The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, 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 to user navigation.

1 Introduction and Motivation

The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous. Current estimates are that there are over 150 million web pages with a doubling life of less than one year. More importantly, the web pages are extremely diverse, ranging from "What is Joe having for lunch today?" to journals about information retrieval. In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions.

However, unlike "flat" document collections, 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. In this paper, we take advantage of the link structure of the Web to produce a global "importance" ranking of every web page. This ranking, called PageRank, helps search engines and users quickly make sense of the vast heterogeneity of the World Wide Web.

1.1 Diversity of Web Pages

Although there is already a large literature on academic citation analysis, there are a number of significant differences between web pages and academic publications. Unlike academic papers which are scrupulously reviewed, web pages proliferate free of quality control or publishing costs. With a simple program, huge numbers of pages can be created easily, artificially inflating citation counts. Because the Web environment contains competing 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. Further, academic papers are well defined units of work, roughly similar in quality and number of citations, as well as in their purpose – to extend the body of knowledge. Web pages vary on a much wider scale than academic papers in quality, usage, citations, and length. A random archived message posting asking an obscure question about an IBM computer is very different from the IBM home page. A research article about the effects of cellular phone use on driver attention is very different from an advertisement for a particular cellular provider. The average web page quality experienced by a user is higher than the quality of the average web page. This is because the simplicity of creating and publishing web pages results in a large fraction of low quality web pages that users are unlikely to read.

There are many axes along which web pages may be differentiated. In this paper, we deal primarily with one - an approximation of the overall relative importance of web pages.

1.2 PageRank

In order to measure the relative importance of web pages, we propose PageRank, a method for computing a ranking for every web page based on the graph of the web. PageRank has applications in search, browsing, and traffic estimation.

Section 2 gives a mathematical description of PageRank and provides some intuitive justification. In Section 3, we show how we efficiently compute PageRank for as many as 518 million hyperlinks. To test the utility of PageRank for search, we built a web search engine called Google (Section 5). We also demonstrate how PageRank can be used as a browsing aid in Section 7.3.

2 A Ranking for Every Page on the Web

2.1 Related Work

There has been a great deal of work on academic citation analysis [Gar95]. Goffman [Gof71] has published an interesting theory of how information flow in a scientific community is an epidemic process.

There has been a fair amount of recent activity on how to exploit the link structure of large hypertext systems such as the web. Pitkow recently completed his Ph.D. thesis on "Characterizing World Wide Web Ecologies" [Pit97, PPR96] with a wide variety of link based analysis. Weiss discuss clustering methods that take the link structure into account [WVS+96]. Spertus [Spe97] discusses information that can be obtained from the link structure for a variety of applications. Good visualization demands added structure on the hypertext and is discussed in [MFH95, MF95]. Recently, Kleinberg [Kle98] 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.

Finally, there has been some interest in what "quality" means on the net from a library community [Til].

It is obvious to try to apply standard citation analysis techniques to the web's hypertextual citation structure. One can simply think of every link as being like an academic citation. So, a major page like http://www.yahoo.com/ will have tens of thousands of backlinks (or citations) pointing to it.

This fact that the Yahoo home page has so many backlinks generally imply that it is quite important. Indeed, many of the web search engines have used backlink count as a way to try to bias their databases in favor of higher quality or more important pages. However, simple backlink counts have a number of problems on the web. Some of these problems have to do with characteristics of the web which are not present in normal academic citation databases.

2.2 Link Structure of the Web

While estimates vary, the current graph of the crawlable Web has roughly 150 million nodes (pages) and 1.7 billion edges (links). Every page has some number of forward links (outedges) and backlinks (inedges) (see Figure 1). We can never know whether we have found all the backlinks of a particular page but if we have downloaded it, we know all of its forward links at that time.

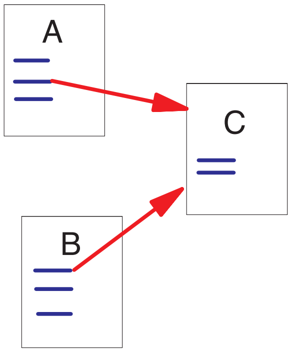

Figure 1: A and B are Backlinks of C

Web pages vary greatly in terms of the number of backlinks they have. For example, the Netscape home page has 62,804 backlinks in our current database compared to most pages which have just a few backlinks. Generally, highly linked pages are more "important" than pages with few links. Simple citation counting has been used to speculate on the future winners of the Nobel Prize [San95]. PageRank provides a more sophisticated method for doing citation counting.

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. For example, if a web page has a link off the Yahoo home page, it may be just one link but it is a very important one. This page should be ranked higher than many pages with more links but from obscure places. PageRank is an attempt to see how good an approximation to "importance" can be obtained just from the link structure.

2.3 Propagation of Ranking Through Links

Based on the discussion above, we give the following intuitive description of PageRank: a page has 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.

2.4 Definition of PageRank

Let u be a web page. Then let F_u be the set of pages u points to and B_u be the set of pages that point to u. Let $N_u = |F_u|$ be the number of links from u and let c be a factor used for normalization (so that the total rank of all web pages is constant).

We begin by defining a simple ranking, R which is a slightly simplified version of PageRank:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

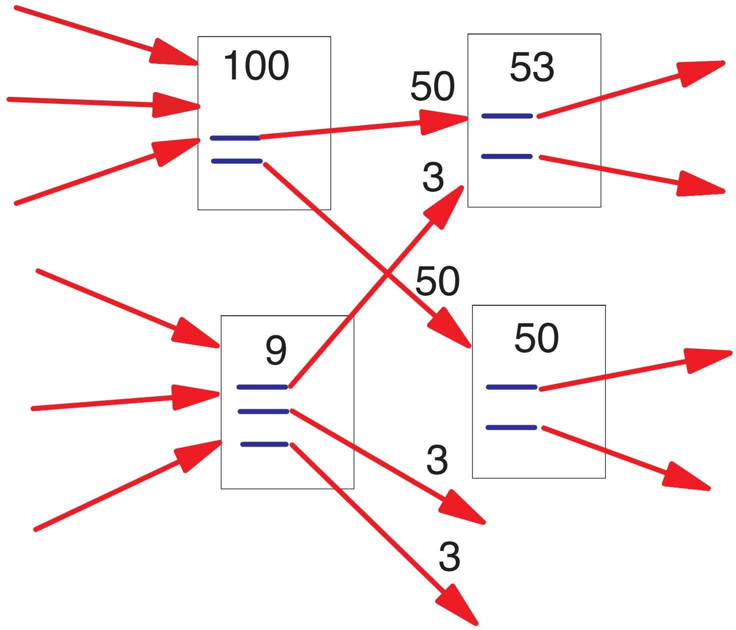

Figure 2: Simplified PageRank Calculation

This formalizes the intuition in the previous section. Note that the rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to. Note that $c < 1$ because there are a number of pages with no forward links and their weight is lost from the system (see section 2.7). The equation is recursive but it may be computed by starting with any set of ranks and iterating the computation until it converges. Figure 2 demonstrates the propagation of rank from one pair of pages to another. Figure 3 shows a consistent steady state solution for a set of pages.

Stated another way, let A be a square matrix with the rows and column corresponding to web pages. Let $A_{u,v} = 1/N_u$ if there is an edge from u to v and $A_{u,v} = 0$ if not. If we treat R as a vector over web pages, then we have $R = cAR$. So R is an eigenvector of A with eigenvalue c. In fact, we want the dominant eigenvector of A. It may be computed by repeatedly applying A to any nondegenerate start vector.

There is a small problem with this simplified ranking function. Consider two web pages that point to each other but to no other page. And suppose there is some web page which points to one of them. Then, during iteration, this loop will accumulate rank but never distribute any rank (since there are no outedges). The loop forms a sort of trap which we call a rank sink.

To overcome this problem of rank sinks, we introduce a rank source:

Definition 1 *Let $E(u)$ be some vector over the Web pages that corresponds to a source of rank. Then, the PageRank of a set of Web pages is an assignment, R', to the Web pages which satisfies*

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u) \qquad (1)$$

such that c is maximized and $||R'||_1 = 1$ ($||R'||_1$ denotes the L_1 norm of R').

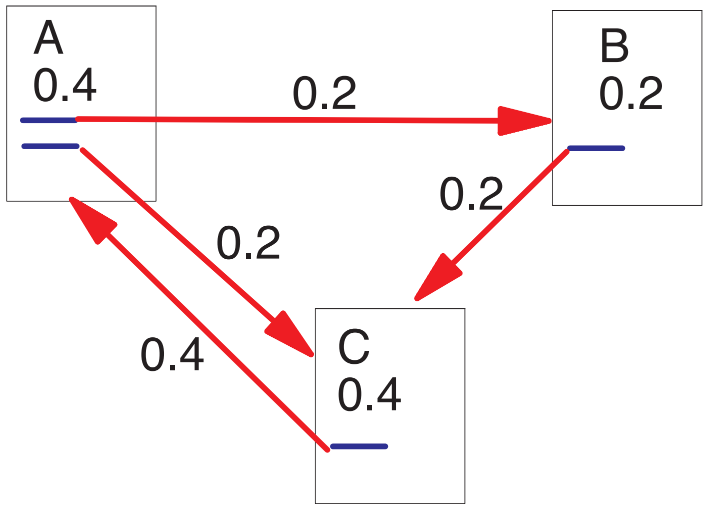

Figure 3: Simplified PageRank Calculation

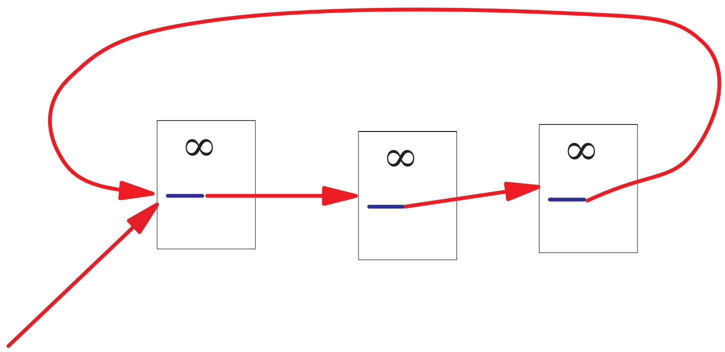

Figure 4: Loop Which Acts as a Rank Sink

where $E(u)$ is some vector over the web pages that corresponds to a source of rank (see Section 6). Note that if E is all positive, c must be reduced to balance the equation. Therefore, this technique corresponds to a decay factor. In matrix notation we have $R' = c(AR' + E)$. Since $||R'||_1 = 1$, we can rewrite this as $R' = c(A + E \times \mathbf{1})R'$ where $\mathbf{1}$ is the vector consisting of all ones. So, R' is an eigenvector of $(A + E \times \mathbf{1})$.

2.5 Random Surfer Model

The definition of PageRank above has another intuitive basis in random walks on graphs. The simplified version corresponds to the standing probability distribution of a random walk on the graph of the Web. Intuitively, this can be thought of as modeling the behavior of a "random surfer". The "random surfer" simply keeps clicking on successive links at random. 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. The additional factor E can be viewed as a way of modeling this behavior: the surfer periodically "gets bored" and jumps to a

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase® Smarter legal research.