Article   Talk

Read   Edit   View history

Search

# PageRank

From Wikipedia, the free encyclopedia

This article is **outdated**. Please update this article to reflect recent events or newly available information. *(February 2014)*

*"Google search algorithm" redirects here. For other search algorithms used by Google, see Google Penguin and Google Panda.*

**PageRank** is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page,[1] 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.[2]

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 best-known.[3][4]
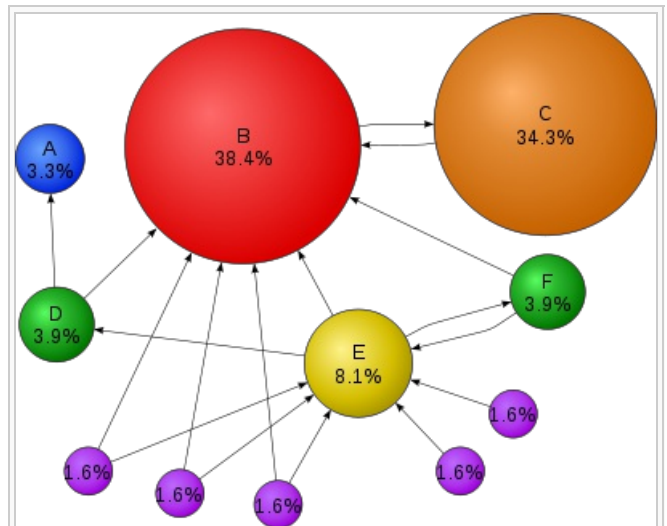


Mathematical **PageRanks** for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random 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.

## Contents

## Description   [edit]

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 $PR(E)$. Other factors like Author Rank can contribute to the importance of an entity.



Cartoon illustrating the 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.

Numerous academic papers concerning PageRank have been published since Page and Brin's original paper.[5] 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.[6]

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, the TrustRank algorithm and the hummingbird algorithm.

## History   [edit]

The idea of formulating a link analysis problem as an eigenvalue problem was suggested in 1976 by Gabriel Pinski and Francis Narin, who worked on scientometrics ranking scientific journals [7] and in 1977 by Thomas Saaty in his concept of Analytic Hierarchy Process which weighted alternative choices.[8]

PageRank was developed at Stanford University by Larry Page and Sergey Brin in 1996[9] as part of a research project about a new kind of search engine.[10] 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.[11] 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:[5] 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.[12]

The name "PageRank" plays off of the name of developer Larry Page, as well as the concept of a web page.[13] The word is a trademark of Google, and the PageRank process has been patented (U.S. Patent 6,285,999 ⮺). 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.[14][15]

PageRank was 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 papers.[5][16]

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.[17] The technology in RankDex would be patented by 1999[18] and used later when Li founded Baidu in China.[19][20] Li's work would be referenced by some of Larry Page's U.S. patents for his Google search methods.[21]
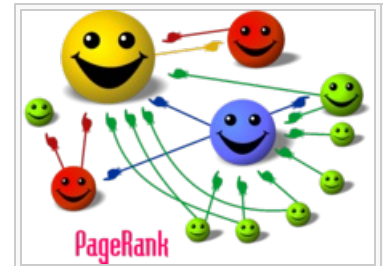
## Algorithm   [edit]

The PageRank algorithm outputs 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  [edit]

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 value 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 first 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 **B$_u$** (the set containing all pages linking to page **u**), divided by the number $L(v)$ of links from page **v**.

### Damping factor  [edit]

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.[5]

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"[5] and claims by other Google employees[22] 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.[5]

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 $p_1, p_2, ..., p_N$ 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 left 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,[23] 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.[24]

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   [edit]

PageRank can be computed either iteratively or algebraically. The iterative method can be viewed as the power iteration method[25][26] or the power method. The basic mathematical operations performed are identical.

### Iterative   [edit]

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 $\epsilon$

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

i.e., when convergence is assumed.

### Algebraic   [edit]

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

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

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   [edit]

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

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

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

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

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] = 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.

## Variations [edit]

### PageRank of an undirected graph [edit]

The PageRank of an undirected [graph] G is statistically close to the degree distribution of the graph G,[27] 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:[28]

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

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 [edit]

There are simple and fast random walk-based [distributed algorithms] for computing PageRank of nodes in a network.[29] 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 $\epsilon$ 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 [edit]

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.[30] 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.[11]

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

### SERP rank [edit]

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),.[31] [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.[32] The PageRank of the HomePage of a website is the best indication Google offers for website authority.[33]

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.[34]

### Google directory PageRank [edit]

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 20, 2011.[35]

### False or spoofed PageRank [edit]

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 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 [edit]

For search engine optimization purposes, some companies offer to sell high PageRank links to webmasters.[36] 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.[36]

### The intentional surfer model [edit]

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*.[37]

## Other uses [edit]

The mathematics of PageRank are entirely general and apply to any graph or network in any domain. Thus, PageRank is now regularly used in bibliometrics, social and information network analysis, and for link prediction and recommendation. It's even used for systems analysis of road networks, as well as biology, chemistry, neuroscience, and physics.[38]

Personalized PageRank is used by Twitter to present users with other accounts they may wish to follow.[39]

Swiftype's site search product builds a "PageRank that's specific to individual websites" by looking at each website's signals of importance and prioritizing content based on factors such as number of links from the home page.[40]

A version of PageRank has recently been proposed as a replacement for the traditional Institute for Scientific Information (ISI) impact factor,[41] and implemented at Eigenfactor as well as at SCImago. 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).[42]

PageRank has been used to rank spaces or streets to predict how many people (pedestrians or vehicles) come to the individual spaces or streets.[43][44] In lexical semantics it has been used to perform Word Sense Disambiguation,[45] Semantic similarity,[46] and also to automatically rank WordNet synsets according to how strongly they possess a given semantic property, such as positivity or negativity.[47]

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 [48] that were used in the creation of Google is *Efficient crawling through URL ordering*,[49] 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](#) 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.[50]

For the analysis of protein networks in biology PageRank is also a useful tool.[51] [52]

In 2005, in a pilot study in Pakistan, *Structural Deep Democracy, SD2*[53][54] was used for leadership selection in a sustainable agriculture group called Contact Youth. SD2 uses *PageRank* for the processing of the transitive proxy votes, with the additional constraints of mandating at least two initial proxies per voter, and all voters are proxy candidates. More complex variants can be built on top of SD2, such as adding specialist proxies and direct votes for specific issues, but SD2 as the underlying umbrella system, mandates that generalist proxies should always be used.

Pagerank has recently been used to quantify the scientific impact of researchers. The underlying citation and collaboration networks are used in conjunction with pagerank algorithm in order to come up with a ranking system for individual publications which propagates to individual authors. The new index known as pagerank-index (Pi) is demonstrated to be fairer compared to h-index in the context of many drawbacks exhibited by h-index.[55]

## nofollow   [edit]

In early 2005, Google implemented a new value, "nofollow",[56] 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[57]—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.[58]

## Deprecation   [edit]

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, saying 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."[59] In addition, The PageRank indicator is not available in Google's own Chrome browser.

The visible page rank is updated very infrequently. It was last updated in November 2013. In October 2014 Matt Cutts announced that another visible pagerank update would not be coming.[60]

PageRank is now one of 200 ranking factors that Google uses to determine a page's popularity. Google Panda is one of the other strategies Google now relies on to rank popularity of pages. Even though PageRank is no longer directly important for SEO purposes, the existence of back-links from more popular websites continues to push a webpage higher up in search rankings.[61]

## See also   [edit]

- EigenTrust — a decentralized PageRank algorithm
- Google bomb
- Google Search
- Google matrix

- Google Panda
- VisualRank - Google's application of PageRank to image-search
- Hilltop algorithm
- Link love
- Methods of website linking
- Power method — the iterative eigenvector algorithm used to calculate PageRank
- Search engine optimization
- SimRank — a measure of object-to-object similarity based on random-surfer model
- Topic-Sensitive PageRank
- TrustRank
- Webgraph
- CheiRank
- Google Penguin
- Google Hummingbird

## Notes [edit]

1. ^ "Google Press Center: Fun Facts". www.google.com. Archived from the original on 2009-04-24.
2. ^ "Facts about Google and Competition". Archived from the original on 4 November 2011. Retrieved 12 July 2014.
3. ^ http://www.google.com/competition/howgooglesearchworks.html. Missing or empty |title= (help)
4. ^ Sullivan, Danny. "What Is Google PageRank? A Guide For Searchers & Webmasters". *Search Engine Land*.
5. ^ *a* *b* *c* *d* *e* *f* Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (PDF). *Computer Networks and ISDN Systems* **30**: 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552.
6. ^ Gyöngyi, Zoltán; Berkhin, Pavel; Garcia-Molina, Hector; Pedersen, Jan (2006), "Link spam detection based on mass estimation", *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06, Seoul, Korea)* (PDF), pp. 439–450.
7. ^ Gabriel Pinski and Francis Narin. "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics". *Information Processing & Management* **12** (5): 297–312. doi:10.1016/0306-4573(76)90048-0.
8. ^ Thomas Saaty (1977). "A scaling method for priorities in hierarchical structures". *Journal of Mathematical Psychology* **15** (3): 234–281. doi:10.1016/0022-2496(77)90033-5.
9. ^ Raphael Phan Chung Wei (2002-05-16). *New Straits Times* (Computimes; 2 ed.). Missing or empty |title= (help); |chapter= ignored (help)
10. ^ Page, Larry, "PageRank: Bringing Order to the Web" at the Wayback Machine (archived May 6, 2002), Stanford Digital Library Project, talk. August 18, 1997 (archived 2002)
11. ^ *a* *b* 187-page study from Graz University, Austria, includes the note that also human brains are used when determining the page rank in Google
12. ^ "Google Technology". Google.com. Retrieved 2011-05-27.
13. ^ David Vise and Mark Malseed (2005). *The Google Story*. p. 37. ISBN 0-553-80457-X.
14. ^ Lisa M. Krieger (1 December 2005). "Stanford Earns $336 Million Off Google Stock". San Jose Mercury News, cited by redOrbit. Retrieved 2009-02-25.
15. ^ Richard Brandt. "Starting Up. How Google got its groove". Stanford magazine. Retrieved 2009-02-25.
16. ^ Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). "The PageRank citation ranking: Bringing order to the Web"., published as a technical report on January 29, 1998 PDF
17. ^ Li, Yanhong (August 6, 2002). "Toward a qualitative search engine". *Internet Computing, IEEE* (IEEE Computer Society) **2** (4): 24–29. doi:10.1109/4236.707687.
18. ^ USPTO, "Hypertext Document Retrieval System and Method", U.S. Patent number: 5920859, Inventor: Yanhong Li, Filing date: Feb 5, 1997, Issue date: Jul 6, 1999
19. ^ Greenberg, Andy, "The Man Who's Beating Google", *Forbes* magazine, October 05, 2009
20. ^ "About: RankDex", *rankdex.com*
21. ^ 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
22. ^ Matt Cutts's blog: Straight from Google: What You Need to Know, see page 15 of his slides.
23. ^ Taher Haveliwala and Sepandar Kamvar. (March 2003). "The Second Eigenvalue of the Google Matrix" (PDF). *Stanford University Technical Report*: 7056. arXiv:math/0307056. Bibcode:2003math......7056N.
24. ^ Gianna M. Del Corso, Antonio Gullí, Francesco Romani (2005). "Fast PageRank Computation via a Sparse Linear System". *Internet Mathematics*. Lecture Notes in Computer Science **2** (3): 118. doi:10.1007/978-3-540-30216-2_10. ISBN 978-3-540-23427-2.
25. ^ Arasu, A. and Novak, J. and Tomkins, A. and Tomlin, J. (2002). "PageRank computation and the structure of the web: Experiments and algorithms". *Proceedings of the Eleventh International World Wide Web Conference,*

*Poster Track*. Brisbane, Australia. pp. 107–117.

26. ^ Massimo Franceschet (2010). "PageRank: Standing on the shoulders of giants". arXiv:1002.2858 [cs.IR].

27. ^ Nicola Perra and Santo Fortunato.; Fortunato (September 2008). "Spectral centrality measures in complex networks". *Phys. Rev. E,* **78** (3): 36107. arXiv:0805.3322. Bibcode:2008PhRvE..78c6107P. doi:10.1103/PhysRevE.78.036107.

28. ^ Vince Grolmusz (2015). "A Note on the PageRank of Undirected Graphs". *Information Processing Letters* **115**: 633–634. Bibcode:http://dx.doi.org/10.1016/j.ipl.2015.02.015.

29. ^ Atish Das Sarma, Anisur Rahaman Molla, Gopal Pandurangan, Eli Upfal (2012). "Fast Distributed PageRank Computation". arXiv:1208.3071 [cs.DC,cs.DS].

30. ^ Google Webmaster central discussion on PR

31. ^ Fishkin, Rand; Jeff Pollard (April 2, 2007). "Search Engine Ranking Factors - Version 2". seomoz.org. Retrieved May 11, 2009.[*unreliable source?*]

32. ^ Dover, D. *Search Engine Optimization Secrets* Indianapolis. Wiley. 2011.

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

34. ^ "Ranking of listings : Ranking - Google Places Help". Google.com. Retrieved 2011-05-27.

35. ^ Google Directory#Google Directory

36. ^ *a b* "How to report paid links". mattcutts.com/blog. April 14, 2007. Retrieved 2007-05-28.

37. ^ Jøsang, A. (2007). "Trust and Reputation Systems". In Aldini, A. *Foundations of Security Analysis and Design IV, FOSAD 2006/2007 Tutorial Lectures.* (PDF) **4677**. Springer LNCS 4677. pp. 209–245. doi:10.1007/978-3-540-74810-6.

38. ^ Gleich, David F. (January 2015). "PageRank Beyond the Web". *SIAM Review* **57** (3): 321–363. doi:10.1137/140976649.

39. ^ Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Bosagh Zadeh WTF: The who-to-follow system at Twitter, Proceedings of the 22nd international conference on World Wide Web

40. ^ Ha, Anthony (2012-05-08). "Y Combinator-Backed Swiftype Builds Site Search That Doesn't Suck". TechCrunch. Retrieved 2014-07-08.

41. ^ Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel.; Rodriguez; Van De Sompel (December 2006). "Journal Status". *Scientometrics* **69** (3): 1030. arXiv:cs.GL/0601030. Bibcode:2006cs........1030B. doi:10.1145/1255175.1255273.

42. ^ Benjamin M. Schmidt and Matthew M. Chingos (2007). "Ranking Doctoral Programs by Placement: A New Method" (PDF). *PS: Political Science and Politics* **40** (July): 523–529. doi:10.1017/s1049096507070771.

43. ^ B. Jiang (2006). "Ranking spaces for predicting human movement in an urban environment". *International Journal of Geographical Information Science* **23** (7): 823–837. arXiv:physics/0612011. doi:10.1080/13658810802022822.

44. ^ Jiang B., Zhao S., and Yin J. (2008). "Self-organized natural roads for predicting traffic flow: a sensitivity study". *Journal of Statistical Mechanics: Theory and Experiment*. P07008 (7): 008. arXiv:0804.1630. Bibcode:2008JSMTE..07..008J. doi:10.1088/1742-5468/2008/07/P07008.

45. ^ Roberto Navigli, Mirella Lapata. "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation". IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 32(4), IEEE Press, 2010, pp. 678–692.

46. ^ M. T. Pilehvar, D. Jurgens and R. Navigli. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, August 4–9, 2013, pp. 1341-1351.

47. ^ Andrea Esuli and Fabrizio Sebastiani. "PageRanking WordNet synsets: An Application to Opinion-Related Properties" (PDF). *In Proceedings of the 35th Meeting of the Association for Computational Linguistics, Prague, CZ, 2007, pp. 424–431.* Retrieved June 30, 2007.

48. ^ "Working Papers Concerning the Creation of Google". *Google.* Retrieved November 29, 2006.

49. ^ Cho, J., Garcia-Molina, H., and Page, L. (1998). "Efficient crawling through URL ordering". *Proceedings of the seventh conference on World Wide Web* (Brisbane, Australia).

50. ^ Burns, Judith (2009-09-04). "Google trick tracks extinctions". BBC News. Retrieved 2011-05-27.

51. ^ G. Ivan and V. Grolmusz (2011). "When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks". *Bioinformatics* (Vol. 27, No. 3. pp. 405-407) **27** (3): 405–7. doi:10.1093/bioinformatics/btq680. PMID 21149343.

52. ^ D. Banky and G. Ivan and V. Grolmusz (2013). "Equal opportunity for low-degree network nodes: a PageRank-based method for protein target identification in metabolic graphs". *PLoS One* (Vol. 8, No. 1. e54204) **8** (1): 405–7. Bibcode:2013PLoSO...854204B. doi:10.1371/journal.pone.0054204. PMID 23382878.

53. ^ "Yahoo! Groups". Groups.yahoo.com. Retrieved 2013-10-02.

54. ^ "CiteSeerX — Autopoietic Information Systems in Modern Organizations". Citeseerx.ist.psu.edu. Retrieved 2013-10-02.

55. ^ Senanayake, Upul; Piraveenan, Mahendra; Zomaya, Albert (2015). "The Pagerank-Index: Going beyond Citation Counts in Quantifying Scientific Impact of Researchers". *PLOS ONE* **10** (8): e0134794. doi:10.1371/journal.pone.0134794. ISSN 1932-6203.

56. ^ "Preventing Comment Spam". *Google.* Retrieved January 1, 2005.

57. ^ "PageRank Sculpting: Parsing the Value and Potential Benefits of Sculpting PR with Nofollow" ⧉. SEOmoz. Retrieved 2011-05-27.
58. ^ "PageRank sculpting" ⧉. Mattcutts.com. 2009-06-15. Retrieved 2011-05-27.
59. ^ Susan Moskwa. "PageRank Distribution Removed From WMT" ⧉. Retrieved October 16, 2009 |chapter= ignored (help)
60. ^ Bartleman, Wil (2014-10-12). "Google Page Rank Update is Not Coming" ⧉. Managed Admin. Retrieved 2014-10-12.
61. ^ "So...You Think SEO Has Changed" ⧉. 19.03.14. Check date values in: |date= (help)

## References  [edit]

- Altman, Alon; Moshe Tennenholtz (2005). "Ranking Systems: The PageRank Axioms" 🔗 (PDF). *Proceedings of the 6th ACM conference on Electronic commerce (EC-05)*. Vancouver, BC. Retrieved 29 September 2014.
- Cheng, Alice; Eric J. Friedman (2006-06-11). "Manipulability of PageRank under Sybil Strategies" 🔗 (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 🔗.
- Haveliwala, Taher; Jeh, Glen; Kamvar, Sepandar (2003). "An Analytical Comparison of Approaches to Personalizing PageRank" 🔗 (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.
- Richardson, Matthew; Domingos, Pedro (2002). "The intelligent surfer: Probabilistic combination of link and content information in PageRank" 🔗 (PDF). *Proceedings of Advances in Neural Information Processing Systems* **14**.

## Relevant patents  [edit]

- Original PageRank U.S. Patent—Method for node ranking in a linked database 🔗—Patent number 6,285,999—September 4, 2001
- PageRank U.S. Patent—Method for scoring documents in a linked database 🔗—Patent number 6,799,176—September 28, 2004
- PageRank U.S. Patent—Method for node ranking in a linked database 🔗—Patent number 7,058,628—June 6, 2006
- PageRank U.S. Patent—Scoring documents in a linked database 🔗—Patent number 7,269,587—September 11, 2007

## External links  [edit]

- Our Search: Google Technology 🔗 by Google
- How Google Finds Your Needle in the Web's Haystack 🔗 by the American Mathematical Society

| v · t · e | Google |
|---|---|
| **Overview** | Alphabet Inc. · History · List of mergers and acquisitions · Products · Criticism · Censorship · Domains · Easter eggs · Don't be evil |
| **Advertising** | AdMob · Adscape · AdSense · AdWords · Analytics · Contributor · Partners · DoubleClick · DoubleClick for Publishers · Wallet |
| **Communication** | Alerts · Apps Script · Calendar · Contacts · Friend Connect · Gmail (history · interface) · Google+ · Groups · Hangouts · Inbox · Sync · Text-to-Speech · Translate · Transliteration · Voice |
| **Software** | Chrome (for Android · for iOS · Chrome Web Store · Apps · Extensions) · Cloud Print · Earth (Sky · Moon · Mars) · Gadgets · Goggles · IME (Pinyin · Japanese) · Photos · Keep · News & Weather · Now · Picasa · OpenRefine · Toolbar |

Categories: Google Search | Search engine optimization | Reputation management | Internet search algorithms | American inventions | Crowdsourcing | Markov models | Link analysis

This page was last modified on 4 September 2015, at 08:28.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.