

William Yip
Dr. Luz Quiroga
LIS 678 Personalized Information Delivery
October 11, 2008

GOOGLE PAGERANK ALGORITHM

Introduction

Many of us are fascinated with the success of the Google search engine since its inception in 1998. What was the landscape of the World Wide Web back then? What made the Google search engine unique among all of the other search engines at the time? The answer lies in Google's PageRank algorithm. In 1998, Sergey Brin and Lawrence Page, two Stanford University doctorate students, presented their search engine prototype called Google at the Seventh International World Wide Web conference (WWW98) in Brisbane, Australia. The Google search engine is based primarily on the PageRank algorithm, which is different from the three traditional information retrieval (IR) models: boolean model, vector space model, and probabilistic model. In their own words, Brin and Page defined PageRank as "an objective measure of its citation importance that corresponds well with people's subjective idea of importance" (Brin & Page, 1998). This paper presents an overview of the unique problems in IR on the web, design goals of the Google search engine, description of the PageRank algorithm, and system architecture of the initial prototype.

Problems

Although the World Wide Web can be considered a very large document collection in a loose sense, it has some unique properties. First of all, webpages do not typically receive the same level of peer reviews as most academic papers and journal articles do in traditional document collections (Page et al, 1998). Or to simply put it: not all webpages are created equal. Quality is of particular concern because webpages can be easily created either manually by hand or automatically by a simple program. A obscure blog entry about President George W. Bush should not be rated as high quality as the official biography released from the White House. Secondly, the amount of content on the web is many times more than all traditional document collections combined (Brin & Page, 1998; Langville & Meyer, 2006). A search engine for the web must then be scalable to index millions of webpages, a portion of which may contain syntax errors and/or dead links. Finally, webpage contents are very dynamic compared to traditional document collection (Langville & Meyer, 2006). Consider the minute-by-minute breaking news that are available on major news websites; and compare this to the relatively static nature of most library collections. The ever-changing web requires a search engine to constantly index web content to provide accurate search results. These are some of the properties of the web that the PageRank algorithm was designed to overcome.

Design Goals

The primary design goal of the PageRank algorithm was to improve the quality of search results. In their paper, Brin and Page illustrated an example that only one out of four commercial engines back in 1997 were able to find itself in their respective top 10 results (Brin & Page, 1998). Recognizing

that searchers typically did not go beyond the first few pages of search results; Brin and Page placed more emphasis on precision rather than recall in their design. The PageRank algorithm looks beyond the classic definition of relevance as a simple measure of how well an IR system had matched queries for information with the information contained in the system's database or databases (Bateman, 1998). In the PageRank algorithm, each webpage is assigned with a score that indicates its importance among all webpages. Google was designed to return the list of relevant documents in descending order of importance.

Another design goal of the PageRank algorithm was scalability. In the Google prototype presented in 1998, the search engine indexed at least 24 million web pages (Brin & Page, 1998; Page et al, 1998). With the dynamic nature of webpages, the search engine index must be continuously updated in a timely fashion. Since the search engine were to handle thousands of queries every minute, the querying process must be optimized. The book "Google's PageRank and Beyond" defines the term *query-independence* in which a ranking process is considered query-independent if "the popularity score for each page is determined off-line, and remains constant (until the next update) regardless of the query" (Langville & Meyer, 2006). In the PageRank algorithm, the ranking and querying processes are independent. When a user performs a search in the Google search engine, the querying process retrieves a list of relevant documents and simply performs a series lookups of the page rankings in a data structure.

The last design goal was to open the opportunities for academic research in search engines and large-scaled web data in general (Brin and Page, 1998). Two examples of these researches were: 1) use of PageRank to find the most important academic web pages (Thelwall, 2003); and 2) use of PageRank to measure of impacts of blogs (Kirchoff et al, 2007). It was Brin's and Page's intention to place more emphasis on academic research in search engine technology, making it easier for non-technical users to use.

PageRank Algorithm

The mathematical definition of the PageRank algorithm is as follow (Brin & Page, 1998):

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

where:

$PR(A)$ represents the PageRank of a webpage A ;

$T1 \dots Tn$ represent a series of webpages that link to A ;

$C(T1)$ represents the number of links going out of webpage $T1$;

d represents a damping factor which can be set between 0 and 1.

In short, the PageRank of a webpage is the collective sum of the PageRanks of each of its backlinks (i.e. webpages that link to the webpage in question). It takes care of both of these situations: 1) a webpage has many backlinks with relatively low PageRanks; and 2) a webpage has a few backlinks that have high PageRanks. Brin and Page made the assumption that if a webpage is linked from an

important webpage (e.g. Yahoo - <http://www.yahoo.com>), the webpage itself must be important. This design is also protected against artificially-inflated rankings because of the difficulty and cost (e.g. advertising) to have a link on an important webpage to an unimportant webpage.

The PageRank algorithm can be described using the “random surfer” model (Brin & Page, 1998; Page et al, 1998; Langville & Meyer, 2006). Suppose a surfer lands on a random webpage and follows the links on the webpage randomly. This process goes on indefinitely. It is inevitable that the surfer will eventually return to webpages that he or she has already visited. The PageRank of a webpage is the probability that a surfer visits the particular webpage. The damping factor d is the probability that a surfer “gets bored” and visits another webpage.

A quick glance of the mathematical equation tells us that the calculation of PageRank is circular. Indeed, the calculation of PageRank is an iterative process. Brin and Page described the process as “convergence”. Initially, it is assumed that every webpage has an identical PageRank (i.e. PageRank of $1/n$ where n is the number of documents in the index). The iterative process provides a mechanism of identifying webpages that have higher importance than others. Figure 1 below illustrates a simplified example of PageRank calculation in a single iteration.

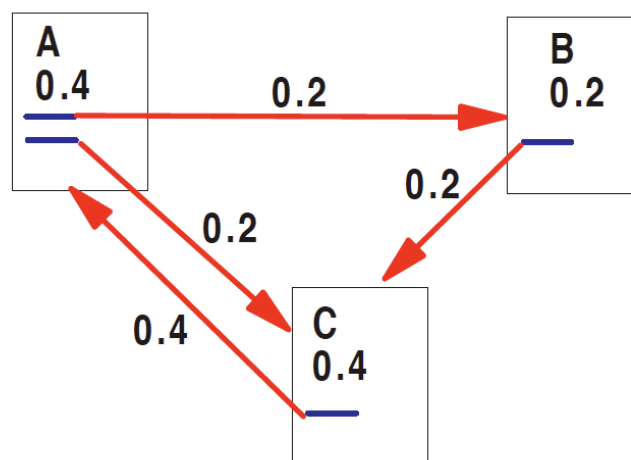


Figure 1 Simplified PageRank Calculation
(Source: The PageRank Citation Ranking: Bringing Order to the Web)

Architecture

The PageRank algorithm was one of many components that made up Google’s initial prototype. Figure 2 below depicts a high-level architecture of the initial prototype. The prototype comprises of three separate processes: crawling, indexing, and searching. This section presents an overview of each of these three processes.

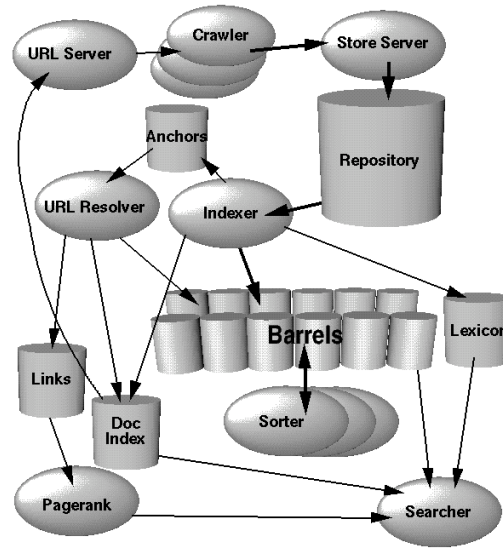


Figure 2 High-Level Google Architecture
 (Source: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*)

Crawling

The Google architecture consists of a series of distributed crawlers that are responsible for downloading webpages. The URL Server first sends a list of URLs to each individual crawlers to download the web content. The content is then forwarded to the Store Server, which compresses the content and stores them in the Repository. To address the issue of scalability, each crawler is designed to keep about 300 simultaneous connections open. In addition, because a Domain Name Server (DNS) lookup is quite time consuming, each crawler maintains its own DNS cache so that a lookup is not necessary.

One of the problems with the PageRank algorithm is “dangling links”. Dangling links are links that point to webpages that do not have any outgoing links. It contradicts with the “random surfer” model because it assumes that the surfer can continuously navigate to another webpage without clicking on the “Back” button. Based on the mathematical equation of the PageRank algorithm, the webpages that these dangling links point to do not have any effect of PageRank of other webpages. As a result, these dangling links are temporarily removed from the URL Server and added back in after the PageRank of all other webpages are calculated.

Indexing

The indexing process in the Google architecture performs several functions. The Indexer first retrieves the webpage content from the Repository and decompresses them. It then parses the webpages and convert them into a series of word occurrences called “hit list”, which are placed in a series of Barrels. A hit list contains a list of words found on a webpage, their corresponding position on the webpage, font and capitalization information. These hit lists, sorted by the Sorter, are used to compute the IR score of each relevant webpage during the searching process. This IR score is then combined with the PageRank for the webpage to determine the final ranking of the

search results. The Lexicon stores the identifiers (i.e. wordID) of every word in the Barrels. In addition to words, the Indexer also parses the links out of each webpage and place them in the Anchor. The links are eventually converted into fully-qualified URLs by the URL Resolver and stored in the Doc Index.

Searching

The searching process in the Google architecture is designed with scalability in mind. It parses a search query by a user, convert each word of the search query into wordIDs, and performs a lookup of relevant documents in the Barrels. When each relevant document is retrieved, a ranking process takes place. The ranking process computes the IR score of each relevant document and combine it with its corresponding PageRank. The IR score is based on a number of factors: position of word within a webpage, proximity of words if multiple words are specified in a query, font size and capitalization of words. The calculation of PageRank of each webpage is performed independently (i.e. query-independent) to speed up the retrieval process.

Conclusion

The Google search engine has evolved a great deal since its initial prototype in 1998. For example, the IR score that determines the final ranking of the search results is made up of hundreds of benchmarks nowadays. As developers are relentless in finding ways to optimize their webpage content for higher ranking, Google continues to evolve their ranking criteria. However, the concept of PageRank has been relatively unchanged. The idea of quality, or subjective importance of a webpage, is crucial. With the enormous amount of information on the web, a searcher must filter out the low-quality webpages from the entire list of “relevant” results. In spite of its simplicity, Google’s PageRank algorithm has facilitated this process, for ten years and counting.

References

Bateman, J. (1998). Changes in Relevance Criteria: A Longitudinal Study. In Proceedings of the 61st Annual Meeting of the American Society of Information Science (Vol. 35, pp. 23-32).

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.

Kirchhoff, L., Bruns, A., & Nicolai, T. (2007). Investigating the impact of the blogosphere: Using PageRank to determine the distribution of attention. In Association of Internet Researchers Conference, Vancouver (pp. 17-20).

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. New Jersey: Princeton University Press.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

Thelwall, M. (2003). Can Google's PageRank be used to find the most important academic Web pages? *Journal of Documentation*, 59(2), 205-217.