Google and the Page Rank Algorithm

Overview

- What is PageRank™ ?
- The Random Surfer Model
- Characteristics of PageRank™
- Computation of PageRank™
- Implementation of PageRank in the Google Search Engine
- The effect of different factors on the PageRank™ values
- Tips for raising your website's PageRank™ value.

PageRank™ - Introduction

- The heart of Google's searching software is PageRank™, a system for ranking web pages developed by Larry Page and Sergey Brin at Stanford University

PageRank™ - Introduction

- Essentially, Google interprets a link from page A to page B as a vote, by page A, for page B.
- BUT these votes don't weigh the same, because Google also analyzes the page that casts the vote.

The original PageRank™ algorithm

$$PR(A) = (1\text{-}d) + d\ (PR(T1)/C(T1) + \ldots + +PR(Tn)/C(Tn))$$

where:

- PR(A) is the PageRank of page A,
- PR(Ti) is the PageRank of pages Ti which link to page A,
- C(Ti) is the number of outbound links on page Ti
- d is a damping factor which can be set between 0 and 1.

PageRank™ algorithm

- It's obvious that the PageRank™ algorithm does not rank the whole website, but it's determined for each page individually. Furthermore, the PageRank™ of page A is recursively defined by the PageRank™ of those pages which link to page A

$PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)$

- The PageRank™ of pages Ti which link to page A does not influence the PageRank™ of page A uniformly.
- The PageRank™ of a page T is always weighted by the number of outbound links C(T) on page T.
- Which means that the more outbound links a page T has, the less will page A benefit from a link to it on page T.

$PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)$

- The weighted PageRank™ of pages Ti is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's PageRank™.

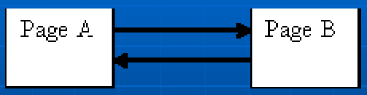
$PR(A) = (1-d) + d * (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$

- After all, the sum of the weighted PageRanks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1. Thereby, the extend of PageRank benefit for a page by another page linking to it is reduced.

How is it Calculated?

- The PR of each page depends on the PR of the pages pointing to it.
- But we won't know what PR those pages have until the pages pointing to them have their PR calculated and so on.
- So what we do is make a guess.

Simple Example



- Each page has one outgoing link. So that means C(A) = 1 and C(B) = 1.

We don't know what their PR should be to begin with, so we will just guess 1 as a safe random number.

- d (damping factor) = 0.85
- PR(A)= (1 – d) + d(PR(B)/1)
- PR(B)= (1 – d) + d(PR(A)/1)

i.e.

- PR(A)= 0.15 + 0.85 * 1 = 1
- PR(B)= 0.15 + 0.85 * 1 = 1

Let's Do It Again with Another Number. Let's try 0 and re-calculate…

- PR(A)= 0.15 + 0.85 * 0
  = 0.15
  = 0.15 + 0.85 *
- PR(B) 0.15
  = 0.2775
- Now we have calculated a "next best guess" so we just plug it in the equation again…

- PR(A)= 0.15 + 0.85 * 0.2775
  = 0.385875
- PR(B)= 0.15 + 0.85 * 0.385875
  = 0.47799375

And again…

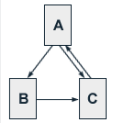
- PR(A)= 0.15 + 0.85 * 0.47799375
  = 0.5562946875
- PR(B)= 0.15 + 0.85 * 0.5562946875
  = 0.622850484375

Principle

- It doesn't matter where you start your guess, once the PageRank calculations have settled down (converged, or a fix-point), the "normalized probability distribution" (the average PageRank for all pages) will be 1.0

Another Example

- We regard a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. The damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5.



PR(A) = 0.5 + 0.5 PR(C)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))
We get the following PageRank™ values for the single pages:
PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

The sum of all pages' PageRanks is 3 and thus equals the total number of web pages.

The Iterative Computation of PageRank

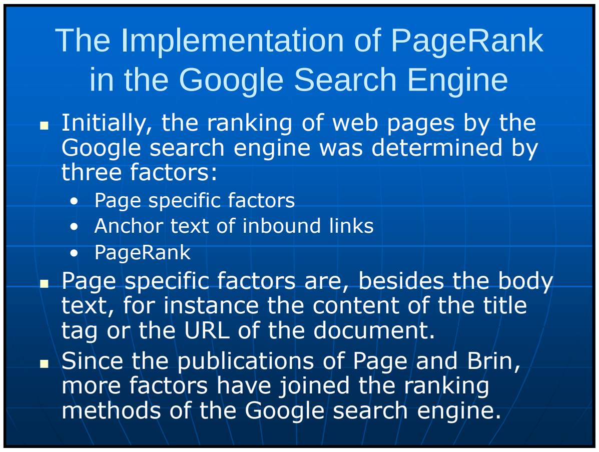
- For the simple three-page example it is easy to solve the according equation system to determine PageRank values. In practice, the web consists of billions of documents and it is not possible to find a solution by inspection.
- Because of the size of the actual web, the Google search engine uses an approximative, iterative computation of PageRank values. This means that each page is assigned an initial starting value and the PageRanks of all pages are then calculated in several computation circles based on the equations determined by the PageRank algorithm. The iterative calculation shall again be illustrated by the three-page example, whereby each page is assigned a starting PageRank value of 1.

The Iterative Computation of PageRank (example)

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

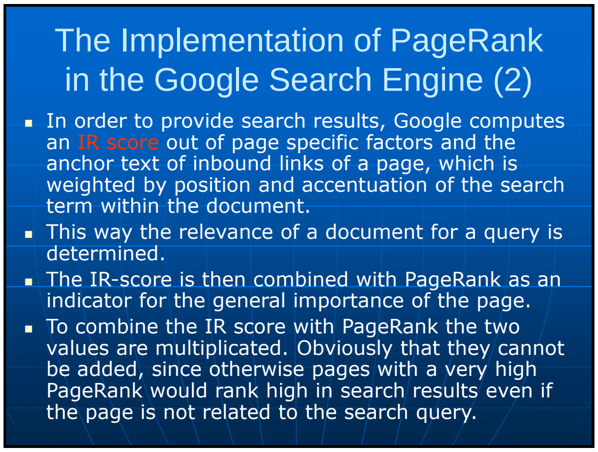
The Iterative Computation of PageRank™ (3)

- We get a good approximation of the real PageRank values after only a few iterations. According to publications of Lawrence Page and Sergey Brin, about 100 iterations are necessary to get a good approximation of the PageRank values of the whole web.
- The sum of all pages' PageRanks still converges to the total number of web pages. So the average PageRank of a web page is 1.



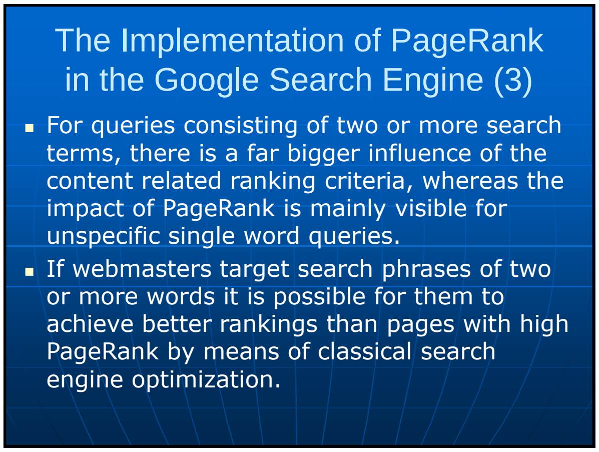
The Implementation of PageRank in the Google Search Engine

- Initially, the ranking of web pages by the Google search engine was determined by three factors:
  - Page specific factors
  - Anchor text of inbound links
  - PageRank
- Page specific factors are, besides the body text, for instance the content of the title tag or the URL of the document.
- Since the publications of Page and Brin, more factors have joined the ranking methods of the Google search engine.



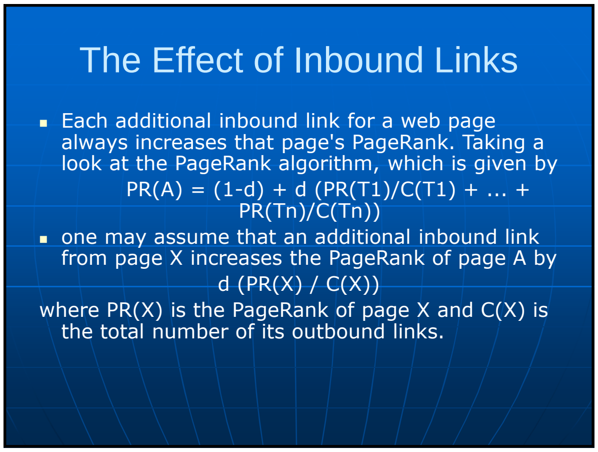
The Implementation of PageRank in the Google Search Engine (2)

- In order to provide search results, Google computes an IR score out of page specific factors and the anchor text of inbound links of a page, which is weighted by position and accentuation of the search term within the document.
- This way the relevance of a document for a query is determined.
- The IR-score is then combined with PageRank as an indicator for the general importance of the page.
- To combine the IR score with PageRank the two values are multiplicated. Obviously that they cannot be added, since otherwise pages with a very high PageRank would rank high in search results even if the page is not related to the search query.



The Implementation of PageRank in the Google Search Engine (3)

- For queries consisting of two or more search terms, there is a far bigger influence of the content related ranking criteria, whereas the impact of PageRank is mainly visible for unspecific single word queries.
- If webmasters target search phrases of two or more words it is possible for them to achieve better rankings than pages with high PageRank by means of classical search engine optimization.



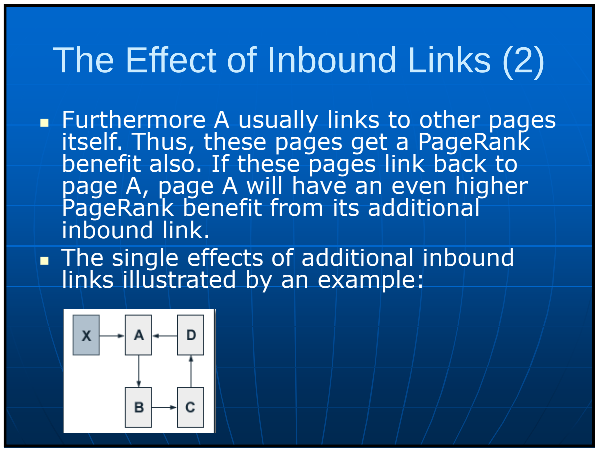
The Effect of Inbound Links

- Each additional inbound link for a web page always increases that page's PageRank. Taking a look at the PageRank algorithm, which is given by

  $PR(A) = (1-d) + d\ (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$
- one may assume that an additional inbound link from page X increases the PageRank of page A by

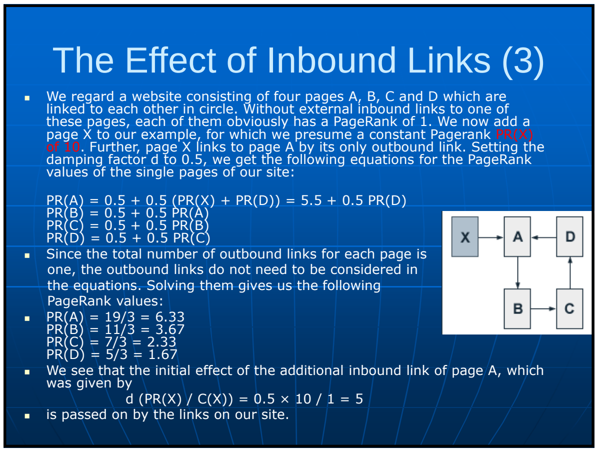
  $d\ (PR(X)\ /\ C(X))$

where PR(X) is the PageRank of page X and C(X) is the total number of its outbound links.



The Effect of Inbound Links (2)

- Furthermore A usually links to other pages itself. Thus, these pages get a PageRank benefit also. If these pages link back to page A, page A will have an even higher PageRank benefit from its additional inbound link.
- The single effects of additional inbound links illustrated by an example:



The Effect of Inbound Links (3)

- We regard a website consisting of four pages A, B, C and D which are linked to each other in circle. Without external inbound links to one of these pages, each of them obviously has a PageRank of 1. We now add a page X to our example, for which we presume a constant Pagerank PR(X) of 10. Further, page X links to page A by its only outbound link. Setting the damping factor d to 0.5, we get the following equations for the PageRank values of the single pages of our site:

  $PR(A) = 0.5 + 0.5\ (PR(X) + PR(D)) = 5.5 + 0.5\ PR(D)$
  $PR(B) = 0.5 + 0.5\ PR(A)$
  $PR(C) = 0.5 + 0.5\ PR(B)$
  $PR(D) = 0.5 + 0.5\ PR(C)$
- Since the total number of outbound links for each page is one, the outbound links do not need to be considered in the equations. Solving them gives us the following PageRank values:
- $PR(A) = 19/3 = 6.33$
  $PR(B) = 11/3 = 3.67$
  $PR(C) = 7/3 = 2.33$
  $PR(D) = 5/3 = 1.67$
- We see that the initial effect of the additional inbound link of page A, which was given by

  $d\ (PR(X)\ /\ C(X)) = 0.5 \times 10\ /\ 1 = 5$
- is passed on by the links on our site.

Influence of the Damping Factor

- The degree of PageRank propagation from one page to another by a link is primarily determined by the damping factor d. If we set d to 0.75 we get the following equations for our above example:

  PR(A) = 0.25 + 0.75 (PR(X) + PR(D)) = 7.75 + 0.75 PR(D)
  PR(B) = 0.25 + 0.75 PR(A)
  PR(C) = 0.25 + 0.75 PR(B)
  PR(D) = 0.25 + 0.75 PR(C)
- The results gives us the following PageRank values:

  PR(A) = 419/35 = 11.97
  PR(B) = 323/35 = 9.23
  PR(C) = 251/35 = 7.17
  PR(D) = 197/35 = 5.63
- First of all, we see that there is a significantly higher initial effect of additional inbound link for page A which is given by

  d × PR(X) / C(X) = 0.75 × 10 / 1 = 7.5

Influence of the Damping Factor (2)

- The PageRank of page A is almost twice as high at a damping factor of 0.75 than it is at a damping factor of 0.5.
- At a damping factor of 0.5 the PageRank of page A is almost four times superior to the PageRank of page D, while at a damping factor of 0.75 it is only a little more than twice as high.
- So, the higher the damping factor, the larger is the effect of an additional inbound link for the PageRank of the page that receives the link and the more evenly distributes PageRank over the other pages of a site.

A Different Notation of the PageRank Algorithm

PR(A) = (1-d) / N + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

- where N is the total number of all pages on the web.
- This version's PageRank™ of a page is the actual probability for a random surfer reaching that page after clicking on many links. The PageRanks then form a probability distribution over web pages, so the sum of all pages' PageRanks will be one.

Tips for raising your website's PageRank™ value

- Add new pages to your website (as many as you can)
- Swap links with websites which have high PageRank™ value
- Raise the number of inbound links (Advertise your website on other sites)

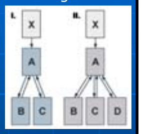
Tips for raising your website's PageRank™ value (2)

- As you can see the sum of PageRanks is the same, unless you create new web pages
- An isolated website can be considered as a mini web, so if you want to raise its PageRank you need to add new pages to it or you can make links to it from outside pages
- The pages that refer to our (isolated) website are like the newly created pages from our point of view

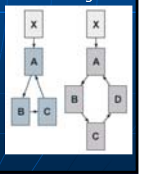
Add new pages to your website

- When you add a new page to your site, be sure to link it to your front page and vice versa as it is shown on the picture
- If you want to reduce your front page's PageRank, then you can make circular references as you see on the second picture ☺

Right



Wrong



The effect of additional pages

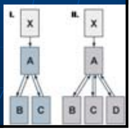
Sub-pages	PageRank of the front page
1	1.000000
2	1.428673
3	1.857347
4	2.286020
5	2.714694
10	4.858060
20	9.144795
50	22.005003
100	43.438648
250	107.739838
500	214.907135
700	300.642426
1000	429.246613

The effect of additional pages

- As you can see:

PageRank ≈ 1+0.428*NumberOfPages

- So, if you add a web page to your website it will increase your page’s rank by ≈0.428. Of course you need to do as it is shown on the picture



Advantages of the Inbound Links

- The PageRank of your page is calculated by adding the PageRanks of the other pages which link to your page, so it doesn’t matter if the ranks are low, because they will still raise your page’s PageRank

References

- http://pr.efactory.de
- http://www.google.com/technology/