имеют свои отличия. В данной работе исследуется поисковый алгоритм, расположенный в домене www.google.ru.

ВЫВОДЫ ПО РАЗДЕЛУ 1

В первой главе рассмотрены устройство и общая схема работы поисковых систем. Приведены принципы определения текстовой релевантности, расчет популярности на основе гиперссылок, а также история поисковых систем Рунет, эволюция алгоритмов Яндекс. Ввиду популярности и перспективности для Рунет Яндекс и Google (www.google.ru) в дальнейшем в работе рассматривается работа этих двух поисковых систем.

ГЛАВА 2. ОПРЕДЕЛЕНИЕ ФАКТОРОВ, УЧАСТВУЮЩИХ В ФОРМУЛЕ РАНЖИРОВАНИЯ

2.1 Группы факторов, влияющих на релевантность

Важнейшей задачей в поисковой оптимизации является не только знание существующих факторов, участвующих в формуле ранжирования, но и умение анализировать изменения, происходящие при ее смене алгоритмов поисковых систем.

Факторы, используемые поисковыми системами при расчете релевантности, можно разбить на три группы: внутренние, внешние и доверие к веб-сайту [16].

К внутренним факторам относятся те, которые характеризуют свойства непосредственно самого документа и веб-сайта в целом. К примеру, наличие слов «происшествие в Челябинске» в тексте документа делает его релевантным запросам «происшествие в Челябинске», «происшествие», «Челябинск», «новости происшествия» и т.д. Свойства веб-сайта в целом – это корректная с точки зрения поисковых систем структура ссылок и HTML-кода страниц, уникальность текстов документов, размещенных на нем, и т.д. Внутренние факторы обеспечивают вклад в текстовое ранжирование, которое учитывается при определении релевантности в любых поисковых системах как в Интернет, так и в системах локального поиска

К внешним факторам относится учет ссылок с других веб-сайтов. Группа факторов называется внешней, поскольку вклад в определение релевантности документа одного веб-сайта дает множество ссылок, расположенных на других веб-сайтах. Такие ссылки называются внешними. Для определения релевантности учитывается анкор (текст) ссылки, которая в простейшей конструкции языка HTML выглядит следующим образом:

анкор, где визуально анкор – это текст, кликая мышью по которому пользователь совершает переход на другой веб-сайт.

Доверие поисковых систем к веб-сайту – один из важнейших факторов при определении релевантности документа запросу. Доверие (авторитетность, «траст»

от англ. trust – доверять) – относительно молодой фактор, который появился всего 2-3 года назад [12]. Смысл его в том, что у веб-сайта, признанного поисковой системой авторитетным, больше шансов занять высокие места в результатах поиска при прочих равных, т.е. получить большую релевантность [26]. Достаточно лишь упоминания слов запроса в документе «трастового» сайта – и он уже на первых страницах результатов поиска. К тому выводу можно прийти, проанализировав, к примеру, веб-сайт Википедии (http://ru.wikipedia.org), который по многим запросам находится на первой странице результатов поиска. При этом многие веб-сайты, расположенные ниже в результатах поиска, обладают куда более скромными данными с точки зрения внешних факторов релевантности.

2.2 Определение факторов, участвующих в формуле ранжирования, методом экспертных оценок

Для определения факторов и их важности использовался метод экспертных оценок [29]. Он позволил объединить знания более 100 экспертов в области поисковой оптимизации для выявления всех возможных факторов, влияющих на релевантность документа запросу [1], а также важность каждого из них в рамках нормированной шкалы. На первом этапе каждому эксперту предлагается перечислить всевозможные факторы, которые могут участвовать в формуле ранжирования. Применяется так называемый мозговой штурм – один из вариантов экспертного оценивания, при котором можно высказывать любые собственные идеи, но наложено одно очень существенное ограничение – нельзя критиковать идеи других экспертов. В итоге, в процессе высказывания идей получается максимально возможное количество факторов.

На втором этапе происходит анализ факторов, полученных на первом этапе. Каждое эксперту предлагается анкета содержащая полный список факторов. Устанавливается шкала от 0 до 1 степени важности фактора, шаг принимается равным 0,1. Каждый эксперт высказывает свое мнение по каждому из факторов, полученных на первом шаге в рамке принятой шкалы. Результирующая оценка получается путем вычисления среднего арифметического результатов анализа

всех экспертов в нормированной шкале с учетом коэффициента доверия к каждому эксперту. Факторы, важность которых в рамках установленной шкалы получилась равной менее 0.1, исключались.

Вопросы в анкете разделяются на два блока:

1. Вопросы, характеризующие степень доверия к эксперту, его профессиональный уровень.

2. Вопросы, характеризующие важность факторов, участвующих в формуле ранжирования.

Обработка результатов и формирование экспертных оценок осуществляется последовательно, начиная с вопросов Блока 1. По каждому эксперту, по каждому вопросу этой группы в зависимости от номера выбранного ответа из таблицы пересчета выбирается коэффициент, характеризующий степень доверия к этому эксперту, и определяется среднее значение коэффициента доверия по всем вопросам Блока 1 для каждого эксперта K_1. В дальнейшем все ответы эксперта умножаются на этот коэффициент.

Вопросы Блока 2 направлены на оценку важности факторов, участвующих в формуле расчета релевантности. Все ответы приводятся к нормированной шкале (от 0.0 до 1.0). Факторы с окончательной оценкой важности менее 0.1 исключаются из окончательного списка.

Таким образом, оценка важности фактора определяется по формуле 14:

$$h_i = \frac{1}{x} \sum_{j=M} \sum_{i=N} K_i p_{ij} \,, \tag{14}$$

где p_{ij} – коэффициент, который определяется на основании анкет в нормированной шквале (от 0.0 до 1.0);

K_i – коэффициент доверия к эксперту;

N – число экспертов;

M – количество слов в анкете.

В результате обработки оценок экспертов получилось три группы критериев и оценка важности каждого из них.

Внутренние факторы поисковой оптимизации

Важность факторов по методу экспертных оценок (в дальнейшем «ключевое слово») в таблице 3:

Таблица 3 – Важность факторов по методу экспертных оценок

Фактор	Оценка
1. Точное вхождение ключевого слово на странице в рамках пассжа	0.9
2. Точное вхождение ключевого слова в тег title	0.9
3. Точное вхождение ключевого слова в тег keywords	0.1
4. Точное вхождение ключевого слова в тег description	0.1
5. Точное вхождение ключевого слова в теги hl-h6	0.7
6. Плотность ключевого слова в тексте страницы – до 5%	0.8
7. Общий объем полезного текста на странице 1000-2000 знаков	0.5
8. Обновление веб-сайта	0.7
9. Выделение ключевого слова жирным шрифтом (теги или <b>)	0.3
10. Точное вхождение ключевого слова в текстах гиперссылок в других документах веб-сайта	0.5
11. Уникальность документа внутри веб-сайта	0.8
12. Уникальность документа и веб-сайта в целом в Интернет	1.0
13. Корректная работа скриптов	0.6
14. Запрет от индексации избыточных и служебных страниц и раделов	0.8
15. «Понятные» URI-адреса веб-страниц	0.4
16. Вынос java-script и ess в отдельные файлы	0.3
17. Главное зеркало для поисковых систем	0.5

Точное вхождение ключевого слова на странице в рамках пассажа

Релевантность документа повышается при наличии ключевого слова в пассаже документа. Поисковые системы делят документ на т.н. пассажи, в которых и осуществляется поиск. В многословных запросах пользователей при прочих равных,

если ключевые слова расположены в одном пассаже документа, он будет считаться более релевантным. Если в базе поисковой системы нет документов, в которых ключевые слова встречаются в одном пассаже, будут искаться документы, в которых слова расположены в разных пассажах. Поисковые системы «понимают» морфологию русского языка, поэтому допустимо изменение словоформ (чисел и падежей). В том случае, если ключевое слово употребляется в документе несколько раз, желательно употреблять его в разных словоформах. Также замечена небольшая разница в результатах поиска в Яндекс при введении букв запроса прописными и строчными буквами. При возможности желательно употребить ключевое слово в разных вариантах.

Точное вхождение ключевого слова в тег title

HTML-тег <title> имеет важнейшее значение в тексте страницы. Он отображает заголовок документа и должен содержать ключевые слова, которые употребляются в нем. Яндекс учитывает 15 слов в <title>, Google – 70 символов. Необходимо избегать бессмысленных повторов ключевых слов. Ввиду ограниченности учета текста в <title> ключевые слова необходимо распределять по разным документам веб-сайта.

Точное вхождение ключевого слова в тег keywords

Тег <keywords> должен содержать ключевые слова, описывающие документ. Влияние текста в этом теге на релевантность крайне мало, однако замечены случаи понижения релевантности при полностью совпадающем тексте для всех или большого числа документов внутри одного веб-сайта.

Точное вхождение ключевого слова в тег description

Тег <description> – описание документа. Его содержимое может выводиться в «сниппете» – выдержке из документа под ссылкой в результатах поиска (рисунок 10).

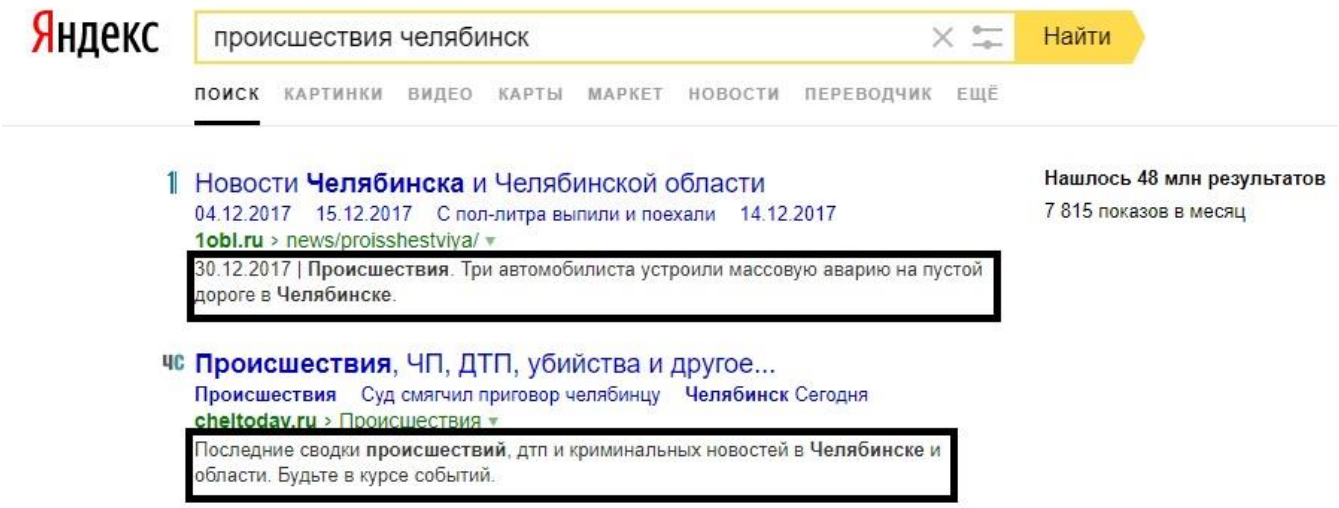

Рисунок 10 – Пример выдержки из документа под ссылкой («сниппет»)

Точное вхождение ключевого слова в теги hl-h6

Один из ключевых тегов в документе, обозначающий иерархию заголовков и подзаголовков на странице. Каждый документ как минимум должен содержать тег <hl> с ключевыми словами. Возможно дублирование текста из тега <title>.

Плотность ключевого слова в тексте страницы – до 5%

Чрезмерное употребление ключевого слова в документе приводи: понижению релевантности вплоть до исключения из результатов поиска. Возможно, сильное снижение релевантности и документ по ключевым словам будет находиться далеко от первых результатов поиска. Причем только по тем ключевым словам, по которым произошел «перебор» по плотности. «Идеального» процента не существует, но не стоит превышать порог в 5%. Перед проведением поисковой оптимизации лучше всего проанализировать первые 10-20 результатов поиска по интересующему ключевому слову и использовать средние значения при написании текста. Необходимо также исходить из простого правила – текст должен быть читаем.

Общий объем полезного текста на странице больше 1000-2000 знаков

Поисковые системы имеют ограничение на объем индексируемого документа. Страницы с большими объемами (3000 и более) текста не повышают релевант-

ность текста, в таких случаях рекомендуется разбивать текст на несколько страниц, связанных ссылками.

Обновление веб-сайта

Появление новых уникальных документов на веб-сайте положительно влияет на релевантность. Причем не обязательно наличие в этих документах ключевых слов, главная задача - показать поисковой системе, что веб-сайт не заброшенный, а постоянно развивается. При этом лучше ежедневно добавлять но одному документу, нежели в один день сразу выложить 30 документов. Под новыми уникальными документами понимаются те, которые содержат новую для поисковой базы информацию, не скопированную из источника ей известного. Текст должен нести смысловую нагрузку (страница, на которой размещено одно - единственное изображение. не является таковой), корректно отформатирован (разбит на абзацы и т.д.).

Выделение ключевого слова жирным шрифтом (теги или <b>)

Выделение ключевых слов в документе жирным шрифтом повышает его релевантность. По не стоит выделять ключевое слово несколько раз или форматировать таким образом целые предложения.

Точное вхождение ключевого слова в текстах гиперссылок в других документах веб-сайта

По аналогии с учетом внешних факторов – учетом анкоров внешних ссылок, при расчете релевантности используются и внутренние ссылки. Таким образом, повышается релевантность документа при наличии ключевых слов в анкорах ссылок внутри веб-сайта. Многие веб-сайты содержат меню «сквозных» (расположенных на всех страницах) ссылок, где на главную страницу указывает ссылка с анкором «главная». Если стоит задача повысить релевантность главной страницы по ключевым словам, то данный анкор недопустим. Эту сквозную ссылку необходимо закрыть от индексации, оставив ее при этом в меню для пользователей. На главную страницу должна вести ссылка с анкором в виде ключевых слов. Хорошо

повышают релевантность ссылки из текста страницы. Огличный пример оптимизации в данном случае – веб-сайт Википедии. Каждая страница ссылается на другую страницу с анкором в виде ключевых слов. При этом гораздо лучше, если все анкоры будут различны (но, естественно, будут содержать ключевое слово).

Уникальность документа внутри веб-сайта

Многие веб-сайты содержат большое количество документов, не имеющих полезной для поисковой системы информации. К примеру, результаты поиска по сайту, версии документов для печати (версии основных документов без графических элементов), метки (теги), списки документов, сортированные по датам, различные сортировки данных в таблицах по столбцам. Все эти данные уже проиндексированы поисковой системой. К примеру, текст новости проиндексирован по постоянному адресу, но выдержка из этой новости встречается в поиске по датам опубликования, метках и т.д. (рисунок 11).

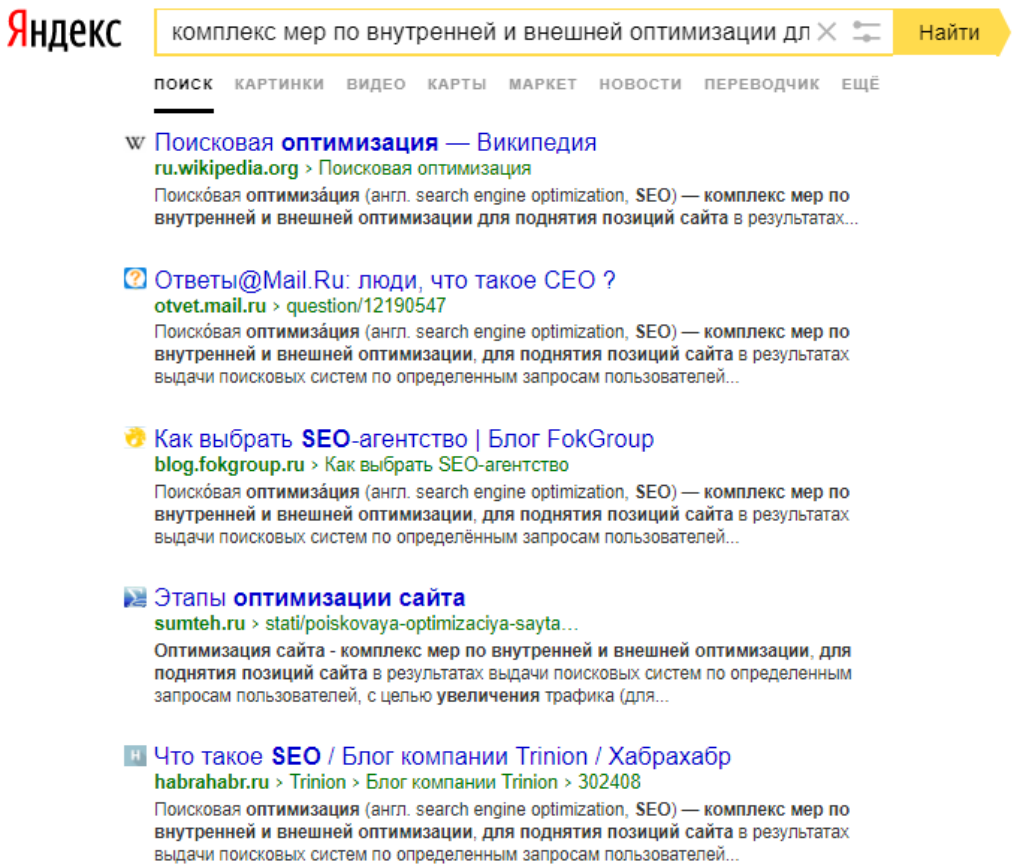

Рисунок 11 – Дублирование информации на веб-сайте

Значительного понижения релевантности замечено не было, но желательно избегать многочисленного повторения документа или его частей внутри веб-сайта.

Уникальность документа и веб-сайта в целом в Интернет

Один из важнейших факторов. Особенно важно для относительно новых (менее 3 лет) веб-сайтов. В случае наличия большого количества документов, скопированных с других веб-сайтов и при этом разрешенных для индексации поисковыми системами, веб-сайты могут полностью исключаться из результатов поиска или становиться нерелевантными к любым запросам. Авторитетные, старые веб-сайты могут размещать неуникальную информацию [32]. К таким можно отнести подавляющее большинство новостных ресурсов, которые копируют новости и пресс-релизы друг у друга и не получают никаких санкций со стороны поисковых систем.

Корректная работа скриптов

Большинство современных веб-сайтов работают на различных «движках» (скриптах), используют различные системы управления содержимым. «Чистых» (не берутся в расчет документы, созданные с помощью подмены адресов через ModRewrite [16] или их аналогами) статических документов HTML (т.е. с расширением .htm или .html) становится все меньше. При этом большинство готовых решений систем управления содержимым неправильно обрабатывают запросы к документам.

Правильная работа скриптов – залог исключения так называемых случайных дублей информации, а также пустых страниц. Допустим, документ с новостью доступен по адресу: http://www.site.ru/news.php?id=l – по URL-адресу можно с большой долей вероятности утверждать, что скрипт работает на языке PHP, а параметр id – идентификатор новости в используемой скриптом базы данных [91]. Тогда, если скрипт больше не использует никаких других параметров, при запросе вида http://www.site.ru/news.php?id=l&par=2 веб-сервер должен возвращать ошибку 404 – не найдено, чтобы поисковый робот не внес эту страницу в свой

индекс. Иначе может возникнуть дублирование информации по адресам:

http://www.site.ru/news.php?id=l

http://www.site.ru/news.php?id=l&par=2

поисковому роботу веб-сервером будет возвращен код 200 – найдено, и в индекс поискового робота могут попасть две одинаковые страницы с разными URL-адресами.

Аналогично должна быть организована проверка корректности значения параметра URL-адреса: в случае, если новость с идентификатором в базе данных не существует, – веб-сервер должен возвращать ответ 404 – не найдено, иначе (а так происходит на 99% веб-сайтах) веб-сервер вернет код 200 – найдено, и поисковый робот проиндексирует веб-страницу, состоящую из графического шаблона для страницы новостей без текстовой информации.

Запрет от индексации избыточных и служебных страниц и разделов

Используя стандарт исключений для роботов, необходимо запретить поисковому роботу индексировать служебные разделы веб-сайта (к примеру, папку со скриптами системы управления http://www.site.ru/admin/), версии документов для печати, различные документы, полученные путем сортировки таблиц по столбцам, а также любые другие документы, не несущие новой текстовой информации [47].

Для запрета индексации отдельных документов используется файл robots.txt, размещенный в корневой папке:

User-agent: Yandex,

Disallow: / # блокирует доступ ко всему сайту

User-agent: Yandex

Disallow: /cgi-bin # блокирует доступ к страницам, начинающимся с

'/cgi-bin'

В поле 'User-agent' указывается поисковый робот: Yandex или Googlebot. Значение 'User-agent’ для остальных поисковых систем можно найти в Интернет.

Затем идет блок с разрешениями ('Allow') или запретами ('Disallow') страниц и

разделов. Недопустимо наличие пустых переводов строки между директивами ‘User-agent’ и 'Disallow1’ ('Allow'), а также между самими ‘Disallow’ ('Allow') директивами.

Кроме того, т.к. директив ‘User-agent' может быть несколько в соответствии со стандартом перед каждой директивой 'User-agent' рекомендуется вставлять пустой перевод строки.

Символ ‘#’ предназначен для описания комментариев. Вес, что находится после этого символа и до первого перевода строки, не учитывается.

Чтобы разрешить доступ поисковому роботу к некоторым частям веб-сайта или индексировать его целиком, используйте директиву 'Allow':

User-agent: Yandex

Allow: /cgi-bin

Disallow: /

запрещает скачивать все, кроме страниц

начинающихся с ‘/cgi-bin'

При «наложении» директив выбирается первая в порядке появления в выбранном User-agent блоке:

User-agent: Yandex

Allow: /cgi-bin

Disallow: /

запрещает скачивать все, кроме страниц

начинающихся с ‘cgi-bin'User-agent: Yandex

User-agent: Yandex

Disallow: /

Allow: /cgi-bin

запрещает скачивать весь сайт

Директивы Allow-Disallow без параметров.

Отсутствие параметров у директивы трактуется следующим образом:

User-agent: Yandex

Disallow: # тоже что и Allow: /

User-agent: Yandex

Allow: # тоже что и Disallow:/

Несмотря на то, что стандарт исключений для роботов не является обязательным для использования, большинством поисковых систем руководствуются содержимым файла robots.txt перед индексированием документов веб-сайта [27]. Однако каждая поисковая система имеет свои особенности при разборе этого файла. Для проведения поисковой оптимизации достаточно вышеперечисленных примеров [15].

«Понятные» URI-адреса веб-страниц

При проектировании системы управления («движка») веб-сайта необходимо использовать замену динамических адресов вида: http://www.site.ru/index.php?a=l&b=2&c=3&d=4

на статические:

http://www.site.ru/al/b2/c3/d4

такой адрес напоминает по структуре вложенные папки.

Почему же необходимо использовать подмену адресов:

1. «Длинные» URI-адреса (адреса с большим количеством параметров скрипта, в примере выше 4 параметра – a, b, c, d) иногда плохо индексируются. Особенно это относится к поисковой системе Google.

2. Поисковые системы подсвечивают слова запроса пользователя, найденные в URI-адресе. При этом используется транслитерация, единого стандарта которой нет. Чтобы проверить, как именно переводятся кириллица в латиницу, необходимо вводить запросы в поисковую систему. При этом одна буква может иметь несколько вариантов транслитерации, к примеру, для 'щ' в Яндекс есть подсветка в URI-адрссах, содержащих ‘sch’ и ‘sh’ (рисунок 12).

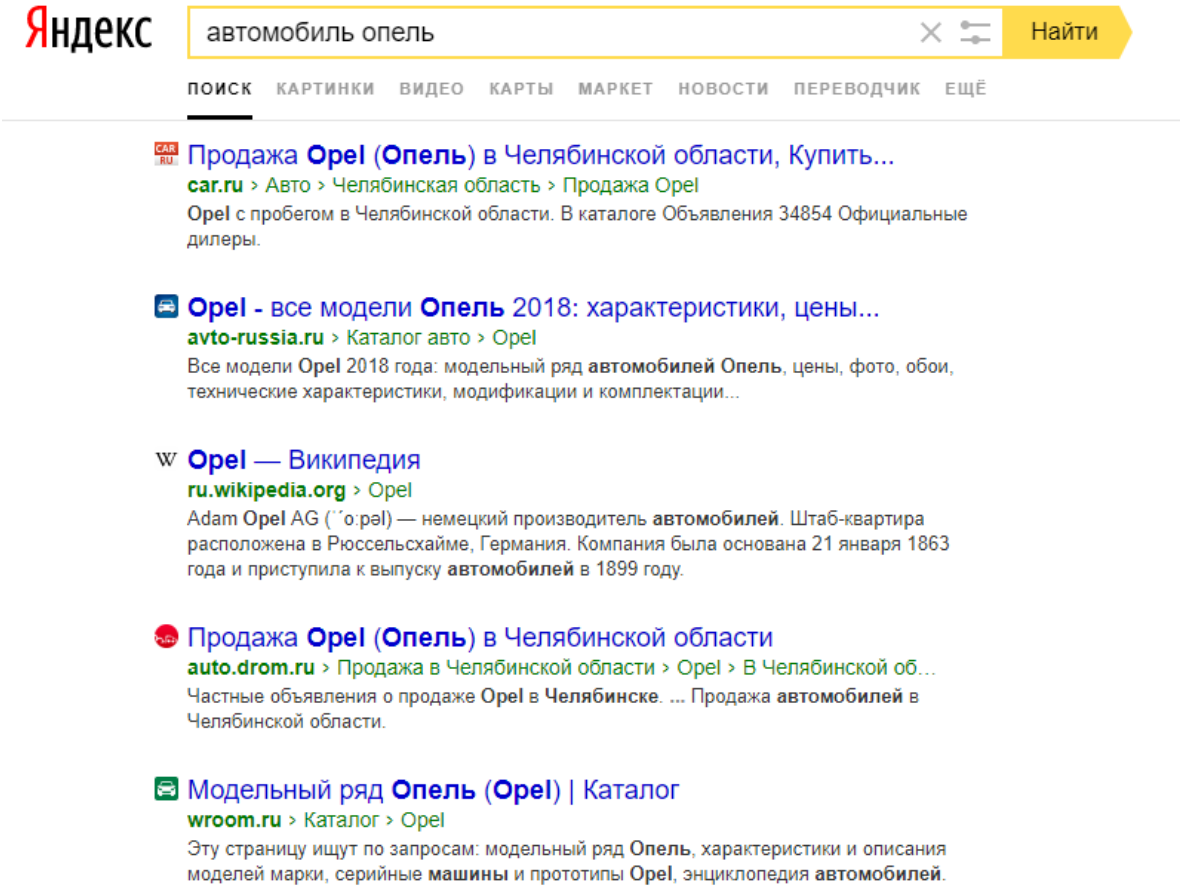

Рисунок 12 – Подсветка слов запросов в UR1 веб-сайта.

Влияние слов запроса в URI-адресе на релевантность было замечено в Яндекс при вводе алгоритма «Снежинск». Влияние это минимально, но подсветка слов запроса делает ссылку более заметной и может повышать CTR (от англ. click-to-ratio), количество переходов, отнесенной к количеству показов ссылки, выраженное в процентах [26].

Для подмены адресов при использовании веб-сервера Apache необходимо загрузить файл под названием «.htaccess» (в начале имени файла должна стоять точка) в папку сервера, в которой будут заменяться адреса. В Случае, если скрипты расположены в разных папках, в каждой из них должен быть свой файл .htaccess [19].

Первые две записи запустят сам модуль:

RewritcEngine on

Options + FoIlowSymlinks

Запись «RewriteEngine off» отменит все последующие команды. Это очень удобно: вместо необходимости комментировать все последующие строки все, что нужно сделать, это установить «off».

Опция «+FoIlowSymlinks» позволит ограничивать использование mod_rewrite для отдельных каталогов, иначе директивы будут действовать на весь сервер.

Следующая запись:

RewriteBase /

«/» является корневым (основным) URI. Если у вас какой-то другой URI, вы можете указать это в данной директиве, однако «/» обычно эквивалентно адресу «http://site.ru».

На некоторых серверах можно легко читать файл .htaccess, просто вводя UR1 следующего формата в поле адреса вашего браузера: http://www.site.ru/.htaccess. Это серьезное упущение защиты, гак как содержание вашего .htaccess может показать важную информацию об установках и настройках веб-сайта человеку, знающему, как эти знания применить против вас.

Для того чтобы защитить файл .htaccess от несанкционированного доступа, необходимо вписать в файл следующее:

RewriteRuIe ^.htaccess$ – [F]

Это правило переводится так:

При попытке обращения к файлу .htaccess система должна произвести код ошибки «HTTP response of 403» или «403 Forbidden – You don't have permission to access /.htaccess on this server» – доступ к файлу .htaccess запрещен.

Конструкция ^.htaccess$ в этом регулярном выражении означает:

^ – якорь начала строки

$ – якорь конца строки

Имя файла должно быть расположено точно между начальным и конечным якорем. Это будет гарантировать то, что только это определенное имя файла и никакое другое сгенерирует код ошибки.

[F] – специальный «запрещающий» флажок (forbidden).

В этом примере, файл «.htaccess» теперь будет состоять из таких строк:

RewriteEngine on

Options +FollowSymlinks

RewriteBase /

RewriteRule ^.htaccess$ – [F]

В итоге содержимое файла «.htaccess» будет выглядеть следующим образом:

RewriteEngine onOptions +FollowSymlinks

RewriteBase /

RewriteRule ^.htaccess$ – [F]

Для «подмены» адреса: http://www.site.ru/cgi-bin/shop.cgi?product1 на http://www.site.ru/shop/productl необходимо использовать следующую конструкцию в файле .htaccess:

RewriteRule ^(.*)shop/(.*)$ $lcgi-bin/shop.cgi?$2

Переменные $1 и $2 составляют так называемые «backreferences». Они связаны с текстовыми группами. Вызываемый URI разбивается на части. Все, что находится перед «shop», плюс все, что находится после «shop/», определяется и хранится в этих двух переменных: S1 и $2.

Для записи вида:

RewriteRuIe ^(.*)shop/(.*)$ $lcgi-bin/shop.cgi?$2 применяется общий синтаксис: RewriteRuIe текущийURL перезаписываемыйURL. Эта директива выполняет «подмену» URI адреса.

Вынос java-script и css в отдельные файлы

Все современные веб-сайты используют Java-script и каскадные таблицы стилей для удобства представления информации в документах. В результате этого объем HTML-кода может значительно увеличиваться. Для уменьшения размера документов, индексируемых поисковым роботом, java-script и каскадные таблицы стилей необходимо выносить в отдельные файлы [18].

Этот можно сделать, включив в HTML-код строки:

<Link href="style.css" rel="stylesheet" type="text/css">

<script language="JavaScript" src="index.js"></script>

Главное зеркало для поисковых систем

Домен с префиксом 'www' (http://www.site.ru) и без (http://site.ru) могут быть разными веб-сайтами для поисковых систем. При проектировании веб-сайта следует указывать главное зеркало, используя файлы 'robots.txt' и '.htaccess' для того, чтобы робот, определяющий зеркала сайтов, сделал все правильно. Т.к. «апдейты» робота-зеркальщика происходят крайне редко, неправильное определение главного зеркала приводит к длительным проблемам в индексации. В этом случае могут быть два варианта:

- определено неверное главное зеркало (к примеру, вместо http://www.site.ru веб-сайт индексируется как http://site.ru).
- проиндексированы и участвуют в поиске оба варианта http://www.site.ru и http://site.ru

Второй вариант гораздо хуже, т.к. все внешние факторы учитываются для каждого в отдельности. Пользователи, ссылаясь на веб-сайт, ставят ссылки вида как http://www.site.ru, так и http://site.ru. В результате релевантность распределяется между двумя веб-сайтами, и позиции в результатах поиска ниже, чем они должны быть. «Склейка» в таком случае может занимать до полугода, а иногда и больше.

Для корректного указания главного зеркала поисковому роботу необходимо прописать:

- В файле ‘robots.txt’:

User-agent: *

Disallow: /admin

…

Disallow: /print

Host: www.site.ru

Директива 'Host' должна идти после директив ‘Disallow’. Между последним ‘Disallow’ и 'Host' не должно быть пустых строк.

- В файле ‘.htaccess’:

Options +FollowSymLinks

RewriteEngine on

RewriteCond %{HTTP_HOST} ^site.ru

RewriteRule (.*) http://www.site.ru/$l [R=301,L]

Теперь роботу – зеркальщику указано главное зеркало как 'www.site.ru’, а также при наборе адресов вида:

http://site.ru/news/pagel

http://site.ru/news/page2

будет осуществляться редирект (перенаправление) на документы главного зеркала:

http://www.site.ru/news/pagel

http://www.site.ru/news/page2

Внешние факторы поисковой оптимизации.

Группа содержит факторы, влияющие на вторую составляющую релевантности – учет анкоров внешних ссылок – ссылок, расположенных на других веб-сайтах.

К примеру, http://site.ru и http://subdomain.site.ru – различные веб-сайты для поисковых систем. Соответственно, ссылка с первого на второй будет считаться для http://subdomain.site.ru внешней, а анкор ссылки будет учитываться при расчете релевантности.

Документ, на котором стоит ссылка на другой домен, называется «донором»; веб-сайт, на который ссылаются, – «акцептором».

Не все ссылки учитываются. Яндекс «понимает» HTML-тег <noindex>, который означает, что часть документа, обрамленную между

<noindex>

Текст, изображения, ссылки

</noindex>

индексировать запрещено. Соответственно, ссылки, заключенные в этот тег, не

индексируются и не учитываются при расчете релевантности.

Текст ссылки – анкор, к примеру, в HTML-коде:

веб-сайт

анкором будет слово «веб-сайт». Это то слово, кликнув мышью по которому, можно осуществить переход на другую страницу.

Множество всех внешних ссылок вебсайта называется «анкор-листом».

Яндекс и Google используют в формуле расчета релевантности т.н. модель ссылочного ранжирования, в которой анкор ссылки значительно влияют на релевантность акцептора. Донор, разместивший на своей странице ссылку вида:

новости

повышает релевантность донора http://www.1obl.ru по запросу «новости».

Веб-мастера, размещающие ссылки преднамеренно для повышения релевантности, заставляют поисковые системы постоянно совершенствовать алгоритмы учета ссылок. Поисковые системы стремятся учитывать только «естественные» ссылки, а влияние искусственных приравнять к нулю.

Рассмотрим форум или гостевую книгу. Пользователи общаются, т.е. задают вопросы и получают на них ответы. Часто в ответах присутствуют ссылки, ведущие на другой ресурс (внешние ссылки), где можно найти полезную информацию по заданному вопросу. Пользователь, разместивший эту ссылку, не задумывается над тем, какой анкор поставить. Анкор может быть в виде URI-адреса, а может содержать любые произвольные слова [15]:

Информацию по вопросу смотри здесь http://www.site.ru/otvety/otvet25.html

Ответ по своему вопросу найдешь здесь.

Более подробно читай на сайте.

В настоящее время существует множество бирж купли-продажи ссылок, в которых на уже созданных страницах, основное текстовое содержание которых меняется крайне редко или вообще никогда, продаются/покупаются ссылки с нужными анкорами. Такие ссылки располагаются в блоках, а размещение оплачивается посуточно. Соответственно, с течением времени содержание таких блоков ме-

няется в отличие от остального содержания страницы-донора.

Поисковые системы отличают ссылки, купленные через такие биржи, и понижают влияние их на релевантность. Наиболее известная из бирж Sape (http://sape.ru) существует с 2005 года, и все эти годы можно было неплохо повысить релевантность по необходимым запросам, покупая в ней ссылки. На сегодняшний момент исследования показывают, что ссылки с Sape и ей подобных практически не работают в момент индексации купленной ссылки поисковой системы. Ссылка начинает работать (т.е. повышать релевантность) после 3-4 месяцев нахождения в базе поисковой системы. В дальнейшем работоспособность таких ссылок под большим вопросом.

Любая ссылка, даже естественная, также повышает релевантность не сразу после попадания в базу поисковой системы. И чем дольше она присутствует в индексе поисковой системы, тем большую релевантность она передает. Точнее сказать, есть временной фактор, т.е. в момент попадания в индекс поисковой системы ссылка передает 10% от максимальной релевантности, которую она способна передать. Этот процент растет, и в какой-то момент ссылка «работает» на 100%. Точного графика передачи релевантности ссылки, естественно, нет, но логика такова [16].

Помимо временного фактора передачи релевантности ссылкой, вес ссылки различны и дают различный вклад в релевантность. Вклад ссылки с авторитетного портала будет больше, нежели вклад ссылки с малоизвестного блога (дневника) никому не известного автора.

Сайты, веб-мастера которых пытаются искусственно повысить релевантность внешними ссылками, распознаются поисковыми системами по анкор-листу. Во-первых, все доноры ссылаются на одну или несколько страниц веб-сайта, т.е. веб-сайт содержит, к примеру, 1000 страниц, а доноры ссылаются всего на 5. В случае веб-сайта с естественными ссылками вероятность такой ситуации практически равна нулю. Во-вторых, если рассматривать анкор-лист для отдельных страниц веб-сайта, т.е. выбрать все ссылки доноров, где акцептором будет являться одна

страница, сразу можно выделить ключевые слова, по которым веб-мастер пытается повысить релевантность. В данном примере веб-мастер пытался влиять на релевантность запросов «свадебные платья оптом» и «вечерние платья оптом». Существует некий порог вхождений ключевых слов в анкор-лист, веб-сайтам, превысившим этот порог, резко понижается релевантность по этим ключевым словам. Этот порог составляет – около 30%.

Стоит отметить, что этот порог был введен в Яндекс относительно недавно, в конце 2009 года, и действовал по принципам «амнистии». Т.е., в момент ввода такого алгоритма учета ссылок в анкор-листе присутствующих в базе поисковой системе веб-сайтов это не коснулось. Но в дальнейшем при изменении анкор-листов этих веб-сайтов, релевантность их запросам при превышении порога жестоко понижалась.

Определить превышение порога очень просто. Если при очередном обновлении поисковой базы Яндекс позиции по некоторым запросам изменились в худшую сторону и очень значительно, и в течение последующих обновлений поисковой базы позиции не возвращаются, скорее всего, это превышение порога.

Важным фактором является динамика изменения количества ссылок в анкор-листе. Классический метод т.н. «раскрутки» (повышения релевантности по необходимым запросам) – размещение веб-мастером ссылок в необходимом количестве (от 10 до нескольких тысяч). Анкор-лист веб-сайта с естественными ссылками пополняется регулярно и, как правило, без резких скачков в количестве ссылок. При искусственной накрутке за короткий промежуток времени в анкор-листе появляется большое количество ссылок, и такая ситуация при этом ранее не была характерна для данного сайта. Т.к. ссылки такого веб-сайта, как правило, покупаются с ежемесячной оплатой, то при исчезновении необходимости высоких позиций в результатах поиска по нужным ключевым словам (заказчик отказался от раскрутки) все ссылки в короткий промежуток так же быстро исчезают, как и появились. Естественно в истории данного сайта для поисковой системы это не останется незамеченным. При резких изменениях в

апкор-листе (на данный момент замечено только при росте количества ссылок) поисковые системы также понижают релевантность, но уже всем любым запросам. С другой стороны, такой алгоритм можно использовать во вред конкурирующему веб-сайту, и такое случается.

Для каждого веб-сайта порог прироста свой, и, как правило, он зависит от возраста сайта и количества уже имеющихся ссылок в анкор-листе. Для нового домена при размещении ссылок лучше придерживаться следующего правила: первые три месяца после индексации веб-сайта поисковой системой – размещать до 100 ссылок в месяц, но не более. Начиная с четвертого месяца, – размещать в количестве до 30% от уже проиндексированных ссылок. Соответственно, старым веб-сайтам с большим количеством ссылок в анкор-листе навредить практически невозможно, с «молодыми» сайтами (до года или с практически пустым анкор-листом) все наоборот.

Таким образом, можно представить внешние факторы поисковой оптимизации и их важность, определенные по методу экспертных оценок [46].

Таблица 4 – Важность внешних факторов поисковой оптимизации

Фактор	Оценка
1. Анкор (разнообразие, естественность)	1.0
2. Качество донора	0.9
3. Разнообразие анкор-листа (порог по ключевым словам)	1.0
4. Динамика изменения анкор-листа (порог по приросту)	1.0

ВЫВОДЫ ПО РАЗДЕЛУ 2

Во второй главе методом экспертных оценок определены факторы поисковой оптимизации и приведено обоснование выбора метода. Факторы позволяют повышать релевантность документов запросам пользователей в поисковых системах. На основе определенных факторов была получена конечная формула релевантности, которая позволяет рассчитывать общую релевантность документа запросам пользователей. Формула учитывает текстовую составляющую, релевантность внутренних и внешних ссылок, понижающие коэффициенты за искусственную накрутку релевантности, а также коэффициент уровня доверия поисковой системы к веб-сайту.

ГЛАВА 3. МЕТОД ПОИСКОВОЙ ОПТИМИЗАЦИИ

Комплекс мер для повышения релевантности веб-сайта запросам пользователей в поисковых системах называется поисковой оптимизацией. В Интернете синонимами этого слова являются: продвижение сайта, раскрутка сайта, сео или seo. Поисковая оптимизация может выполняться как в процессе разработки веб-сайта, так и на уже существующем веб-сайте. Причем, как правило, второй вариант наиболее распространен. Каждый специалист по поисковой оптимизации имеет свои особенности работы, к тому же алгоритмы расчета релевантности периодически меняются, что требует внесения изменений [14].

3.1 Составление семантического ядра

За исключением проблем с индексацией (не все страницы по падают в индекс поисковой системы) или санкций, наложенных поисковыми системами, поисковая оптимизация начинается с подбора ключевых слов.

Этот стартовый этап – один из самых важных. Желая повысить посещаемость через результаты выдачи поисковых систем и оптимизировать для этого сайт, можно на самой ранней стадии – составлении семантического ядра (подборе ключевых слов) совершить ошибки, и дальнейшая поисковая оптимизация не принесет желаемого результата.

Перед подбором ключевых слов необходимо задать вопрос: по каким словам может искать пользователь информацию об оптимизируемом веб-сайте? Необходимо проанализировать разделы сайта, и, используя статистику запросов Яндекс (http://wordstat.yandex.ru/), определить, каким запросам он может отвечать? К примеру, для интернет-магазина книг можно подобрать как общетематические запросы (купить книгу, купить книгу онлайн, интернет магазин книг, книги с доставкой), так и конкретные названия (купить Войну и Мир, купить сборник стихов Пушкина). Если веб-сайт нацелен на пользователей определенного региона, необходимо указать его по ссылке «уточнить регион», В случае сезонного характера услуг, которым отвечают ключевые слова (к примеру, «заказ деда мороза»), необходимо воспользоваться ссылкой «по месяцам» [21].

При подборе ключевых слов для веб-сайтов коммерческой направленности необходимо учитывать, что не все запросы пользователей могут привести к заказу услуг через веб-сайт. К примеру, по запросу «ремонт» пользователь может искать информацию о строительных материалах, технологиях ремонта, статьи о том, как сделать ремонт собственными силами, но это не означает, что он ищет подрядчика на оказание услуг. Как правило, для веб-сайтов, присутствующих в результатах поиска по коммерческим запросам, уже проведена поисковая оптимизация. Поэтому при сомнениях, принесет ли тот или иной запрос заказ услуг через веб-сайт, в большинстве случаев достаточно внимательно посмотреть на веб-сайты в результатах поиска. Если преобладают порталы с тематическими статьями, пресс-релизами, и переходя по ссылкам из результатов поиска на большей части страниц нет явного предложения коммерческих услуг - с большой долей вероятности потенциальных клиентов такой запрос не принесет. В примере с ремонтом, заказ услуг через сайт принесут запросы: «ремонт квартир» + город или регион («ремонт квартир в Москве»), «косметический ремонт квартир», «заказ на ремонт квартир» и т.д.

При составлении семантического ядра (списка запросов для поисковой оптимизации) на основании общих запросов («новости Челябинск», «новости Челябинская область» и т.д.) составляют расширенный список из многословных запросов:

Новости Челябинска и Челябинской области

Новости 74

Новости происшествия Челябинск

Новости сегодня Челябинск

Главные новости Челябинской области

затем необходимо воспользоваться правой колонкой (рисунок 13). В ней перечислены запросы, которые вводили пользователи после первоначального запроса.

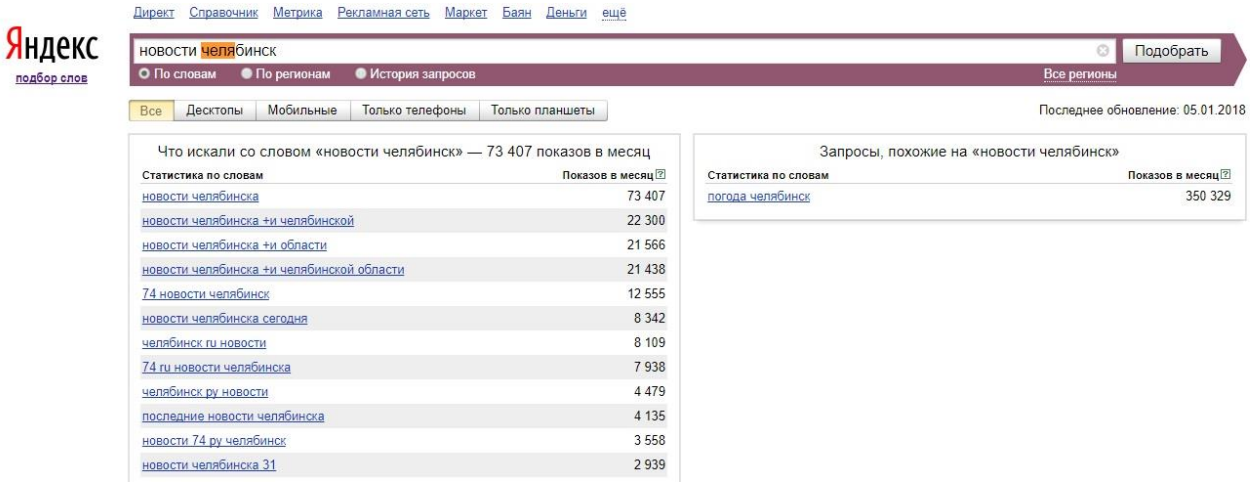

Рисунок 13 – Уточняющие запросы

Т.е. пользователь не нашел ответа в результатах поиска «новости Челябинск» и уточняет свой запрос. Из данного списка к семантическому ядру можно добавить запрос: погода Челябинск.

3.2 Распределение ключевых слов по страницам

После подбора ключевых слов их необходимо распределить по страницам.Вопрос распределения слов по страницам неразрывно связан с возможностью написания текста для страницы, чтобы он был логичным, читаемым, удобным для восприятия пользователем [37]. Возможна группировка слов, например, по тематике новостей, датам и т.д.

3.3 Внутренняя оптимизация

После распределения слов по страницам, для них необходимо провести внутреннюю оптимизацию.

Тег <title> – вписываем ключевые слова учитывая, что Яндекс воспринимает 150 символов текста в данном теге, а также то. что содержимое данного тега будет отображаться в результатах поиска в виде ссылки. Соответственно, необходимо сделать содержание лаконичным и в то же время привлекательным для пользователя. Необходимо также избегать простого повтора ключевых слов, для первой группы слов содержимое <title> может быть таким:

Главные новости Челябинска и области

Тег <keywords> – перечисляем все ключевые слова, разделяя их запятой.

Тег <description> – описание документа – одно предложение, характеризующее содержание страницы. В данном примере может быть следующее:

Информационный сайт Медиахолдинга ОТВ

Тег <h1> – необходимо вставить наиболее часто запрашиваемые 1-2 ключевых слова, в данном случае <h1> может быть таким:

Свежие новости Челябинска

Необходимо помнить, что <hl> – это заголовок основного текста страницы, т.е. он должен быть оформлен несколько большим шрифтом, чем остальной текст, и вставлять в него большой текст не рекомендуется из соображений удобства визуального восприятия страницы.

Основной текст страницы необходимо разбивать на абзацы (тег текст страницы), он должен содержать хотя бы но одному вхождению всех ключевых слов группы.

Общий объем текста для написания должен быть таким, чтобы ключевые слова можно было употребить хотя бы но одному разу и не в ущерб восприятию текста. Можно также оценить страницы веб-сайтов, находящихся в результатах поиска по интересующим ключевым словам, и взять средние значения по частоте употреблений ключевых слов и объему текста на странице.

В данном примере можно использовать иерархию заголовков <h1> – <h6> и HTML-код документа будет таким;

<html>

<head>

<title>Главные новости Челябинской области</title>

<keywords>сайт областного телеканала, прямой эфир, свежие новости Челябинска и Челябинской области</keywords>

<descriptian>Видеоролики и прямой эфир. Самые свежие новости Челябинска и Челябинской области. Афиша. Радио «Business FM». Проекты ОТВ. </description>

</head>

<body>

<h1> Свежие новости Челябинска</h1>

текст страницы с ключевыми словами

«новости Челябинск»

«погода Челябинск»

<h2>Происшествия Челябинск</h2>

текст страницы – с ключевым словом «происшествия»

<h3>Политика Челябинск</h3>

текст страницы – с ключевым словом «политика»

</body>

</html>

В конце текста страницы необходимо добавить абзац с контактами холдинга, а так же данные об свидетельстве регистрации СМИ. В анкорах же необходимо использовать следующие слова: новости, лонгриды, телепередачи, проекты, прямой эфир.

3.4 Указание главного зеркала, настройки скриптов

В случае если оптимизируемый веб-сайт уже проиндексирован, необходимо определить главное зеркало, по адресу которого индексируется веб-сайт. Для этого необходимо проверить, какие страницы проиндексированы поисковой системой Яндекс (рисунок 14).

На основании этого составляем файл robots.txt и размещаем его в корневой директории:

User-agent: *

Host: 1obl.ru

в данном случае главное зеркало индексируется без «www».

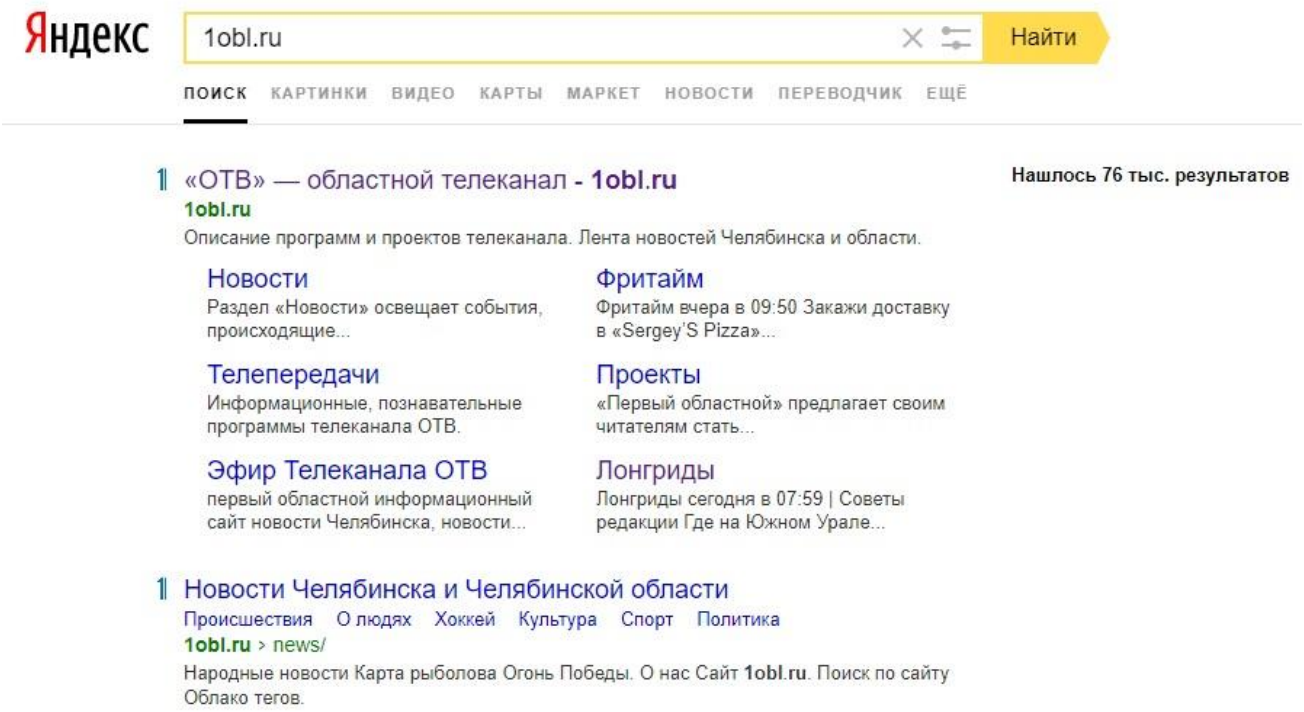

Рисунок 14 – Страницы, проиндексированные Яндекс

В случае, когда веб-сайт не проиндексирован необходимо определиться с главным зеркалом и прописать директиву Host в файле robots.txt. На основании этого можно использовать структуру внутренних ссылок на веб-сайте с «www» или без, а также настроить редирект в файле .htacccss:

RewriteRule Mitaccess$ – [F]

RewriteEngine on

RewriteCond %{HTTP_HOST} ^m-master.ru [NC]

RewriteRule ^(.*)$ http://www.m-master.ru/$l [L,R=301]

Необходимо запретить от индексации все страницы, не несущие полезной информации: результаты поиска по сайту, версии документов для печати (версии основных документов без графических элементов), метки (теги), списки документов, сортированные по датам, различные сортировки данных в таблицах по столбцам. Сделать это можно следующим образом в robots.txt:

User-agent: *

Disallow: /print/*

Disallow: /pagel.htm

…

Host: 1obl.ru

Директивы «Disallow» должны быть расположены между «User-agent» и «HOS»

В случае динамического веб-сайта и «длинных» адресов страниц (с большим числом параметров), к примеру:

http://www.site.ru/index.php?a=l&b=2&c=3&d=4

необходимо при помощи ModRewrite (для веб-сервера Apache) реализовать статические адреса следующего вида:

http://www.site.ru/l/2/3/4

ITO главная стр;

Одной из распространенных ошибок является то, что главная страница доступна по нескольким адресам:

http://www.site.ru/

http://www.site.ru/index.html

http://www.sitc.ru/index.php

В таком случае необходимо реализовать редирект на первый вариант главной страницы со вторых двух.

Для исключения возможности индексации одной страницы по двум и более адресам необходимо отсылать запрашивающему некорректный вариант роботу ошибку 404 – не найдено.

Если веб-сайт меняет адресацию страниц (это зачастую происходит при смене «движка» веб-сайта), необходимо настроить редирекг со старых страниц на новые адреса, т.е. вместо

http://www.site.ru/old_page.php

теперь на сайте размещена

http://www.site.ru/new_page.php

тогда при запросе старого варианта необходимо отсылать заголовок 301 (Moved Permanently – перемещено навсегда), все внешние ссылки, если таковые

имеются, будут учитываться для новой страницы. Проще говоря, если перенести страницу на новый адрес с полностью сохраненным содержимым, то через некоторое время в результатах поиска на тех же позициях по ключевым словам отобразится новая страница вместо старой. Если редирект не прописывать, то внешние ссылки учитываться не будут, и позиции в результатах поиска будут ниже[45].

Для повышения релевантности запросов с помощью внутренних ссылок необходимо указывать ключевые слова в анкорах на других страницах. Поисковые системы легко различают ссылки меню, ведь их анкоры и расположение повторяются на всех или большинстве страниц веб-сайта. Поэтому необходимо проставлять ссылки с нужными анкорами внутри основного текста страницы, если, конечно, это гармонирует с контекстом. Тексты анкоров должны быть разными, и чем большим будет это разнообразие, тем лучше. Можно менять словоформы, добавлять «разбавочные» слова (т.е. к двухсловному запросу добавлять третье слово), менять регистр слов.

Одним из важных параметров достижения высоких позиций в поисковых системах является уникальность текстовой информации, размещенной на страницах веб-сайта. Особенно важно это в случае создания нового сайта, т.к. в случае неуникальюго содержания веб-сайт может быть вообще не проиндексирован, и не будет участвовать в результатах поиска [42].

Существуют программы для автоматической проверки текстов на уникальность в поисковых системах. Среди них:

http://copyscane.com/

http://www.antiplagiat.ru/

Все эти системы позволяют проверить уникальность документа в режиме он-лайн.

3.5 Определение внешне-ссылочной конкуренции

После внутренней оптимизации веб-сайта god выбранные ключевые слова необходимо определить т.н. конкуренцию по внешним ссылкам. Проще говоря, оценить, каким количеством ссылок с анкорами в виде ключевых слов обладают

веб-сайты, занимающие позиции на первой странице результатов поиска. Задача заключается в том, чтобы найти как можно больше ссылок на веб-сайты из первой десятки Яндекс и среди всей ссылочной массы подсчитать количество анкоров, которые содержат ключевое слово.

В 2007 году Яндекс запретил просмотр внешних ссылок, найденных поисковым роботом для любого веб-сайта. Сегодня можно лишь посмотреть внешние ссылки веб-сайта, зарегистрированного в Яндекс.Вебмастер (http://webmaster.yandex.ru) для этого необходим доступ по протоколу FTP, соответственно ссылочную массу для веб-сайтов, конкурирующих по интересующим запросам, посмотреть не удастся.

Решение на сегодняшний день выглядит следующим образом – внешние ссылки можно посмотреть через поисковые системы Yahoo, Altavista и Alexa, а затем проверить найденные ссылки на индексацию в Яндекс.

Для сбора внешних ссылок рекомендуется использовать программу Yazzle (http://www.yazzle.ru), которая в автоматическом режиме делает запросы к поисковым системам, проверяет наличие ссылки на странице, определяет анкор и множество других параметров. Программа позволяет задавать фильтры по ключевым словам в анкорах и выводит суммарные значения по фильтру. Необходимо отметить, что поисковая система Yahoo, которая находит большую часть ссылок, показывает максимум 1000 найденных ссылок. Поэтому при анализе высококонкурентных запросов необходимо учитывать этот факт [39].

Т.к. необходимо оценить лишь порядок количества ссылок с нужным анкором, то проверять индексацию их в Яндекс вовсе не обязательно.

Вышеперечисленные поисковые системы индексируют Рунет медленней, чем Яндекс, поэтому для большей точности следует умножить количество анкоров на 2-3, и мы получим приблизительное количество ссылок с нужным анкором, которое необходимо проставить на оптимизированную страницу нашего веб-сайта для попадания в десятку.

3.6 Источники внешних ссылок

Если ваш веб-сайт интересен и полезен, то со временем внешние ссылки будут появляться естественным путем, т.е. пользователи будут ссылаться с форумов, блогов, собственных веб-сайтов. Этот процесс может быть долгим. Поэтому встает задача в кратчайшие сроки обзавестись ссылочной массой для попадания на первые страницы результатов поиска Яндекс и Google.

Если существует возможность проставить внешние ссылки на дружественных, партнерских ресурсах, то это необходимо сделать. Для попадания на первые страницы низко- и среднеконкурентньтх запросов требуется от 1 до нескольких тысяч внешних ссылок, поэтому найти такое количество дружественных веб-сайтов, где можно разместить свою ссылку, в большинстве случаев невозможно. Тогда необходимо воспользоваться размещением ссылок на платной основе. Несмотря на то, что представители поисковых систем заявляют о том, что они отрицательно относятся в покупным ссылкам, но не учитывать они их не могут. Ссылок, размещенных на платной основе в Рунет, по словам одного из представителей Яндекс, имеется свыше 90%. Наилучший вариант покупки ссылок на сегодняшний день – это размещение ссылок «навсегда», т.е. размещение оплачивается один раз, и вебмастер или система, через которую куплена ссылка, гарантирует ее размещение на время жизни сайта-донора.

Так же в современных реалиях весьма эффективным является размещение ссылок в социальных сетях. Например, ссылки на «происшествия Челябинск» есть смысл «раскидывать» в группы (паблики, страницы) с соответствующей тематикой. Так же можно создать свои страницы в социальных сетях и регулярно размещать гиперссылки [33].

При размещении внешних ссылок необходимо делать это как можно более естественно, т.е. прирост ссылочной массы должен быть равномерным без резких скачков.

В случае проведения поисковой оптимизации веб-сайта, расположенного на новом домене, необходимо размещать не более 50 – 100 внешних ссылок ежеме-

сячно в течение первых 3 – 4 месяцев. Затем можно размещать до 30% от количества уже проиндексированных ссылок в каждый последующий месяц. Если проводится поисковая оптимизация домена, уже имеющего ссылочную массу, то допустимо размещать до 30% от общего количества уже проиндексированных ссылок в каждый месяц.

ВЫВОДЫ ПО РАЗДЕЛУ 3

В третьей главе описаны этапы проведения поисковой оптимизации веб-сайта: от составления семантического ядра, оптимизации текстов и структуры до определения внешнессылочной конкуренции и наращивания ссылочной массы.

ГЛАВА 4. ПОИСКОВАЯ ОПТИМИЗАЦИЯ ВЕБ-САЙТА 1OBL.RU

Практическая реализация работы представлена поисковой оптимизацией веб-сайта «Первый областной». Необходимость в поисковой оптимизации портала возникла в 2016 году после анализа источников посетителей – количество поискового трафика могло быть существенно увеличено.

4.1 Анализ источников посетителей и постановка задачи поисковой отимизации сайта

Свыше 30% трафика портала (рисунок 15) в 2016 году составляли переходы по ссылкам на сайтах, а переходы из поисковых систем всего 22% [48].

Рисунок 15 – Источник трафика сайта 1obl.ru

Количество переходов на октябрь 2016 года с Яндекс и Google составило 159 тыс. в месяц с каждой из них (рисунок 16).

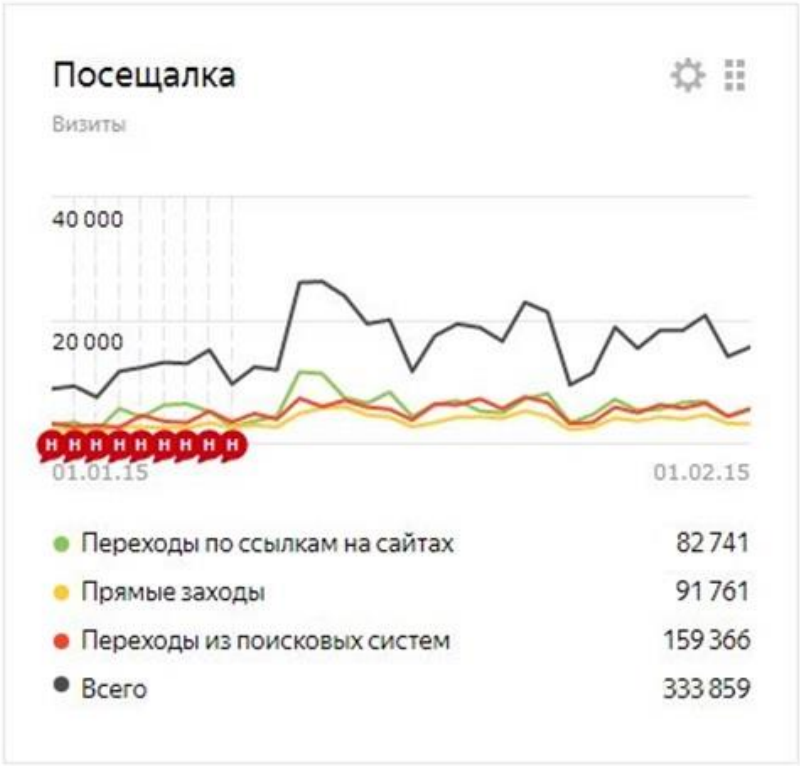

Рисунок 16 – Переходы на сайт 1obl.ru с «поисковиков»

По результатам анализа структуры сайта, целевой аудитории, а также запросов, по которым возможно получение дополнительного трафика из поисковых систем, было принято решение подбирать не высокочастотные и общие запросы («происшествие Челябинск», «губернатор Челябинской области» и т.д.), а сосредоточиться на поисковой оптимизации всей ленты новостей сайта. Целевыми запросам являются имена первых лиц Челябинска и области, а также отдельные проекты канала ОТВ и сайта 1obl.ru (конкурс караоке «Поют все», проект «Национальный интерес», и др.).

Была поставлена задача оптимизации внутренних страниц и внутренней ссылочной структуры под поисковые запросы.

4.2 Анализ текущей оптимизации портала

В ходе анализа текущей структуры сайта были выявлены недостатки в HTML-коде страниц. На страницах с телепередачами телеканала ОТВ отсутствовали ключевые слова в важных для поисковых систем тегах. Поисковым роботом Яндекс было проиндексировано около 25 тыс. страниц, часть из которых являлась

версиями для печати страниц. Такие страницы не содержат уникальной информации для поисковой системы, т.к. являются дубликатами с отсутствующими элементами дизайна. Другие существующие страницы роботом Яндекса вовсе не были проиндексированы, так как информации в описании настолько мало, что робот «не видит» эту страницу (количество символов менее 500) [22].

В ходе анализа структуры внутренних гиперссылок было установлено, что портал содержит порядка 26 тыс. страниц в форматах, индексируемых поисковыми системами.

В ходе анализа уникальности текстовых документов с помощью сайта advego.ru было установлено, что документов с уникальностью 90% и выше на портале около 80% от общего числа. Уникальность остальных документов находилась в пределах от 65 до 90%. Такой процент неуникальных документов обусловлен тем, что в новостных статьях на сайте присутствуют цитаты, которые могут быть и в новостях на других порталах. Аналогичная ситуация, например, существует среди юридических порталов и веб-сайтов, которые содержат множество нормативных документов.

В данном случае существует два варианта решения проблемы. В первом случае, если веб-сайт обладает большим уровнем доверия с точки зрения поисковой системы, можно размещать неуникальные документы без изменений. Но при этом общий прирост неуникальных документов, по сравнению с приростом уникальных, должен находиться в соотношении один к двум, и более. Второй вариант решения проблемы состоит в написании уникальной информации, а так же текстов, соответствующих требованиям поисковиков относительно количества символов (более 500). Таким образом, повышается общая уникальность страницы.

Для выбора варианта необходимо определить два параметра: примерное соотношение уникальных документов ко всем документам на сайте («невидимых» для поисковика), а также, что наиболее важно, процент индексации веб-сайта. К примеру, веб-сайт имеет 1000 страниц, разрешенных для индексации, а в базе поисковой системы хранится всего 400. Низкий процент индексации (менее 90%) в те-

чение 1 – 2 месяцев после размещения документов на веб-сайте может говорить о том, что поисковый робот игнорирует некоторые документы.

Процент индексации страниц сайта «Первый областной» поисковыми системами превышал 90%, и дополнительного придания уникальности документам не требовалось.

В структуру внутренних гиперссылок портала также должны быть внесены изменения. Анкоры содержали недостаточное количество ключевых слов, поэтому страницы-доноры передавали акцепторам меньшую релевантность. При рассмотрении множества внутренних ссылок на отдельно взятую страницу был выявлен большой процент ссылок с одинаковыми анкорами, что также понижало релевантность.

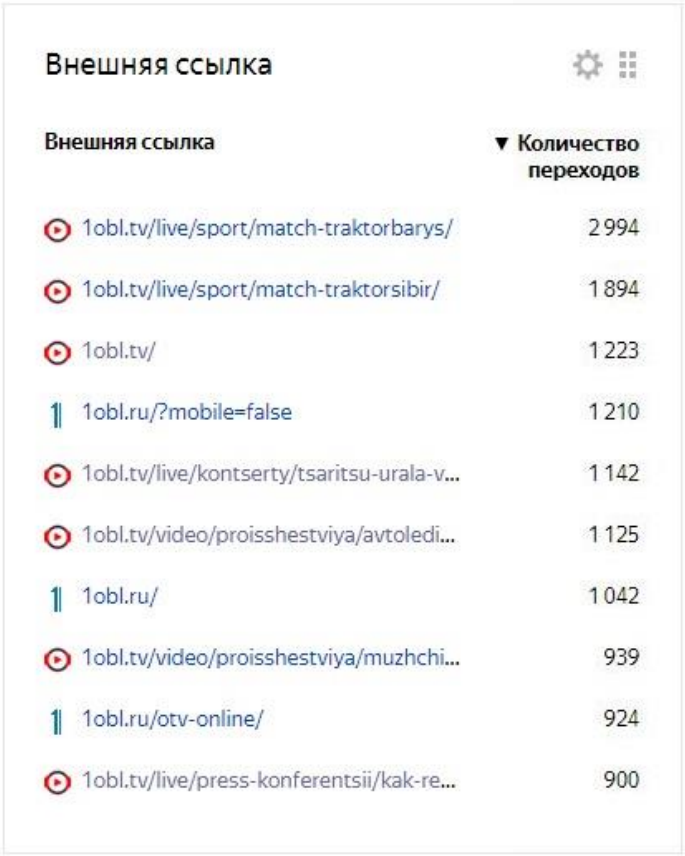

Рисунок 17 – Список самых посещаемых внешних ссылок

Все внешние ссылки были естественными, т.е. пользователи проставляли их добровольно. Все ссылки имели разнообразные анкоры (рисунок 17), около 50% из них ссылались на главную, остальные – на различные внутренние страницы.

Сайт имел большое количество внешних ссылок с главных страниц и в совокупности всего перечисленного обладал достаточным уровнем доверия с точки зрения поисковых систем.

С учетом этого дополнительной ссылочной массы дл оптимизации страниц каталога по целевым запросам не требовалось. Была поставлена задача повышения текстовой и внутриссылочной релевантности.

4.3 Оптимизация портала

Исходя из поставленных задач поисковой оптимизации страниц сайта 1obl.ru были внесены изменения в структуру администрирования. Количество обязательных тегов и Keywords для каждой страницы было увеличено до 5 (больше возможно, меньше запрещено). А так же был интегрирован сервис Яндекс.Дзен (рисунок 18).

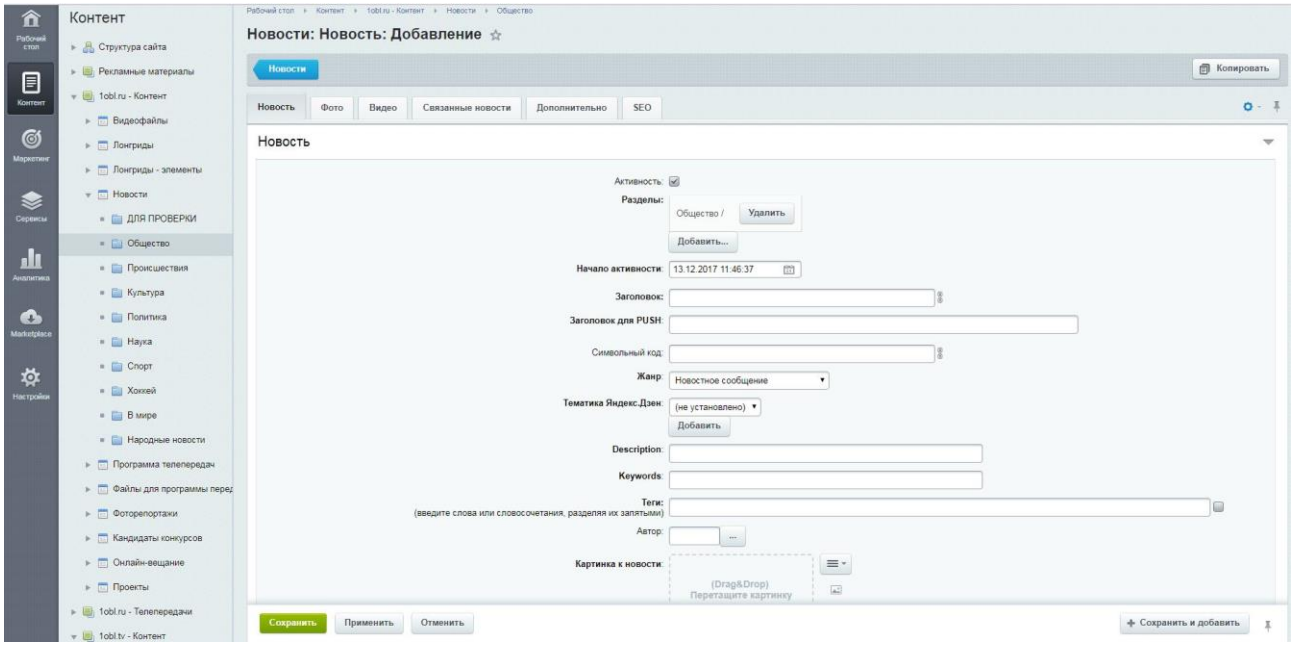

Рисунок 18 – Внешний вид добавляемой новости

Дзен – это сервис персональных рекомендаций Яндекса. Он составляет подборку новостей, постов из блогов и других интернет-публикаций, которые могут быть интересны пользователю. Публикации отбираются на основе истории посещенных страниц и указанных предпочтений.

На экране рекомендации отображаются в виде карточек, при открытии новой

вкладки. Для того чтобы увидеть все рекомендации, пользователю необходимо прокрутить ленту с карточками вниз.

Чтобы прочитать публикацию, пользователь нажимает рекомендацию в ленте – текст открывается в новой вкладке. При этом лента рекомендаций остается на вкладке Яндекс.Дзен, пользователь может к ней вернуться, чтобы просмотреть другие интересные для него публикации.

ВЫВОДЫ ПО РАЗДЕЛУ 4

В результате комплекса работ по поисковой оптимизации посещаемость сайта в октябре 2017 года составила 520 327, относительно прошлого23,3 %посетителей месяц, т.е. в 6,1 раза больше, чем в 2016 году (рисунок 19):

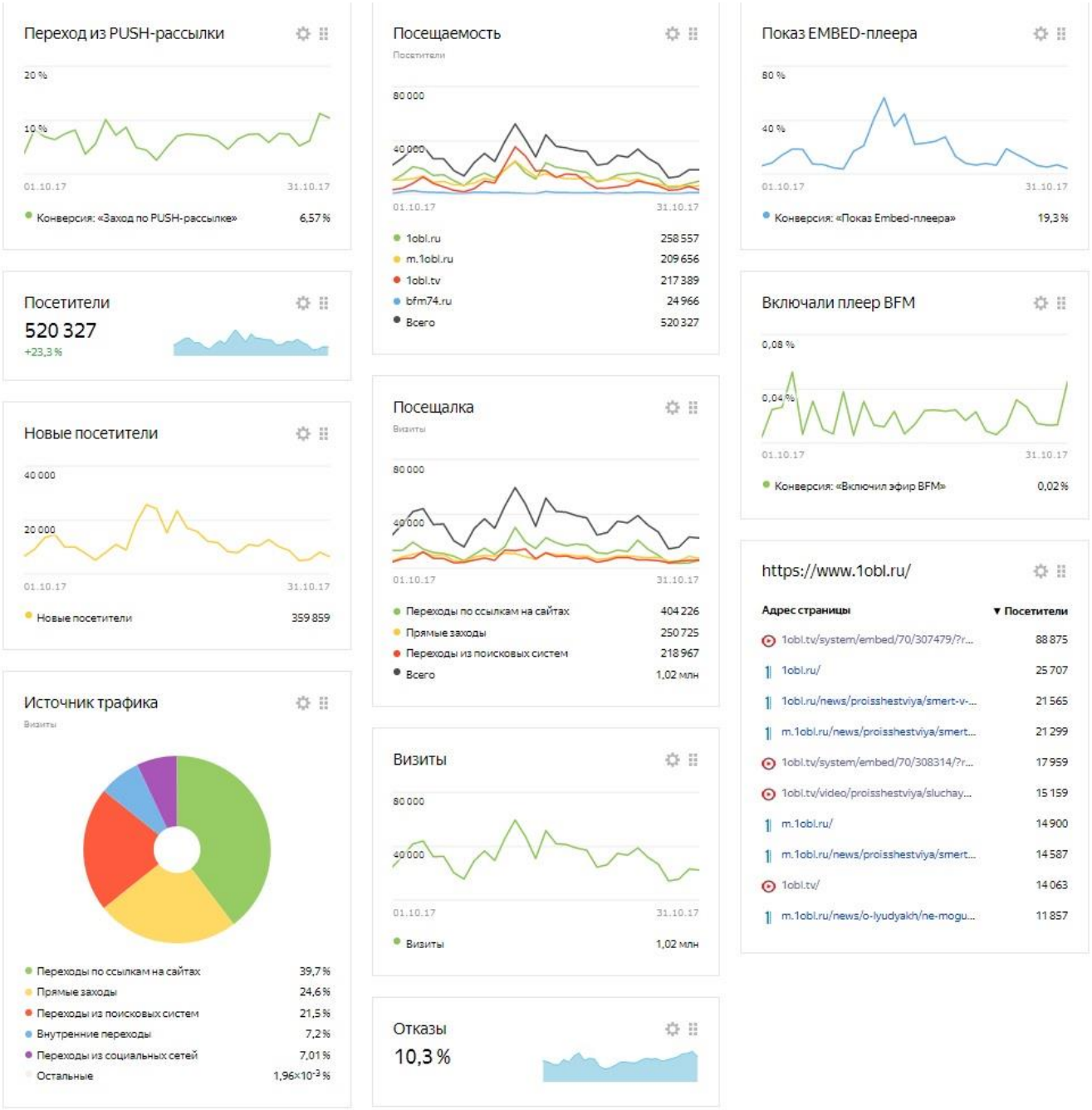

Рисунок 19 – Посещаемость сайта 1obl.ru за октябрь 2017 года

Трафик с Яндекс и Google изменился с 159 тыс. в месяц на 250 тыс. (рисунок 20):

Рисунок 20 – Трафик с Яндекс и Google

Если рассматривать точки входа, т.е. страницы, на которые попадают пользователи при переходе на сайт, то значительно возросло количество внутренних страниц (рисунок 21):

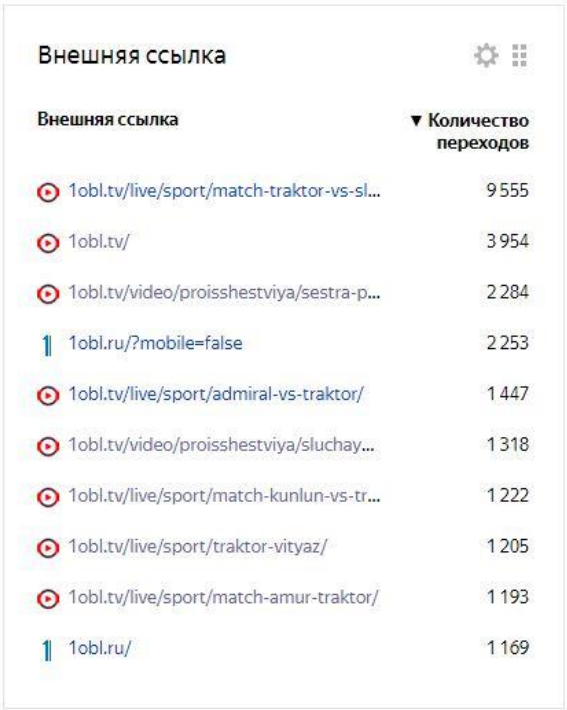

Рисунок 21 – Список посещаемых внешних ссылок на октябрь 2017

Сравнивая количества переходов на не главные страницы каталога портала в октябре 2016 и 2017 годов из результатов поиска Яндекс и Google, получим, что данные показатели возросли более чем в 3 и 1,5 раза соответственно.

На основе этого можно сделать вывод о правильной постановке задачи поисковой оптимизации, которая подразумевала повышение релевантности по низко- и среднечастотным запросам пользователей в поисковых системах. Данный метод отлично подходит для оптимизации веб-сайтов и порталов, содержащих большое количество документов (от 1000).

ЗАКЛЮЧЕНИЕ

В результате проведенного исследования на основе метода экспертных оценок поставлена и решена задача определения факторов, влияющих на ранжирование.

К основным результатам, полученным в работе, относятся:

1. Результаты анализа поисковых систем Рунет, позволяющие определить основные принципы ранжирования документов.

2. Результаты анализа подходов при определении релевантности текстовых документов, а также определение ссылочной релевантности, позволивших выявить основные принципы ранжирования результатов поиска.

3. Полученная система критериев, участвующих в формулах ранжирования поисковых систем, на основе метода экспертных оценок. Для каждого критерия определена важность в рамках шкалы от 0 до 1 с шагом 0,1.

4. Разработан метод поисковой оптимизации на основе факторов, участвующих в формулах ранжирования поисковых систем, позволяющий повышать посещаемость веб-сайтов.

5. Применение предложенного в диссертации метода для проведения поисковой оптимизации веб-сайта «Первый областной», который позволил повысить посещаемость 1obl.ru с 159 тысяч в месяц на 250 тысяч в месяц.

Научная новизна полученных результатов состоит в разработке метода поисковой оптимизации, а так же в исследовании алгоритмов ранжирования поисковых систем.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ашманов, И.С., Иванов, А.А. Продвижение сайта в поисковых системах. – М.: Вильяме, 2007. – 304 с.
2. Ашманов, И.С., Иванов, А.А. Оптимизация и продвижение сайтов в поисковых системах. – СПб.: Питер, 2009. – 400 с.
3. Байков, В. Д. Интернет. Поиск информации. Продвижение сайтов. – СПб.: БХВ-Петербург, 2000. – 288 с.
4. Бенкен, Е. PHP, MySQL, XML. Программирование для Интернета. – СПб.: БХВ-Петербург, 2007. – 336 с.
5. Булакнна, М.Б., Лысенко, Д.С., Чайников, Ю.С. Увеличение посещаемости портала «Российское образование» на основе анализа поисковых запросов. – СПб., 2007. – с. 366-369.
6. Гридиия, Е.Г., Лысенко, Д.С. Использование поисковых систем для увеличения посещаемости. – М.:Бином, 2009. – с. 43-48.
7. Губин, М.В. Модели и методы представления текстового документа в системах информационного поиска: дис. канд. физ.-мат. наук. –СПб., 2005.
8. Гусев, В.С. Аналитика веб-сайтов. Использование аналитических инструментов для продвижения в Интернет. – М.: Диалектика, Вильямс, 2008. – 176 с.
9. Дамашке, Г. PHP и MySQL = PHP & MySQL. – М.: НТ Пресс, 2008. – 320 с.
10. Дорнфест, Р., Бош, П., Калишейн, Т.. Секреты Google. Трюки и тонкая настройка – Б.м.: Русская Редакция, 2008. – 748 с.с
11. Евдокимов, Н.В. Раскрутка Web-сайтов. Эффективная Интернет-коммерция. – М.: Вильямс, 2007. – 160 с.
12. Евдокимов, Н.В., Лебединский, И.А. Раскрутка веб-сайта. Практическое руководство. – М.: Вильямс, 2011. – 288 с.
13. Евдокимов, Н.В. Основы контентной оптимизации. Эффективная интернет-коммерция и продвижение сайтов. М.: Вильямс, 2007. – 160 с.
14. Зандстра, М. PHP. Объекты, шаблоны и методики программирования – М.: Вильямс, 2011. – 560 с.

15. Зуев, М.Б., Маурус, П.А., Прокофьев, Л.Г. Продвижение сайтов в поисковых системах. Спасательный круг для малого бизнеса. – М.:Бином, 2007. – 304 с.
16. Интернет-маркетинг па 100%. – СПб.: Питер, 2009. – 240 с.
17. Клифтон Бр. Google Analytics. Профессиональный анализ посещаемости веб-сайтов. = Advanced Web Metrics with Google Analytics. - М.: Вильямс, 2009. – 400 с.
18. Колисничсико, Д.Н. Поисковые системы и продвижение сайтов в Интернете. – М.: Диалектика, 2007. – 272 с.
19. Кошик, А Веб-аналитика. Анализ информации о посетителях веб¬сайтов. = Web Analytics: An Hour A Day. – М.: Диалектика, Вильямс, 2009. – 464 с.
20. Кузнецов, М.С., Симдянов, И.Д., PHP. Практика создания Web-сайтов. – СПб.: БХВ-Петербург, 2008. – 1244 с.
21. Ландэ Дм. Поисковые системы: Поле – семантика – 2004.
22. Коваленко, А.В. Разработка автоматизированной системы создания Интернет представительства организации – Пенза: Информационно-издательский центр ПГУ, 2007. – с. 187-190.
23. Мидоу Ч. Анализ информационно-поисковых систем. – М.: Мир, 1970. – 368 с.
24. Орлов, А.И. Экспертные оценки – М.: Мир, 2002.
25. Севостьянов, И.И. Поисковая оптимизация. Практическое руководство по продвижению сайта в Интернете. – Б.м.: Питер, 2010. – 240 с.
26. Сегалович, И.А., Маслов, М.С., Зеленков, Ю.В. Цели и результаты программы научных стипендий Яндекса. – М.: 2005. – с. 7-17.
27. Сергеев, А.П. Раскрутка сайтов и основы электронной коммерции. Краткое руководство. – М.: Диалектика, 2005. – 256 с.
28. Сирович Дж., Дари Кр. Поисковая оптимизация на PHP для профессионалов. Руководство разработчика по SEO. – М,: Диалектика, Вильяме, 2008. – 352 с.
29. Скляр Д., Трахтенберг А. PHP. Рецепты программирования PHP Cookbook. – Спб.: БХВ-Петербург, 2007. – 736 с.

30. Солтон Дж. Динамические библиотечно-поисковые системы. – М.: Мир, 1979. – 558 с.
31. Суэринг Ст., Конверс Т., Парк Дж. PHP и MySQL. Библия программиста. – М.: Диалектика, 2010. – 912 с.
32. Тероу Ш. Видимость в Интернете. Поисковая оптимизация сайтов. = Search Engine Visibility. – Б.м.: Сим вол-Плюс, 2009. – 288 с.
34. Уайт Э., Камаль Э. Дж. Статистические методы работы с электронными документами в библиотечной сфере, или Э-метрики. – М.: Омега-Л, 2006. – 393 с.
35. Фролов, И.Л., Перелыгин, В.А., Самойлов, Е.Э. Разработка, дизайн, программирование и раскрутка web-сайта. – М.: Триумф, 2009. – 302 с.
36. Энж Э., Спенсер Ст., Фишкин Р., Стрикчиола Дж. SEO – искусство раскрутки сайтов. = The Art of Seo. – СПб.: БХВ-Петербург, 2011. –592 с.
37. Яковлев, А.А. Раскрутка и продвижение сайтов: основы, секреты, трюки. – СПб.: БХВ-Петербург, 2007. – 336 с.
38. Яковлев, А.М., Ткачев, В.В. Раскрутка сайтов. Основы, секреты, трюки. – СПб.: БХВ-Петербург, 2010. – 352 с.
39. https://yandex.ru/blog/company/10095.
40. https://yandex.ru/company/technologies/datacentcr/.
41. https://yandex.ru/company/technologies/matrixnct.
42. https://yandex.ru/company/technologies/regions/.
43. http://credibility.stanford.edu/guidelines/index.html.
44. https://devaka.ru/articles/chto-takoe-pagerank.
45. http://devaka.ru/articles/trust-and-authority.
46. http://download.yandex.ru/company/03_yandex.pdf.
47. http://help.yandex.ru/webmaster/?id=996567.
48. http://httpd.apache.org/docs/1.3/mod/mod_rewrite.
49. http://www.codeisart.ru/part-l-shingles-algorithm-for-web-documents/.
50. http://www.seonews.ru/glossary/detail/8908.php.
51. http://www.1obl.ru

ПРИЛОЖЕНИЕ

Оценка важности факторов поисковой оптимизации

АНКЕТА ЭКСПЕРТА

Блок 1

		Оценка в нормированной шкале
1.	Возраст эксперта	
	• 9-16 лет	0,4
	• 17-21 год	0,5
	• 22-30 лет	0,8
	• 31-40 лет	1,0
	• 41-50 лет	1,0
	• 51-60 лет	0,9
	• Свыше 60 лет	0,8
2.	Род занятий	
	• школьник	0,2
	• абитуриент	0,4
	• студент	0,6
	• преподаватель	0,8
	• администратор	0,9
	• научный работник	0,8
	• IT специалист	1,0
	• другой род занятий	0,5
3.	Знание предметной области	
	• Специалист другой области	0,1
	• Низкое	0,3
	• Среднее	0,5
	• Высокое	0,7
	• Очень высокое	1.0
4.	Проводили ли Вы поисковую оптимизацию веб-сайтов?	
	• В первый раз	0,2
	• Иногда	0,5
	• Периодически	0,7
	• Очень часто	1,0

РЕЗУЛЬТАТ ОБРАБОТКИ АНКЕТЫ ЭКСПЕРТА

Блок 2

Фактор	0.0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	Ре-зуль-
Точное вхождение ключевого слова на странице в рамках пассажа	0	0	0	0	0	0	1	0	1	2	105	
Точное вхождение ключевого слова в тег keywords	0	0	0	0	0	0	1	0	1	14	93	
Точное вхождение ключевого слова в description	18	60	29	1	0	0	1	0	0	0	0	
Точное вхождение ключевого слова в h1-h6	15	65	23	4	1	0	1	0	0	0	0	
Плотность ключевого слова в тексте страницы до 5%	0	0	0	0	1	0	1	5	90	10	2	
Общий объем полезного текста на странице 1000-2000 знаков	0	1	1	1	3	30	43	30	0	0	0	
Обновлене веб-сайта	1	1	3	22	35	42	5	0	0	0	0	
Выделение ключевого слова жирным шрифтом	0	0	0	1	1	45	25	25	12	0	0	
Точное вхождение ключевого слова в текстах гиперс-сылок в других документах веб-сайта	0	0	1	0	1	0	1	7	22	42	35	
Уникальность документа внутри веб-сайта	0	0	0	0	0	0	0	1	0	1	107	
Уникальность документа и веб-сайта в целом в Интернете	0	0	0	0	1	0	0	1	0	1	0	
Корректная работа скрипттов	0	0	0	1	1	4	35	46	12	10	0	
Запрет от индексации избыточных служебных страниц	0	0	0	0	1	0	1	5	23	39	34	
«Понятные» URL-адреса веб-страниц	1	4	13	23	23	23	22	0	0	0	0	
Вынос java-скриптов и css в отдельные файлы	1	4	5	2	23	35	28	11	0	0	0	
Главное зеркало для систем	1	0	3	1	12	32	23	24	11	2	0	
Соответствие HTML-кода страницы стандарту	74	23	9	0	0	0	0	0	0	0	0	
Уровень вложенности страницы	99	5	3	2	0	0	0	0	0	0	0	
Возраст страницы	88	3	12	5	1	0	0	0	0	0	0	
Количество разрешенного для индексации контента	100	4	5	0	0	0	0	0	0	0	0	
Ключевые словав URL-страницы	97	1	6	4	1	0	0	0	0	0	0	
Частота обновлення текста страницы	91	6	5	5	2	0	0	0	0	0	0	
Ключевые слова в тегах alt картинок	99	6	2	1	1	0	0	0	0	0	0	
Наличие уникальных картинок в тексте страницы	98	7	2	0	1	1	0	0	0	0	0	
Корректная орфография текста страницы	96	10	1	0	2	0	0	0	0	0	0	
Анкор (разнообразие, естественность)	0	0	0	0	0	0	0	0	1	2	106	
Качество донора	0	0	0	0	0	0	0	2	12	26	69	
Разнообразие анкор-листа (по ключевым словам)	0	0	0	0	0	0	1	1	0	1	107	
Динамика изменения анкор-листа (по приросту)	0	0	0	0	0	0	1	1	1	3	104	
Траст	0	0	0	0	0	0	1	1	2	1	105	
Возраст домена	0	0	0	1	0	1	5	10	91	1	0	
Наличие в Яндекс-каталог	0	2	0	1	4	33	41	15	14	0	0	
Наличие в каталоге DVOZ	0	0	0	1	2	38	44	13	11	0	0	

Количество экспертов – 109, из них с уровнем доверия 0,8 – 5; 0,9 – 34; 1,0 – 70.