Webology, Volume 2, Number 3, October, 2005

Home	Table of Contents	Titles & Subject Index	Authors Index

editorial

Alireza Noruzi

Hyperlinks and Their Roles in Web Information Retrieval

A web page generally includes elements such as text, hyperlink, image, etc. Hyperlink represents a relationship between two web pages or just between sections of the same page. Understanding the hyperlink structure is fundamental to understanding the Web connectivity structure, because hyperlinks have been used in web indexing and information retrieval, as well as page ranking. If the Web were a car, hyperlinks would be the engine, because without them, we are not going anywhere.

In the context of the Web generally, there are three types of hyperlink:

- links to another page in the same site which connect pages within the site, also called navigational links ("Click here to return to the homepage", "Return", etc). This is a fundamental requirement for ease of navigation of sites by end-users. It is also necessary for indexing websites by search engine robots;
- links from other sites, called backlinks or inlinks. Search engines use backlinks for deciding which web resources to add to their collections (i.e., which web pages to crawl), and how to rank the resources matching a user query. Backlinks are singled out as more important than navigational links. The creation of a backlink indicates the human judgment that the creator of page A, by including a backlink to page B, has in some measure conferred *authority* on B and represents the "*latent human judgment*" (Kleinberg, 1999). For instance, Google relies heavily on inbound relevant backlinks to rank a site (Acharya et al., 2005); and
- links to email addresses (e-mail mailto link).

From the point of view of a link-based search engine there are three types of hyperlink (Figure 1).

Figure 1. Hyperlink types

Types of hyperlinks	Attribute	Quality
Text link	Keywords in anchor text	Best
	Keywords in title	Good
Graphic link	no text, only image link	Poor
Affiliate link	link to database, then page	Poor

A text link includes a URL and a set of words that are known as anchor text. An example of a simple *text link* is <a href="http://www.ABC.com">underlined anchor text</a>.

Search engine robots prefer the simplicity of the text link to any other form of linking. In fact, the *anchor text* is very important in the eyes of the search engines and is assigned more weight within the search engine algorithms than ordinary body text. The reason search engines assign more weight to the anchor text is that search engines assume that a web author would only link to a page that s/he deems important. If s/he deems a page important, so should the search engines. So web authors make sure to link to important pages and should use right keywords in the text link for the page s/he is linking to. The idea of associating anchor text with the page the text points to was first implemented in the *World Wide Web Worm* (McBryan, 1994) and later has been applied by several search engines.

Search engines cannot see the graphic and if the graphic is really a graphical representation of a word - such as *About Us* - then how does a search engine know that? In all cases where graphics are used, it is wise to use an alternative text tag within the source code of the image. For instance, the alternative text would be embedded into the above example as:

<a href="http://www.ABC.com/aboutus.html"><img src="/aboutus.jpg" alt="*Short Text Explaining Where the Graphic Link will Take Visitor*"></a>

The ALT tag (alternative) gives search engines as well as screen readers the ability to assign a meaning to the graphic, in place of the anchor text. Hence, it is important to use specific keywords that apply to the page you are linking to in the alternative tag area (Schwartz, 2003).

So, link context is an important aspect of link analysis. Link context analyzes how close a link appears on a page to keywords within the text of that page. Search engines are more likely to give greater weight to a content site than a site with nothing but optimized doorway pages (Ward, 2001). All links are not created equal, but some are more important than others. The concept is simple. *Graphic* and *affiliate* links are generally for marketing and sales. These links can be "bought" so the search engines do not value them. But a text link requires a web author to inspect the target site and then create a hyperlink. It is visible to everyone. The implication is that the web author has validated the theme and quality of the target web site/page through the hyperlink. When the link text (anchor text), body or title includes keywords, the link has a higher value and complements proper page optimization (Vidals, 2001). An example of a text link with keywords in title is <a href="http://www.IFLA.org/" title="International Federation of Library Associations and Institutions (IFLA)">IFLA</a>

To fully describe text links, we need to go back to the early days of the Web. The Web first started by providing basic information with very few graphics on plain HTML pages. When these web pages wanted to reference another information resource related to their own information, they created text links pointing to this other information resource. So the Web was built on text links (see the first web page). Text links have the greatest importance when they point to a web resource. Therefore, text links are important for promoting a website because:

- they provide a direct link for visitors to follow based on anchor text enriched with keywords or title, and provide the traffic without the server side redirection;
- they serve as a '*vote*' or an increase in popularity that all major search engines, such as Google, use when ranking relevant web pages in response to a web user searching for terms/phrases related to the web page/site. "Google interprets a link from page 'A' to page 'B' as a vote by page 'A' for page 'B'. But Google considers more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves *important* weigh more heavily and help to make other pages *important*" (Google, 2005).

Different search engines use different ranking algorithms and their exact form is kept a secret. But Google have filed a US Patent Application entitled "Information retrieval based on historical data" on March 31, 2005, which reveals a great deal of how it ranks websites. It revealed many of the ranking criteria by publishing this patent with 63 claims. This patent claims a method includes: "determining an inception date corresponding to the document, and scoring the document based, at least in part, on the weights assigned to the links associated with the document [which is] based on at least one of a date of appearance of the link, a date of a change to the link, a date of appearance of anchor text associated with the link, a date of a change to anchor text associated with the link, a date of appearance of a linking document containing the link, and a date of a change to a linking document containing the link." In addition, according to this patent, "information relating to a manner in which anchor text changes over time may be used to generate (or alter) a score associated with a document. For example, changes over time in anchor text associated with links to a document may be used as an indication that there has been an update or even a change of focus in the document."

Google considers the anchor text as part of the web page it refers to, as well as part of the page it is actually on. "Because anchor text is often considered to be part of the document to which its associated link points" (Acharya et al., 2005). The designers of Google's PageRank algorithm (Page et al., 1998) reasoned that the anchor text serves as an extremely succinct summary of the web page it refers to. From a link-based search engine's point of view, each backlink to a website is interpreted as a vote of confidence in that website - the more votes, the higher the ranking.

Text links are based on the most basic form of what the Web was created on and will always be around. Search engines treat text links as the most credible sign of popularity. In fact, the text used to describe a backlink can affect the target site's rankings. These three links all point to the same site (enriched with keywords or title) but use different text:

- Webology journal
- Online journal
- Click here

Search engines can include anchor text associated with backlinks to the page, as this approach has several advantages, including: (i) anchors often provide more accurate descriptions of web pages than the pages themselves; (ii) anchors may exist for images, programs, and other objects that cannot be indexed by a text-based search engine. In addition, even though the text of the document itself may not match the search terms, if the document is linked by documents whose *titles* or *backlink anchor text* match the search terms, the document will be considered a match (Page, 2001).

It can be concluded that search engines consider that any words used by other sites to describe a site is particularly relevant even if the keywords are not used in the backlinked site/page (the linked target destination). In other words, the foreign language text links allow the linked sites to have a chance to be retrieved as relevant results in response to a search query. Many search engines judge the linking page partly based on the quality of the linked page, and if many sites backlinking to a site use keywords in their *link text*, search engines will raise its ranking for those keywords. Ultimately, backlinks from popular websites with a higher ranking, have a higher weight then backlinks from smaller unknown websites.

Articles in This Issue

We have three articles in this issue, two of them are application-oriented, concerned with aspects of the World Wide Web (websites and search engines), and the third one is a comparative study of *License Agreements* of societal and commercial publishers.

Peter Williams, Karen Dennis & David Nicholas: *An Evaluation of the Websites of Charities and Voluntary Organisations Providing Support for Young People: Case Study: Drugscope* . This study examines the usage, usability and impact of a charitable website 'Drugscope', using a range of methods to evaluate the site, including "Inspection, examining the extent to which the site met recognised quality criteria; formal usability tests, including information retrieval tasks; an online user survey and computer log analysis." It is concluded that "the site is very well organised for retrieving information." This research can also be seen as a pattern for study of usability and accessibility of websites.

Haidar Moukdad & Hong Cui: *How Do Search Engines Hhandle Chinese Queries*? This article explores the characteristics of the Chinese language and how queries in this language are handled by different search engines, comparing Google and AlltheWeb with two Chinese search engines (Sohu, and Baidu). The researchers argued that the number of Internet users in China has increased and the Chinese language became the second most often used online language following English. It is concluded that "major search engines designed for English do not handle Chinese queries as well as search engines specifically designed for Chinese do. Research on other languages has reached similar results, pointing to the need for new approaches to IR issues on the Web and for serious investigations of the feasibility of developing super search engines capable of handling a multitude of languages with equal degrees of effectiveness and efficiency."

B.M. Meera & Anuradha K.T.: *Contractual Solutions in Electronic Publishing Industry: A Comparative study of License Agreements*. The Internet is revolutionizing information access. Libraries and information centers are engaged in signing different License Agreements for access to electronic information resources. This study compares the clauses of the license agreements among publishers of commercial databases (i.e. Web of Science; Engineering Village; Springer Journals) and societal databases (i.e. MathSci Net; SciFinder Scholar; ACM Journals). It is observed that "the licensors' rights are well protected compared to that of licensees' rights."

References

- Acharya, A., Cutts, M., Dean, J., Haahr, P., Henzinger, M., Hoelzle, U., Lawrence, S., Pfleger, K., Sercinoglu, O., & Tong, S. (2005). *Information retrieval based on historical data*. United States Patent Application 2005/0071741, Kind Code A 1.
- Google (2005). Our search: Google technology, PageRank explained. Retrieved September 15, 2005 from http://www.google.com/technology/index.html
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* , 46(5), 604-32.
- McBryan, O.A. (1994). GENVL and WWWW: Tools for taming the Web. In: *Proceedings of the First International Conference on the World Wide Web*, CERN, Geneva, May 25-27, 1994, pp. 1-13.
- Page, L. (2001). *Method for node ranking in a linked database*. United States Patent Application 2001/6285999, Kind Code B 1.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the Web* . Technical report, Stanford University, Stanford, CA, 1998.
- Schwartz, B. (2004, July 14). Internal linking structure elements strategy. Retrieved September 15, 2005 from http://www.rustybrick.com/seo_articles_3.php
- Vidals, G. (2001). Strategic link analysis. Retrieved September 15, 2005 from http://www.positionresearch.com/research/link_analysis.html
- Ward, E. (2001, December 19). How search engines use link analysis. A special report from the *Search Engine Strategies 2001 Conference*, November 14-15,

Dallas, Texas. Retrieved September 15, 2005 from http://searchenginewatch.com/searchday/article.php/34711_2158431

Bibliographic information of this note for citing:

Noruzi, A. (2005). "Editorial: Hyperlinks and Their Roles in Web Information Retrieval." *Webology*, **2**(3), editorial 5. Available at: http://www.webology.org/2005/v2n3/editorial5.html