**PageRank: анализ потоков**

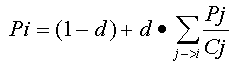
Евгений Трофименко

В первой части статьи было установлено, что итерационные методы не имеет смысла применять для расчетов PageRank, учитывающих окружение сайта и "входящий" PR. Поэтому мы будем рассчитывать PageRank страниц не в численном виде, а виде функций от входящего PR. Это позволит выделить ту компоненту PageRank, которая увеличивается по мере раскрутки, и отделить "остатки" в виде констант, величина которых порядка единицы.

Повторение: функциональный метод расчета PageRank

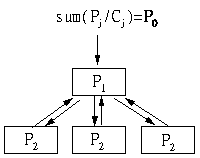
Задача: рассчитать стабильные значения PageRank, не применяя итерационных методов. Рассмотрим уравнение (1) внимательнее - в нем нет никаких особенностей, которые требуют применения итераций. Наоборот, PR каждой страницы определяется как функция PR других страниц. Предположим, что мы достигли стационарного 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.

Различные случаи: два типа страниц

Итак, начнем рассмотрение самого простого случая - сайт состоит из одной головной страницы и некоторого количества подчиненных страниц. Ссылки извне направлены на головную страницу.

Случай 1: "метла"

С головной страницы (PageRank=P1) есть ссылки на N эквивалентных подчиненных страниц (PageRank=P2). Подчиненные страницы не связаны между собой, на каждой из них есть одна ссылка на головную страницу.

|  |  |
| --- | --- |
|  | Система уравнений: (N>=1)  P1=0.15+0.85\*(P0+N\*P2) - на страницу вида 1 ссылаются N страниц вида 2, на каждой из которых 1 ссылка  P2=0.15+0.85\*(P1/N) - на страницу вида 2 ссылается одна страница вида 1, на ней N ссылок |

Решая систему, находим зависимости P1, 2 от P0, N.

Решение системы: P1=0.15/(1-0.85^2)+0.85\*0.15/(1-0.85^2)\*N+0.85^2/(1-0.85^2)\*P0 =>

P1=0.541+0.459\*N+3.063\*P0 P2=0.541+0.459/N+2.604\*P0/N

Итак, мы видим, что:

PageRank главной страницы (P1) увеличился не на P0, но на 3P0. Т.е, мы получаем реальный выигрыш за счет взаимного влияния страниц

P2 обратно пропорционален числу страниц. Очевидный вывод.

P1 прямо пропорционален числу страниц. Следовательно, увеличивая количество страниц на сайте, можно сконцентрировать большой PageRank на главной странице. При этом P2 по мере увеличения количества страниц стремится к 0.541, а не к 1; и отдает часть PageRank главной странице

Средний PR

Рассчитаем суммарное значение PageRank по нашему сайту-

PR=P1+N\*P2=0.541+N+5.667\*P0

Это показывает, что:

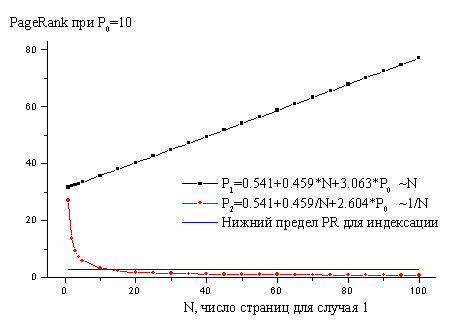
Средний PageRank при большом количестве страниц близок к 1.

PageRank на сайте увеличился не на P0, а на 5.667=0.85/0.15 P0. Это произошло из-за существования обратной связи между главной и второстепенными страницами, т.е., волна PR как бы "откатилась обратно".

Однако, при обращении в ноль P0=0 средний PR не становится равным 1. Это происходит из-за того, что мы отказались от условия нормировки, но при "отрыве" сайта от внешней среды должны его снова применить. Отказ от нормировки позволяет сохранить связь с внешним PageRank, при этом мы помним о необходимости перенормировки из-за "стягивания одеяла на себя".

Можно ли получить большой PageRank только за счет массы сайта?

Судя по формуле, можно получить любое значение PageRank главной страницы только за счет увеличения количества страниц, ведь P1 линейно увеличивается при увеличении N. При этом все страницы сайта должны быть проиндексированы Google. Однако, вспомним о том, что существует минимальное значение PR, достичь которого необходимо для индексации документа Google. Но ведь PageRank второстепенных страниц (P2) уменьшается с N! Посмотрим на график:



Если число страниц на сайте окажется слишком большим, то второстепенные страницы просто не проиндексируются! И никакого сверхвысокого PR достигнуть не удастся. Можно попытаться:

Постепенно добавлять страницы Поставив несколько ссылок на другой странице с хорошим PR, дождаться индексации части страниц. Затем добавить еще несколько страниц. И так далее. Но: на полное добавление потребуется очень много времени!

Увеличить внешний PR Как видно, P2~P0/N. Т.е., увеличивая P0, можно добиться индексации второстепенных страниц. Но чем больше число страниц, тем больше придется работать на увеличение внешнего PR! На это тоже требуется много времени!

Использовать оба метода сразу

Ставить ссылки с второстепенных страниц друг на друга. Это мы рассмотрим далее: случай 2.

Итак, несмотря на явную зависимость P1 от N, нет возможности быстро увеличить свой PageRank - а это означает относительную устойчивость в распределении сил. К тому же, если сайт действительно очень хороший, и на него ссылаются "добровольно" - он всегда обгонит "плохой сайт" (за счет более полной индексации страниц в том числе). Алгоритм PageRank как бы подталкивает к постепенному развитию сайта и постепенному добавлению страниц.

Случай 2: "кольцо на палочке"

Добавим к предыдущему случаю одну ссылку на "соседа":

С головной страницы (PageRank=P1) есть ссылки на N эквивалентных подчиненных страниц (PageRank=P2). Подчиненные страницы связаны между собой: есть одна ссылка на соседнюю подчиненную страницу, на каждой из них есть одна ссылка на головную страницу.

|  |  |
| --- | --- |
|  | Система уравнений: (N>=2)  P1=0.15+0.85\*(P0+N\*P2/2) - на страницу вида 1 ссылаются N страниц вида 2, на каждой из которых 2 ссылки  P2=0.15+0.85\*(P1/N+P2/2) - на страницу вида 2 ссылается одна страница вида 1, на которй N ссылок; и одна страница вида 2, на которой 2 ссылки |

Решение системы: P1=0.403+0.298\*N+2.287\*P0 P2=0.702+0.596/N+3.380\*P0/N

Видно, что ситуация примерно такая же, как и в случае 1, однако теперь PageRank второстепенных страниц увеличился, а главной страницы - уменьшился. Это касается как зависимости от N, так и зависимости от P0.

В данном случае мы немного выигрываем в PR второстепенных страниц - это поможет их проиндексировать быстрее, но снизит PageRank главной страницы, ради которой все и затевалось. А что будет в предельном случае - когда все страницы плотно связаны?

Случай 3: "камушек"

С головной страницы (PageRank=P1) есть ссылки на N эквивалентных подчиненных страниц (PageRank=P2). Подчиненные страницы попарно связаны между собой: на каждой из них есть ссылки на всех N-1 соседей. На каждой из них есть одна ссылка на головную страницу.

|  |  |
| --- | --- |
|  | Система уравнений:  P1=0.15+0.85\*(P0+N\*P2/N) - на страницу вида 1 ссылаются N страниц вида 2, на каждой из которых N ссылок (N-1 на соседей и одна на страницу 1)  P2=0.15+0.85\*(P1/N+(N-1)\*P2/N) - на страницу вида 2 ссылается одна страница вида 1, на которй N ссылок; и N-1 страниц вида 2, на каждой из которых N ссылок |

Решение системы: P1=(0.85\*P0+1)/(1-(0.85^2/(0.15N+0.85)))-0.85/(0.15N+0.85-0.85^2)

Для простоты рассмотрим предельные значения P1, 2 при стремлении N к бесконечности: P1->1+0.85\*P0-x/N P2->1+y\*P0/N, где x, y порядка единицы.

Видно, что фактически, внешний P0 влияет только на главную страницу сайта, но слабее, чем в предыдущих случаях. Зато увеличивается до единицы PageRank втростепенных страниц.

Случай 4 (общий)

С головной страницы (PageRank=P1) есть ссылки на N эквивалентных подчиненных страниц (PageRank=P2). Подчиненные страницы связаны между собой: на каждой из них есть ссылки на M соседних подчиненных страниц (M<=N). На каждой из них есть одна ссылка на головную страницу.

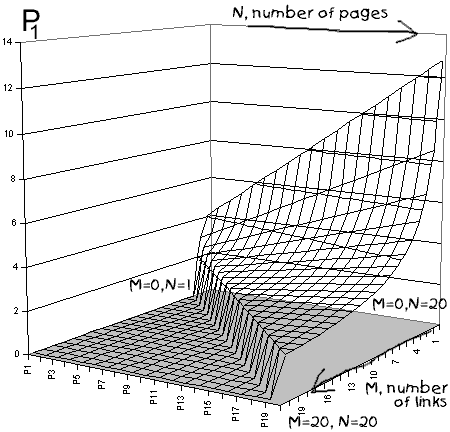
|  |  |
| --- | --- |
|  | Система уравнений: (M<=N)  P1=0.15+0.85\*(P0+N\*P2/(M+1)) - на страницу вида 1 ссылаются N страниц вида 2, на каждой из которых M+1 ссылок (M на соседей и одна на страницу 1)  P2=0.15+0.85\*(P1/N+M\*P2/(M+1)) - на страницу вида 2 ссылается одна страница вида 1, на которй N ссылок; и M страниц вида 2, на каждой из которых M+1 ссылок |

Решение системы: P1=(0.15+0.85P0)/(1-(0.85^2/(0.15M+1)))+0.15\*0.85\*N/(0.15M+1-0.85^2)

Из этой зависимости видно, что при увеличении числа ссылок (M) между второстепенными страницами PageRank главной страницы все слабее зависит от N и от P0. Это происходит из-за передачи большой части PageRank в область второстепенных страниц, откуда "возвращается обратно" малая его часть.

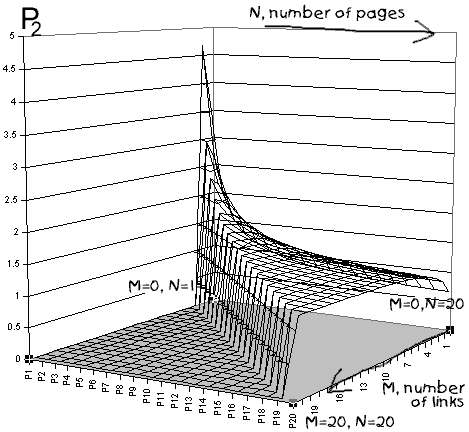
Для наглядности рассмотрим зависимости PagaRank от числа страниц (N) и числа ссылок на соседние второстепенные страницы (M) при входящем P0=1 (скачать графики в Excel):

PageRank главной страницы (P1)



Видно, что самый быстрый рост P1 происходит при увеличении числа страниц происходит при отсутствии ссылок между второстепенными страницами (M=0). При наличии нескольких ссылок возрастание P1 становится медленным. Если рассматривать изменение P1 при фиксированном количестве страниц N (например, N=20), видно, что при увеличении числа ссылок M происходит резкое уменьшение PageRank главной страницы.

PageRank второстепенных страниц (P2)



Второстепенные страницы имеют высокий PageRank только при небольшом их числе. При отсутствии ссылок на соседей (M=0) P2 уменьшается несколько быстрее, чем в случае M=N, и достигает меньших предельных значений. При фиксированном числе страниц (N=20) и увеличении числа ссылок (M) P2 немного возрастает.

**Промежуточные выводы**

Подходы к развитию сайта для полной индексации Google и выигрыша в PageRank главной страницы:

Постепенно добавлять страницы

Увеличивать внешний PageRank

Поставить ссылки с второстепенных страниц друг на друга, а после индексации убрать их.

Замечание. Многие поисковики учитывают PageRank-подобные критерии при ранжировании, но не обязательно существование нижней границы PageRank для индексации.