*This is a summary of the original PageRank white paper, which can be found*[***at this link***](http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf).

**Abstract**

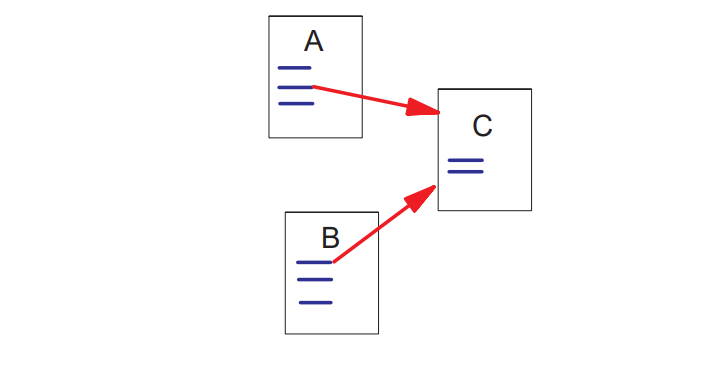
* PageRank is a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.
* We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and user navigation.

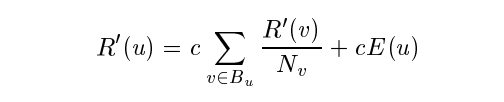
1. Introduction and Motivation

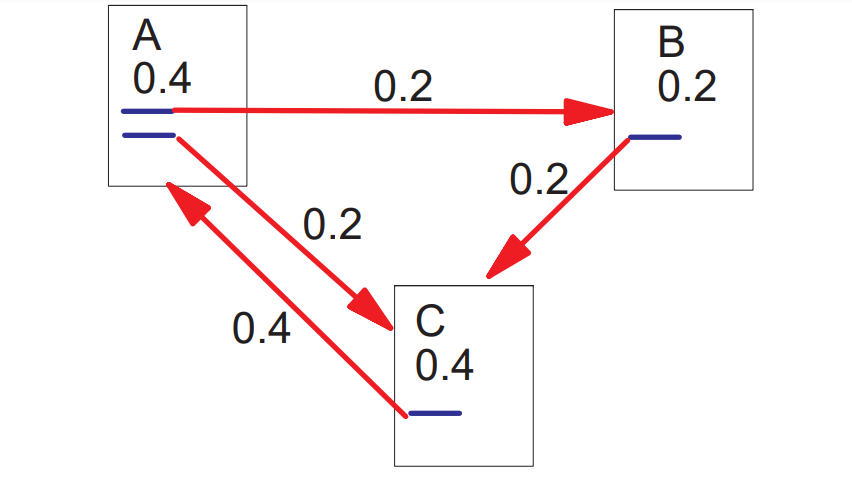
* The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous. More importantly, the web pages are extremely diverse, ranging from "What is Joe having for lunch today?" to journals about information retrieval.
* The World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages, such as link structure and link text.
* We use PageRank, which helps search engines and users quickly make sense of the vast heterogeneity of the World Wide Web.
  1. Diversity of Web Pages
     1. There are several significant differences between web pages and academic publications.
     2. Unlike academic papers which are scrupulously reviewed, web pages proliferate free of quality control or publishing costs.
     3. the Web environment contains competing for profit-seeking ventures, attention-getting strategies evolve in response to search engine algorithms. For this reason, any evaluation strategy which counts replicable features of web pages is prone to manipulation.
     4. The average web page quality experienced by a user is higher than the quality of the average web page.
  2. PageRank
     1. PageRank is a method for computing a ranking for every web page based on the graph of the web.
     2. PageRank has applications in search, browsing, and traffic estimation.

2. A Ranking for Every Page on the Web

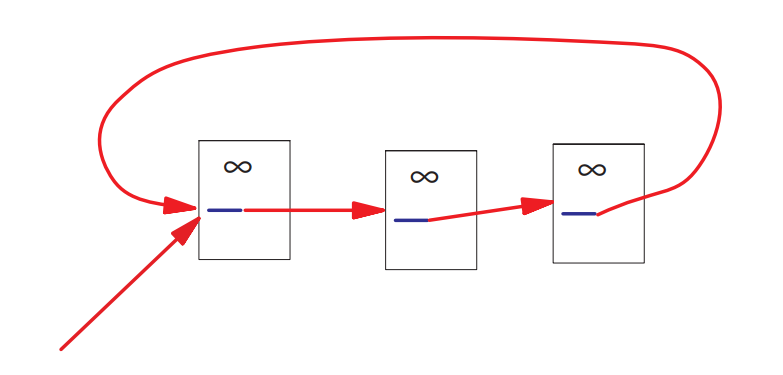
1. Related Work
   1. Goffman has published an interesting theory of how information flow in a scientific community is an epidemic process.
   2. There has been a fair amount of recent activity on how to exploit the link structure of large hypertext systems such as the web.
   3. Weiss discusses clustering methods that take the link structure into account.
   4. Kleinberg has developed an interesting model of the web as Hubs and Authorities, based on an eigenvector calculation on the co-citation matrix of the web.
2. Link Structure of the Web
3. The current graph of the crawlable Web has roughly 150 million nodes (pages) and 1.7 billion edges (links).
4. Every page has several out edges and in edges.
   1. Out edges are the forward links on the web page
   2. In edges are the backlinks on the web page.

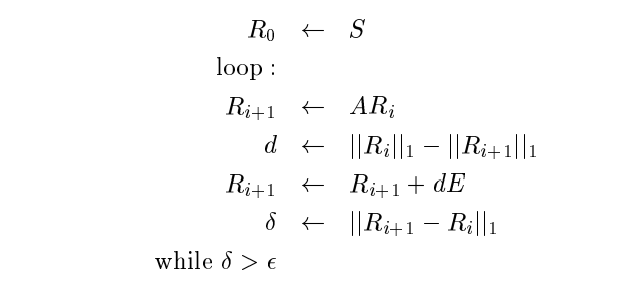


1. The Netscape home page has 62,804 backlinks in our current database compared to most pages that have just a few backlinks.
2. Generally, highly linked pages are more important" than pages with few links.
3. The reason that PageRank is interesting is that there are many cases where simple citation counting does not correspond to our common-sense notion of importance. 1. For example, if a web page has a link to the Yahoo home page, it may be just one link but it is a very important one.
4. PageRank is an attempt to see how good an approximation to importance" can be obtained just from the link structure.
5. Propagation of Ranking Through Links
   1. A page has a high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks.
6. Definition of PageRank
   1. Let E(u) be some vector over the Web pages that correspond to a source of rank. Then, the PageRank of a set of Web pages is an assignment, R0, to the Web pages which satisfies such that c is maximized and ||R|| =1 where E(u) is some vector over the web pages that correspond to a source of rank.
   2. 
   3. Simplified PageRank Calculation



* 1. Loop Which Acts as a Rank Sink



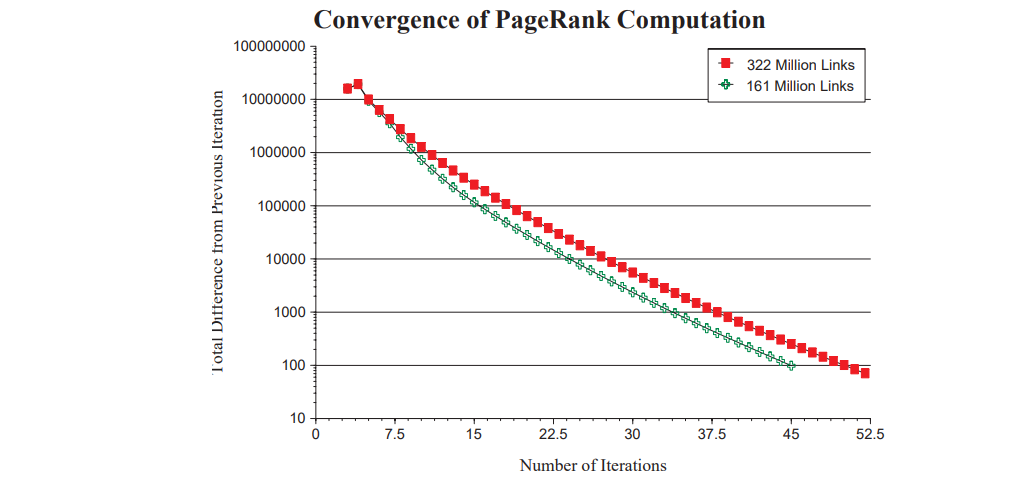
1. Random Surfer Model
   1. The simplified version of PageRank corresponds to the standing probability distribution of a random walk on the graph of the Web.This can be thought of as modeling the behavior of a random surfer.
   2. The random surfer simply keeps clicking on successive links at random.
   3. However, if a real Web surfer ever gets into a small loop of web pages, it is unlikely that the surfer will continue in the loop forever. Instead, the surfer will jump to some other page.
2. Computing PageRank
   1. The computation of PageRank is fairly straightforward if we ignore the issues of scale.
   2. 
3. Dangling Links
   1. Dangling links are simply links that point to any page with no outgoing links.
   2. They affect the model because it is not clear where their weight should be distributed, and there are a large number of them.
   3. Dangling links do not affect the ranking of any other page directly. We simply remove them from the system until all the PageRanks are calculated.

3. Implementation

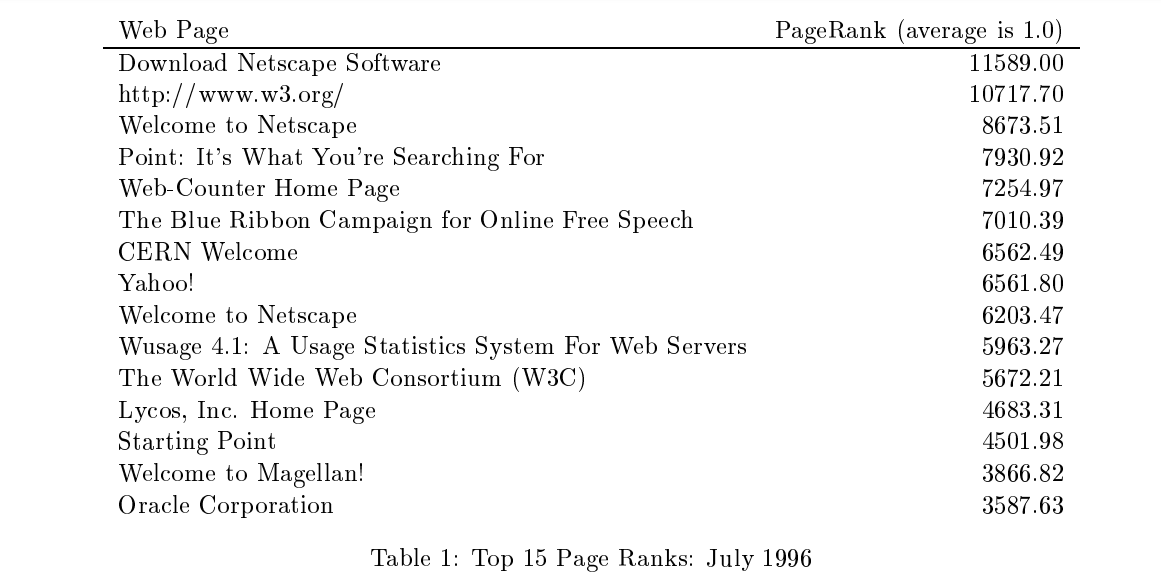
* To implement PageRank, the web crawler simply needs to build an index of links as it crawls.
* To index our current 24 million page database in about five days, we need to process about 50 web pages per second. Since there about 11 links on an average page (depending on what you count as a link) we need to process 550 links per second.
  1. PageRank Implementation
     1. We convert each URL into a unique integer and store each hyperlink in a database using the integer IDs to identify pages.
     2. First, we sort the link structure by Parent ID.
     3. Then dangling links are removed from the link database.
     4. Then we need to make an initial assignment of the ranks.We believe that careful choice of the initial assignment and a small finite number of iterations may result in excellent or improved performance.
     5. Memory is allocated for the weights for every page.
     6. After the weights have converged, we add the dangling links back in and recompute the rankings.

4. Convergence Properties

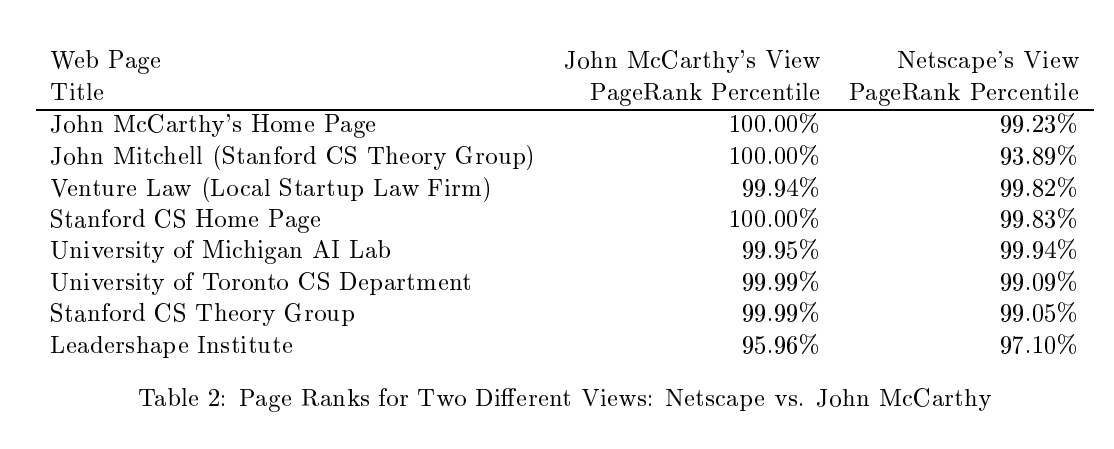
* PageRank on a large 322 million link database converges to a reasonable tolerance in roughly 52 iterations. The convergence on half the data takes roughly 45 iterations.
* One of the interesting ramifications of the fact that the PageRank calculation converges rapidly is that the web is an expander-like graph.
* To understand this better, we give an overview of the theory of random walks on graphs.
  1. A random walk on a graph is a stochastic process where at any given time step we are at a particular node of the graph and choose an outedge uniformly at random to determine the node to visit at the next time step.
  2. To relate all this to the PageRank computation, note that it is essentially the determination of the limiting distribution of a random walk on the Web graph.



5. Searching with PageRank

* A major application of PageRank is searching. We have implemented two search engines that use PageRank.
  1. The first one is a simple title-based search engine.
  2. The second search engine is a full-text search engine called Google.
* Title Search
  1. To test the usefulness of PageRank for search we implemented a search engine that used only the titles of 16 million web pages.
  2. The search engine finds all the web pages whose titles contain all of the query words. Then it sorts the results by PageRank. This search engine is very simple and cheap to implement.
* Rank Merging
  1. For more specific searches where recall is more important, the traditional information retrieval scores over full-text and the PageRank should be combined.
  2. Our Google system does this type of rank merging.
  3. Rank merging is known to be a very difficult problem, and we need to spend considerable additional effort before we will be able to do a reasonable evaluation of these types of queries.
* Some Sample Results
  1. 
* Common Case
  1. One of the design goals of PageRank was to handle the common case for queries well.
  2. It is important to note that the goal of finding a site that contains a great deal of information about the search query is a very different task than finding the common case search query.
* Subcomponents of Common Case
  1. PageRank can also represent a collaborative notion of authority or trust.
     1. A user might prefer a news story simply because it is linked directly from the New York Times home page. Of course, such a story will receive quite a high PageRank simply because it is mentioned by a very important page.
     2. This seems to capture a kind of collaborative trust.

6. Personalized PageRank

* An important component of the PageRank calculation is E.
  1. E is a vector over the Web pages which is used as a source of rank to make up for the rank sinks such as cycles with no outedges.
* E turns out to be a powerful parameter to adjust the page ranks.
* We have performed most experiments with an E vector that is uniform overall web pages with ||E|| = 0:15. This corresponds to a random surfer periodically jumping to a random web page. This is a very democratic choice for E since all web pages are valued simply because they exist.
* 
* we show the resulting page rank percentiles for an assortment of different pages. Pages related to computer science have a higher McCarthy-rank than Netscape-rank and pages related to computer science at Stanford have a considerably higher McCarthy-rank.
* Such personalized page ranks may have several applications, including personal search engines.These search engines could save users a great deal of trouble by efficiently guessing a large part of their interests given simple input such as their bookmarks or home page.
* Manipulation by Commercial Interests
  1. personalized PageRanks are virtually immune to manipulation by commercial interests.
  2. This immunity to manipulation is an extremely important property. This kind of commercial manipulation is causing search engines a great deal of trouble and making features that would be great to have very difficult to implement.
     1. Fast updating of documents is a very desirable feature, but it is abused by people who want to manipulate the results of the search engine.

7. Applications

* Estimating Web Traffic
  1. We used the counts of web page accesses from NLANR proxy cache and compared these to PageRank.
  2. The NLANR data was from several national proxy caches over several months and consisted of 11,817,665 unique URLs with the highest hit count going to Altavista with 638,657 hits.
  3. It is extremely difficult to compare these datasets analytically for several different reasons.
  4. we did see some interesting trends in the data.
     1. There seems to be a high usage of pornographic sites in the cache data.
     2. These sites generally had low PageRanks.
  5. Some sites have very high usage, but low PageRank such as netscape.yahoo.com.
* PageRank as Backlink Predictor
  1. One justification for PageRank is that it is a predictor for backlinks.
  2. we explore the issue of how to crawl the web efficiently, trying to crawl better documents first.
  3. We found on tests of the Stanford web that PageRank is a better predictor of future citation counts than citation counts themselves.
* User Navigation: The PageRank Proxy
  1. We have developed a web proxy application that annotates each link that a user sees with its PageRank.
  2. It is quite useful, because users receive some information about the link before they click on it.
  3. This proxy is very helpful for looking at the results from other search engines, and pages with large numbers of links.
  4. The proxy can help users decide which links in a long listing are more likely to be interesting. Or, if the user has some idea where the link they are looking for should fall in the importance" spectrum, they should be able to scan for it much more quickly using the proxy.
* Other Uses of PageRank
  1. The original goal of PageRank was a way to sort backlinks.
  2. It turns out this view of the backlinks ordered by PageRank can be very interesting when trying to understand your competition.
  3. PageRank can help the user decide if a site is trustworthy or not.

8. Conclusion

* PageRank is a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure.
* Using PageRank, we can order search results so that more important and central Web pages are given preference.
* Th intuition behind PageRank is that it uses information that is exernal to the Web pages themselves - their backlinks, which provide a kind of peer review.
* PageRank could be used to separate a small set of commonly used documents that can answer most queries.
* We have found several applications for PageRank in addition to search which includes traffic estimation, and user navigation.
* Overall, our experiments with PageRank suggest that the structure of the Web graph is very useful for a variety of information retrieval tasks.