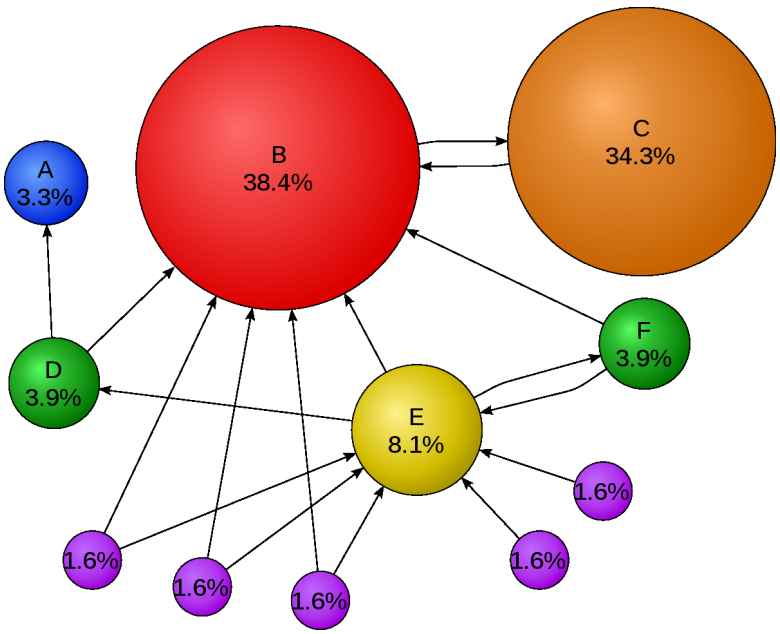
PageRank

PageRank is an algorithm used by the Google web search engine to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

> PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.
>
> —Facts about Google and Competition [1]

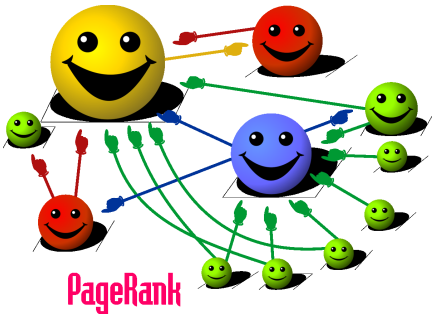
It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the most well-known. Google uses an automated web spider called Googlebot to actually count links and gather other information on web pages.



Mathematical **PageRanks** for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.

Description

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element *E* is referred to as the *PageRank of E* and denoted by $PR(E)$.



Cartoon illustrating basic principle of PageRank. The size of each face is proportional to the total size of the other faces which are pointing to it.

A PageRank results from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself. If there

are no links to a web page, then there is no support for that page. The value of incoming links is known as "Google juice", "link juice" or "Pagerank juice".[*citation needed*]

Numerous academic papers concerning PageRank have been published since Page and Brin's original paper. In practice, the PageRank concept may be vulnerable to manipulation. Research has been conducted into identifying falsely influenced PageRank rankings. The goal is to find an effective means of ignoring links from documents with falsely influenced PageRank.Wikipedia:No original research

Other link-based ranking algorithms for Web pages include the HITS algorithm invented by Jon Kleinberg (used by Teoma and now Ask.com)[*citation needed*], the IBM CLEVER project, and the TrustRank algorithm.

History

The idea of formulating a link analysis problem as a eigenvalue problem was probably first suggested in 1976 by Gabriel Pinski and Francis Narin, who worked on scientometrics ranking scientific journals. PageRank was developed at Stanford University by Larry Page and Sergey Brin in 1996 as part of a research project about a new kind of search engine.[2] Sergey Brin had the idea that information on the web could be ordered in a hierarchy by "link popularity": a page is ranked higher as there are more links to it.[3] It was co-authored by Rajeev Motwani and Terry Winograd. The first paper about the project, describing PageRank and the initial prototype of the Google search engine, was published in 1998: shortly after, Page and Brin founded Google Inc., the company behind the Google search engine. While just one of many factors that determine the ranking of Google search results, PageRank continues to provide the basis for all of Google's web search tools.

The name "PageRank" plays off of the name of developer Larry Page, as well as the concept of a web page. The word is a trademark of Google, and the PageRank process has been patented (U.S. Patent 6,285,999 [4]). However, the patent is assigned to Stanford University and not to Google. Google has exclusive license rights on the patent from Stanford University. The university received 1.8 million shares of Google in exchange for use of the patent; the shares were sold in 2005 for $336 million.

PageRank has been influenced by citation analysis, early developed by Eugene Garfield in the 1950s at the University of Pennsylvania, and by Hyper Search, developed by Massimo Marchiori at the University of Padua. In the same year PageRank was introduced (1998), Jon Kleinberg published his important work on HITS. Google's founders cite Garfield, Marchiori, and Kleinberg in their original paper.

A small search engine called "RankDex" from IDD Information Services designed by Robin Li was, since 1996, already exploring a similar strategy for site-scoring and page ranking. The technology in RankDex would be patented by 1999[5] and used later when Li founded Baidu in China.[6][7] Li's work would be referenced by some of Larry Page's U.S. patents for his Google search methods.[8]

Algorithm

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

Simplified algorithm

Assume a small universe of four web pages: **A**, **B**, **C** and **D**. Links from a page to itself, or multiple outbound links from one single page to another single page, are ignored. PageRank is initialized to the same value for all pages. In the original form of PageRank, the sum of PageRank over all pages was the total number of pages on the web at that time, so each page in this example would have an initial PageRank of 1. However, later versions of PageRank, and the remainder of this section, assume a probability distribution between 0 and 1. Hence the initial value for each page is 0.25.

The PageRank transferred from a given page to the targets of its outbound links upon the next iteration is divided equally among all outbound links.

If the only links in the system were from pages **B**, **C**, and **D** to **A**, each link would transfer 0.25 PageRank to **A** upon the next iteration, for a total of 0.75.

$$PR(A) = PR(B) + PR(C) + PR(D).$$

Suppose instead that page **B** had a link to pages **C** and **A**, page **C** had a link to page **A**, and page **D** had links to all three pages. Thus, upon the next iteration, page **B** would transfer half of its existing value, or 0.125, to page **A** and the other half, or 0.125, to page **C**. Page **C** would transfer all of its existing value, 0.25, to the only page it links to, **A**. Since **D** had three outbound links, it would transfer one third of its existing value, or approximately 0.083, to **A**. At the completion of this iteration, page **A** will have a PageRank of 0.458.

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

In other words, the PageRank conferred by an outbound link is equal to the document's own PageRank score divided by the number of outbound links **L( )**.

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

In the general case, the PageRank value for any page **u** can be expressed as:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

i.e. the PageRank value for a page **u** is dependent on the PageRank values for each page **v** contained in the set $\mathbf{B_u}$ (the set containing all pages linking to page **u**), divided by the number $L(v)$ of links from page **v**.

Damping factor

The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor *d*. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85.

The damping factor is subtracted from 1 (and in some variations of the algorithm, the result is divided by the number of documents (*N*) in the collection) and this term is then added to the product of the damping factor and the sum of the incoming PageRank scores. That is,

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

So any page's PageRank is derived in large part from the PageRanks of other pages. The damping factor adjusts the derived value downward. The original paper, however, gave the following formula, which has led to some confusion:

$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

The difference between them is that the PageRank values in the first formula sum to one, while in the second formula each PageRank is multiplied by *N* and the sum becomes *N*. A statement in Page and Brin's paper that "the sum of all PageRanks is one" and claims by other Google employees[9] support the first variant of the formula above.

Page and Brin confused the two formulas in their most popular paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine", where they mistakenly claimed that the latter formula formed a probability distribution over web pages.

Google recalculates PageRank scores each time it crawls the Web and rebuilds its index. As Google increases the number of documents in its collection, the initial approximation of PageRank decreases for all documents.

The formula uses a model of a *random surfer* who gets bored after several clicks and switches to a random page. The PageRank value of a page reflects the chance that the random surfer will land on that page by clicking on a link. It can be understood as a Markov chain in which the states are pages, and the transitions, which are all equally probable, are the links between pages.

If a page has no links to other pages, it becomes a sink and therefore terminates the random surfing process. If the random surfer arrives at a sink page, it picks another URL at random and continues surfing again.

When calculating PageRank, pages with no outbound links are assumed to link out to all other pages in the collection. Their PageRank scores are therefore divided evenly among all other pages. In other words, to be fair with pages that are not sinks, these random transitions are added to all nodes in the Web, with a residual probability usually set to d = 0.85, estimated from the frequency that an average surfer uses his or her browser's bookmark feature.

So, the equation is as follows:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where $p_1, p_2, \cdots, p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.

The PageRank values are the entries of the dominant eigenvector of the modified adjacency matrix. This makes PageRank a particularly elegant metric: the eigenvector is

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

where **R** is the solution of the equation

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

where the adjacency function $\ell(p_i,p_j)$ is 0 if page p_j does not link to p_i, and normalized such that, for each j

$$\sum_{i=1}^{N} \ell(p_i,p_j) = 1,$$

i.e. the elements of each column sum up to 1, so the matrix is a stochastic matrix (for more details see the computation section below). Thus this is a variant of the eigenvector centrality measure used commonly in network analysis.

Because of the large eigengap of the modified adjacency matrix above, the values of the PageRank eigenvector can be approximated to within a high degree of accuracy within only a few iterations.

As a result of Markov theory, it can be shown that the PageRank of a page is the probability of arriving at that page after a large number of clicks. This happens to equal t^{-1} where t is the expectation of the number of clicks (or random jumps) required to get from the page back to itself.

One main disadvantage of PageRank is that it favors older pages. A new page, even a very good one, will not have many links unless it is part of an existing site (a site being a densely connected set of pages, such as Wikipedia).

Several strategies have been proposed to accelerate the computation of PageRank.

Various strategies to manipulate PageRank have been employed in concerted efforts to improve search results rankings and monetize advertising links. These strategies have severely impacted the reliability of the PageRank concept[*citation needed*], which purports to determine which documents are actually highly valued by the Web community.

Since December 2007, when it started *actively* penalizing sites selling paid text links, Google has combatted link farms and other schemes designed to artificially inflate PageRank. How Google identifies link farms and other PageRank manipulation tools is among Google's trade secrets.

Computation

PageRank can be computed either iteratively or algebraically. The iterative method can be viewed as the power iteration method or the power method. The basic mathematical operations performed are identical.

Iterative

At $t = 0$, an initial probability distribution is assumed, usually

$$PR(p_i; 0) = \frac{1}{N}.$$

At each time step, the computation, as detailed above, yields

$$PR(p_i; t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j; t)}{L(p_j)},$$

or in matrix notation

$$\mathbf{R}(t+1) = d\mathcal{M}\mathbf{R}(t) + \frac{1-d}{N}\mathbf{1}, \quad (*)$$

where $\mathbf{R}_i(t) = PR(p_i; t)$ and $\mathbf{1}$ is the column vector of length N containing only ones.

The matrix $\mathcal{M}$ is defined as

$$\mathcal{M}_{ij} = \begin{cases} 1/L(p_j), & \text{if } j \text{ links to } i \\ 0, & \text{otherwise} \end{cases}$$

i.e.,

$$\mathcal{M} := (K^{-1}A)^T,$$

where A denotes the adjacency matrix of the graph and K is the diagonal matrix with the outdegrees in the diagonal.

The computation ends when for some small ϵ

$$|\mathbf{R}(t+1) - \mathbf{R}(t)| < \epsilon\,,$$

i.e., when convergence is assumed.

Algebraic

For $t \to \infty$ (i.e., in the steady state), the above equation (*) reads

$$\mathbf{R} = d\mathcal{M}\mathbf{R} + \frac{1-d}{N}\mathbf{1}. \qquad (**)$$

The solution is given by

$$\mathbf{R} = (\mathbf{I} - d\mathcal{M})^{-1}\frac{1-d}{N}\mathbf{1},$$

with the identity matrix $\mathbf{I}$.

The solution exists and is unique for $0 < d < 1$. This can be seen by noting that $\mathcal{M}$ is by construction a stochastic matrix and hence has an eigenvalue equal to one as a consequence of the Perron–Frobenius theorem.

Power Method

If the matrix $\mathcal{M}$ is a transition probability, i.e., column-stochastic with no columns consisting of just zeros and $\mathbf{R}$ is a probability distribution (i.e., $|\mathbf{R}| = 1, \mathbf{E}\mathbf{R} = 1$where $\mathbf{E}$is matrix of all ones), Eq. (**) is equivalent to

$$\mathbf{R} = \left(d\mathcal{M} + \frac{1-d}{N}\mathbf{E}\right)\mathbf{R} =: \widehat{\mathcal{M}}\mathbf{R}. \qquad (***)$$

Hence PageRank $\mathbf{R}$is the principal eigenvector of $\widehat{\mathcal{M}}$. A fast and easy way to compute this is using the power method: starting with an arbitrary vector $x(0)$, the operator $\widehat{\mathcal{M}}$ is applied in succession, i.e.,

$$x(t+1) = \widehat{\mathcal{M}}x(t),$$

until

$$|x(t+1) - x(t)| < \epsilon.$$

Note that in Eq. (***) the matrix on the right-hand side in the parenthesis can be interpreted as

$$\frac{1-d}{N}\mathbf{I} = (1-d)\mathbf{P}\mathbf{1}^t,$$

where $\mathbf{P}$ is an initial probability distribution. In the current case

$$\mathbf{P} := \frac{1}{N}\mathbf{1}.$$

Finally, if $\mathcal{M}$ has columns with only zero values, they should be replaced with the initial probability vector $\mathbf{P}$. In other words

$$\mathcal{M}' := \mathcal{M} + \mathcal{D},$$

where the matrix $\mathcal{D}$ is defined as

$$\mathcal{D} := \mathbf{P}\mathbf{D}^t,$$

with

$$\mathbf{D}_i = \begin{cases} 1, & \text{if } L(p_i) = 0 \\ 0, & \text{otherwise} \end{cases}$$

In this case, the above two computations using $\mathcal{M}$ only give the same PageRank if their results are normalized:

$$\mathbf{R}_{\text{power}} = \frac{\mathbf{R}_{\text{iterative}}}{|\mathbf{R}_{\text{iterative}}|} = \frac{\mathbf{R}_{\text{algebraic}}}{|\mathbf{R}_{\text{algebraic}}|}.$$

PageRank MATLAB/Octave implementation

```
% Parameter M adjacency matrix where M_i,j represents the link from 'j'
 to 'i', such that for all 'j' sum(i, M_i,j) = 1
% Parameter d damping factor
```

```
% Parameter v_quadratic_error quadratic error for v
% Return v, a vector of ranks such that v_i is the i-th rank from [0,
1]

function [v] = rank(M, d, v_quadratic_error)

N = size(M, 2); % N is equal to half the size of M
v = rand(N, 1);
v = v ./ norm(v, 2);
last_v = ones(N, 1) * inf;
M_hat = (d .* M) + (((1 - d) / N) .* ones(N, N));

while(norm(v - last_v, 2) > v_quadratic_error)
      last_v = v;
      v = M_hat * v;
      v = v ./ norm(v, 2);
end

endfunction

function [v] = rank2(M, d, v_quadratic_error)

N = size(M, 2); % N is equal to half the size of M
v = rand(N, 1);
v = v ./ norm(v, 1);   % This is now L1, not L2
last_v = ones(N, 1) * inf;
M_hat = (d .* M) + (((1 - d) / N) .* ones(N, N));

while(norm(v - last_v, 2) > v_quadratic_error)
      last_v = v;
      v = M_hat * v;
        % removed the L2 norm of the iterated PR
end

endfunction
```

Example of code calling the rank function defined above:

```
M = [0 0 0 0 1 ; 0.5 0 0 0 0 ; 0.5 0 0 0 0 ; 0 1 0.5 0 0 ; 0 0 0.5 1 0];
rank(M, 0.80, 0.001)
```

This example takes 13 iterations to converge.

The following is a proof that rank.m is incorrect. It's based on the first graphic example. My understanding is that rank.m uses the wrong norm on the input, then continues to renormalize L2, which is unnecessary.

```
% This represents the example graph, correctly normalized and
accounting for sinks (Node A)
% by allowing it to effectively random transition 100% of time,
```

```
including to itself.
% While RANK.m doesn't actually handle this incorrectly, it does not
show exactly how one should
% handle sink nodes (one possible solution would be a SELF-TRANSITION
of 1.0), which does not
% give the correct result.

test_graph = ...
[  0.09091   0.00000   0.00000   0.50000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   1.00000   0.50000   0.33333   0.50000   0.50000
   0.50000   0.50000   0.00000   0.00000;
   0.09091   1.00000   0.00000   0.00000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.33333   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.00000   0.50000   0.50000
   0.50000   0.50000   1.00000   1.00000;
   0.09091   0.00000   0.00000   0.00000   0.33333   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000;
   0.09091   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
   0.00000   0.00000   0.00000   0.00000 ]

pr = rank(test_graph, 0.85, 0.001)   % INCORRECT is not normalized.

%  0.062247
%  0.730223
%  0.650829
%  0.074220
%  0.153590
%  0.074220
%  0.030703
%  0.030703
% 0.030703
%  0.030703
%  0.030703

pr / norm(pr,1)    % CORRECT once normalized. I still don't know why
the L2 normalization happens ( v = v/norm(v, 2))
```

```
%   0.032781
%   0.384561
%   0.342750
%   0.039087
%   0.080886
%   0.039087
%   0.016170
%   0.016170
%   0.016170
%   0.016170
%   0.016170

pr = rank2(test_graph, 0.85, 0.001) % CORRECT, only requires input PR
normalization (make sure it sums to 1.0)

%   0.032781
%   0.384561
%   0.342750
%   0.039087
%   0.080886
%   0.039087
%   0.016170
%   0.016170
%   0.016170
%   0.016170
%   0.016170
```

Efficiency

Depending on the framework used to perform the computation, the exact implementation of the methods, and the required accuracy of the result, the computation time of the these methods can vary greatly.

Variations

PageRank of an undirected graph

The PageRank of an undirected graph G is statistically close to the degree distribution of the graph G, but they are generally not identical: If R is the PageRank vector defined above, and D is the degree distribution vector

$$D = \frac{1}{2|E|}\begin{bmatrix} deg(p_1) \\ deg(p_2) \\ \vdots \\ deg(p_N) \end{bmatrix}$$

where $deg(p_i)$ denotes the degree of vertex p_i, and E is the edge-set of the graph, then, with $Y = \frac{1}{N}\mathbf{1}$, by:

$$\frac{1-d}{1+d}\|Y - D\|_1 \leq \|R - D\|_1 \leq \|Y - D\|_1,$$

that is, the PageRank of an undirected graph equals to the degree distribution vector if and only if the graph is regular, i.e., every vertex has the same degree.

Distributed Algorithm for PageRank Computation

There are simple and fast random walk-based distributed algorithms for computing PageRank of nodes in a network. They present a simple algorithm that takes $O(\log n/\epsilon)$ rounds with high probability on any graph (directed or undirected), where n is the network size and ϵ is the reset probability ($1-\epsilon$ is also called as damping factor) used in the PageRank computation. They also present a faster algorithm that takes $O(\sqrt{\log n}/\epsilon)$ rounds in undirected graphs. Both of the above algorithms are scalable, as each node processes and sends only small (polylogarithmic in n, the network size) number of bits per round. For directed graphs, they present an algorithm that has a running time of $O(\sqrt{\log n/\epsilon})$, but it requires a polynomial number of bits to processed and sent per node in a round.

Google Toolbar

The Google Toolbar's PageRank feature displays a visited page's PageRank as a whole number between 0 and 10. The most popular websites have a PageRank of 10. The least have a PageRank of 0. Google has not disclosed the specific method for determining a Toolbar PageRank value, which is to be considered only a rough indication of the value of a website.

PageRank measures the number of sites that link to a particular page.[10] The PageRank of a particular page is roughly based upon the quantity of inbound links as well as the PageRank of the pages providing the links. The algorithm also includes other factors, such as the size of a page, the number of changes, the time since the page was updated, the text in headlines and the text in hyperlinked anchor texts.

The Google Toolbar's PageRank is updated infrequently, so the values it shows are often out of date.

SERP Rank

The search engine results page (SERP) is the actual result returned by a search engine in response to a keyword query. The SERP consists of a list of links to web pages with associated text snippets. The SERP rank of a web page refers to the placement of the corresponding link on the SERP, where higher placement means higher SERP rank. The SERP rank of a web page is a function not only of its PageRank, but of a relatively large and continuously adjusted set of factors (over 200),. Search engine optimization (SEO) is aimed at influencing the SERP rank for a website or a set of web pages.

Positioning of a webpage on Google SERPs for a keyword depends on relevance and reputation, also known as authority and popularity. PageRank is Google's indication of its assessment of the reputation of a webpage: It is non-keyword specific. Google uses a combination of webpage and website authority to determine the overall authority of a webpage competing for a keyword.[11] The PageRank of the HomePage of a website is the best indication Google offers for website authority.[12]

After the introduction of Google Places into the mainstream organic SERP, numerous other factors in addition to PageRank affect ranking a business in Local Business Results.

Google directory PageRank

The Google Directory PageRank was an 8-unit measurement. Unlike the Google Toolbar, which shows a numeric PageRank value upon mouseover of the green bar, the Google Directory only displayed the bar, never the numeric values. Google Directory was closed on July 20th, 2011.[13]

False or spoofed PageRank

In the past, the PageRank shown in the Toolbar was easily manipulated. Redirection from one page to another, either via a HTTP 302 response or a "Refresh" meta tag, caused the source page to acquire the PageRank of the destination page. Hence, a new page with PR 0 and no incoming links could have acquired PR 10 by redirecting to the Google home page. This spoofing technique, also known as 302 Google Jacking, was a known vulnerability. Spoofing can generally be detected by performing a Google search for a source URL; if the URL of an entirely different site is displayed in the results, the latter URL may represent the destination of a redirection.

Manipulating PageRank

For search engine optimization purposes, some companies offer to sell high PageRank links to webmasters. As links from higher-PR pages are believed to be more valuable, they tend to be more expensive. It can be an effective and viable marketing strategy to buy link advertisements on content pages of quality and relevant sites to drive traffic and increase a webmaster's link popularity. However, Google has publicly warned webmasters that if they are or were discovered to be selling links for the purpose of conferring PageRank and reputation, their links will be devalued (ignored in the calculation of other pages' PageRanks). The practice of buying and selling links is intensely debated across the Webmaster community. Google advises webmasters to use the nofollow HTML attribute value on sponsored links. According to Matt Cutts, Google is concerned about webmasters who try to game the system, and thereby reduce the quality and relevance of Google search results.

The intentional surfer model

The original PageRank algorithm reflects the so-called random surfer model, meaning that the PageRank of a particular page is derived from the theoretical probability of visiting that page when clicking on links at random. A page ranking model that reflects the importance of a particular page as a function of how many actual visits it receives by real users is called the *intentional surfer model*. The Google toolbar sends information to Google for every page visited, and thereby provides a basis for computing PageRank based on the intentional surfer model. The introduction of the nofollow attribute by Google to combat Spamdexing has the side effect that webmasters commonly use it on outgoing links to increase their own PageRank. This causes a loss of actual links for the Web crawlers to follow, thereby making the original PageRank algorithm based on the random surfer model potentially unreliable. Using information about users' browsing habits provided by the Google toolbar partly compensates for the loss of information caused by the nofollow attribute. The SERP rank of a page, which determines a page's actual placement in the search results, is based on a combination of the random surfer model (PageRank) and the intentional surfer model (browsing habits) in addition to other factors.

Other uses

A version of PageRank has recently been proposed as a replacement for the traditional Institute for Scientific Information (ISI) impact factor, and implemented at eigenfactor.org [14]. Instead of merely counting total citation to a journal, the "importance" of each citation is determined in a PageRank fashion.

A similar new use of PageRank is to rank academic doctoral programs based on their records of placing their graduates in faculty positions. In PageRank terms, academic departments link to each other by hiring their faculty from each other (and from themselves).

PageRank has been used to rank spaces or streets to predict how many people (pedestrians or vehicles) come to the individual spaces or streets. In lexical semantics it has been used to perform Word Sense Disambiguation[15] and to automatically rank WordNet synsets according to how strongly they possess a given semantic property, such as positivity or negativity.

A dynamic weighting method similar to PageRank has been used to generate customized reading lists based on the link structure of Wikipedia.

A Web crawler may use PageRank as one of a number of importance metrics it uses to determine which URL to visit during a crawl of the web. One of the early working papers that were used in the creation of Google is *Efficient crawling through URL ordering*, which discusses the use of a number of different importance metrics to determine how deeply, and how much of a site Google will crawl. PageRank is presented as one of a number of these importance metrics, though there are others listed such as the number of inbound and outbound links for a URL, and the distance from the root directory on a site to the URL.

The PageRank may also be used as a methodology [16] to measure the apparent impact of a community like the Blogosphere on the overall Web itself. This approach uses therefore the PageRank to measure the distribution of attention in reflection of the Scale-free network paradigm.

In any ecosystem, a modified version of PageRank may be used to determine species that are essential to the continuing health of the environment.

For the analysis of protein networks in biology PageRank is also a useful tool.

nofollow

In early 2005, Google implemented a new value, "nofollow", for the rel attribute of HTML link and anchor elements, so that website developers and bloggers can make links that Google will not consider for the purposes of PageRank—they are links that no longer constitute a "vote" in the PageRank system. The nofollow relationship was added in an attempt to help combat spamdexing.

As an example, people could previously create many message-board posts with links to their website to artificially inflate their PageRank. With the nofollow value, message-board administrators can modify their code to automatically insert "rel='nofollow'" to all hyperlinks in posts, thus preventing PageRank from being affected by those particular posts. This method of avoidance, however, also has various drawbacks, such as reducing the link value of legitimate comments. (See: Spam in blogs#nofollow)

In an effort to manually control the flow of PageRank among pages within a website, many webmasters practice what is known as PageRank Sculpting—which is the act of strategically placing the nofollow attribute on certain internal links of a website in order to funnel PageRank towards those pages the webmaster deemed most important. This tactic has been used since the inception of the nofollow attribute, but may no longer be effective since Google announced that blocking PageRank transfer with nofollow does not redirect that PageRank to other links.

Deprecation

PageRank was once available for the verified site maintainers through the Google Webmaster Tools interface. However on October 15, 2009, a Google employee confirmed that the company had removed PageRank from its *Webmaster Tools* section, explaining that "We've been telling people for a long time that they shouldn't focus on PageRank so much. Many site owners seem to think it's the most important metric for them to track, which is simply not true." In addition, The PageRank indicator is not available in Google's own Chrome browser.

The visible page rank is updated very infrequently.

On 6 October 2011, many users mistakenly thought Google PageRank was gone. As it turns out, it was simply an update to the URL used to query the PageRank from Google.

Google now also relies on other strategies as well as PageRank, such as Google Panda.

Notes

[1] http://www.google.com/competition/howgooglesearchworks.html

[2] Page, Larry, "PageRank: Bringing Order to the Web" (http://web.archive.org/web/20020506051802/www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1997-0072?1), Stanford Digital Library Project, talk. August 18, 1997 (archived 2002)

[3] 187-page study from Graz University, Austria (http://www.google-watch.org/gpower.pdf), includes the note that also human brains are used when determining the page rank in Google

[4] http://www.google.com/patents/US6285999

[5] USPTO, "Hypertext Document Retrieval System and Method" (http://www.google.com/patents?hl=en&lr=&vid=USPAT5920859&id=x04ZAAAAEBAJ&oi=fnd&dq=yanhong+li&printsec=abstract#v=onepage&q=yanhong li&f=false), U.S. Patent number: 5920859, Inventor: Yanhong Li, Filing date: Feb 5, 1997, Issue date: Jul 6, 1999

[6] Greenberg, Andy, "The Man Who's Beating Google" (http://www.forbes.com/forbes/2009/1005/technology-baidu-robin-li-man-whos-beating-google_2.html), *Forbes* magazine, October 05, 2009

[7] "About: RankDex" (http://www.rankdex.com/about.html), *rankdex.com*

[8] Cf. especially Lawrence Page, U.S. patents 6,799,176 (2004) "Method for scoring documents in a linked database", 7,058,628 (2006) "Method for node ranking in a linked database", and 7,269,587 (2007) "Scoring documents in a linked database"2011

[9] Matt Cutts's blog: Straight from Google: What You Need to Know (http://www.mattcutts.com/blog/seo-for-bloggers/), see page 15 of his slides.

[10] Google Webmaster central (http://www.google.com/support/forum/p/Webmasters/thread?tid=4aeb4d5fce33350b&hl=en) discussion on PR

[11] Dover, D. *Search Engine Optimization Secrets* Indianapolis. Wiley. 2011.

[12] Viniker, D. *The Importance of Keyword Difficulty Screening for SEO*. Ed. Schwartz, M. Digital Guidebook Volume 5. News Press. p 160–164.

[13] https://en.wikipedia.org/wiki/Google_Directory#Google_Directory

[14] http://www.eigenfactor.org

[15] Roberto Navigli, Mirella Lapata. "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation" (http://www.dsi.uniroma1.it/~navigli/pubs/PAMI_2010_Navigli_Lapata.pdf). IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 32(4), IEEE Press, 2010, pp. 678–692.

[16] http://de.scientificcommons.org/23846375

References

- Altman, Alon; Moshe Tennenholtz (2005). "Ranking Systems: The PageRank Axioms" (http://stanford.edu/~epsalon/pagerank.pdf) (PDF). *Proceedings of the 6th ACM conference on Electronic commerce (EC-05)*. Vancouver, BC. Retrieved 2008-02-05.
- Cheng, Alice; Eric J. Friedman (2006-06-11). "Manipulability of PageRank under Sybil Strategies" (http://www.cs.duke.edu/nicl/netecon06/papers/ne06-sybil.pdf) (PDF). *Proceedings of the First Workshop on the Economics of Networked Systems (NetEcon06)*. Ann Arbor, Michigan. Retrieved 2008-01-22.
- Farahat, Ayman; LoFaro, Thomas; Miller, Joel C.; Rae, Gregory and Ward, Lesley A. (2006). "Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization". *SIAM Journal on Scientific Computing* **27** (4): 1181–1201. doi: 10.1137/S1064827502412875 (http://dx.doi.org/10.1137/S1064827502412875).

- Haveliwala, Taher; Jeh, Glen and Kamvar, Sepandar (2003). "An Analytical Comparison of Approaches to Personalizing PageRank" (http://www-cs-students.stanford.edu/~taherh/papers/comparison.pdf) (PDF). *Stanford University Technical Report*.
- Langville, Amy N.; Meyer, Carl D. (2003). "Survey: Deeper Inside PageRank". *Internet Mathematics* **1** (3).
- Langville, Amy N.; Meyer, Carl D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press. ISBN 0-691-12202-4.
- Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). *The PageRank citation ranking: Bringing order to the Web* (http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1999-66&format=pdf&compression=).
- Richardson, Matthew; Domingos, Pedro (2002). "The intelligent surfer: Probabilistic combination of link and content information in PageRank" (http://www.cs.washington.edu/homes/pedrod/papers/nips01b.pdf) (PDF). *Proceedings of Advances in Neural Information Processing Systems* **14**.

Relevant patents

- Original PageRank U.S. Patent—Method for node ranking in a linked database (http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=6,285,999)—Patent number 6,285,999—September 4, 2001
- PageRank U.S. Patent—Method for scoring documents in a linked database (http://patft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=/netahtml/PTO/srchnum.htm&r=1&f=G&l=50&s1=6,799,176.PN.&OS=PN/6,799,176&RS=PN/6,799,176)—Patent number 6,799,176—September 28, 2004
- PageRank U.S. Patent—Method for node ranking in a linked database (http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=/netahtml/PTO/search-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,058,628.PN.&OS=pn/7,058,628&RS=PN/7,058,628)—Patent number 7,058,628—June 6, 2006
- PageRank U.S. Patent—Scoring documents in a linked database (http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=/netahtml/PTO/search-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,269,587.PN.&OS=pn/7,269,587&RS=PN/7,269,587)—Patent number 7,269,587—September 11, 2007

External links

- Our Search: Google Technology (http://www.google.com/technology/) by Google
- How Google Finds Your Needle in the Web's Haystack (http://www.ams.org/featurecolumn/archive/pagerank.html) by the American Mathematical Society

Article Sources and Contributors

PageRank *Source*: http://en.wikipedia.org/w/index.php?oldid=577159611 *Contributors*: -Midorihana-, 1-555-confide, 10metreh, 121a0012, 133u, 2008UEFA, 2620:0:1002:100D:9118:941:9583:9BA2, 345Kai, A. B., AKA MBG, ALargeElk, Aapo Laitinen, Aaronasterling, Aaronhill, Academic Challenger, Adambro, Adsandy, Affmark1, Afluent Rider, AgadaUrbanit, Ageekgal, Akashvedi, Al E., Alabalababayaga, AlanUS, Alcides, Alerante, Alexwg, Alma Pater, Alnokta, Alon, Aminto, Amitkn, Ams80, Anandnadaar, Andreas Kaufmann, Andrei Stroe, Anis.wiki12, Anomalocaris, Anshulsood9, Ant2101, Anthonyhcole, Apolkadot, Appraiser, Areldyb, Arivbran, ArmadilloFromHell, Arvindn, Asafe, Asgsoft, Audaciter, Aude, Audriusa, Ausref, Avsa, AxelBoldt, Banus, Barek, Bbatsell, Beetstra, Beland, Biars, Bill Slawski, Billsmithaustin, Binjiangwiki, BlaiseFEgan, Blaisorblade, Blase40, BlueYellowRed, Bluerhythm, Bobmutch, Bonadea, Booles, Brat32, Brewcrewer, Brockert, Bruce404, Bryan Derksen, Bugkai, Burtonator, CWii, Calabe1992, Camster342, Can't sleep, clown will eat me, Canadian Monkey, Cantaloupe2, Capricorn42, Carmitsp, Cartercole, Casey Abell, Chato, Clausen, Cody.feilding.nz, Collonell, Cometstyles, Commander, CommonsDelinker, Compfreak7, Computerjoe, Conversion script, Coolbuddy 459, Courcelles, Cumbrowski, Cybercobra, D.scain.farenzena, Damian Yerrick, Dan D. Ric, DanKeshet, Daniel Simanek, Daniel.Cardenas, Dante51763, Dave Runger, David Eppstein, DavidAViniker, DavidWBrooks, Davidpairey, Dawn Bard, Dcadenas, DeadEyeArrow, Deathphoenix, Dejitarob, Delsworld, Deror avi, Dimator, Dimitar petrov, Dirkbb, Discospinster, Dispenser, Dngrogan, Doc z, Doddy Wuid, Dorward, Doulos Christos, DrQuincy, Dreadengineer, Droll, Ds825, Dspradau, Durova, Dwo, Dysprosia, ESkog, Earth, Ebelular, Ebraminio, Ecopetition, Edward, Efe, Efitu, Efren, Elangokp, ElizabethFong, Emanuelvianna, Emile Barker, EncMstr, Engunneer, Ento, Er.punit, Ercan1334, Erin Lox, Esteban Zissou, Esthr, Eugman, Euryalus, Evenmadderjon, Evil saltine, FML, Fadesga, Fattyjwoods, Favonian, Ferengi, Fmccown, Fnielsen, Foot, Forderud, Fourthords, Fran Rogers, Fred Bauder, FrummerThanThou, Funnyfarmofdoom, Furrykef, Fvw, G J Lee, Gaius Cornelius, Galwhaa, Gamkiller, Gangaz, Gargaj, Gary King, Garyzx, Gdsdong, Gennaro Prota, George124, Ghodsnia, Giftlite, Gingerginger, Gnix, Goldenrowley, Gomm, Gousiosg, Gpanterov, GraemeL, Graham87, Greatlijo, Greyabernethy, Grolmusz, Gurchzilla, Gurubrahma, Guruweb, Gwernol, Haakon, Habibrubel, Hadal, HaeB, Hagedis, Hankwang, Haosusays, Harald Hansen, Harej, Haveatom, Hayabusa future, Headbomb, Helix84, Hgranqvist, Hiteshhoney, Hm2k, Hopelessless, HorsePunchKid, Hroshaan, Hu12, Hubsauthorities, Hughesey, Husond, Iapetus, Ibjhb, Ibnuasad, Ikescs, Ilovedar, Infobulletin, IngSoc BigBrother, Inky, IronGargoyle, It writer, J, J.delanoy, J04n, JGXenite, JJ Harrison, JLaTondre, JaGa, Jakehallbite, James435, Jamesday, JameyBM, Jamie Mercer, Jarble, Jarix, Jasonwlh314, Jay Vravos, Jean.julius, JeffMHoward, Jehochman, Jenozky25, Jeremy Visser, Jeyraul, Jim1138, Jimregan, Jmchuff, Jnode, JoanneB, John DiMatteo, Johnbibby, Johnuniq, JonathanBennaim, Jonyyb, Josang, JoshuaZ, Jpbowen, Jpgordon, Jsnow, Jugander, Junkinbomb, Just Another Dan, JustAnotherJoe, Justin W Smith, Jérôme, Karnpatel18, Kbdank71, Kemrin, Kencf0618, Kenniejyoung, Kesla, Kevpartner, Khalid hassani, Kharri1073, KingsOfHearts, Knownot, Knudvaneeden, Korg, Kuru, Kylemcinnes, LaMenta3, Laura SIMMONS, Legoktm, Lethe, Lightdarkness, LilHelpa, LinguistAtLarge, Linkexperts, Linkspamremover, Lisaedesign, LittleDan, Lizorkin, Logan, Lomonline, Lsmll, Lumingz, Luna Santin, LutzL, MECU, MER-C, Macrakis, Macshiva, Madhero88, Maghnus, Magioladitis, Mahdiiranpoor, Mandarax, Manishearth, MansonP, Marcelivan, MarkSweep, Martarius, Martin Jensen, Materialscientist, Matt Crypto, MaxVeers, Maximus Rex, Mblumber, McGeddon, Mean as custard, Mehnaz Khan, MelbourneStar, Mentisock, MiauImut, Michael Frind, Michael Hardy, Michael Martinez, Michaelas10, Michaeldsuarez, Mickeynguyen2107, Midgrid, Mido321, Mikeshaws, Mindmatrix, Mion, MissDanni, Misza13, Mo ainm, Mosmof, Mqchen, Mr. Random, Mr.Technology, Ms2ger, MuZemike, MustafaeneS, Nabla, NathanHurst, Necenzurat, Nekroshorume, Neurolysis, NeuronExMachina, Nichtich, Nik Gibbs, Nikaggar, Nochargebacks, Nongbell, Not a dog, Notheruser, Noveltysystems, NuclearWarfare, OSUKid7, Ohnoitsjamie, Oli Filth, Omegadeluxesupreme, Ortolan88, Pagerank10, Panther.ru, Paps34c, Paragon12321, Parkerh, Parthg88, Pataya1, Patelmiteshb, Paul Matthews, Paulomi333, Pbfy0, Pde, Pedrito, Pellucidity, Penbat, Penmachine, Pgluck, Phatalbert, Phil Boswell, PierreSenellart, Pifpafpouf26, Pigsonthewing, Plrk, Pmc, Postcrypto, Poweroid, Pro Plans Don, Prodego, Pruneau, Pseen, Pseudomonas, Psychonaut, Puekai, Pxma, Quintote, Qwertyus, Ram18y, Rasmus Faber, Rbraunwa, Redvers, Reedy, Reinyday, Retodon8, RexNL, Rheun, Rhobite, Rhomboid, Rich Farmbrough, Rickington, Rjwilmsi, Robin Sharaya, Robykiwi, Rotational, RoySmith, Royg73, Rsm99833, Rtaytug, RubySS, Ruud Koot, SDSandecki, Sadek409, SamWhited, Samratkafle, Sbluen, Schalling, Schneelocke, Schnolle, Scott McNay, Scottgallagher55, Sdp desk, Seoplus1, Serenity11pal, Sfan00 IMG, ShAd0w N1nJa, Sharcho, Shepelyansky, Showdown, Shreevatsa, Shubhransu, Shuffdog, SideScape, Simon Lacoste-Julien, Singingwolfboy, SiobhanHansa, SirQuill, Sivasankar1984, Skittleys, SkyMachine, SnappingTurtle, Snoyes, Soda drinker, Spicerider1, Spidey104, Spiff666, SpuriousQ, SqueakBox, Stannered, Steeev, Steel, Steevven1, Stephen, Stephenb, SteveSEOUK, Stevenj, Stochata, Stratocracy, Stringerace, StuffOfInterest, Subaru, Sumit41bct063, Syed.ali1996, SyntaxError55, TNeloms, Taa, Tangotango, Taw, Tbhotch, Teit22, Teoden44, TerriersFan, Tgr, ThaddeusB-public, The Anome, The Master of Mayhem, The Thing That Should Not Be, Theoldanarchist, Theroadislong, Thingg, Thorpe, Thumperward, Thv, Tintinobelisk, Toastyman, Tobias Bergemann, Tom-, TomDubai, Toytoy, Tracy Hall, Travelbird, Tree Biting Conspiracy, Tregoweth, Trigger hurt, Trivialist, Triwbe, Troller 69, Trumpsternator, Turkingside, Ultimatesouvik, Usb10, Uttaddmb, VINNIEs, Vagnerist, Vary, Veinor, Venterrqua, Versageek, Vietnam visa advisor, Vishalm 2710, Vogtadi, Vrenator, Vroo, WODUP, WaldoJ, Wavelength, Wdavies, Webranker345, Webrescue, Whimsley, Whosasking, WikiDan61, Wikiklrsc, William M. Connolley, Willsmith, Wmahan, WouterBolsterlee, Woz2, Wwwjscom, Wysprgr2005, X7q, Xevfgv123456, Xs08, Yagibear, Yboord028, Yintan, Youandme, Zanetu, Zealotgi, ZeroOne, Zfeinst, Zhoog, Zidonuke, ZimZalaBim, Zizzybaluba, ZombieDance, Zzuuzz, Милан Јелисавчић, 1263 anonymous edits

Image Sources, Licenses and Contributors

Image:PageRanks-Example.svg *Source*: http://en.wikipedia.org/w/index.php?title=File:PageRanks-Example.svg *License*: Public domain *Contributors*: en:User:345Kai, User:Stannered

File:PageRank-hi-res.png *Source*: http://en.wikipedia.org/w/index.php?title=File:PageRank-hi-res.png *License*: unknown *Contributors*: Cwbm (commons), FML, Mayhaymate, McGeddon, Microsoftstorepromocode, Paulo Cesar-1

License