Link Analysis in Web Information Retrieval

Monika Henzinger
Google Incorporated
Mountain View, California
monika@google.com

Abstract

The analysis of the hyperlink structure of the web has led to significant improvements in web information retrieval. This survey describes two successful link analysis algorithms and the state-of-the art of the field.

1 Introduction

The goal of information retrieval is to find all documents relevant for a user query in a collection of documents. Decades of research in information retrieval were successful in developing and refining techniques that are solely word-based (see e.g., [2]). With the advent of the web new sources of information became available, one of them being the *hyperlinks* between documents and records of user behavior. To be precise, *hypertexts* (i.e., collections of documents connected by hyperlinks) have existed and have been studied for a long time. What was new was the large number of hyperlinks created by independent individuals. Hyperlinks provide a valuable source of information for web information retrieval as we will show in this article. This area of information retrieval is commonly called *link analysis*.

Why would one expect hyperlinks to be useful? A hyperlink is a reference of a web page B that is contained in a web page A. When the hyperlink is clicked on in a web browser, the browser displays page B. This functionality alone is not helpful for web information retrieval. However, the way hyperlinks are typically used by authors of web pages can give them valuable information content. Typically, authors create links because they think they will be useful for the readers of the pages. Thus, links are usually either navigational aids that, for example, bring the reader back to the homepage of the site, or links that point to pages whose content augments the content of the current page. The second kind of links tend to point to high-quality pages that might be on the same topic as the page containing the link.

Based on this motivation, link analysis makes the following simplifying assumptions:

- A link from page A to page B is a recommendation of page B by the author of page A.
- If page A and page B are connected by a link the probability that they are on the same topic is higher than if they are not connected.

Copyright 2000 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

Link analysis has been used successfully for deciding which web pages to add to the collection of documents (i.e., which pages to *crawl*), and how to order the documents matching a user query (i.e., how to *rank* pages). It has also been used to categorize web pages, to find pages that are related to given pages, to find duplicated web sites, and various other problems related to web information retrieval.

The idea of studying "referrals" is, however, not new. A subfield of classical information retrieval, called bibliometrics, analyzed citations (see, e.g., [19, 14, 29, 15]). The field of sociometry developed algorithms [20, 25] very similar to the PageRank and HITS algorithms described below. Some link analysis algorithms can also be seen as collaborative filtering algorithms: each link represents an opinion and the goal is to mine the set of opinions to improve the answers to individuals.

This paper is structured as follows. We first discuss graph representations for the web (Section 2). In Section 3 we discuss two types of connectivity-based ranking schemata: a *query-independent* approach, where a score measuring the intrinsic quality of a page is assigned to each page without a specific user query, and a *query-dependent* approach, where a score measuring the quality and the relevance of a page to a given user query is assigned to some of the pages. In Section 4 other uses of link analysis in web information retrieval are described.

2 A Graph Representation for the Web

In order to simplify the description of the algorithms below we first model the web as a graph. This can be done in various ways. Connectivity-based ranking techniques usually assume the most straightforward representation: The graph contains a node for each page u and there exists a directed edge (u, v) if and only if page u contains a hyperlink to page v. We call this directed graph the *link graph* G.

Some algorithms make use of the undirected *co-citation graph*: As before each page is represented by a node. Nodes u and v are connected by an undirected edge if and only if there exists a third node x linking to both u and v.

The link graph has been used for ranking, finding related pages, and various other problems. The co-citation graph has been used for finding related pages and categorizing pages.

3 Connectivity-Based Ranking

3.1 Query-Independent Connectivity-Based Ranking

In *query-independent* ranking a score is assigned to each page without a specific user query with the goal of measuring the intrinsic quality of a page. At query time this score is used with or without some query-dependent criteria to rank all documents matching the query.

The first assumption of connectivity based techniques immediately leads to a simple query-independent criterion: The larger the number of hyperlinks pointing to a page the better the page. The main drawback of this approach is that each link is equally important. It cannot distinguish between the quality of a page pointed to by a number of low-quality pages and the quality of a page that gets pointed to by the same number of high-quality pages. Obviously, it is therefore easy to make a page appear to be high-quality – just create many other pages that point to it.

To remedy this problem Brin and Page [5, 26] invented the PageRank measure. The PageRank of a page is computed by weighting each hyperlink proportionally to the quality of the page containing the hyperlink. To determine the quality of a referring page, they use its PageRank recursively. This leads to the following definition of the PageRank $R(p)$ of a page p:

$$R(p) = \epsilon/n + (1 - \epsilon) \cdot \sum_{(q,p) \in G} R(q)/outdegree(q),$$

where

- ϵ is a dampening factor usually set between 0.1 and 0.2;
- n is the number of nodes in G; and
- $outdegree(q)$ is the number of edges leaving page p, i.e., the number of hyperlinks on page q.

Alternatively, the PageRank can be defined to be the stationary distribution of the following infinite random walk $p_1, p_2, p_3, \ldots$, where each p_i is a node in G: Each node is equally likely to be the first node p_1. To determine node p_{i+1} with $i > 0$ a biased coin is flipped: With probability ϵ node p_{i+1} is chosen uniformly at random from all nodes in G, with probability $1 - \epsilon$ node p_{i+1} is chosen uniformly at random from all nodes q such that edge (p_i, q) exists in G.

The PageRank is the dominant eigenvector of the probability transition matrix of this random walk. This implies that when PageRank is computed iteratively using the above equation, the computation will eventually converge under some weak assumptions on the values in the probability transition matrix. No bounds are known on the number of iterations but in practice roughly 100 iterations suffice.

The PageRank measure works very well in distinguishing high-quality web pages from low-quality web pages and is used by the Google[1] search engine.

The PageRank algorithm assigns a score to each document independent of a specific query. This has the advantage that the link analysis is performed once and then can be used to rank all subsequent queries.

3.2 Query-Dependent Connectivity-Based Ranking

In *query-dependent* ranking a score measuring the quality and the relevance of a page to a given user query is assigned to some of the pages.

Carriere and Kazman [11] proposed an indegree-based ranking approach to combine link analysis with a user query. They build for each query a subgraph of the link graph G limited to pages on the query topic. More specifically, they use the following query-dependent *neighborhood graph*. A *start set* of documents matching the query is fetched from a search engine (say the top 200 matches). This set is augmented by its *neighborhood*, which is the set of documents that either point to or are pointed to by documents in the start set. Since the indegree of nodes can be very large, in practice a limited number of predecessors (say 50) of a document are included. The neighborhood graph is the subgraph of G induced by the documents in the start set and its neighborhood. This means that each such document is represented by a node u and there exists an edge between two nodes u and v in the neighborhood graph if and only if there is a hyperlink between them. The indegree-based approach then ranks the nodes by their indegree in the neighborhood graph. As discussed before this approach has the problem that each link counts an equal amount.

To address this problem, Kleinberg [21] invented the *HITS* algorithm. Given a user query, the algorithm first iteratively computes a *hub* score and an *authority* score for each node in the neighborhood graph[2]. The documents are then ranked by hub and authority scores, respectively.

Nodes, i.e., documents that have high authority scores are expected to have relevant content, whereas documents with high hub scores are expected to contain hyperlinks to relevant content. The intuition is as follows. A document which points to many others might be a good hub, and a document that many documents point to might be a good authority. Recursively, a document that points to many good authorities might be an even better hub, and similarly a document pointed to by many good hubs might be an even better authority. This leads to the following algorithm.

(1) Let N be the set of nodes in the neighborhood graph.

[1] `http://www.google.com/`

[2] In the HITS algorithm the neighborhood graph is slightly modified to exclude edges between nodes on the same host. The reason is that hyperlinks within the same host might be by the same author and hence might not be a recommendation.

(2) For every node n in N, let $H[n]$ be its hub score and
$A[n]$ its authority score.
(3) Initialize $H[n]$ to 1 for all n in N.
(4) While the vectors H and A have not converged:
(5) For all n in N, $A[n] := \sum_{(n',n)\in N} H[n']$
(6) For all n in N, $H[n] := \sum_{(n,n')\in N} A[n']$
(7) Normalize the H and A vectors.

Since this algorithm computes the dominant eigenvectors of two matrices, the H and A vectors will eventually converge, but no bound on the number of iterations is known. In practice, the vectors converge quickly.

Note that the algorithm does not claim to find *all* valuable pages for a query, since there may be some that have good content but have not been linked to by many authors or that do not belong to the neighborhood graph.

There are two types of problems with this approach: First, since it only considers a relatively small part of the web graph, adding edges to a few nodes can potentially change the resulting hubs and authority scores considerably. Thus it is easier for authors of web pages to manipulate the hubs and authority scores than it is to manipulate the PageRank score. See [23] for a more extensive discussion of this problem. A second problem is that if the neighborhood graph contains more pages on a topic different from the query, then it can happen that the top authority and hub pages are on this different topic. This problem was called *topic drift*. Various papers [7, 8, 4] suggest the use of edge weights and content analysis of either the documents or the anchor text to deal with these problems. In a user study [4] it was shown that this can considerably improve the quality of the results.

A recent paper by Lempel and Moran [23] gives anecdotal evidence that a variant of the indegree-based approach achieves better results than the HITS algorithm. They compute the stationary distribution of a random walk on an auxiliary graph. This corresponds to scaling the indegree of a node u in the link graph by the relative size of u's connected component in the co-citation graph and the number of edges in u's component in the auxiliary graph. Basically, each link is weighted and the quality of a page is the sum of the weights of the links pointing to it. However, more experimental work is needed to evaluate this approach.

3.3 Evaluation of Query-Dependent Rankings

Amento, Terveen, and Hill [1] evaluated different link-based ranking criteria on a graph similar to the neighborhood graph. They start from a seed-set of relevant pages for a given query and their goal is to rank them by quality using various criteria.

The seed-set has the property that no url in the seed-set is the prefix of another one. They consider these urls to be *root urls* of *sites*: all pages which contain the root url as prefix belong to the site of this root url. Then they perform a neighborhood expansion using link and text similarity heuristics and restricting the expansion to pages on the above sites. For their analysis they use either this graph or a *site graph*, where all pages on a site are collapsed to one node. Note that the set of nodes in the site graph is fully determined by the seed-set and the neighborhood expansion is used only to determine the edges in the site graph.

They use five link-based metrics (in-degree, out-degree, HITS authority score, HITS hub score, and PageRank) and some other metrics to rank the root urls by either using the score assigned to the root url (in the pages-based graph) or to the site (in the site graph). Interestingly, the ranking on the site graph outperformed the ranking on the pages-based graph. Furthermore, there is a large overlap and correlation in the rankings of the in-degree, HITS authority score, and PageRank metric and these three metrics perform roughly equally well. They also outperform the other metrics together with another simple metric that counts the number of pages on a site that belong to the graph.

Note, however, that they perform the PageRank computation on a small graph, while the PageRank computation described before was performed on the whole link graph and the resulting PageRank values will most likely differ considerably.

4 Other Uses of Link Analysis in Web Information Retrieval

Apart from ranking, link analysis can also be used for deciding which web pages to add to the collection of web pages, i.e., which pages to crawl. A *crawler* (or *robot* or *spider*) performs a traversal of the web graph with the goal of fetching high-quality pages. After fetching a page, it needs to decide which page out of the set of uncrawled pages to fetch next. One approach is to crawl the pages with highest number of links from the crawled pages first. Cho et al. propose to visit the pages in the order of PageRank [10].

Link analysis was also used for a search-by-example approach to searching: given one relevant page find pages related to it. Kleinberg [21] proposed using the HITS algorithm for this problem and Dean and Henzinger [12] show that both the HITS algorithm and a simple algorithm on the co-citation 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 with are related.

Extensions of the HITS and PageRank approaches were used by Rafiei and Mendelzon to compute the reputation of a web page [27] and by Sarukkai to predict personalized web usage [28].

Almost completely mirrored web hosts cause problems for search engines: they waste space in the index data structure and might lead to duplicate results. Bharat et al. [3] showed that a combination of IP address analysis, URL pattern analysis, and link structure analysis can detect many near-mirrors. The idea is that near-mirrors exhibit as very similar link structure within the host as well as to the other hosts.

Chakrabarti et al. [9] made first steps towards using the link structure for web page categorization.

In [17, 18] PageRank-like random walks were performed on the web to sample web pages almost according to the PageRank distribution and the uniformly 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. [6] and Ding et al. [13] classify web pages based on their geographical scope by analyzing the links that point to the pages.

5 Conclusions

The main use of link analysis is currently in ranking query results. Other areas were link analysis has been shown to be useful are crawling, finding related pages, computing web page reputations and geographic scope, prediction link usage, finding mirrored host, categorizing web pages, and computing statistics of web pages and of search engines.

However, research of the hyperlink structure of the web is just at its beginning and a much deeper understanding needs to be gained.

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 on Research and Development in Information Retrieval (SIGIR'00)*, pages 296–303.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] K. Bharat, A. Z. Broder, J. Dean, and M. Henzinger. A comparison of techniques to find mirrored hosts on the World Wide Web. *Workshop on Organizing Web Space (WOWS)* in conjunction with *ACM Digital Library '99*. To appear in the *Journal of the American Society for Information Science,* 2000.

[4] 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.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World Wide Web Conference* 1998, pages 107–117.

[6] O. Buyukkokten, J. Cho, H. García-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of Web pages. *Proc. of the ACM SIGMOD Workshop on the Web and Databases (WebDB'99)*, 1999.

[7] S. Chakrabarti, B. Dom, D. Gibson, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Experiments in topic distillation. In *ACM–SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*.

[8] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International World Wide Web Conference* 1998, pages 65–74.

[9] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1998, pages 307–318.

[10] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. In *Proceedings of the Seventh International World Wide Web Conference* 1998, pages 161–172.

[11] J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through connectivity. In *Proceedings of the Sixth International World Wide Web Conference* 1997, pages 701–711.

[12] J. Dean and M. R. Henzinger. Finding related Web pages in the World Wide Web. In *Proceedings of the 8th International World Wide Web Conference* 1998, pages 389–401.

[13] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of Web resources. *Proceedings of the 26th International Conference on Very Large Databases (VLDB'00)*, 2000.

[14] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178, 1972.

[15] E. Garfield. *Citation Indexing.* ISI Press, 1979.

[16] T. Haveliwala. Efficient computation of PageRank. Technical Report 1999-31, Stanford University, 1999.

[17] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring search engine quality using random walks on the Web. In *Proceedings of the 8th International World Wide Web Conference* 1999, pages 213–225.

[18] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *Proceedings of the Ninth International World Wide Web Conference* 2000, pages 295–308.

[19] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14, 1963.

[20] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39-43, March 1953.

[21] 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.

[22] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: Measurements, models and methods. Invited survey at the *International Conference on Combinatorics and Computing*, 1999.

[23] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In *Proceedings of the Ninth International World Wide Web Conference* 2000, pages 387–401.

[24] D. S. Modha and W. S. Spangler. Clustering hypertext with applications to Web searching. *Proceedings of the ACM Hypertext 2000 Conference, San Antonio, TX*, 2000. Also appears as IBM Research Report RJ 10160 (95035), October 1999.

[25] M. S. Mizruchi, P. Mariolis, M. Schwartz, and B. Mintz. Techniques for disaggregating centrality scores in social networks. In N. B. Tuma, editor, *Sociological Methodology*, pages 26–48. Jossey-Bass, San Francisco, 1986.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. *Stanford Digital Library Technologies*, Working Paper 1999-0120, 1998.

[27] D. Rafiei, and A. Mendelzon. What is this page known for? Computing Web page reputations. In *Proceedings of the Ninth International World Wide Web Conference* 2000, pages 823–836.

[28] R. Sarukkai. Link prediction and path analysis using Markov chains. In *Proceedings of the Ninth International World Wide Web Conference* 2000, pages 377–386.

[29] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Amer. Soc. Info. Sci.*, 24, 1973.