RANKING THE PAGES OF THE WORLD WIDE WEB

DÁNIEL FOGARAS *

Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Phone: +36 1 463 2585, Fax: +36 1 463 3157, E-mail: fd@cs.bme.hu

Abstract

Query search engines are fundamental tools in locating documents related to Web surfers' interests. A Web search engine enumerates no more than few hundreds of documents for any key word search query. The quality of a search engine largely depends on its ranking algorithm, the heuristics applied for selecting the hit list from the pages containing the key word. This extended abstract discusses general issues about ranking algorithms, then focuses on PageRank, the ranking algorithm implemented in Google. We provide an alternative formulation of PageRank and discuss some further extensions derived from our formulation. Finally, the performance of the modified algorithms is evaluated and compared with other well-known ranking schemes in novel graph based experiments computed on a collection of approximately one million documents downloaded from the Irish domain.

I. Introduction

The World Wide Web is the most extensively developing database with over three billion pages in various languages, topics, styles and lengths. Finding information relevant to one's interest seems hopeless without *datamining* tools supporting navigation in the jungle of Web pages. Suppose we are looking up web pages containing a *query string* for example "travel". The phrase is submitted to the query search engine like `www.google.com` that maintains a local database containing a possible outdated copy of the whole Web. The pages were downloaded by a so called *crawler* that automatically explores the Web by following the hyperlinks of already accessed pages. See [1, 2] for details about downloading and refreshing the entire Web, and [3, 4] for architectures of web search engines. After the query string is submitted, the search engine extracts from its local database the *query set*—the set of pages in which the query string occurs. Notice that the size of a query set is likely to reach tens of thousands or even higher; for instance the phrase "travel" occurs in approximately 70 million documents. Then non-negative weights, *ranks* are assigned to each page of the query set, measuring the quality and the relevance of pages in the query. Finally, the *hit list* of length 100-1000 is returned to the user containing the addresses of the top ranked pages.

Ranking algorithms are fundamental component of search engines, since ranks determine which parts of query sets will be explored by the users. Millions of travelers visit the top hit `www.travelocity.com`, but they rarely find the less useful `www.traveleye.com` occurring in the 800th position of the list. In the rest of the introduction the basic approaches to ranking algorithms are discussed.

- **Textual content vs. hyperlinks.** The first type of ranking algorithms evaluates scores from the *textual content* of Web pages. Other ranking algorithms employ the *hyperlink content* of Web pages. The existence of a hyperlink implies that one page recommends the other to visit, so it can be regarded as a vote for the linked page. The hyperlink structure turned out to be a rich source of information for ranking algorithms [5, 6, 7, 8]. In compared with the textual content, hyperlinks are more homogeneous not relaying on languages and authors' styles.

*Research is supported by grants OTKA T 42559 and T 042706 of the Hungarian National Science Fund.

- **On-line vs. off-line.** The *on-line* ranks are assigned to the pages, just after the query string was entered to the search engine [9]. While *off-line* or *static* ranks are computed on the whole database of documents in advance [5]. Then the same overall scores of pages are applied in each key word search query. On-line strategies need to be very fast and simple, since the response time is limited to fractions of a second, while off-line strategies can be more time-consuming.
- **Authority vs. hub scores.** The pages with most influential content or descriptive information are referred to as *authorities*. The *hub* pages, on the other hand, are high quality collections of hyperlinks serving good starting points for exploring some topic. Distinguishing between hub and authority scores of a page first appears in the seminal paper of [9].

II. PageRank and its extensions

Now, we turn to the formulation of PageRank the link-based, off-line, authority scores applied in Google. Our description of PageRank provides an alternative explanation to the original random surfer model [5], furthermore we introduce some new PageRank variants that are natural consequences of the novel path-counting formulation of PageRank. The formal proof of the equivalence of the original PageRank and the ranking scheme introduced below can be found in [10], and a slightly different proof in [11] with further applications of the equivalence. For the numerical computation of PageRank scores we refer to [5, 12].

The *PageRank score* of a page p is defined as the weighted in-degree of the search graph of the Web, i.e., it is the weighted sum of all edges pointing to p. The *search graph* $G_S(V, E_S, w)$ has vertices corresponding to Web pages and each directed edge $u \to v \in E_S$ represents a search path through hyperlinks from page u with target page v. (Since the search paths may contain loops, the graph can have infinitely many edges.) Finally w assigns a real weight to each edge of the search graph. The weight $w(u \to v)$ corresponding to a search path $u \to v$ is the product of the following terms:

- $d(1-d)^{\ell(u\to v)}$, where $\ell(u \to v)$ denotes the length of the search path $u \to v$, and the *damping factor* d denotes an overall parameter; (The most common setting is $d = 0.15$, see [5].)
- the inverses of the numbers of outgoing links for all pages occurring in the path $u \to v$;
- *start probability* $s(u)$ of the first node of the search path that is an overall parameter of page *u*. In the simplest setting $s(u)$ is identical over the pages.

Intuitively, PageRank algorithm will give credit for popular, authoritive pages that are end points of a large amount of high weight search paths, where the weight of a path models the value of the path for searching. The iterative algorithm for computing PageRank scores is detailed in [5].

Further variations

From the above definition of PageRank, we see that the rank of each page is evaluated as the weighted sum of search paths with common end vertex. Applying this fact to modified graphs with different parameters yields meaningful novel extensions of PageRank; further extensions can be found in [10].

The first variations apply PageRank to compute overall hub scores for each Web page. Recall that a page should possess large hub score, if it serves as a valuable starting page for browsing in the sense that a large amount of content can be accessed within a few clicks. This desire is fulfilled by the Reverse PageRank scoring scheme introduced below.

- **Reverse PageRank (RPR)** reverses all the hyperlinks occurring in the local database, and then evaluates PageRank on the reversed link structure. It follows from the above formulation of PageRank that the RPR score of a page equals to the weighted sum of all search paths departing from the page, thus RPR performs as an overall hub ranking algorithm of Web pages.

  The start nodes in the reversed structure turns back to target pages in the original hyperlink structure. Therefore, the parameter $s(u)$ of a page u translates to target weight that is given for a search paths for hitting the target u. In the simplest case we set $s(u)$ identical over the Web pages treating equally all the pages as targets of search paths.

- **Popular Reverse PageRank** algorithm first computes PageRank scores with uniform start weights. Then the PR scores are used as target weights of pages in Reverse PageRank computation. By the assumption that ordinary PageRank measures the quality of pages, popular RPR will be raised for those pages from which a large amount of high quality content can be accessed within short click streams.
- **Product PageRank** score of each page is defined as the product of PageRank and reverse PageRank values. The product of the two values acts as a trade-off between hub and authority scores. Web pages possessing high product PageRank are both valuable hubs and authorities, so the numbers of in-coming and out-going paths are both large. We believe that such pages play an important role in maintaining the connectivity of the Web graph as verified in the next section by numerical experiments.

III. Comparing ranking algorithms

Finally we face the problem of evaluating and comparing the performance of the above introduced ranking algorithms. Unfortunately, there is no standardized method for qualifying the output of ranking algorithms; typical approaches are based on user studies and query examples, see for example the detailed survey [8]. For query examples of the above introduced PageRank algorithms we refer to [10].

In this extended abstract we propose *centrality analysis* as a graph-based tool to provide quantitative justification of the above introduced ranking algorithms. The key idea is to measure the centrality of top ranked pages in the hyperlink structure of the Web. According to our knowledge, such connectivity based experiments are novel in the information retrieval literature. The techniques were inspired by the measurements of [13] for different purposes.

The hyperlink structure of the Web is modeled by the *Web graph* with vertices corresponding to the Web pages and directed edges corresponding to the hyperlinks. The *distance* from page p to q is the smallest number of clicks on hyperlinks leading from p to q. Following [13] we refer to the average of all the distances over all pairs of pages as the *diameter* of the web graph. Furthermore the *domination* of a set C is defined as the average distance from C to all the other web pages. To avoid the problem of infinite distances in any of the above definitions, the average is computed as the harmonic mean of distances.

The numerical experiments were conducted on the Web graph of 986,207 Irish pages crawled from the `.ie` domain in October, 2002. We believe that the structure and diversity of this domain is similar to that of the whole WWW. In our first centrality experiments the dominations of subsets of top ranked pages with sizes $100 - 3000$ are evaluated. Large domination means that the ranking algorithm is able to find graph theoretically good sets of hubs. The results are depicted on the left of Fig. 1 for the above introduced hub ranking algorithms RPR and popular RPR. According to the results RPR scores outperform either PR or out-degree rank, the simplest hub ranking that counts the number of hyperlinks on each page.

In our second type of experiments we successively removed the sets of top ranked pages with sizes $100 - 1000$ and evaluated the increasing diameter of the remaining structure. The decay of the connectivity testifies the ability of the ranking algorithm to filter web pages that are typical intermediate pages of search paths. Such pages may also serve as good shortcuts when browsing the Web. According to the right part of Fig. 1 Product PR turns out to be the strongest destructor fulfilling our assumptions of the previous section. Besides the PR variations *degree rank* is also depicted referring to the ranking algorithm that counts all the edges adjacent to the given vertex. Similar experiment was introduced in [13] for the subsets of largest degree vertices and for randomly chosen subsets. They concluded that the web graph is extremely fragile for the removal of highest degree pages. Our results is also interesting from this point of view showing that the connectivity of the Web graph can be even more destroyed by removing the pages with top Product PR scores.

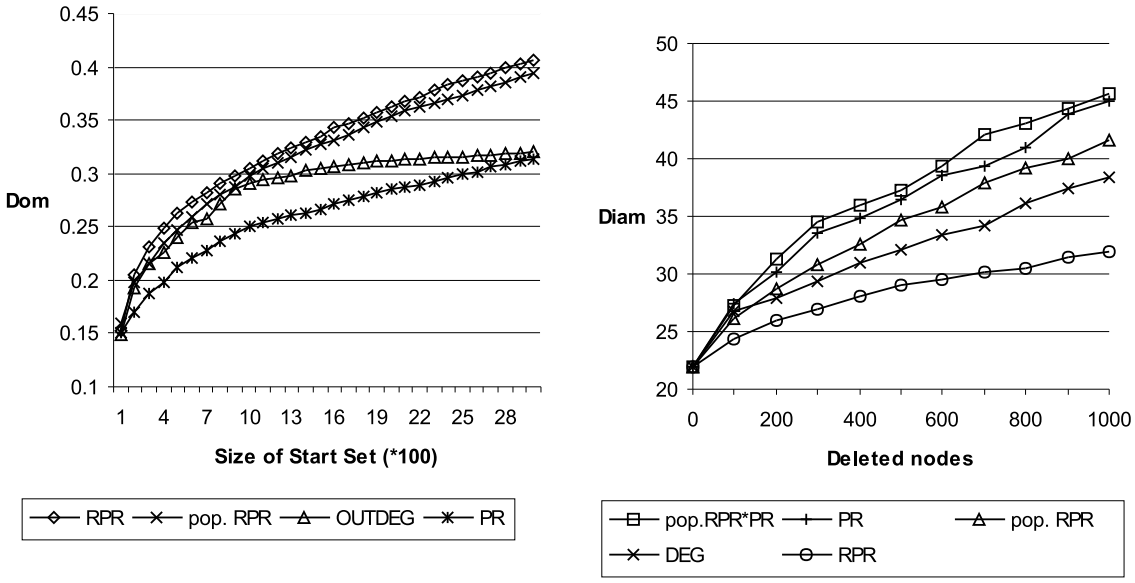

Figure 1: Domination and the increased diameter after the removal of top ranked pages.

IV. Acknowledgment

I wish to thank Katalin Friedl, András Benczúr for valuable discussions and for improving the level of this manuscript. I am also glad to András Lőrincz for supporting the research in its early stage.

References

[1] J. Cho and H. Garcia-Molina, "Synchronizing a database to improve freshness," in *Proceedings of the International Conference on Management of Data*, pp. 117–128, 2000.

[2] E. G. Coffman, Z. Liu, and R. R. Weber, "Optimal robot scheduling for web search engines," Tech. Rep. RR-3317, INRIA, 1997.

[3] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the web," *ACM Transactions on Internet Technology (TOIT)*, 1(1):2–43, August 2001.

[4] L. Huang, "A survey on web information retrieval technologies," Tech. Rep., ECSL, 2000.

[5] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[6] B. D. Davison, A. Gerasoulis, K. Kleisouris, Y. Lu, H. ju Seo, W. Wang, and B. Wu, "Discoweb: Applying link analysis to web search," in *Proceedings of the 8th World Wide Web Conference, Toronto, Canada*, 1999.

[7] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," in *9th International World Wide Web Conference*, 2000.

[8] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas, "Finding authorities and hubs from link structures on the world wide web," in *10th International World Wide Web Conference*, pp. 415–429, 2001.

[9] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, 46(5):604–632, 1999.

[10] D. Fogaras, "Where to start browsing the web?," in *Proceedings of the Innovative Internet Computing Systems Workshop (I2CS)*, LNCS. Springer, 2003.

[11] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the 12th World Wide Web Conference (WWW)*, 2003.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Tech. Rep., Stanford Digital Library Technologies Project, 1998.

[13] R. Albert, H. Jeong, and A. Barabási, "Error and attack tolerance of complex networks," *Nature*, 406:378–382, 2000.