

Hyperlink Analysis on the World Wide Web

Monika Henzinger
Google Inc
Freigutstr. 12
8002 Zurich, Switzerland
monika@google.com

ABSTRACT

We give a short survey of the use of hyperlink analysis in web search engine ranking. Additionally we sketch other applications of hyperlink analysis in the web space.

Categories and Subject Descriptors

H.5 [Information Interfaces and Presentation]: Hypertext/Hypermedia; H.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Hyperlink Analysis, World Wide Web

1. INTRODUCTION

Pre-web information retrieval¹ showed that retrieval techniques based only on statistics of words in the document text perform well on relatively homogenous collections of high-quality papers, like collections of news paper articles and collections of scientific articles. With the web a huge number of documents of various quality levels became available. Still the first web search engines employed the “classic” text-only retrieval strategies as these were the dominant methods in the literature. However, it soon became clear that they performed very poorly in this setting for the following reasons: (1) Some authors of web pages deliberately tried to manipulate the information retrieval strategies to list their pages first. For text-only strategies this was often straightforward: For example, simply repeating terms could improve the ranking of the page on the results page. (2) There were often a very large number of documents containing the query terms so that it became impossible for the user to review all such documents. Thus, the result set needed to be restricted to the most authoritative documents on the query topic. (3) The users of web search engines do not issue long, carefully crafted queries. The average query consists of about 2.5 terms and 80% of the queries do not use any operators[22].

¹See e.g. [20]

The web, however, has additional information that most previously studied collections did not have: Web pages can reference each other using hyperlinks. Both, Brin and Page [4] and Kleinberg [16] proposed independent of each other to exploit the hyperlink structure of the web to improve the quality of search results of web search engines. As they and various papers since then showed the hyperlink structure is very useful in finding the most authoritative documents for a query.

Why are hyperlinks useful? The hyperlink functionality (bringing the web surfer from one page to the next) by itself is not useful for finding the most authoritative documents. However, the way web page authors tend to use hyperlinks can make them valuable. Authors usually create hyperlinks to either help their readers navigate their site or to augment the content of their page by the material in the referenced pages. In the latter case the authors tend to reference authoritative pages that are on the same topic as the source of the hyperlink. Thus, this kind of hyperlink can be seen as a *recommendation* of the page that the hyperlink points. The basic idea behind using hyperlinks for web information retrieval is simply that the more hyperlinks of the latter type a web page has the more authoritative it is. The approach by Brin and Page simply adds recursion to this idea: The *PageRank* score of a page depends on the number of hyperlinks and of the *PageRank* scores of the pages containing the hyperlinks.

Even though this idea seems simple, it is very powerful: PageRank has been used in the Google search engine since its inception and a large field of research of the analysis of the hyperlink structure of the web and its use for web information retrieval has developed.

2. ASSUMPTIONS

There are different approaches for analyzing the structure of the hyperlinks. Common to them is that they all assume that authors tend to create hyperlinks in a specific way. Specifically, they make either one or both of the following assumptions:

- Assumption 1. A hyperlink from page *A* to page *B* is a recommendation of *B* by the author of *A*.
- Assumption 2. If page *A* contains a hyperlink to page *B* then the two pages are on related topics.

Note that Assumption 1 basically implies that all hyperlinks are created to augment the content of the page and none are created to navigate the site. This does obviously not hold. However, Assumption 1 still seems to be “close enough” to the truth so that techniques that are based on it work well in practice.

The main use of hyperlink analysis in Web information retrieval is in *ranking* search results. When a user sends a query to a search engine, the search engine determines all documents match all terms². *Ranking* is the process of ordering these documents in decreasing order of relevance, that is, so that the documents that best answer the query are on top. We will describe other uses of hyperlink analysis in Section 4

3. HYPERLINK ANALYSIS FOR RANKING IN WEB SEARCH ENGINES

Any ranking algorithm that is based purely on information controlled by the author, like the text of the page, is susceptible to manipulation by the page author. The strength of hyperlink analysis comes from the fact that the ranking of a page is dependent on the content of other pages and, thus, hopefully dependent on information *not* controlled by the creator of the page. However, the longer hyperlink analysis has been used, the more have creators (or companies hired by the creators) started to manipulate the hyperlink structure pointing to their pages. Wu and Davison [24] recently showed how to identify common forms of this manipulation.

Hyperlink-based ranking algorithms come in two classes:

- *query-independent* algorithms, which assign a score to each page independent of a specific query; and
- *query-dependent* algorithms, which assign a score to each page dependent on a specific query

The advantage of query-independent schemes is that the hyperlink analysis need to be done only once, namely right after the collection of web pages was compiled. The analysis assigns a score to a document and this score is then used for all subsequent user queries. Query-dependent schemes on the other side require a hyperlink analysis for each query. However, the advantage of query-dependent schemes is that they can take the query into account, for example, by performing the hyperlink analysis only on the pages that are related to the query. Thus, unlike the query-independent schemes, query-dependent schemes do not require a combination with query-dependent approaches like the above text-only ranking.

Both schemes model the web in the following way as a graph: Each page is modeled as a node in the graph. There is a directed edge from page *A* to page *B* if and only if page *A* has a hyperlink to page *B*. Query-dependent schemes restrict the nodes in the graph to pages that are related to the query, while query-independent schemes use all available pages.

²Some search engines also return documents that contain only some of the query terms.

3.1 Query-independent ranking

The goal of query-independent ranking is to assign a score to each page that measures the intrinsic quality or authority of the page. At query time, this score needs to be combined with some query-dependent ranking criteria to rank all documents matching the query. For example, it can be used together with the “classic” text-only ranking algorithm.

According to the assumptions behind hyperlink analysis a page *A* to which many hyperlinks point is likely to be more authoritative than a page to which few hyperlinks point. However, an approach that just counts the hyperlinks can be easily manipulated and it ignores the fact that the pages pointing to *A* also can be different quality. Brin and Page [4] proposed instead the following recursive approach: The authority score, called *PageRank*, of a page *A* depends on the authority scores of the pages pointing to *B*. More specifically, the PageRank $R(A)$ of page *A* is defined as

$$R(A) = \delta/n + (1 - \delta) \sum_{\text{hyperlink } (B,A)} R(B)/\text{outdegree}(B),$$

where δ is a constant usually chosen between 0.1 and 0.2, n is the number of pages in the collection, and $\text{outdegree}(B)$ is the number of hyperlinks on page *B*.

This formula implies that each page *A* collects some PageRank from every page pointing to *A* and each page *B* “distributes” an equal amount of its PageRank to all the pages it references. Note that each page generates one linear equation. To compute the PageRank for all pages on the web a huge set of linear equations, one per page, needs to be solved.

Even though the formula is fairly simple and can be obviously improved upon in various ways it works surprisingly in distinguishing high-quality from low-quality web pages, i.e., in determining authoritative pages.

3.2 Query-dependent ranking

In query-dependent ranking each page is assigned a score that measures its quality as well as its relevance to the user query. This score can but does not have to be combined with further query-dependent approaches. The basic idea is to build for each query a small subgraph of the whole web graph, called a *neighborhood graph*, and perform hyperlinks analysis on it. How to construct this graph is crucial as only documents contained in the graph will be returned for the user query. Ideally, this graph contains the most authoritative pages on the query topic and few pages on other topics.

Carriere and Kazman [6] propose the following approach to construct a neighborhood graph:

- A *start set* of documents matching the query is determined. Specifically, they propose to choose the top 200 matches returned by a search engine for the query.
- This start set is augmented by its “neighborhood”, which is the set of documents that either contain hyperlinks to a document in the start set or are referenced by documents in the start set.

- Each document in the start set and its neighborhood is modeled by a node in the neighborhood graph. Two nodes are considered to be *unaffiliated* if they are on different web hosts. There exists an edge from node A and node B if and only if A and B are unaffiliated and document A has a hyperlink to document B . Edges between affiliated nodes are not allowed to limit the manipulation of the score.

Given the neighborhood graph various ranking approaches were suggested. Carriere and Kazman proposed simply to count the number of hyperlinks in the neighborhood graph. Computing the PageRank on the neighborhood graph produces a very similar ranking [1], mostly because the neighborhood graph is usually small (on the order of a thousand nodes) and the impact of the global web is lost.

Kleinberg proposed another approach of ranking pages in the neighborhood graph. It assumes that a topic can be roughly divided into pages with good content, called *authorities*, and pages with good references, called *hubs*. To determine the authorities and hubs the algorithm iteratively computes hub and authority scores for each node. The computation of the hub scores depends on the authority scores computed in the previous iteration and the computation of the authority scores depends on the hub scores of the previous iteration. The intuition is that high authority nodes should be pointed to by many hubs and especially by the good hubs. Good hub nodes on the other side should point to high authority nodes. This leads to the following algorithm.

1. Construct the neighborhood graph with node set V and edge set E .
2. For every node A in V , initialize its hub score $Hub[A]$ to 1 and its authority score $Aut[A]$ to an arbitrary value.
3. While the vectors Hub and Aut have not converged:
 4. For all A in V , $Aut[A] = \sum_{(B,A) \in E} Hub[B]$
 5. For all A in V , $Hub[A] = \sum_{(A,B) \in E} Aut[B]$
 6. Normalize the Hub and Aut vectors to 1.

Using linear algebra it is straightforward to show that the Hub and Aut vectors will eventually converge, but no bound on the number of iterations required for convergence is known. In practice, they converge quickly.

As mentioned before the success of this approach is very dependent on the quality of the neighborhood graph. Specifically, the neighborhood graph needs to contain high-quality pages on the query topic. If the majority of the pages is on a different topic then *topic drift* can occur where the top authority pages belong to the different topic. Another disadvantage is that adding a few edges to the neighborhood graph can considerably change the top authority and hub pages since the neighborhood graph is usually fairly small. Thus, there is the risk of manipulation of the results by adding just a few edges.

Many improvements and generalizations have been suggested to the above two algorithms (see e.g. [2, 3, 18, 8, 23, 25]).

4. OTHER APPLICATIONS OF HYPERLINK ANALYSIS

Crawling is the process of collecting pages from the web. The software which performs this process is called *crawler*. All search engines need to crawl the web since unlike in classic information retrieval no collection of documents is given to them. Instead, starting from a small set of *seed pages* crawlers follow the hyperlinks contained in the already crawled pages to find new pages to visit and to extract hyperlinks from. This process continues until no more hyperlinks to unvisited pages are found or a desired number of web pages has been visited. Usually the latter is the case.

Since a crawler usually does not visit all possible pages, it has to determine in which order to visit the yet unvisited pages to which a hyperlink was discovered. The goal is to preferentially crawl “high quality” web pages. Hyperlink analysis provides a means to estimating the quality of web pages. The first assumption of hyperlink analysis implies that pages pointed to by many other pages are of higher quality than pages pointed to by fewer pages. Thus, the crawler can simply keep track of how often it has seen a hyperlink to each page and preferentially crawl pages with a large count. Alternatively, a PageRank score can be computed using all the hyperlinks found so far and then the un-crawled pages with high PageRank can be preferentially be crawled [9].

Hyperlink-analysis was also used for a search-by-example approach to search: given one page find pages related to it. Kleinberg [16] proposed using the HITS algorithm for this problem and Dean and Henzinger [10] show that both the HITS algorithm and a simple algorithm on the co-citation graph perform very well. The idea behind the latter algorithm is that frequent co-citation is a good indication of relatedness and thus the edges with high weight in the co-citation graph tend to connect nodes which are related. Recently, Fogaras and Rácz [12] presented hyperlink-based similarity algorithms that are scalable to huge graphs.

Extensions of Kleinberg’s algorithm and PageRank were used by Rafiei and Mendelzon to compute the reputation of a web pages [19] and by Sarukkai to predict personalized web used [21].

Gibson et al. [13] and Kumar et al [17] analyzed the graph structure of the web to find “web communities”, groups of web pages on the same or closely-related topics.

Chakrabarti et al [7] made first steps towards using the hyperlink structure for web page categorization.

Henzinger et al [14, 15] used PageRank-like random walks to sample web pages almost according to the PageRank distribution and the uniform distribution, respectively. The goal was to compute various statistics on the web pages and to compare the quality, respectively the number, of the pages in the indices of various commercial search engines.

Buyukkokten et al. [5] and Ding et al. [11] classified web

pages based on their geographical scope by analyzing the links to point to the pages.

5. REFERENCES

- [1] B. Amento, L. Terveen, and W. Hill. Does authority mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd International ACM SIGIR Conference in Research and Development in Information Retrieval (SIGIR 00)*, pages 296–303, New York, USA, 2000. ACM Press.
- [2] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 111–104, 1998.
- [3] P. Boldi, M. Santini, and S. Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th International World Wide Web Conference (WWW2005)*, pages 557–566, May 2005.
- [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual (web) search engine. *Computer Networks and ISDN Systems (Proceedings of WWW7)*, 30(1-7):107–308, 1998.
- [5] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In *WebDB (Informal Proceedings)*, pages 91–96, 1999.
- [6] J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through connectivity. In *Proceedings of the Sixth International World Wide Web Conference*, pages 701–711, Santa Clara, California, April 1997.
- [7] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 65–74, 1998.
- [8] S. Chakrabarti, B. Dom, R. P., S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems (Proceedings of WWW7)*, 30(1-7):65–74, 1998.
- [9] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. *Computer Networks and ISDN Systems (Proceedings of WWW7)*, 30(1-7):161–172, 1998.
- [10] J. Dean and M. R. Henzinger. Finding related web pages in the world wide web. *Computer Networks and ISDN Systems (Proceedings of WWW8)*, 31:1467–1479, 1999.
- [11] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB'00)*, pages 545–556, 2000.
- [12] D. Fogaras and B. Rácz. Scaling link-based similarity search. In *Proceedings of the 14th International World Wide Web Conference (WWW2005)*, pages 641–650, May 2005.
- [13] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. 9th ACM Conf. on Hypertext and Hypermedia*, 1998.
- [14] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring search engine quality using random walks on the web. *Computer Networks (Proceedings of WWW8)*, 31:1291–1303, 1999.
- [15] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. *Computer Networks (Proceedings of WWW9)*, 33(1-6):295–308, 2000.
- [16] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, January 1998.
- [17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks (Proceedings of WWW8)*, 31(11-16):1481–1493, 1999. Available online at <http://www8.org/w8-papers/-4a-search-mining/trawling/trawling.html>.
- [18] A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2005.
- [19] D. Rafiei and A. Mendelzon. What is this page known for? computing web page reputations. *Computer Networks and ISDN Systems (Proceedings of WWW9)*, 33(1-6):823–836, 2000.
- [20] G. Salton. *The SMART System – Experiments in Automatic Document Processing*. Prentice Hall, 1971.
- [21] R. Sarukkai. Link prediction and path analysis using markov chains. *Computer Networks and ISDN Systems (Proceedings of WWW9)*, 33(1-6):377–386, 2000.
- [22] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large AltaVista query log. *ACM SIGIR Forum*, 33(1):6–12, 1999.
- [23] A. C. Tsoi, G. Norini, F. Scarselli, M. Hagenbuchner, and M. Maggini. Adaptive ranking of web pages. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, pages 356–365, May 2003.
- [24] B. Wu and B. D. Davison. Identifying link farm spam pages. In *Proceedings of the 14th International World Wide Web Conference (WWW2005)*, pages 820–829, May 2005.
- [25] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma, and E. A. Fox. Link fusion: A unified link analysis framework for multi-type interrelated data objects. In *Proceedings of the 13th International World Wide Web Conference (WWW2004)*, pages 319–327, May 2004.