**PageRank: начала анализа**

Евгений Трофименко

**Введение в PageRank**

Как известно, количество информации в сети Интернет растет очень быстро, чего нельзя сказать о ее качестве. Пользователь в поисках нужной информации может провести всю жизнь, если только случайно не наткнется на искомый материал; единственный выход для него - воспользоваться поисковиками, которые хранят информацию об адресах и содержимом веб-страниц. Поисковые машины, которые помогают пользователю, пытаются решить проблему - как среди сотен однотипных документов выбрать лучший?

В настоящее время используются текстовые и ссылочные критерии ранжирования страниц при поиске. Первые определяют уместность ("релевантность") документа исходя из наличия слов запроса в тексте и заголовках страницы. Однако, наличие большого количества документов может обесценить изощренные механизмы расчета релевантности, основанные только на содержимом страницы. Это и произошло, когда люди поняли, какую выгоду они получают от целевых посетителей, которых бесплатно предоставляют поисковики. Качество поиска испортилось, количество документов возросло - "релевантный" документ стало очень легко создать.

В целях улучшения качества поиска часть работы по определению "хороших", "важных" документов косвенно возложили на вебмастеров сети. Размещая ссылку на внешний сайт, создатель как бы рекомендует его посетителям своего сайта - именно эту особенность интернета решили использовать для улучшения качества поиска. Повышенная значимость документа определяется, таким образом, с учетом ссылок извне на сайт, содержащий этот документ.

Ссылочные критерии ранжирования помогли несколько исправить положение. Такой критерий достаточно трудно подделать - на это требуется добрая воля других вебмастеров, которые заботятся о качестве своих ресурсов и не будут "продвигать" недостойные сайты. Таким образом, ставка была сделана на саморегуляцию интернета, но новичков такой порядок не устраивал - их просто так никто не пускал в "клуб известных сайтов". И когда новые правила игры были осознаны, поисковики постепенно начали проигрывать.

Однако, как учесть цитируемость ресурса? Ссылки ведь тоже бывают разные. Количество внешних ссылок на сайт не годится для представления цитируемости - с появлением бесплатных хостингов количество ссылок очень легко увеличить. Но важность таких ссылок ничтожна по сравнению со ссылками с известных ресурсов. PageRank и есть такой параметр важности, он выражает цитируемость страницы.

**Что такое PageRank и зачем он нужен?**

Слово PageRank буквально можно перевести как "ранг страницы". Само название определяет алгоритм расчета цитируемости, разработанный и используемый by Sergey Brin & Larry Page, разработчиками поисковой системы Google. Русские аналоги - Взвешенный Индекс Цитирования (ВИЦ у Яндекса), есть аналог и у Апорта, Рамблер планирует ввести учет цитируемости осенью 2002 года. В дальнейшем будем употреблять обозначения цитируемость и PR наравне с PageRank.

Цитируемость -это число, которое рассчитывается для каждой веб-страницы отдельно, и определяется цитируемостью ссылающихся на нее страниц. Своего рода замкнутый круг.

В чем основная идея? Нужно найти жизненный критерий, выражающий важность страницы. В качестве такого критерия была выбрана теоретическая посещаемость страницы. Была построена модель путешествия пользователя по сети путем перехода по ссылкам. При этом есть вероятность того, что посетителю сайт надоест и он закроет броузер и начнет со случайной страницы (допустим, вероятность этого равна 0.15 на каждом шаге). Соответственно, с вероятностью 0.85 он продолжит путешествие, кликнув на одну из доступных на странице ссылок (все ссылки при этом равноправны). Продолжая путешествие до бесконечности, он побывает на цитируемых страницах много раз, а на нецитируемых - меньше.

Таким образом, PageRank веб-страницы был определен как вероятность нахождения пользователя на этой веб-странице; при этом, конечно, сумма вероятностей по всем веб-страницам сети равна единице - где-то он должен обязательно быть!

Из модели следуют три вывода. Во-первых, PageRank нормируется по всем документам сети. Правда, сами величины, в общем-то, относительны, поэтому при расчетах часто нормируют не на единицу по сумме всех страниц, а на единичный усредненный PR (т.е. суммарный по N страницам PageRank равен N, а в среднем - единица). Пугаться этого не следует, просто PR выражен уже не в единицах вероятности, а в относительных единицах.

Во-вторых, PR передается не полностью, есть "затухание". Поэтому длинные цепочки ссылок на сайте малополезны. С человеческой точки зрения то же самое выражает известное правило "трех кликов".

В-третьих, каждая страница изначально имеет ненулевой PR, но очень маленький.

Относитесь с осторожностью к расчетам PageRank, если-

PR рассчитывается для совокупности страниц без учета "внешнего" PR. PageRank - величина, которая не имеет физического смысла в отрыве от Глобальной сети. Точнее, такой PR - это совсем новый PR.

Выявляются закономерности о "сохранении среднего PR" или проводятся нормировки по ограниченному набору страниц. PageRank определен и действует в глобальном масштабе.

Аналогия

Представьте себе озеро (сайт), в которое впадают ручьи и речки (потоки посетителей, пусть "теоретических"). Количество потоков может быть любым, но река приносит много воды, а ручей мало. Поэтому в свое озеро нужно направлять мощные потоки. Какая-то часть воды "уходит в песок", остальное вытекает из вашего озера и впадает в другие озёра. Часть воды испаряется.

В этом смысле рассмотрение распределения PageRank по страницам сайта в отрыве от внешних источников PageRank аналогично переливанию из пустого в порожнее. По внешнему виду сухого русла сложно представить силу потока в реке. Дождь дает очень мало воды - это и есть PageRank сайта, на который никто не ссылается.

Замечания

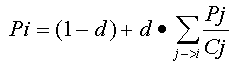
PageRank - не единственный ссылочный критерий ранжирования. Он учитывает только наличие ссылки, но не учитывает текст в ссылке, и текст ссылающегося документа.

Алгоритм "выдавливает" наверх в поиске те документы, которые и без поисковика наиболее популярны. Однако введение такого алгоритма при поиске существенно ужесточает конкуренцию, если это поисковик масштаба Google.

**Расчет PageRank**

Итак, будем рассматривать PageRank страницы как вероятность попадания пользователя на страницу, выраженную в относительных единицах.

PageRank (Pi) страницы i выражается как {1}



где: d -т.н. "damping factor", параметр затухания. Принимается равным 0.85-0.9. Выражает вероятность того, что пользователь, зашедший на страницу, будет продолжать путешествие и переходить по ссылкам. Pi - PageRank интересующей нас страницы i j - обозначение страниц, на которых есть ссылки на i-ю Pj - PageRank страницы j, ссылающейся на i-ю. Сj - Число ссылок на странице j. 1/Сj - Вероятность того, что пользователь, находящийся на странице j, из Сj доступных ему ссылок выберет именно ссылку на нашу страницу i. d\*Pj/Сj - поток "теоретической посещаемости", который дойдет до страницы i со страницы j. Суммирование идет по всем страницам, ссылающимся на i-ю. (1-d) - минимальный PageRank страницы. Он не равен нулю за счет того, что пользователь регулярно выбирает новый сайт в качестве стартовой точки.

Однако, на PageRank наложено ограничение:



где N - общее количество веб-страниц в Интернет.

Т.е., средний PageRank равен единице. Ограничение это следует из нормировки вероятности пребывания пользователя по всей сети - сумма вероятностей по всем страницам равна единице. Таким образом, Вероятностьi=PageRanki/число страниц в сети

Отметим, что значение PageRank, равное единице, только кажется большим. Количество страниц в сети (N) очень велико, и вероятность 1/N - чрезвычайно мала.

Решая систему уравнений, можно найти PageRank всех страниц в Интернет. Расчет можно вести разными методами:

Итерационный метод

Матричный метод

Функциональный метод

Итерационный метод расчета PageRank

Метод наиболее часто используется. Он состоит в численном решении системы уравнений:

Выбираем геометрию сайта, расстановку ссылок, систему уравнений

Задаемся начальными значениями PageRank для каждой страницы. Они могут быть любыми.

Рассчитываем новый набор значений PageRank по уравнению (1) исходя из имеющегося набора значений

Рассчитываем средний PageRank по всему набору страниц, и делим PR каждой страницы на полученную величину. В результате средний PR становится равным единице.

Если набор значений PageRank изменился по сравнению с исходным набором шага 3, возвращаемся к шагу 3. Если нет, то расчет заканчиваем.

При исследовании влияния геометрии сайта на распределение PageRank удобно представить структуру ссылок в виде матрицы:

|  |  |
| --- | --- |
| 0-ссылки нет 1-ссылка есть | На какую страницу указывает ссылка |
| На какой странице находится ссылка | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | 1 | 2 | 3 | 4 | | 1 | 0 | 1 | 0 | 0 | | 2 | 0 | 0 | 1 | 0 | | 3 | 0 | 0 | 0 | 1 | | 4 | 1 | 0 | 0 | 0 | |

В таблице выше представлен сайт из четырех страниц, на котором ссылки замкнуты в "кольцо". Страница 1 ссылается на 2 (1- есть ссылка, 0-ссылки нет), 2 на 3, 3 на 4, 4 обратно на 1. Представление структуры сайта в таком виде удобно, в частности для расчетов.

Для того, чтобы поэкспериментировать с различными структурами сайтов, можно скачать заготовки в MS Excel для 10 страниц (30 итераций) и 30 страниц (90 итераций). Распределение PageRank по страницам рассчитывается сразу и представлено в желтой строке.

**Матричный метод расчета PageRank**

По уравнению 1:

Нижеприведенную "матрицу связей" можно умножить на вектор значений PageRank m-го шага итерации, полученный вектор умножить на d , прибавить единичный вектор, умноженный на (1-d) и получить следующее приближение вектора PageRank с номером m+1, который нужно пронормировать (чтобы сумма проекций вектора PR была равна N). При навыках работы с математическими программами (например, Mathcad) этот способ может быть более удобным.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| 1 | 0 | 1/3 | 1/3 | 1/3 |
| 2 | 0 | 0 | 1/2 | 1/2 |
| 3 | 0 | 0 | 0 | 1 |
| 4 | 1 | 0 | 0 | 0 |

Здесь страница 1 ссылается на 2, 3, 4; страница 2 - на 3 и 4; страница 3 на 4, а 4 на 1. Представленная матрица содержит значения Mij=1/Cj->i, т.е. значение в каждой ячейке разделено на общее количество ссылок Cj на странице j.

Недостатки численных и итерационных методов

Фактически, оба приведенные выше метода являются разными формулировками итерационного метода расчета значений PageRank. Они требуют работы с конкретными численными значениями PageRank. Методы использованы для расчетов в работах [3,4].

Однако, рассмотрим реальную ситуацию. Для того, чтобы воплотить в жизнь свои знания о распределении PageRank, необходима индексация ваших страниц. В случае Google, ваш сайт не будет проиндексирован (либо придется ждать индексации очень долго) до достижения некоего порогового значения PageRank. В любом случае, на ваш сайт должны существовать ссылки, хотя бы одна. Это значит, что ваш сайт не оторван от "внешнего мира", и существует ненулевой "входящий PageRank", направленнный извне на ваш сайт.

Из этого рассуждения следует, что:

Расчеты PR "в отрыве" от окружения сайта неточны для каждой страницы вашего сайта - они проделаны для нулевого входящего PageRank

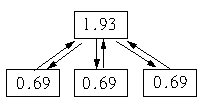
Правило нормировки не работает в пределах вашего сайта (но работает в пределах глобального набора проиндексированных страниц, т.е. в рамках Интернет по версии Google)

Никакой численный расчет не может применяться в динамике - ведь входящий PageRank изменяется по мере раскрутки сайта (если вы дочитали до этого места, вероятно, раскруткой своего сайта будете заниматься так же упорно). Соответственно, меняется во времени PR каждой страницы.

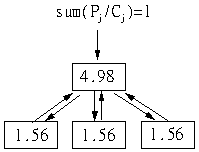
Стоит помнить о том, что по своей сути PageRank - это поток (поток теоретической посещаемости). Соответственно, расматривая свой сайт как "маленькую вселенную", вы не учитываете потоки извне. Если применить аналогию, такой сайт похож на высохшее озеро, на дне которого осталось несколько луж, и вы рассчитываете, в какой из них будет больше воды.

Посмотрим, что происходит при увеличении входящего PageRank.

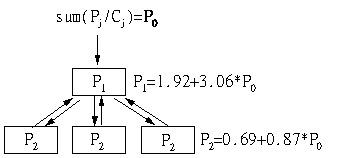
Вот простейший сайт из четырех страниц, ссылок извне нет-



А здесь входящий PageRank равен единице-



Но нам скоро станет лень рассчитывать PageRank при каждом "воображаемом" изменении внешнего PageRank (P0). Поэтому рассмотрим общий случай и выразим PR страниц как функции от P0-

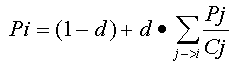


В дальнейшем будем рассчитывать PageRank страниц как функции от входящего PR. Это позволит выделить ту компоненту PageRank, которая увеличивается по мере раскрутки, и отделить "остатки" в виде констант, величина которых порядка единицы. А солипсистскими методами расчета пользоваться на будем - мы ведь не одни в Интернете...

**Функциональный метод расчета PageRank**

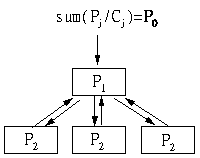
Задача: рассчитать стабильные значения PageRank, не применяя итерационных методов. Рассмотрим уравнение (1) внимательнее - в нем нет никаких особенностей, которые требуют применения итераций. Наоборот, PR каждой страницы определяется как функция PR других страниц. Предположим, что мы достигли стационарного состояния, и PageRank страниц не меняется. Остается только записать уравнения для PR каждой из страниц и решить систему.

{1}



Итак, будем рассчитывать PageRank страниц сайта как функцию от внешнего, "входящего" PageRank. Для этого нужны: уравнение (1) и представление об эквивалентности страниц одного типа. Пример-

На сайте, который приведен ниже, 3 нижних страницы эквивалентны между собой во всех смыслах. Соответственно, все они будут иметь одинаковый PageRank (P2). Головная страница отличается от них и имеет PR=P1.



Запишем уравнения для страниц вида 1 и вида 2:

P1=0.15+0.85\*(P0+3P2) - на страницу вида 1 ссылаются 3 страницы вида 2, на каждой из которых есть одна ссылка.

P2=0.15+0.85\*(P1/3) - на страницу вида 2 ссылается страница вида 1, на которой есть 3 ссылки.

Решая эту систему, получаем-

P1=0.15\*(1+3\*0.85)/(1-0.85^2)+0.85/(1-0.85^2)\*P0=1.92+3.06\*P0 P2=0.69+0.87\*P0

Этим методом хотя и сложнее пользоваться, но он обладает одним хорошим качеством, которого нет у итерационных методов - общностью.

Откуда берется PageRank?

Поль Дирак выдвигал предположение, что существует оптимальное расстояние, с которого лучше всего наблюдать женское лицо. Действительно: на нулевом расстоянии, равно как и на бесконечном, удовольствие от созерцания стремится к нулю. В то же время, на промежуточном расстоянии оно явно не нулевое. Значит, между нулевым и бесконечным расстоянием существует максимум функции Удовольствие=f(Расстояние)

Давайте рассмотрим, хотя бы с помощью итерационного метода, два граничных случая связности сайтов.

Случай 1: Все страницы в Интернете замкнуты в "кольцо" - на каждой есть только одна ссылка на соседа, и только одна входящая ссылка. Результат: PageRank равен единице для всех страниц.

Случай 2: Все страницы в Интернете перелинкованы друг с другом - на каждой из N страниц есть ссылки на всех N-1 соседей, и столько же входящих ссылок (N-1). Результат: PageRank равен единице для всех страниц.

Откуда же берется большой PageRank?

Ответ: из неоднородности распределения ссылок по страницам. Дело в том, что все страницы сети были эквивалентны, что привело к одинаковому значению PageRank. Но если в однородном Интернете две страницы "обменяются ссылками", их PageRank увеличится. А у всего остального Интернета - чуть-чуть, но уменьшится. Таким образом, те, кто обмениваются ссылками, "стягивают одеяло на себя".

Надо сказать, что приведенный выше функциональный метод чуть-чуть неточен. Дело в том, что он не учитывает изменения среднего PageRank сети при появлении рассмотренного сайта. На сайте средний PageRank не равен единице, в отличие от Интернета, поэтому после проведенного расчета нужно пересчитать PR всех страниц в сети:

PRinew=PRiold\*(Средний PR в интернете без вашего сайта)/(Средний PR в интернете, включая ваш сайт)

Но, поскольку суммарный PR по Интернету никто не знает, делать этого мы не будем. В любом случае эти изменения мизерные, но именно они и являются тем самым "стягиванием одеяла на себя".

**Промежуточные выводы**

Мало смысла в расчете PageRank страниц без учета "внешнего" PageRank

Нормировка PageRank на единицу работает только в глобальном масштабе, но не в пределах одного сайта

Значения PageRank порядка единицы очень малы и неинтересны для анализа. Основной интерес представляет передача потока PageRank от одной страницы к другой

Продолжение, в котором рассмотрены частные случаи и различные случаи иерархии страниц сайта - PageRank: анализ потоков.

**Список литературы**

Larry Page PageRank: Bringing Order to the Web

Александр Садовский Растолкованный PageRank, перевод старого варианта статьи [4]

Ian Rogers The Google Page Rank Algorithm and How It Works Обзор ошибок старого варианта статьи [4]

Chris Ridings PageRank Explained, новый вариант статьи

Артем Шкондин PageRank: Больше ссылок хороших и важных