Министерство образования Республики Беларусь

Учреждение образования

Белорусский государственный университет

информатики и радиоэлектроники

УДК 621.316.7

Ромейко

Олег Александрович

ПОИСКОВАЯ ОПТИМИЗАЦИЯ САЙТОВ С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА КОНТЕНТА

АВТОРЕФЕРАТ

на соискание степени магистра техники и технологии

по специальности 1-39 81 03 Информационные радиотехнологии

_____О.А. Ромейко

Научный руководитель

Светлана Юрьевна Михневич

кандидат ф-м. наук

_____

Минск 2015

ВВЕДЕНИЕ

В первые дни интернета, поисковая оптимизация была не очень сложной задачей. Для высокого ранга в поисковой системе достаточно было поместить ключевое слово в название и добавить мета-теги. После этого сайт уже можно было с легкостью находить на самых высоких позициях в поисковых системах.

Все это привело к манипулированию «ключевыми словами», повышению частоты и/или плотности ключевых слов на странице, повторению ключевого слова мелкими буквами или тем же цветом, что и фон страницы. Для спамминга использовались meta-тэги, тэг title, подписи к графике, которые могут быть совершенно не относящимися к тематике сайта ключевыми словами.

Поисковые системы проанализировали эти трюки и предложили веб-мастерам использовать еще один принцип, который повлияет на позиции сайта в результатах поиска. Это количество входящих ссылок на сайт.

Случилось это примерно в то же время, когда родился Page Rank Google. Этот алгоритм гласит, что каждая ссылка на страницу, представляет собой «вес» для этой страницы, так что чем больше ссылок странице, тем больше доверия к ней со стороны поисковой системы. Сосредоточив всё внимание на количество входящих ссылок появилась еще одна форма спама. Вебмастера начали покупать ссылки или производить обмен ссылками для того чтобы увеличить их количество. Очевидно, почему бы не сделать веб-страницу, которая будет содержит 1000 ссылок более авторитетной, чем имея 100 ссылок, если учесть, что ссылки можно просто купить.

Поэтому еще раз поисковые системы вынуждены были пересмотреть свои критерии для измерения важности веб-сайте. На этот раз они сосредоточили внимание на релевантности веб-сайта с которого идёт ссылка на другой. Если они были связаны, это считалось хорошим признаком в поисковых системах. Если нет, то это было не так хорошо, и иногда даже потенциально вредными.

Поисковые системы постоянно пытаются менять настройки своих алгоритмов так, чтоб SEO специалисты не остались впереди по игре.

Исследуемый алгоритм делает попытку решить проблему оптимизации веб-приложения не пользуясь помощью SEO специалистов. Система, реализованная на основе данного алгоритма, позволяет в автоматизированном режиме выдавать рекомендации, которые помогут оптимизировать приложение.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Вместе с появлением и развитием поисковых систем в середине 1990-х появилась поисковая оптимизация. Обычно, чем выше позиция сайта в результатах поиска, тем больше заинтересованных посетителей переходит на него с поисковых систем. При анализе эффективности поисковой оптимизации оценивается стоимость целевого посетителя с учётом времени вывода сайта на указанные позиции и конверсии сайта. Поэтому множество компаний заинтересованы в поисковой оптимизации. Так по оценке Российской ассоциации электронных коммуникаций рынок поисковой оптимизации в 2014 году составил 10,24 млрд. руб. В 2015 году прогнозируется рост рынка на 19 %.

На заре интернета, поисковая оптимизация была достаточно простой задачей. Для высокого ранга в поисковой системе достаточно было поместить ключевое слово в название и добавить мета-теги. После этого сайт уже можно было с легкостью находить на самых высоких позициях в поисковых системах.

Все это привело к манипулированию ключевыми словами, увеличению плотности ключевых слов на странице, множественному повторению ключевых слов маленькими буквами или тем же цветом, что и фон страницы. Для спаминга использовали meta-тэги, тэг title, подписи к изображениям, которые могут быть совершенно не относящимися к тематике сайта ключевыми словами.

Основа поисковой оптимизации — ключевые слова. Пользователи поисковых систем находят нужный сайт, вводя в строке поиска нужное слово или словосочетание, и поисковые системы, выполняя заказ пользователя, принимаются за поиск нужных слов и предложений в проиндексированных ими сайтах. Чем более текстовой контент сайта, по мнению поисковой системы соответствует запросу, тем выше в результатах поиска система разместит ссылку на ресурс. Здесь и кроется причина и следствие: основной объект приложения усилий специалистов по поисковой оптимизации — позиция сайта в результатах поиска по определенным ключевым словам и словосочетаниям.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В работе представлен алгоритм оптимизации, плотно работающий с ключевыми словами. Разработан прототип сервиса, который выдает рекомендации по оптимизации сайта на основе адреса и требуемых запросов. Реализация этого алгоритма максимально просто и быстро может помочь рядовым пользователям добиться повышения ранга своей страницы.

Реализуемый алгоритм считается алгоритмом «белой» оптимизации, т.е. алгоритмом в котором не применяются запрещённые и недобросовестные методы продвижения, не нарушаются правила поисковых систем. Веб-страницы, продвигаемые при помощи алгоритма, полезны как для пользователей интернета, так и для поисковых машин, и самих SEO специалистов.

Основным компонентом функционирования веб-приложения является работа с поисковыми системами. В данном случае в качестве ведущей информационной единицы выступает адрес сайта, иными словами, подразумевается работа с самой страничкой пользователя и текстом на ней.

В качестве источников данных о ранжировании сайтов и данных о ранке сайта, выступает API google.com.

Создание архитектуры подразумевает выделение общих абстрактных компонентов (слоев) и описание функциональности каждого слоя.

Создание приложения в данном случае подразумевает создание двух компонентов: сервера данных – модуля, отвечающего за сбор и унификацию данных из источников, реализацию алгоритма в серверной части и клиентской части в виде веб-морды в браузере, а также обеспечение взаимодействия между указанными модулями.

Алгоритм оптимизации построен на поиске и анализе релевантной страницы по запросу. Анализ происходит путем парсинга страниц: релевантной, TOP-10 и той, которую ввел пользователь для оптимизации.

Анализируется абсолютно весь контент в выбранных тегах. Последующие рекомендации выводятся путем сравнения анализированных данных с релевантной страницы, TOP-10 и страницы пользователя.

Рассмотрим более подробно алгоритм. Структура алгоритма представлена на рисунке 1.

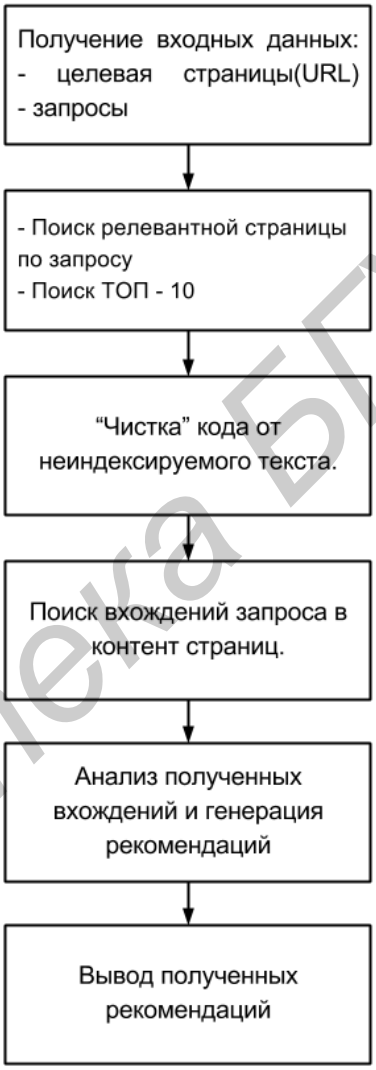

Рис. 1 – Структура алгоритма оптимизации

Работа алгоритма построена следующим образом: на входе - адрес целевой страницы и запросы, по которым мы хотим подняться в ТОП поисковой системы.

Получив начальные данные, для каждого запроса находим релевантную страницу, т.е. страницу которая наиболее точно отражает информацию о запросе, и ТОП – 10 страниц по запросу.

Для целевой страницы, релевантной страницы и каждой страницы из ТОП-10 проводим своеобразную «чистку». Из кода убираем все лишнее, т.е. скрипты, комментарии, текст внутри noindex. Также нужно убрать все стоп-слова (слова и частицы, которые не индексируются поисковыми системами).

После проведения «чистки», проанализируем каждую из страниц. Нам нужно узнать следующее:

точное вхождение запроса;

словоформу запроса (без окончаний с сохранением порядка слов), с указанием количества вхождения каждой словоформы;

точное вхождение каждого слова из запроса;

словоформу каждого слова запроса (без окончания), с указанием количества вхождения каждой словоформы для каждого слова;

точное вхождение фразы с пропуском одного любого из слов (порядок слов во фразе сохраняется);

вхождение словоформы фразы с пропуском одного любого из слов (порядок слов во фразе сохраняется), с указанием количества вхождения каждой словоформы;

точное вхождение фразы с добавлением между словами фразы в любом месте одного случайного слова;

вхождение словоформы фразы с добавлением между словами фразы в любом месте одного случайного слова, с указанием количества вхождения каждой словоформы;

точное вхождение фразы с заменой одного из слов фразы на случайное слово (замена происходит в любом слове за исключением первого и последнего);

вхождение словоформы фразы с заменой одного из слов фразы на случайное слово (замена происходит в любом слове за исключением первого и последнего) , с указанием количества вхождения каждой словоформы.

Имея данные анализа страниц можно получить рекомендации, путем анализа усредненных значений релевантной страницы, данных по ТОП-10 и целевой страницы.

Пройдя все стадии работы алгоритма, мы получаем список из рекомендаций по увеличению ранга страницы в поисковой системе. Примерные выходные данные алгоритма, выглядят так:

1. You need to add**"request"** in <body>...</body> 37 times.
2. You need to delete **"request"** in <a>... 5 times.
3. You need to add**"request"** in <li>... 3 times.

Для демонстрации принципов архитектуры в целом, скорость разработки сервиса и уровень абстракции являются одними из наиболее важных факторов. Поэтому для демонстрации работы алгоритма использовалось самое простое приложение.

В общем случае на выбор среды и инструментов реализации приложения влияют также такие характеристики как функциональность, качество, скорость и простота реализации приложения.

ЗАКЛЮЧЕНИЕ

Изучены основные принципы работы поисковых систем *Яндекс* и *Google*, важные понятия и алгоритмы, а также отличия продвижения сайта в этих поисковых системах.

Рассмотрен в практической части продвижения алгоритм внутренней оптимизации сайта, в том числе, правильная оптимизация страницы и ее текстовой составляющей. Предложен модифицированный алгоритм правильной оптимизации, позволяющий продвигать сайты, не прибегая к услугам SEO специалистов.

Разработана простая, гибкая и современная информационная система, основанная на открытых технологиях и продуктах, в которой был реализован один из алгоритмов оптимизации.

Использование информационной системы позволит упростить оптимизацию сайтов для рядовых пользователей, а также значительно увеличить количество объективных статистических показателей, характеризующих работу сайтов.