

Оптимизация расчета ссылочной популярности и учета ее при ранжировании результатов поиска

Трофименко Евгений Александрович
info@promosite.ru

Аннотация

Проблема поисковых алгоритмов, учитывающих наличие внешних ссылок на документ или сайт, состоит в возможности искусственного увеличения ссылочной популярности путем обмена ссылками, участия в ссылочных фермах. Для решения проблемы накруток обычно используют индивидуальные меры: исключение сайтов и ферм из индекса, наложение фильтров на исходящие ссылки и т.п., что требует участия человека-модератора. Кроме того, масса промежуточных случаев (тематические кольца, обмен ссылками в узких темах), могут быть ошибочно отнесены в категорию накрутчиков.

В работе предложена идея по разделению индекса ссылочной популярности (PageRank, SiteRank) на независимые части, соответствующие «добровольной» и «обменной» цитируемости с тем, чтобы в алгоритме ранжирования учитывать их с разными весами. Предложенный подход позволяет количественно и алгоритмически определять степень вовлеченности в системы ссылочной накрутки.

Введение

Алгоритмы поисковых систем по ранжированию веб-документов, учитывающие наличие ссылок на других документах, подвержены внешним влияниям. Влияние на результаты ранжирования со стороны владельцев сайтов может осуществляться с помощью обмена ссылками с другими сайтами, участия в ссылочных фермах, создания ссылок на свои сайты в гостевых книгах, каталогах, форумах, создания сети поддерживающих основной сайт ресурсов, обменивающихся ссылками и ссылающимися на основной сайт.

Для решения проблемы накрутки ссылочной популярности обычно используют такие меры, как: исключения сайтов из индекса, наложение фильтра на исходящие ссылки с сайтов. Однако, эти действия требуют ручной проверки ссылочных ферм и отдельных сайтов. Кроме того, ссылочная накрутка может остаться незамеченной при следующих условиях:

- 1) Малые масштабы накрутки (число сайтов, участвующих в обмене ссылками)
- 2) «Слив» ссылочной популярности с заранее «накачанных» ресурсов
- 3) Небольшая степень перелинковки (сайты могут ссылаться друг на друга не по алгоритму «каждый на каждого», а разреженно)

Кроме того, ошибки человека-модератора могут возникать в случаях, если:

- 1) В узких тематических областях естественным образом возникает перелинковка сайтов между собой
- 2) Веб-кольца для обмена посетителями могут вызвать подозрения
- 3) Обмен видимой пользователю статической рекламой со ссылками

В общем случае, почти любой обмен ссылками предполагает договоренность между ссылающимися сайтами. Следовательно, ценность таких ссылок в алгоритме ранжирования должна быть более низкой, нежели ценность «добровольных», односторонних ссылок.

В условиях, когда около 27% всех ссылок в русскоязычном Интернете (по данным Яндекса) являются обменными (т.е., в обмен вовлечено около 14% хостов) невозможно просто исключить взаимные ссылки из рассмотрения. Кроме того, обмен ссылками, даже и договорной, не всегда является накруткой – многие владельцы сайтов обмениваются ссылками с действительно качественными ресурсами в своей тематике и не заслуживают штрафных санкций.

При учете ссылочной популярности отдельных документов (хостов) часто в виде ее количественной меры используют взвешенную цитируемость, или PageRank. Алгоритм расчета PageRank документа предполагает учет цитируемости ссылающихся на него документов. Однако в алгоритме PageRank смешиваются все виды ссылок – односторонние и взаимные. Отсюда возникают следующие возможности для накрутки ссылочной популярности путем создания ссылочных ферм и массового обмена ссылками.

Невозможность разделить разные компоненты PageRank ведет к необходимости принятия резких мер – сайт либо полностью принимается поисковой системой, либо полностью отвергается ей. Кроме того, в этой деятельности особую роль играет человеческий фактор.

Поэтому задачей является разработка алгоритмического метода количественной оценки вклада «добровольной» и «обменной» цитируемости в общую цитируемость.

В данной работе предлагается метод количественной оценки цитируемости хостов (SiteRank), позволяющий разделить долю цитируемости, полученную

путем специальных действий (обмена ссылками и т.п.) и долю цитируемости, полученную за счет добровольных односторонних ссылок.

В дальнейшем эти ранги страниц можно использовать в алгоритмах ранжирования с разными весами при учете ссылочного ранжирования.

Идея исследования

Идея исследования – в разделении общей системы ссылок между хостами в Интернете на 2 подсистемы, не связанные между собой ссылками. Первая подсистема состоит из только лишь обменных ссылок. Вторая подсистема состоит из всех остальных ссылок.

Две подсистемы хостов могут пересекаться, т.е., один и тот же хост может находиться и в подсистеме «односторонних» ссылок (в ссылочную матрицу будут входить только односторонние ссылки) и в подсистеме «обменных» ссылок (в ссылочную матрицу будут входить только обменные ссылки). Это важный момент: это позволит не рассматривать хост либо как «только лишь накрученный» либо «абсолютно чистый».

Гипотеза 1: «добровольные, односторонние» ссылки ставятся в случае действительно качественного контента сайта и его уместности в контексте ссылающегося сайта. Поэтому вероятность перехода по такой ссылке должна быть выше. «Взаимные, обменные» ссылки привлекают к себе меньше внимания посетителя в силу их расположения на сайте (в каталоге ссылок) и меньшей уместности в контексте ссылающегося сайта.

Поэтому вероятность перехода посетителя по ссылкам разного типа должна быть разной, и, соответственно, при расчете pagerank нужно использовать разные значения dumping factor (d).

Гипотеза 2 (следствие из 1): поскольку вероятность перехода по «добровольной» ссылке выше, чем по «обменной», должно происходить естественное «перекачивание» посетителей из подсистемы хостов с обменными ссылками в подсистему хостов с добровольными ссылками. Таким образом, вероятность посещения сайта из подсистемы «добровольных ссылок» должна быть выше, чем подсистемы «обменных ссылок».

Гипотеза 3: даже при одинаковой вероятности перехода по обменной и добровольной ссылкам ценность второй для алгоритмов ранжирования выше, т.к. в первом случае выше вероятность того, что ссылки поставлены по предварительной договоренности. Отсюда следует целесообразность учета

«добровольной» цитируемости в алгоритме ранжирования с более высоким весом.

Методы, алгоритмы и эксперименты

Методы

Последовательность проведения исследования представлена ниже:

- 1) Описать уравнения расчета рангов
- 2) Выбрать константу d (dumping factor) для расчетов
- 3) Разделить подсистемы добровольных и взаимных ссылок
- 4) Рассчитать значения SiteRank для общей системы и двух добровольных и взаимных ссылок
- 5) Сделать выводы

0. Использованные данные

Использовался хост-граф номер 1. Данные по 4.9 млн. хостов, из которых около 500 тыс. известных Яндексу (скачанных), из которых около 250 тыс. имеют внешние ссылки на другие хосты (т.е., не являются «висящими»).

1. Уравнения расчета ранга SiteRank

Т.к. данные были получены по ссылкам между хостами, рассчитывались значение не PageRank (по ссылкам между документами), а SiteRank (между хостами). При этом в уравнениях каждый хост представляет собой одну «страницу», на которой есть ссылки вовне и на которую есть ссылки извне. Для расчета использовалась система уравнений:

$$SR(i)=(1-d)+d*\text{SUM}\{SR(j)/C(j)\}$$

Где $SR(i)$ – SiteRank хоста, $SR(j)$ – SiteRank ссылающегося хоста j , $C(j)$ – число внешних ссылок с сайта j , сумма по всем ссылающимся хостам.

Физический смысл SiteRank в этой системе из N уравнений – число находящихся на хосте пользователей при условии, что всего в Интернете «ходят» N пользователей. Это позволяет легко сравнивать значения SiteRank для нескольких подсистем с разным числом сайтов в них.

Расчет проводился с помощью итераций. Перед расчетом из матриц удалялись висящие страницы и ссылки на висящие страницы. Такая чистка матрицы проводилась несколько раз (до 6), т.к. после удаления ссылок на «висящие»

хосты появлялись новые «висящие» хосты. Таким образом, при расчете не требуется использование нормировок.

2. Выбор значения dumping factor

В литературе приводился пример значения dumping factor = 0.85. Для работы необходимы реальные значения этого параметра, который по смыслу соответствует вероятности перехода посетителя по одной из ссылок со страницы при условии, что он находится на этой странице.

Для оценки реальных значений d использовалась статистика по отношению числа просмотров страниц к числу посетителей (хит/хост). Было сделано предположение, что число внутренних ссылок на сайтах гораздо больше числа внешних ссылок.

Если посетитель зашел на сайт, он просматривает 1 страницу. Вторую страницу он просмотрит с вероятностью d , третью – с вероятностью d^2 , и т.д., таким образом, получаем ряд:

$\text{Hit}/\text{host} = 1 + d + d^2 + d^3 + \dots + d^N$, который при $d < 1$ сходится к значению $= 1/(1-d)$

Таким образом, при $d=0.75$ $\text{hit}/\text{host}=4$, при $d=0.85$ $\text{hit}/\text{host}=6.66$, при $d=0.9$ $\text{hit}/\text{host}=10$. Значение $d=0.85$ использовалось как разумное при дальнейших расчетах.

3. Разделение подсистем добровольных и взаимных ссылок

Использовался следующий алгоритм разделения матрицы на подсистемы:

1. По выданной матрице входящих ссылок после отбрасывания висящих страниц строилась матрица исходящих ссылок.
2. По каждой ссылке принималось решение – есть ли у нее парная обратная ссылка:
3. Если обратная ссылка есть – обе ссылки удаляются из матриц входящих и исходящих ссылок и добавляются в матрицу обменных ссылок для соответствующих хостов. Если обратной ссылки нет – ссылка остается в исходной матрице.
4. После прохода по всем ссылкам в исходной матрице мы получаем: в качестве матрицы добровольных (односторонних) ссылок – исходная матрица, в качестве матрицы обменных ссылок – новая созданная матрица (она симметрична). Один хост может содержаться в обеих матрицах.
5. После проведения разделения из каждой матрицы отдельно удаляются висящие ссылки.

Таким образом, мы разделяем ссылки на две подсистемы. В результате исходная матрица содержала 248 тыс. не-висящих хостов, матрица взаимных

ссылок – 190 тыс. не-висящих хостов, матрица односторонних ссылок – 171 тыс. не-висящих хостов.

4. Расчет значений SiteRank

Расчет значений SiteRank по каждой из матриц проводился итерационным методом, стартовое значение – 1. Сходимость оценивалась по норме $|SR-E|$, где SR – вектор SiteRank, E – единичный вектор.

По общей матрице ссылок: 50 итераций, точность $|SR-E|$ до 5 знака

По матрице взаимных ссылок: 69 итераций, точность $|SR-E|$ до 5 знака

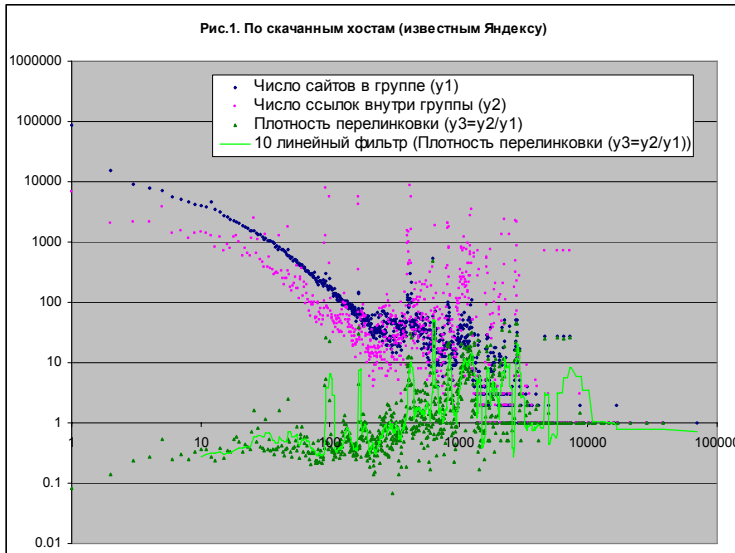
По матрице односторонних ссылок: 28 итераций, точность $|SR-E|$ до 6 знака

Эксперименты

Перед обсуждением результатов необходимо показать наличие явных проблем с существованием линкообменников и ссылочных ферм. Для этого было сделано следующее:

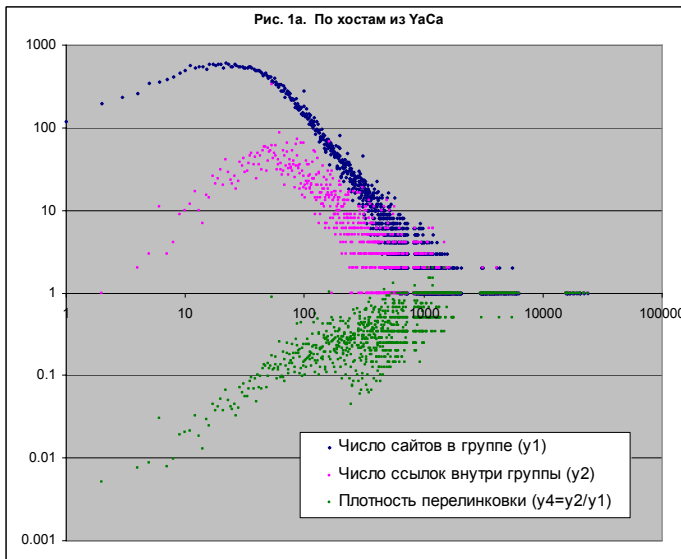
По всем хостам в исходной матрице (* после удаления висящих ссылок) было рассчитано число входящих ссылок. Число входящих ссылок дискретизировалось и все сайты были объединены в группы по числу входящих ссылок. На [рис. 1](#) по горизонтальной оси отложено число входящих ссылок и по вертикальной оси представлены: (1) число сайтов в группе, имеющих это число входящих ссылок, (2) число входящих ссылок с хостов, попавших в эту же группу и (3) отношение 1 и 2.

Видно, что на плотности перелинковки и числе входящих ссылок внутри группы наблюдаются отчетливые пики. Эти пики, вероятно, должны соответствовать ссылочным фермам вида «каждый на каждого». При этом большинство хостов, участвующих в ссылочной ферме, должно попасть в одну группу по числу входящих ссылок и сместить средние параметры.

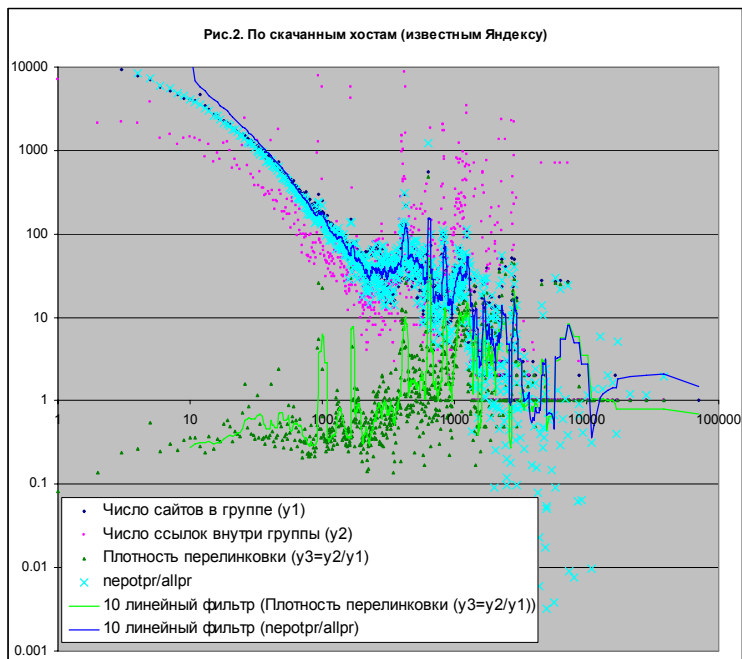


Как говорилось выше, эти дискретные группы приходится выделять вручную и вручную же исключать из индекса или накладывать фильтры.

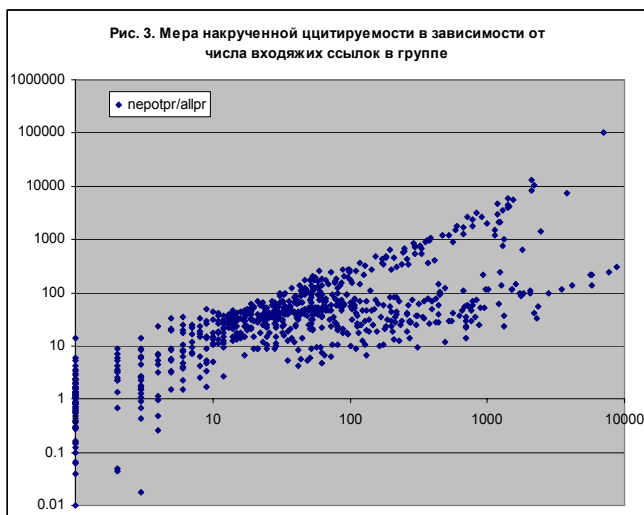
Для сравнения можно привести аналогичный график, построенный по сайтам, зарегистрированным в Яндекс-Каталоге ([рис. 1а](#)), на котором видно отсутствие подобных пиков перелинковки:



После того, как для части хостов были рассчитаны все 3 значения SiteRank (первый, **allpr** – в общей матрице, второй, **nepotpr** – в матрице обменных ссылок, третий, **purepr** – в матрице односторонних ссылок), стало очевидно, что целесообразно в качестве меры «накрученной» ссылочной популярности использовать отношение **nepotpr/allpr**. Ниже представлен рис. 2, на котором изображено и значение суммы **nepotpr/allpr** внутри группы. Видно, что значение этой меры (**nepotpr/allpr**) повышено в тех же группах по числу входящих ссылок, которые являются подозрительными по наличию ссылочных ферм.



Для более наглядного представления посмотрим на рис. 3, на котором представлены величины **nepotpr/allpr**, по горизонтальной оси отложено число входящих ссылок на сайт из группы.



Видно, что есть определенная корреляция между числом подозрительных по накрутке ссылок и выбранной мерой – отношением SiteRank по матрице взаимных ссылок к SiteRank по общей матрице. При этом есть и вторая ветвь диаграммы **nepotpr/allpr**, которая лежит значительно ниже первой. В данном случае есть большое влияние случайных факторов, например, отдельные хосты со слишком низким allpr могут сильно завязать общую сумму по группе, и, возможно, существуют другие, лучшие меры для вычисления «накрученной» цитируемости.

Однако и эта мера хорошо иллюстрирует общий вывод – наличие корреляций между числом ссылок в группах, подозрительных по автоматическому линкообмену и SiteRank, рассчитанному по матрице взаимных ссылок. Однако, SiteRank – количественная мера, для расчета которой не требуется ручное выделение ссылочных ферм.

Выводы и обсуждение результатов

Рассчитанные значения рангов SiteRank по подсистеме взаимных ссылок может использоваться как количественная мера «накрученной ссылочной популярности».

При ранжировании результатов поиска и расчете относительной важности ссылок с различных сайтов ранги, рассчитанные по матрицам односторонних и обратных ссылок можно использовать с разными весами.

Литература

- 1) Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry
The PageRank Citation Ranking: Bringing Order to the Web.
<http://dbpubs.stanford.edu:8090/pub/1999-66>

Optimization of link popularity determination and its application in search results ranking

Trofimenko Evgeny Alexandrovich
info@promosite.ru

Search engine ranking algorithms which use inbound links to documents, may be influenced by creation of artificial links and raising link popularity with links exchange, link farms etc.

To solve this problem some non-algorithmic methods are used: ban of some web sites or whole link farms, filtering of outbound links, which require human moderator to revise all these sites. Additionally, there are lot of intermediate cases between fair and unfair use of links exchange (web rings, small groups of linked thematic web sites), which may cause mistakes.

The suggested idea consists in dividing of common PageRank (SiteRank) into two parts: “fair rank” (including only one-side links) and “exchange rank” (including only links exchange) for further use them in ranking algorithm with different weights. The approach allows to quantitatively determine a rate of artificially raised rank.