электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.021

Обзор алгоритмов ранжирования Google и Yandex

***Дьякова Н.А.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Системы обработки информации и управления»

Научный руководитель: Самохвалов Э.Н., к.т.н, профессор
кафедры «Системы обработки информации и управления»
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
bauman@bmstu.ru

Введение

На сегодняшний день те или иные поисковые системы востребованы практически для всех пользователей Интернета. Поисковые системы связывают пользователя, который стремится найти необходимую информацию и владельца сайта, желающего получить заинтересованного пользователя.

Поисковики ищут информацию, которую запрашивает пользователь, не в сети, а в ее базе данных. Из нее поисковая машина выдает пользователю результат на его поисковый запрос в виде страницы со списком сайтов, который называется "поисковой выдачей". Любой владелец сайта хочет, чтобы его сайт был на вершине этого списка. Процесс сортировки списка сайтов в результатах поисковой выдачи от наиболее соответствующих запросу (то есть релевантных) к менее подходящим называется "ранжированием". Специалисты по разработке поисковых систем стремятся к тому, чтобы пользователь в ответ на запросы получал ссылки на самые подходящие сайты, содержащие именно ту информацию, которую он ищет.

Что такое ссылочное ранжирование (PageRank)

Google PageRank (Google PR) является одним из методов, который Google использует для определения релевантности страницы. Важные страницы получают более высокий PageRank и, скорее всего, появляются в верхней части результатов поиска. PageRank (PR) измеряется по шкале от 0 до 10. PageRank основан на обратных ссылках. Чем больше качество обратных ссылок, тем более высокий PageRank. Улучшение рейтинга страницы Google очень важно, если вы хотите улучшить свой рейтинг в поисковых системах.

Поисковики анализируют структуру ссылок веб-страниц друг на друга. Так они узнают авторитет страниц среди тех, кто создает сайты и ссылается на другие сайты.

Идея является автоматизированной вариацией индекса цитируемости. Это значит, что кого больше цитируют, тот является авторитетным, и его труды полезнее человечеству. А тот, на кого реже ссылаются, людям менее интересен. Первыми эту идею применили в 1998 году создатели Google – Сергей Брин и Ларри Пейдж, тогда еще аспиранты Стэнфордского университета. Именно ссылочный ранг страницы стал основным принципом ранжирования результатов поиска в Google, что привело к резкому отрыву от конкурентов по качеству поиска и стало одной из основных причин доминирования Google в мировом Интернете. Они назвали его PageRank, упомянув в названии фамилию одного из создателей – Larry Page.

Как вычисляется ранг страницы

В теории идея вычисления ранга страницы такова: берем матрицу всех ссылок всех страниц Интернета друг на друга. Далее присвоим всем страницам одинаковый вес (ранг). Затем, начиная с какого-либо угла данной матрицы, посчитаем вес страниц и ссылок так: в случае если на данную страницу ссылается много страниц, то ее ранг повышается согласно классической формуле расчета PageRank:

$$PR = (1 - d) + d \sum_{i-1}^{n} \frac{PRi}{Ci}$$

где

PR — PageRank рассматриваемой страницы,

d — коэффициент демпфирования (вероятность, что пользователь, посетивший страницу, перейдет по одной из содержащихся на этой странице ссылок), в классической формуле обычно он равен 0,85.

PR_i — PageRank i-й страницы, ссылающейся на рассматриваемую страницу,

C_i — общее число ссылок на i-й странице.

Рассмотрим простой пример. Google присваивает каждой новой веб-странице первоначальный вес PageRank. Пусть в нашем примере, как видно на рисунке 1, начальный PageRank будет равен 1. Если мы создаем две новые страницы продукта, страницу A и страницу B, у каждой из этих страниц будет начальный PageRank, равный 1. Ссылка со страницы B на страницу A фактически будет голосовать за важность страницы A, и что голосование должно увеличить PageRank страницы A на 2 - начальный PageRank

плюс стоимость голосования страницы B. Голосование за страницу B это стоимость его PageRank, называемая мощностью ранжирования.

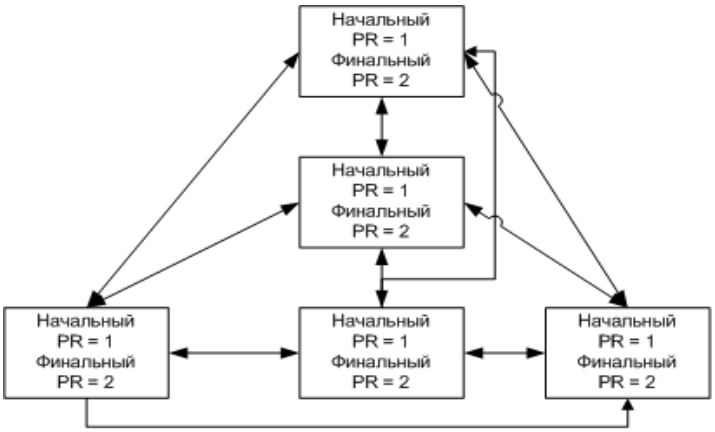

Рис.1. Вариант 1 стратегии навигации

Если мы добавляем новую страницу C и страницы B также связанные с ней, PageRank страницы A упадет от 2 до 1,5 в то время как PageRank страницы C увеличится с 1 до 1,5.

Добавление дополнительных ссылок со страницы B либо страницы A или страницу C не изменит ситуацию, так как мощность ранжирования распределяет только одну ссылку со страницы B к странице A. Вторая ссылка не будет добавлять дополнительную мощность ранжирования.

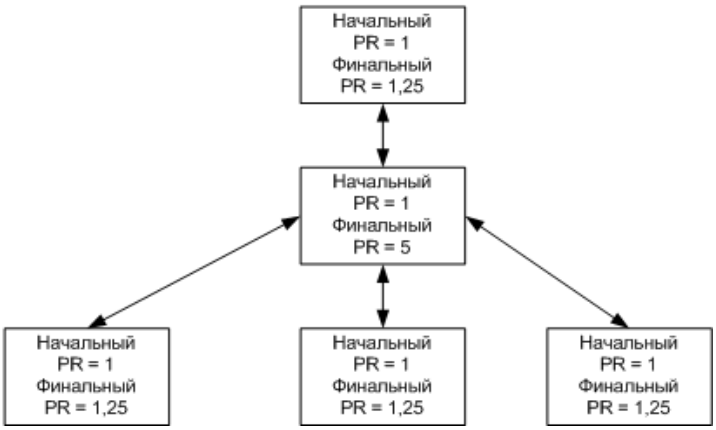

Рис.2. Вариант 2 стратегии навигации

С помощью этой простой модели мы можем теперь начать тестировать некоторые SEO-тактики внутренних ссылок. Рассмотрим простой сюжет из двух сценариев, добавим PageRank каждой странице, чтобы определить, какая тактика будет работать лучше для данной цели.

Например, давайте представим, что наш сайт имеет пять страниц, в том числе домашнюю страницу, страницу категории и три страницы продукта. Какая стратегия навигации будет лучшей, если нашей целью является повысить ранг вашей страницы категории?

Соединение каждой страницы даст PageRank страницы категории, равный 2, (см. рис.1). Связывание страницы товаров только со страницей категории, как на рисунке 2, приведет к PageRank = 5 на странице категории, что делает его лучшим выбором.

Алгоритмы ранжирования результатов поиска являются секретом поисковых систем. Поисковые системы работают автоматически. Полный ручной контроль результатов поиска и ранжирования по всем запросам невозможен, поскольку крупные поисковые системы работают с миллиардами страниц и сотнями миллионов запросов. Поэтому есть возможность продвигать сайты в зону видимости пользователя при помощи приемов, которые вводят в заблуждение роботов поисковиков либо иными способами вредят работе поисковых систем.

Главные параметры, которые учитывают поисковые машины для сортировки результатов поиска, являются известными. Оказывая влияние на эти параметры, можно поднять позиции сайта в результатах выдачи поисковиков. Влияние на характеристики сайта для продвижения ссылок на него в зону видимости поисковых систем по запросам называется поисковой оптимизацией.

Алгоритм Google Panda

Алгоритм Панда, главной целью которого является повышение качества результатов поиска, был запущен 23 февраля 2011 года. Главной задачей этого алгоритма было очищение выдачи от сайтов низкого качества. Так что же это означает для владельцев веб-сайтов?

Panda смотрит на объем контента на странице, скорее всего, делая простой подсчет слов. Благодаря Panda, "тонкие" страницы контента не так же хороши, как "толстые" страницы контента. Это означает, что на вашем сайте должно быть приличное количество текста.

Panda также учитывает оригинальность контента.

Алгоритм Google Penguin

Алгоритм Пингвин был запущен 24 апреля 2012 года. В отличие от Панды, этот алгоритм нацелен на борьбу с неестественными обратными ссылками**.** Алгоритм Пингвин

сосредоточен на связях. Пингвин наказывает сайты с неестественными профилями связи. Это могут быть сайты с ссылками из «ссылочных ферм» или других платных ссылок, множество ссылок с сайтов, которые не имеют отношения к веб-сайту, ссылки на низкокачественные сайты, и ссылки, которые неестественно оптимизированы ключевыми словами. Также санкции фильтра Google Penguin можно получить за дублирование контента и большое количество рекламы на главной странице.

Алгоритм Google Hummingbird (Колибри)

Google запустил алгоритм Hummingbird 26 сентября 2013 года. Алгоритм Hummingbird умеет обрабатывать длинные поисковые фразы, предлагая пользователю сайты, которые обеспечивают наиболее ценные ответы на запрос.

Алгоритм Колибри был разработан для того, чтобы лучше понимать запросы пользователей. Теперь, когда пользователь вводит запрос «Где можно вкусно поесть в Москве», поисковая система понимает, что под словом «где» пользователь подразумевает рестораны и кафе. До введения Hummingbird Google возвращал результаты, которые были сосредоточены на ключевых словах в вопросе. Напротив, Hummingbird анализирует вопрос, определяет его цель или смысл, а затем дает соответствующие ответы.

Алгоритм ранжирования Яндекса

Алгоритм ранжирования проводит лемматизацию (приведение словоформы к ее первоначальной словарной форме) слов документа и запроса, поэтому ему не важно в какой форме будет использоваться слово или его синонимы. Расчет релевантности документа запросу (Score) проводится по формуле:

$$Score = W_{single} + W_{pair} + k_1 * W_{AllWords} + k_2 * W_{Phrase} + k_3 * W_{HalfPhase}$$

где

W_{single} — вклад слов из запроса в документе

W_{pair} — вклад пар слов из запроса в документе

W_{Phrase} — вклад текста запроса целиком

$W_{HalfPhase}$ — вклад всех слов из запроса

Рассмотрим слагаемые более подробно.

1. Вклад слов из запроса рассчитывается по формуле:

$$W_{single} = \log(p) * (TF_1 + 0.2 * TF_2)$$

$$TF_1 = \frac{TF}{TF + k_1 + k_2 * DocLength}, k_1 = 1, k_2 = 1/350$$

$$TF_2 = \frac{Hdr}{1 + Hdr}$$

$$p = 1 - \exp\left(-1.5 * \frac{CF}{D}\right)$$

где

TF — число вхождений леммы в документ

DocLength — длина документа в словах

Hdr — сумма весов слова за форматирование

CF — количество вхождений леммы в коллекцию

D — количество документов в коллекции

2. Учет пар слов. В данном алгоритме учитывается также вхождение пар слов запроса в документ. Вес пары вычисляется по формуле:

$$W_{pair} = 0.3 * (\log(p_1) + \log(p_2)) * \frac{TF}{1 + TF}$$

p_1, p_2 — рассчитываются так же, как и для W_{single}

3. Учет всех слов. Существенным параметром помимо перечисленных является наличие в рассматриваемом документе всех слов запроса, которые вычисляются по формуле:

$$W_{AllWords} = 0.2 * \sum \log(p_i) * 0.03^{N_{miss}}$$

где Nmiss – число слов запроса, которые отсутствуют в документе.

4. Учет запроса целиком:

$$W_{Phrase} = 0.1 * \sum \log(p_i) * \frac{TF}{1 + TF}$$

5. Учет части запроса:

$$W_{HalfPhrase} = 0.02 * \sum \log(p_i) * \frac{TF}{1 + TF}$$

Таким образом, данный алгоритм способен отличать релевантные документы от нерелевантных.

Фильтры Яндекса

Если Google берет в качестве названий алгоритмов животных, то у Яндекса

названия алгоритмов совпадают с городами, например Магадан, Находка, Арзамас, Снежинск. У Яндекса, как и у Google, есть свои фильтры, предназначенные для исключения из результатов поиска сайтов, которые не несут полезную информацию конечным пользователям.

АГС 17 и АГС 30 (фильтры Яндекса) - это гранатометы автоматического действия, предназначенные для поражения живой силы и огневых средств противника. Фильтры АГС анализируют сайты на уникальность, полезность и интересность, и на основе этих данных решают, надо ли индексировать данный ресурс. Фильтр АГС 17 был внедрен в 2009 году. Главная цель этого фильтра – это борьба с дублированием контента. Фильтр АГС 30, выпущенный в 2010 году - улучшенная модель АГС 17. АГС 30 ищет дублируемый контент на чужих сайтах и удаляет найденные сайты из результатов поиска.

Основная задача фильтров АГС – борьба с сайтами-сателлитами, созданными не для людей, а для заработка путем размещения на них ссылок на основной сайт (который продвигается с помощью сателлитов) или оплаченных ссылок с бирж. Сателлиты не обладают нужной информацией для пользователей, обычно содержат дублированный контент, их можно назвать интернет-мусором.

Фильтры АГС работают автоматически, периодически проводится перепроверка всех сайтов. Если качество сайта улучшается и он становится полезным для пользователей, раннее наложенные ограничения снимаются.

Заключение

Таким образом, алгоритм Panda фильтрует некачественный контент, Penguin обесценивает сайты, построенные на естественных ссылках, а Hummingbird обеспечивает человеческий поиск ответов на вопросы. Фильтры Яндекса тоже не отстают от Google и эффективно борятся с неоригинальным контентом. У всех алгоритмов поиска одна цель – заставить веб-разработчиков создавать качественный, интересный и полезный контент. Google непрерывно продолжает работу над повышением качества результатов поиска. Веб-мастера должны создавать контент, который поможет людям найти решение их проблем, и тогда первые места в списке выдачи им гарантированы.

Список литературы

1. Ашманов И., Иванов А. Оптимизация и продвижение сайтов в поисковых системах. СПб.: Питер, 2011. 464 с.: ил.
2. Google Caffeine и новые факторы ранжирования. Режим доступа: http://www.cy-pr.com/articles/seo/453 (дата обращения 01.11.14)

3. Новый алгоритм Гугла (Google Penguin), как вывести сайт из под фильтра (вебспам и черное SEO). Режим доступа: http://ruslansavchenko.com/seo/google-penguin.html (дата обращения 01.11.14).
4. Как и за что можно загнать сайт под фильтр Google Penguin. Режим доступа: http://firelinks.ru/raskrutka-i-seo/165-filtr-google-penguin.html (дата обращения 01.11.14).
5. Google Hummingbird - новый алгоритм ранжирования. Режим доступа: http://texterra.ru/blog/kak-izmenitsya-internet-marketing-posle-zapuska-printsipialno-novogo-poiskovogo-algoritma-google-hum.html (дата обращения 01.11.14).
6. Алгоритм текстового ранжирования Яндекса на РОМИП-2006. Режим доступа: http://download.yandex.ru/company/03_yandex.pdf (дата обращения 01.11.14)
7. Что такое АГС от Яндекса. Режим доступа: http://www.bigfozzy.com/Articles/Based/Terminology/what-is-ags.php (дата обращения 01.11.14).