Problems of modern serps
Dovbenko A.
Проблемы современной поисковой выдачи
Довбенко А. В.

Довбенко Алексей Викторович / Dovbenko Aleksey – аспирант,
кафедра теоретических основ информатики, факультет прикладной математики, информатики и механики,
Воронежский государственный университет, г. Воронеж

Аннотация: *в работе произведен анализ основных актуальных проблем поисковой выдачи, а также возможные варианты их решения. На заре развития поисковых машин основными проблемами было хранение данных, их анализ и быстрота поиска, с развитием технологий эти проблемы становились всё менее и менее актуальными. Сейчас уже любой человек без особого удара по бюджету может спокойно хранить терабайты данных, и выполнять поиск по ним стандартными средствами баз данных путем простой оптимизации. Но с решением проблем технической стороны начали появляться проблемы качественные, накапливаемая информация становиться менее актуальной и в этой статье будут изложены одни из популярных проблем, а также их возможные решения, которые сейчас, к сожалению, либо ещё не внедрены, либо находятся на этапе внедрения. Статья будет полезна всем, кто, так или иначе, работает с поисковыми машинами.*

Abstract: *the paper made analysis of the main actual problems of search results, as well as possible solutions. In the early days of search engines were the main problems of data storage, analysis and retrieval speed with advances in technology, these problems become less and less relevant. Already, anyone without much impact on the budget can safely store terabytes of data, and search for them by standard database means by simple optimization. But the decision of the technical side of the problem of problems began to appear quality, accumulated information becomes less relevant and this article will outline some of the most popular problems and their possible solutions, which are now unfortunately, or not yet implemented, or are in the implementation phase. The article will be useful to all those who in one way or another, work with search engines.*

Ключевые слова: *поисковая машина, анализ информации, индексация, SEO.*
Keywords: *search engine, information, analysis, indexing, SEO.*

1. Дублирующая информация

С каждым днем сайтов становиться все больше и больше и в идеале каждый сайт должен обладать уникальной информацией, но это далеко не так. Как правило, большинство сайтов копирует информацию у других, причем копирует полностью, пытаясь таким образом вылезти в топ выдачи, не редко получается даже так, что сайт с оригинальной статьей находиться на второй странице поиска, а сайт, который её скопировал на первой. На данный момент поисковые боты начинают уметь просто определять «ворованные» статьи (от простого хранения хэша, до упрощенного анализа текста), но всё это легко обходиться благодаря вставкам в парсер логики унификации текста (например, заменить русскую «с» на английскую “c”).

Как упрощенное решение — это перевод текста в транслит, перевод всех символов в нижний регистр и убирание всех знаков препинаний, соответственно так мы получаем уникальный отпечаток информации и можем сверять его полностью (по хэшу) или частично, находя одинаковые элементы в тексте. Таким образом, можно отсечь кучу парсеров, безусловно, со всеми не справиться, т. к. есть куда более изощренные способы агрегирования чужой информации, такие как генерация статьи на основе нескольких других. Но основную проблему такой подход вполне способен решить.

2. Устаревшая информация

Пожалуй, одна из основных текущих проблем современных поисковых движков. В связи бурным развитием форумов, новостных сайтов и социальных сетей мы имеем страницы, с информацией, которая давно уже не актуальна. Например - если ввести запрос на покупку редкого животного или товара по своему городу, то с большой вероятностью нам покажут объявления, которым более 5, а то и 10 лет, очевидно, что по этим объявлениям уже бесполезно звонить. Тоже самое можно наблюдать при поиске новостей, к примеру, по поиску информации о каком-то человеке или событии, в топе будет популярная новость за прошлые года, а не более новая. Основная проблема в том, хранить только свежую информацию тоже нельзя, так как основная часть поисковых запросов часто ориентирована на статическую информацию (события в истории, научные стати и т. п.), которая в свою очередь может не обновляться с момента создания. На данный момент у поисковой машины принуждают владельцев сайтов удалять из индекса неактуальную информацию (путем добавления мета-тэгов, запрещающих индексирование страницы), из-за ограничения количества индексируемых страниц в зависимости от тИЦ и PR (величина индекса цитирования ресурса в интернете, для Яндекса и Google соответственно).

То есть ответственность на выдачу лежит на владельцах сайтов, что в корне неверно по нескольким причинам:

- многим достаточно того количества индексируемых страниц которых дает поисковая машина;
- требует технических навыков, которых может не быть у владельцев простых форумов, построенных на простых CMS;
- крупным ресурсам имеющих постоянную аудиторию или раскручивающихся на базе социальных сетей или любой другой системы кроме поисковых машин, они просто не обращают внимания на то, есть у них неактуальная информация в поисковой выдаче или нет;
- некоторым это даже выгодно, т. к. человек зайдет на их сайт, пусть и не найдет нужную информацию, убрав же из индекса эту устаревшую информацию они могут лишиться неплохого процента пользователей, приходящих с низкочастотных поисковых запросов.

Лучшим решение этой проблемы будет определение тематики сайта, и давать старой информации определенный срок жизни, например, объявления - месяц, форумы - год и т. д., таким образом можно без особых усилий «очистить» выдачу от устаревшей информации. На данный момент поисковые системы путем инструментов для SEO предлагают определить тематику сайта, но опять же решения этой проблемы лежит на стороне пользователей, что опять же некорректно. Единственное видимое решение этой проблемы — это определение тематики сайта путем анализа сайта на содержимое и с помощью нейронной сети присваивать сайту ту или иную категорию, и учитывать её в выдаче информации и длительности её хранения.

3. Повышение рейтинга инертной и популярной тематики

Не критичная, но довольно часто встречаемая проблема. При появлении какого-то очень актуального события все поисковые запросы, по ключевым словам, связанным с ними выдают информацию, связанную с ним. Как пример, когда вышел фильм «Мальчишник в Вегасе», запросы, по словам «мальчишник» или «Вегас», непременно вели на рецензии или обзоры этого фильма, удивительно, что на текущий момент у всех крупных поисковых систем эта проблема легко решается путем использования поискового языка запросов. Основная проблема в том, что он не обладает интерфейсом понятным для пользователя. Поисковый языка запросов у ИПС имеет символы подстановки, в принципе человек, который работал с регулярными выражениями, найдет его очень простым, в то время как рядовой пользователь вряд ли запомнит хотя бы пару знаков, которые помогут ему отфильтровать результаты. В связи с этим ИПС разрабатывают более дружественный интерфейс, заменяя основные и часто требующиеся символы подстановки простыми формами, которые куда более понятны пользователю.

Безусловно, основная ответственность за решение этой проблемы в основном лежит на UI-дизайнерах, что подчеркивает, что ИПС далеко не полностью зависит от технической стороны.

4. Ресурсы и сервисы с малой текстовой информацией

Тоже одна из проблем современных ИПС. Зачастую хостинг картинок, видео, сайты фотографов имеют очень много медийной информации, но крайне мало текстовой, в связи с чем, поисковые машины предлагают пользоваться мета тегами для описания содержимого картинки или размещать медиа ресурсы на отдельных страницах, описывая содержимое в title, но в современных тенденциях дизайна и дружественного пользовательского интерфейса более привычно показывать медийный элемент в pop-up окне, нежели открывать новую вкладку, а описание мета информации часто просто забывается (важно учитывать, что многие изображения и видео загружаются простыми пользователями). В идеале для таких сайтов ИПС должна иметь уже заготовленный набор ключевых слов, путем анализа содержимого. Возможно в будущем можно будет при малых затратах на ресурсы анализировать изображения и видео самостоятельно описывая его, на данный момент похожую систему разработала компания Microsoft под названием CaptionBot (https://www.captionbot.ai/) система относительно быстро анализирует изображение и кратко описывает что изображено на картинке. Но пока данная система далека от совершенства, довольно часто система не может описать изображения или описывает в корне неправильно. Плюс система ещё довольно медленная для работы с большим объемом данных. Так что текущее решение - описывать картинку самому самое результативное на данном этапе. Но в последнее время, судя по активности многих компаний над работой распознавания, что изображено на картинке следует ожидать подобного решения в ближайшем будущем.

5. «Черное» SEO

Пожалуй, самая трудно-решаемая проблема из общего списка. Не смотря на довольно жесткую борьбу с обманным путем поднятия ресурсов (причем ресурсов часто с содержащую мошеннические схемы или вредоносный код) путем бана ресурса, полного исключения из поисковой выдачи, предупреждением при переходе, «черное» SEO развивается бурными темпами. Методы обман ИПС становятся всё изощреннее вот примеры нескольких методов:

- Скрытый текст, наполненный огромным количеством ключей. Используется слишком мелкий шрифт и/или одинаковый цвет текста и фона, что делает его невидимым для людей, но доступным для

индексации поисковыми роботами. Подобные черные методы оптимизации относятся к классическим. Поисковые системы давно научились успешно их выявлять и наказывать виновных [3].

- Клоакинг (англ. cloaking – маскировка). Этот метод также относится к black SEO и подразумевает отображение нескольких видов контента: один (интересный, полезный) – для пользователей, другой (корявый, с ключевыми запросами) – для поисковых ботов. При этом каждый из них видит только предназначенный ему текст [3].
- Создание дорвеев (от англ. doorway – входная дверь) – черное SEO, использующее одностраничники или полноценные веб-ресурсы, часто автоматически сгенерированные, наполненные низкокачественным, бессмысленным контентом с высокой плотностью низко частотных и средне частотных запросов, по которым они быстро продвигаются в топ выдачи. Далее настраивается перенаправление трафика (автоматически или с помощью ссылок и баннеров) на веб-ресурс, «заточенный» уже под людей и приносящий основную прибыль. Этот метод актуален до сих пор, т. к. его довольно трудно распознать [2, 3].
- Линкфарминг (от англ. link farm – ферма ссылок) – черные способы продвижения, основанные на создании сети веб-ресурсов с целью «разведения/выращивания» ссылок, указывающих друг на друга. Таким образом, происходит взаимное наращивание ссылочной массы. Элементы сети обычно не имеют вразумительного содержания [1, 3].
- Метод сателлитов (от англ. satellite – спутник). Подобное black SEO основывается на строительстве сети веб-ресурсов, каждый из которых ссылается на основной продвигаемый сайт. Вся сеть продвигается с целью размещения основного сайта на первом месте и захватом всех позиций в ТОП 10 сателлитами (по определенным запросам, чаще всего коммерческим) [2, 3].

И это только несколько примеров, причем каждый из них имеет множество вариаций выполнения, что очень сильно затрудняет написание общего алгоритма обнаружения использования черного SEO. В основном поисковики делают ставку на анализ текста и обнаружение, что текст, сгенерированный, далее идет анализ всех ссылок с этого сайта, и попытка обнаружить целую сеть. Но опять же абсолютной системы защиты от черного SEO нет и вряд ли когда будет.

В статье рассмотрены основные проблемы поисковой выдачи, как мы видим большинство из них вполне возможно решить, но для каждой проблемы необходимо проводить должные исследования, стоит заметить, что представленные в статье решения носят исключительно предполагаемый характер, и, возможно, при боевом тестировании окажутся не столь эффективными как ожидалось.

Литература

1. Link Farm. [Electronic resource]. URL: https://en.wikipedia.org/wiki/Link_farm/ (date of access: 12.12.2016).
2. Сателлит. [Электронный ресурс]. Режим доступа: http://seowikipedia.su/index.php/%D0%A1%D0%B0%D1%82%D0%B5%D0%BB%D0%BB%D0%B8%D1%82/ (дата обращения: 12.12.2016).
3. 7 приемов черного SEO. [Электронный ресурс]. Режим доступа: http://shakin.ru/seo/black-grey- and-white-seo.html (дата обращения: 12.12.2016).