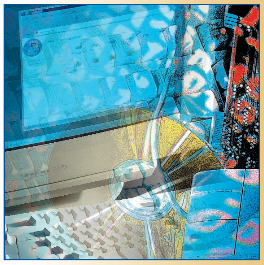

Survey

Hyperlink Analysis for the Web

Hyperlink analysis algorithms allow search engines to deliver focused results to user queries. This article surveys ranking algorithms used to retrieve information on the Web.

Monika R. Henzinger
Google Inc.

Information retrieval is a computer science subfield whose goal is to find all documents relevant to a user query in a given collection of documents. As such, information retrieval should really be called document retrieval. Before the advent of the Web, IR systems were typically installed in libraries for use mostly by reference librarians. The retrieval algorithm for these systems was usually based exclusively on analysis of the words in the document.

The Web changed all this. Now each Web user has access to various search engines whose retrieval algorithms often use not only the words in the documents but also information like the hyperlink structure of the Web or markup language tags.

How are hyperlinks useful? The hyperlink functionality alone—that is, the hyperlink to Web page *B* that is contained in Web page *A*—is not directly useful in information retrieval. However, the way Web page authors use hyperlinks can give them valuable information content. Authors usually create hyperlinks they think will be useful to readers. Some may be navigational aids that, for example, take the reader back to the site's home page; others provide access to documents that augment the content of the current page. The latter tend to point to high-quality pages that might be on the same topic as the page containing the hyperlink. Web information retrieval systems can exploit this information to refine searches for relevant documents.

Hyperlink analysis significantly improves the relevance of the search results, so much so that all major Web search engines claim to use some type of hyperlink analysis. However, the search engines do not disclose details about the type of hyperlink analysis they perform—mostly to avoid manipulation of search results by Web-positioning companies.

In this article, I discuss how hyperlink analysis can be applied to ranking algorithms, and survey other ways Web search engines can use this analysis.

Hyperlink Analysis on the Web

Hyperlink analysis algorithms make either one or both of the following simplifying assumptions:

- *Assumption 1.* A hyperlink from page *A* to page *B* is a recommendation of page *B* by the author of page *A*.

1089-7801/00/$10.00 ©2001 IEEE

- *Assumption 2*. If page *A* and page *B* are connected by a hyperlink, then they might be on the same topic.

The two main uses of hyperlink analysis in Web information retrieval are *crawling* and *ranking*. Other uses of hyperlink analysis include computing the geographic scope of a Web page, finding mirrored hosts, and computing statistics of Web pages and search engines; and are discussed in the sidebar, "Uses of Hyperlink Analysis in Web Information Retrieval" on page 48.

Collecting Web Pages

Crawling is the process of collecting Web pages. Web information retrieval is different from classic information retrieval in that the collection is not simply "given" to a Web search engine, but the search engine has to "find" the documents for the collection. The crawling process usually starts from a set of source Web pages. The Web crawler follows the source page hyperlinks to find more Web pages. Search engine developers use the metaphor of a spider "crawling" along the Web creating hyperlinks. This process is repeated on each new set of pages and continues until no more new pages are discovered or until a predetermined number of pages have been collected. The crawler has to decide in which order to collect hyperlinked pages that have not yet been crawled. The crawlers of different search engines make different decisions, and so collect different sets of Web documents. For example, a crawler might try to preferentially crawl "high quality" Web pages. To do this, it would put all discovered but uncrawled pages into a priority queue ordered by quality.

The power of hyperlink analysis comes from the fact that it uses the content of other pages to rank the current page.

Hyperlink analysis provides a means for judging the quality of pages. For example, the first assumption of hyperlink analysis algorithms implies that pages pointed to by many pages are of higher quality than pages pointed to by fewer pages. This means that the number of hyperlinks to a given page can be used to measure its quality. Alternatively, PageRank (described later in this article) can be used as such a measure.[1]

Ranking Returned Documents

When a user sends a query to a search engine, the search engine returns the URLs of documents matching all or one of the terms, depending on both the query operator and the algorithm used by the search engine. *Ranking* is the process of ordering the returned documents in decreasing order of relevance, that is, so that the "best" answers are on the top. Ranking that uses hyperlink analysis is called *connectivity-based ranking*.

Classic information retrieval usually used ranking algorithms based solely on the words in the documents. One such algorithm is the *vector space model* introduced by Salton and associates.[2] It considers a high-dimensional vector space with one dimension per term. Each document or query is represented as a *term vector* in this vector space. Entries of terms occurring in the document are positive, and entries of terms not occurring in the document are zero. More specifically, the entry of the term is usually a function that increases with the frequency of the term within the document and decreases with the number of documents in the collection containing the term. The idea is that the more documents the term appears in, the less characteristic the term is for the document, and the more often the term appears in the document, the more characteristic the term is for the document. The term vectors of documents might be normalized to one to account for different document lengths. The similarity between a document and a query is usually computed by the dot-product of their term vectors. For a given query, this assigns a nonnegative score to each document. To answer a query, the documents with positive scores are returned in decreasing order of score.

Why do classical information retrieval techniques not work well on the Web? Many Web page authors have a commercial interest in their pages ranking high for certain queries. Thus, to make their pages rank higher, they will modify them in many ways. These attempts sometimes go so far as to add text in an invisible font to manipulate the ranking algorithm. For example, if the vector space model were used for ranking, adding 1,000 repetitions of the word "car" would help the ranking of a given page for the query "car." There are even so-called Web-positioning companies that make money by advising their clients how to manipulate search engine rankings.

Any algorithm that is based purely on the content of a page is susceptible to this kind of manipulation. The power of hyperlink analysis comes from the fact that it uses the content of other pages to rank the current page. Hopefully, these pages were created by authors independent of the author of the original page, thus adding an unbiased factor to the ranking.

Connectivity-Based Ranking

Connectivity-based ranking schemes can be partitioned into two classes:

- *query-independent* schemes, which assign a score to a page independent of a given query; and
- *query-dependent* schemes, which assign a score to a page in the context of a given query.

Query-independent ranking schemes assign a score to a document once and then use this score for all subsequent queries. Query-dependent ranking schemes require a hyperlink analysis for each query but tailor their hyperlink analysis specifically to the query.

To simplify the description of the algorithms presented here, we must first model a collection of Web pages as a graph. We can do this in various ways. Connectivity-based ranking techniques usually assume the most straightforward representation: each Web page in the collection is modeled by a node in the graph. If page *A* contains a hyperlink to page *B*, then there exists a directed edge (*A*, *B*) in the graph. If *A* does not have a hyperlink to *B*, there is no directed edge (*A*, *B*). We call this directed graph the *link graph* *G*. This is illustrated in Figure 1.

Some algorithms use an undirected *co-citation graph*. In an undirected co-citation graph, nodes *A* and *B* are connected by an undirected edge if and only if there exists a third page *C* hyperlinking to both *A* and *B* (see Figure 2). We say that *A* and *B* are co-cited by *C*.

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

Query-Independent Ranking

Query-independent ranking aims to measure the intrinsic quality of a page. To this end, a score is assigned to each page independent of a specific user query. At query time, this score is used with or without some query-dependent ranking criteria to rank all documents matching the query.

The first assumption of connectivity-based techniques leads to a simple criterion: the more hyperlinks pointing to a page, the better the page. The main drawback of this approach is that it does not distinguish between the quality of a page pointed to by a number of low-quality pages and the quality of a page pointed to by the same number of high-quality pages. Obviously, this means you can make a page rank high by simply creating many other pages that point to it.

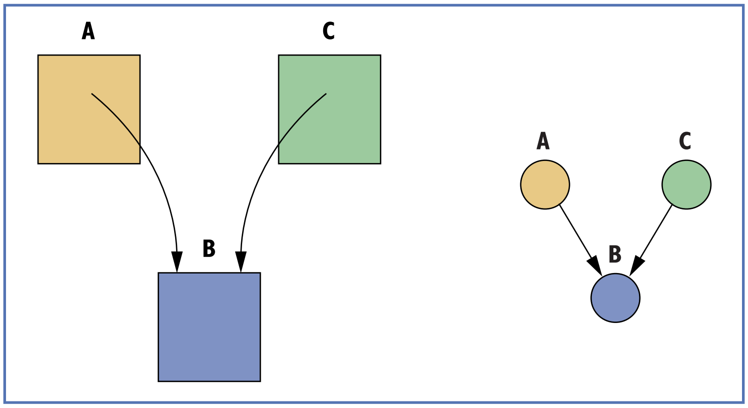

Figure 1. A set of Web pages and the corresponding link graph. If Web page *A* links to Web page *B*, there exists a directed edge (*A*, *B*) in the graph.

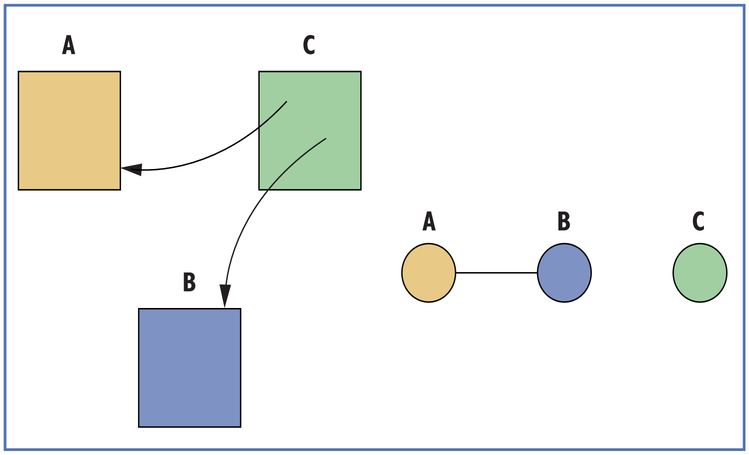

Figure 2. A set of Web pages and the corresponding co-citation graph. If Web page *C* links to both *A* and *B*, then *A* and *B* are connected by an undirected edge in the graph and are said to be co-cited by *C*.

Brin and Page created the PageRank algorithm to remedy this problem.[3,4] They compute the PageRank of a page by weighting each hyperlink to the page proportionally to the quality of the page containing the hyperlink. To determine the quality of a referring page, they use its PageRank recursively, with an arbitrary initial setting of the PageRank values. More specifically, the PageRank $R(A)$ of a page *A* can be defined as

$$R(A) = \in /n + (1 - \in) \cdot \sum_{(B,A)\in G} R(B) / outdegree(B),$$

where

Uses of Hyperlink Analysis in Web Information Retrieval

The hyperlink structure of the Web can be used to analyze more than the quality of a Web page. It can also be used to find Web pages similar to a given Web page or Web pages of interest to a given geographical region. These and other applications of hyperlink analysis in Web information retrieval are described below.

Search-by-Example

A search-by-example approach to Web information retrieval looks for pages related to a given page. For example, given www.nytimes.com, find www.washingtonpost.com and www.wsj.com. Both the HITS algorithm and a simple algorithm on the co-citation graph perform very well.[1,2] The idea behind the latter is that frequent co-citation indicates relatedness, especially when the co-citations occur close to each other on the page. Thus, pages with many co-citations that are close to each other on the page containing the co-citation tend to be related.

Mirrored Hosts

The *path* of a Web page is the part of the URL following the host, that is, after the third slash. For example, in the URL http://www.google.com/about.html, www.google.com is the host and /about.html is the path. Two hosts, H_1 and H_2, are *mirrors* if and only if for every document on H_2, there is a highly similar document on H_2 with the same path, and vice versa.

Mirrors exhibit a very similar hyperlink structure both within the host and among the mirror host and other hosts. Mirrored Web hosts waste space in the index data structure and can lead to duplicate results. Combining hyperlink analysis with IP address analysis and URL pattern analysis can detect many near-mirrors.[3]

Web Page Categorization

Hyperlink analysis can also be used to compute statistics about groups of Web pages, like their average length, the percentage that are in French, and so on. PageRank-like random walks can be performed on the Web to sample Web pages in an almost uniform distribution.[4] These almost random samples can then be used to measure various properties of Web pages, but also to compare the number of the pages in the indices of various commercial search engines. For example, this technique was used in November 1999 to measure that roughly 47 percent of all Web pages belonged to the .com domain.

Geographical Scope

Whether a given Web page is of interest only for people in a given region or is of nation- or worldwide interest is an interesting problem for hyperlink analysis. For example, a weather-forecasting page is interesting only to the region it covers, while the Internal Revenue Service Web page may be of interest to U.S. taxpayers throughout the world. A page's hyperlink structure also reflects its range of interest.[5,6] Local pages are mostly hyperlinked to by pages from the same region, while hyperlinks to pages of nationwide interest are roughly uniform throughout the country. This information lets search engines tailor query results to the region the user is in.

References

1. J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms*, ACM Press, New York, Jan. 1998, pp. 668-677.
2. J. Dean and M.R. Henzinger, "Finding Related Web Pages in the World Wide Web," *Proc. Eighth Int'l World Wide Web Conf.*, Elsevier Science, New York, 1999, pp. 389-401.
3. K. Bharat et al., "A Comparison of Techniques to Find Mirrored Hosts on the World Wide Web," *J. American Soc. for Information Science*, Vol. 51, No. 12, Nov. 2000, pp. 1,114-1,122.
4. M.R. Henzinger et al., "On Near-Uniform URL Sampling," *Proc. Ninth Int'l World Wide Web Conf.*, Elsevier Science, Amsterdam, 2000, pp. 295-308.
5. O. Buyukkokten et al., "Exploiting Geographical Location Information of Web Pages," *Proc. ACM SIGMOD Workshop on the Web and Databases* (WebDB 99), INRIA, Philadelphia, 1999, pp. 91-96.
6. J. Ding, L. Gravano, and N. Shivakumar, "Computing Geographical Scopes of Web Resources," *Proc. 26th Int'l Conf. Very Large Databases* (VLDB 00), Morgan Kaufmann, San Francisco, 2000, pp. 545-556.

- ϵ is a constant usually set between 0.1 and 0.2;
- n is the number of nodes in G, that is, the number of Web pages in the collection; and
- $outdegree(B)$ is the number of edges leaving page B, that is, the number of hyperlinks on page B.

This formula shows that the PageRank of a page A depends on the PageRank of a page B pointing to A. Since the PageRank definition introduces one such linear equation per page, a huge set of linear equations need to be solved in order to compute PageRank for all pages.

The PageRank measure effectively distinguishes high-quality Web pages from low-quality Web pages, and it is used by the Google search engine (http://www.google.com/).

Query-Dependent Ranking

In *query-dependent* ranking, an algorithm assigns a score that measures the quality and relevance of a selected set of pages to a given user query. The basic idea is to build a query-specific graph, called a *neighborhood graph*, and perform hyperlink analysis on it. Ideally, this graph will contain only pages on the query topic.

Carriere and Kazman propose the following approach for building a neighborhood graph[5]:

- A *start set* of documents matching the query is fetched from a search engine (say, the top 200 matches).
- The start set is augmented by its *neighborhood*, which is the set of documents that either hyperlinks to or is hyperlinked to by documents in the

start set (see Figure 3). Since the *indegree* (that is, the number of documents hyperlinking to a document in the start set) of nodes can be very large, in practice a limited number of these documents (say, 50) is included.

- Each document in both the start set and the neighborhood is modeled by a node. There exists an edge from node *A* to node *B* if and only if document *A* hyperlinks to document *B*. Hyperlinks between pages on the same Web host can be omitted since the authors might be affiliated and thus the hyperlink might not express a recommendation.

Various ranking schemes can now be used on the neighborhood graph. As with query-independent ranking schemes, an indegree-based approach[5] ranks the nodes in the neighborhood graph by the number of documents hyperlinking to them. Again, in this approach, all hyperlinks are considered equal.

Neighborhood graphs typically consist of thousands of nodes (that is, they are relatively small). Computing the PageRank on a neighborhood graph produces a ranking similar to that produced by indegree-based ranking.[6]

Another approach to ranking pages in the neighborhood graph assumes that a topic can be roughly divided into pages with good content on the topic, called *authorities*, and directory-like pages with many hyperlinks to pages on the topic, called *hubs*.[7]

Kleinberg's hyperlink-induced topic search (HITS) algorithm tries to determine good hubs and authorities.[7] Given a user query, the algorithm iteratively computes hub and authority scores for each node in the neighborhood graph, and then ranks the nodes by those scores. Nodes with high authority scores should be good authorities, and nodes with high hub scores should be good hubs. The algorithm presumes that a document that points to many others is a good hub, and a document that many documents point to is a good authority. Recursively, a document that points to many good authorities is an even better hub, and a document pointed to by many good hubs is an even better authority. This gives us the following recursive algorithm:

```
Let $N$ be the set of nodes in the neighborhood graph.
For every node $A$ in $N$, let $Hub$[A] be its hub score and $Aut[A]$ its authority score.
Initialize $Hub$[A] to 1 for all $A$ in $N$.
While the vectors $Hub$ and $Aut$ have not converged:
    For all $A$ in $N$, $Aut[A] := \sum_{(B, A) \in N} H[B]$
    For all $A$ in $N$, $Hub[A] := \sum_{(A, B) \in N} A[B]$
    Normalize the $Hub$ and $Aut$ vectors.
```

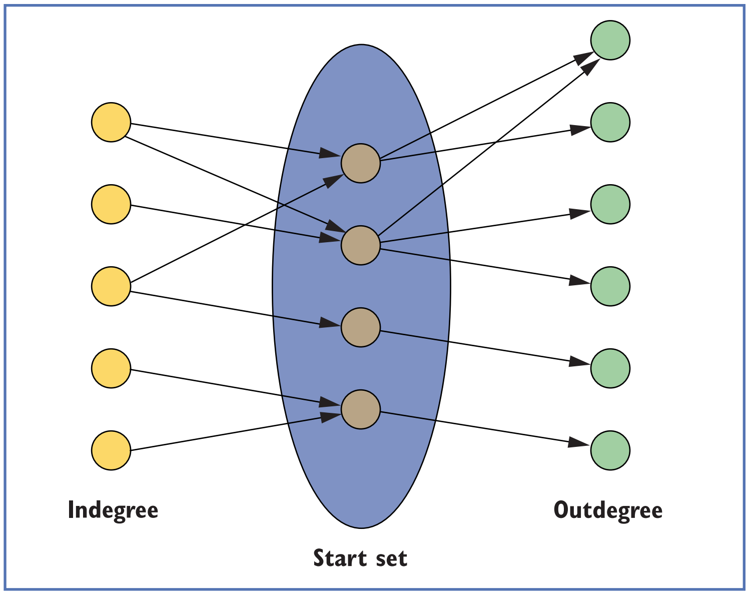

Figure 3. A start set and its neighborhood. In the indegree approach, the start set is augmented by its neighborhood, which is the set of documents that either hyperlinks to or is hyperlinked to by documents in the start set.

Elementary linear algebra shows that the *Hub* and *Aut* 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 high-quality pages for a query, since there may be some that do not belong to the neighborhood graph or that do belong to the neighborhood graph but have not been hyperlinked to by many pages.

A page that points to many others should be a good hub, and a page that many pages point to should be a good authority.

There are two problems with the HITS algorithm:

- Since it considers only a relatively small part of the Web graph, adding edges to a few nodes can change the resulting hubs and authority scores considerably.[8] Thus it is relatively easy to manipulate these scores. As mentioned earlier, manipulation of search engine rankings is a serious problem on the Web.
- If the majority of pages on a neighborhood

graph is on a topic different from the query topic, the top-ranked authority and hub pages might be on the different topic. This problem is called *topic drift*. Adding weights to edges based on text in the documents or their anchors alleviates this problem considerably.[9-11]

To the best of my knowledge, the HITS algorithm is not currently used in a commercial search product.

Conclusion

Research into the hyperlink structure of the Web is just beginning, and more exciting applications (such as those outlined in the sidebar, “Uses of Hyperlink Analysis in Web Information Retrieval”) can be expected in the future.

When Web search engines started to use hyperlink analysis, Web-positioning companies started to manipulate hyperlinks by creating hyperlinks that try to boost the ranking of their clients’ pages, making it harder for Web search engines to return high-quality results. To offset this, Web search engines need to design more sophisticated techniques—and to keep them secret since any disclosure leads to more manipulation attempts.

Acknowledgment

An earlier version of this article, “Link Analysis in Web Information Retrieval,” appeared in the *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, September 2000.

References

1. J. Cho, H. García-Molina, and L. Page, “Efficient Crawling through URL Ordering,” *Proc. Seventh Int’l World Wide Web Conf.*, Elsevier Science, New York, 1998, pp. 161-172.
2. G. Salton et al., “The SMART System—Experiments in Automatic Document Processing,” Prentice-Hall, Englewood Cliffs, N.J., 1971.
3. S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” *Proc. Seventh Int’l World Wide Web Conf.*, Elsevier Science, New York, 1998, pp. 107-117.
4. L. Page et al., “The PageRank Citation Ranking: Bringing Order to the Web,” Stanford Digital Library Technologies, Working Paper 1999-0120, Stanford Univ., Palo Alto, Calif., 1998.
5. J. Carriere and R. Kazman, “Webquery: Searching and Visualizing the Web through Connectivity,” *Proc. Sixth Int’l World Wide Web Conf.*, Elsevier Science, New York, 1997, pp. 701-711.
6. B. Amento, L. Terveen, and W. Hill, “Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents,” *Proc. 23rd Int’l ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 00), ACM Press, New York, 2000, pp. 296-303.
7. J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” *Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms*, ACM Press, New York, 1998, pp. 668-677.
8. R. Lempel and S. Moran, “The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect,” *Proc. Ninth Int’l World Wide Web Conf.*, Elsevier Science, New York, 2000, pp. 387-401.
9. S. Chakrabarti et al., “Experiments in Topic Distillation,” *Proc. ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 98), Post-Conference Workshop on Hypertext Information Retrieval for the Web.
10. S. Chakrabarti et al., “Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text,” *Proc. Seventh Int’l World Wide Web Conf.*, Elsevier Science, New York, 1998, pp. 65-74.
11. K. Bharat and M. Henzinger, “Improved Algorithms for Topic Distillation in Hyperlinked Environments,” *Proc. 21st Int’l ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 98), ACM Press, New York, 1998, pp. 111-104.

Monika R. Henzinger is the director of research at Google Inc., a next-generation Web search engine. From 1996 to the fall of 1999 she was a member of the research staff at Digital Systems Research Center in Palo Alto, California, and, prior to that, an assistant professor at Cornell University. Her research interests include information retrieval, Web analysis, and algorithms. Henzinger received a PhD from Princeton University in 1993. In 1995 she received an NSF Career award.

Readers can contact Henzinger at monika@google.com.