УДК 519.7

ГВОЗДИНСКИЙ А.Н., канд.техн.наук., проф. ХНУРЭ (Харьков)
МИЛЬЦЕВА Л.А., студент ХНУРЭ (Харьков)

ОЦЕНКА ОПТИМИЗАЦИИ САЙТОВ ПОД ЗАПРОСЫ ПОИСКОВЫХ СИСТЕМ

Сегодня количество сайтов в Интернете растёт неимоверными темпами и ориентироваться в этом потоке не всегда легко. Сегодня на многие сайты клиент попадает посредством поисковых систем и как следствие большие маркетинговые усилия тратятся на то, чтобы увеличить свои позиции в этих поисковых системах. Данная статья посвящена алгоритмам для оценки качества оптимизации под запросы поисковых систем.

Today amount of sites in the Internet of rastt by incredible rates and the oriented in this stream easily not always. Today on many sites a client gets by means of the searching systems and as a result large marketings efforts are outlaid on that, to increase the positions in these searching systems. This article

is devoted algorithms for the estimation of quality of optimization under the queries of the searching systems.

Введение. Оптимизация сайтов под запросы поисковых систем – процесс долгий и требует глубокого знания принципов, по которым поисковые системы определяют релевантность страниц по ключевым словам.Возникает необходимость в том, чтобы помочь оптимизаторам в их работе, создав некоторую систему по оценке качества их работы. Такая система должна анализировать содержание отдельных страниц или всего сайта целиком, давая цифровую оценку всему комплексу мер по оптимизации на сайте в целом и так же отдельным её параметрам. Подобная система будет полезной как и профессиональным оптимизаторам, позволяя проверять результаты своей работы, так и новичку, указывая на то, что следовало бы изменить на сайте в рамках проведения оптимизации.

Данная статья посвящена алгоритмам для оценки качества оптимизации под запросы поисковых систем.

Постановка задачи. Предлогаемый алгоритмы является общим для большинства поисковых систем, эту задачу облегчает то, что принципы работы основных ведущих поисковых **систем** в общем-то схожи.

Мы ограничимся параметрами, которые можно определить, анализируя только содержание отдельно взятого сайта, то есть набор страниц, связанных между собой структурой ссылок, расположенных в Интернете под одним доменным именем. Таким образом, мы сразу отметаем такой очень важный фактор при проведении оптимизации, как ссылки, ведущие на ваш сайт. Для того, чтобы учитывать количество и качество ссылок, ведущих на определённую страницу, необходимо либо сканировать все веб сайты в сети Интернет (как это делают все серьёзные веб поисковые системы), либо мы должны обратиться к какой-либо из поисковых систем, которые иногда предоставляют доступ к различной информации, в том числе и количеству ссылок на ту или иную страницу. Первый случай не приемлем по техническим причинам, а второй привязывает нас к какой то конкретной поисковой системе, что, как мы уже условились, для нас не допустимо.

Оценка оптимизации отдельной страницы. В идеале, чтобы определить, насколько хорошо та или иная страница оптимизирована для, скажем, поисковой системы Google, необходимо воспользоваться алгоритмом подсчёта релевантности страницы самой поисковой системы Google. То есть задача оценки оптимизированости одной страницы фактически сводиться к подсчёту релевантности документа по конкретным ключевым словам.

Общий принцип определения релевантности документов – это моделирование ситуации, когда документ просматривается некоторым абстрактным веб-пользователем и, отталкиваясь от этой модели определять релевантность документа. Действительно, ключевое слово или ключевая фраза тем более релевантны на странице, чем больше вероятность того, что данное слово/фраза будет замечена пользователем. Так, например, в первую очередь пользователь обращает внимание на заголовок документа, затем на название его разделов, далее на слова, выделенные жирным шрифтом и т.д.

Также необходимо учитывать тот факт, что все серьёзные поисковые системы борются с так называемым поисковым спамом. Поисковым спамом называют такое явление, когда различные держатели сайтов намеренно создают страницы, которые с точки зрения пользователя не представляют большой актуальности и специально созданы для того, чтобы занять хорошие позиции в поисковых системах. Самой примитивной разновидностью поискового спама являются страницы, на которых много раз повторяется одна и та же фраза..

Релевантность одного слова на странице. Опишем формулу, для описания численной релевантности единичной встречаемости слова в тексте. Под встречаемостью слова мы будем подразумевать единичное упоминание слова в документе.

$$\text{Relevancy}(x) = H(x)S(x)p(x) \quad (1)$$

где x – какая-либо встречаемость слова на странице, S(x) – площадь, которую слово занимает при данном появлении на странице, p(x) – коэффициент релевантности, который зависит от позиции появления слова на странице и H(x) – некоторый коэффициент, отвечающий за то выделено ли слово на странице.

Так, H(x) можно определить следующим образом

$$H(x) = \begin{cases} 1.3, \text{если x выделено отличным от основного текста цветом или шрифтом} \\ 1, \text{иначе} \end{cases} \quad (2)$$

Таким образом, мы получаем, что выделенные слова релевантнее простого текста.

Далее определим p(x). Данная функция учитывает то, как на релевантность встречаемости слова влияет его позиция на странице. Если обратится к нашей модели поведения веб-пользователя, данная величина отображает как вероятность того, что слово будет замечено пользователем, зависит от позиции данного слова на странице. Естественно, что если слово находиться в верхней части страницы, то данная вероятность велика, и чем ниже это слово стоит, тем вероятнее, что пользователь до данного места в тексте просто не дойдёт и перестанет читать страницу. Таким образом величина p(x) отражает то, на сколько быстро читатель теряет интерес к тексту по мере его чтения. Данная величина зависит от огромного числа параметров, в том числе и от индивидуальных особенностей человека. Поэтому логично воспользоваться нормальным распределением:

$$p(x) = e^{-\alpha_1 pos(x)^2} \quad (3)$$

где pos(x) это позиция данного слова в документе, а α_1 константа, отвечающая за скорость потери интереса человека к текущему документу. Необходимо отметить, что при выборе такой функции релевантность слов, находящихся в конце больших документов фактически сводится к нулю. Это легко объяснить с точки зрения здравого смысла: если человек дочитал документ, допустим, до 30-й страницы, то, скорее всего для него задача, стоит ли данный документ читать или нет уже решена положительно и релевантность этих слов для поиска сводится к нулю.

S(x) достаточно легко определяется математически как площадь, занимаемая словом на странице – то есть произведение высоты на ширину.

В тривиальном алгоритме подсчёт релевантности целой страницы был бы простым суммированием релевантности каждой встречаемости слова на странице, однако в таком случае, оптимизаторам следовало бы действовать по принципу - “Чем больше, тем лучше”, то есть перенасыщать страницу ключевыми словами. В реальной жизни необходимо знать меру в использовании ключевых слов. Алгоритм, описывающий релевантность отдельно взятой страницы по ключевым словам должен учитывать отношение встречаемости ключевого слова к общему количеству слов и, если данное отношение слишком большое, то релевантность страницы должна быть меньше.

Постараемся выразить написанное выше математическим языком.

$$Relevancy_p(P, x) = \sum_{k=1}^{n} Relevancy(x_k) - U(P, x) \quad (4)$$

Данная формула считает релевантность страницы P по ключевому слову x. x_i это встречаемости слова x на странице P, а n – общее количество таких встречаемостей. U(P,x) это своеобразное “наказание” за чрезмерную переоптимизированность. Из формул 1 и 3 следует, что сумма $\sum_{k=1}^{n}$Relevancy(x_k) является конечной. Обозначим для краткости

$$\eta(P,x) = \frac{\sum_{k=1}^{n} Relevancy(x_k)}{\sum_{k=1}^{M} Relevancy(y_k)} \quad (5)$$

тогда U(P,x) можно определить следующим образом.

$$U(P,x) = \alpha_2(\frac{\sum_{k=1}^{n} Relevancy(x_k)}{\sum_{k=1}^{M} Relevancy(y_k)})^{1+\alpha_3} = \alpha_2(\eta(P,x))^{1+\alpha_3} \quad (6)$$

где M это общее количество слов на странице, а . α_2 и α_3 это некоторые константы, для которых выполняются следующее условия:

$$\alpha_2 \geqslant \sum_{k=1}^{M} Relevancy(y_k), \alpha_3 \geqslant 0 \quad (7)$$

это максимальная релевантность, которую можно получить на данной странице. Таким образом η(P,X) обозначает степень оптимизированности данной страницы по данному ключевому слову. Если данная величина слишком велика, то это означает, что страница чрезмерно оптимизирована и маловероятно, что она несёт ценную информацию Допустим, если текст страницы состоит только из ключевого слова, то в этом случае U(P,x) = α_2 и $Relevancy_p(P,x) \leq 0$.Так, допустим, в тексте данной статьи, посвящённой оптимизации сайтов, слово “оптимизация” и производные от него, занимают около 2% от общего количества слов. Данное соотношение будет выполняться и для других научный работ, однако, для страниц, в сети Интернет оно может быть значительно больше (допустим, если страница содержит много картинок и мало текста). Если мы смогли определить оптимальное значение η(P,x) (например, статистически, анализируя страницы в Интернете), то константы α_2 и α_3 должны быть подобраны таким образом, чтобы $Relevancy_p(P,x)$ достигала своего максимума при данном значении η(P,x).Таким образом, умея оценивать релевантность страниц по определённым ключевым словам, мы можем перейти к задаче оценки её оптимизации. В первую очередь, мы должны оценить релевантность страницы по данному ключевому слову или фразе. Далее мы должны выделить потенциальные ошибки, которые допустил оптимизатор при проведении оптимизации. Например, если значение U(P,X), то оптимизатор должен об этом знать. Собранные данные должны быть представлены в виде отчёта и могут быть использованы оптимизатором для дальнейшей работы.

Оценка оптимизированности целого сайта. До сих пор, мы говорили об оценки оптимизации одной страницы. Оценка же оптимизации целого сайта – намного более интересная задача. Если при определении релевантности конкретной страницы опираться только на её содержимое, то систему будет очень легко обмануть. Допустим, если я знаю, что релевантность страницы по данному слову считается по алгоритму указанному выше, то я легко создам страницу, которая получит максимальную релевантность, но будет иметь случайное или почти случайное содержание. Одним из примеров подобных страниц, с которым мне приходилось встречаться на практике, это страницы, сформированные из запросов к самим поисковым системам. Допустим, я хочу создать страницу, хорошо оптимизированную по словосочетанию “поиск работы”. Это очень распространённое словосочетание, и по подобному запросу я найду очень много страниц в любой поисковой системе. Для того чтобы сформировать новую страницу на данную тему, мне достаточно ввести его допустим в поисковую систему Яндекс и результат скопировать на свою страницу. Полученная страница, будет очень

хорошо оптимизирована по данному словосочетанию. Данный процесс легко автоматизировать, и в результате можно получить большой сайт большим количеством страниц, оптимизированных под различные популярные фразы. Как уже упоминалось ранее, подобное явление называется поисковым спамом.

К счастью, проблема поискового спама уже давно и эффективно решена благодаря тому, что современные поисковые системы в значительной мере учитывают количество и качество ссылок, для каждой страницы.

Алгоритм учёта ссылочной структуры. В основе алгоритма учёта ссылочной структуры лежит подсчёт величины “важности” каждой страницы в сети Интернет.

Основная идея подсчёта ранга страницы, заключается в том, что чем больше ссылок идёт на конкретную страницу, тем популярнее эта страница, и как следствие наиболее вероятно, что она несёт полезную информацию. В качестве обоснования алгоритма, используется модель, в которой рассматривается некий абстрактный пользователь Интернета. Пользователь заходит на какую-то случайную страницу в Интернете и далее начинает переход по ссылкам. Ранг страницы в данной модели выражается как величина прямо пропорциональная вероятности того, что человек попадёт на данную страницу и выражается следующей формулой:

$$PR(x) = d + (1 - d)\sum_{k=1}^{n}\frac{PR(k)}{N(k)} \quad (8)$$

где d (0, 1), некая константа, отвечающая за минимальный ранг страницы, k пробегает по всем страницам, которые ссылаются на данную страницу, а N(k) это общее количество ссылок на k-й странице. Согласно данной формуле, каждая ссылка на данную страницу увеличивает её ранг прямо пропорционально рангу ссылающейся страницы и обратно пропорционально количеству ссылок со ссылающейся страницы.

Если вернуться к модели, то d это вероятность того, что пользователь сети Интернет, зашёл на страницу напрямую, а не перешёл на неё по некоторой ссылке. Далее, находясь на некоторой странице A, с вероятностью 1 - d пользователь перейдёт по ссылке страницы данной страницы на страницу B с вероятностью (1 - d). $\frac{1}{N(A)}$Соответственно, если PR(A) это вероятность того, что пользователь находится на странице A, то (1 - d) $\frac{PR(A)}{N(A)}$ это вероятность того, что пользователь попадёт на страницу B через ссылку на странице A. Просуммировав по всем ссылкам на страницу B и добавив d мы и получаем формулу 8.

Подобный подход делает более релевантными при поиске те страницы, на которые больше ссылаются с других страниц. Более того, нужно, чтобы ссылки шли со страниц с хорошим рангом, иначе их значение будет значительно меньше.

Учёт ссылочной структуры при оценки оптимизации. Таким образом пользуясь формулами 9 и 8, мы можем окончательную формулу для определения релевантности страницы по поисковому запросу:

$$Relevancy_p^{pr}(P, x) = PR(P)Relevance_p(P, x) \quad (9)$$

То есть, поисковая релевантность прямо пропорциональна рангу страницы и релевантности слова на данной странице.В связи с указанной выше формулой, работа оптимизаторов сводится к двум основным направлениям:

1. Получение ссылок на оптимизируемый сайт с внешних источников.
2. Организация внутренней структуры ссылок оптимизируемого сайта.

Как было определенно выше, мы ставим перед собой задачу оценки оптимизации в рамках одного конкретного сайта. То есть мы не учитываем внешние ссылки и работаем только с внутренними.

Если обратиться к формуле 8, то видно, что подсчёт ранга страницы, даже в рамках одного сайта – задача достаточно трудоёмкая и вручную почти не осуществимая. Из-за этого оптимизаторы, при организации структуры ссылок сайта, вынуждены во многом полагаться на интуицию. Система, осуществляющая пересчёт рангов страниц сайта основываясь на его структуре ссылок была бы крайне полезной сама по себе. В качестве же фактора оценки оптимизации конкретного сайте можно использовать суммарный ранг всех страниц сайта. Данная величина очень показательна, и может варьироваться в зависимости от того, какую структуру сайта вы выбрали.

Для осуществления процесс подсчёта ранга страниц необходимо применять итерационный подход. Сначала для каждой страницы присваивается начальный ранг PR(x) = d и затем ранг каждой страницы пересчитывается по формуле 8 до тех пор, пока ранги страниц не достигнут устойчивого значения. Полученные значения рангов страниц будут ценны для оптимизатора как сами по себе, так и в виде суммарного ранга всех страниц, так как являются показателем того, на сколько успешны или не успешны оказались усилия по оптимизации сайта.

Список литературы: 1.Google PageRank - A Survey" – статья посвященная алгоритму подсчета PageRank и эффектам связанным с ним, Markus Sobek, http://pr.efactory.de/ **2.** "The Anatomy of a Large-Scale Hypertextual Web Search Engine" – статья описывающая работу и устройство поисковой системы Google, Sergey Brin, http://www-db.stanford.edu/pub/papers/google.pdf **3.**"Google History" – история Google на официальном сайте, автор не указан, http://www.google.com/corporate/history.html история Google . **4.**"Танцы Google по флоридски" – статья посвященная оптимизации сайтов, автор не указан, http://www.searchengines.ru/articles/004584.html . **5.** "Успешный сайт для Google за 12 месяцев" – перевод статьи посвященной оптимизации сайтов, Brett Tabke, http://www.searchengines.ru/articles/004523.html

Поступила в редколлегию 13.06.2009