Link Proximity Analysis - Clustering Websites by Examining Link Proximity

Bela Gipp[1,2], Adriana Taylor[1], Jöran Beel[1,2]

[1] UC Berkeley, Berkeley, California, USA
[2] Otto-von-Guericke University, Computer Science/ITI/VLBA-Lab, Magdeburg, Germany
{gipp, aitaylor, beel}@berkeley.edu

Abstract. This research-in-progress paper presents a new approach called Link Proximity Analysis (LPA) for identifying related web pages based on link analysis. In contrast to current techniques, which ignore intra-page link analysis, the one put forth here examines the relative positioning of links to each other within websites. The approach uses the fact that a clear correlation between the proximity of links to each other and the subject-relatedness of the linked websites can be observed on nearly every web page. By statistically analyzing this relationship and measuring the amount of sentences, paragraphs, etc. between two links, related websites can be automatically, identified as a first study has proven.

Keywords: Web page, Website, clustering, Network Analysis, Link Analysis, Citation Proximity Analysis

1 Introduction

Most modern search engines offer a "find similar pages" function which returns web pages similar to a given one. Would it not be useful if an author's knowledge of his subject matter could be used to identify similar pages?

Websites usually address a specific topic, and each section addresses a particular facet of that topic. Embedded hyperlinks operate in a parallel manner; the closer two links are to each other, the more likely it is that they have a similar theme. Since web pages necessarily reflect the course of human cognition, the adjacency of links is also an indication of the author's conception of their relatedness, and likely of the user's perception of their similarity, as well. The approach presented here seeks to exploit this by analyzing link proximity to identify related web pages.

2 Related Work

The main strategies for assessing the similarity of hypertext documents are text-based, user-based, and link-based analyses [9, 10]. While their synchronous operation is the end goal, the focus of this paper is on improving the last.

Link-based techniques have the advantage of circumventing text-based methods' dependency on a document's language, ambiguous nomenclature, synonyms and homonyms [5]. Furthermore, the application of customary measures of likeness, sc., cosine and extended Jaccard similarity, has been straightforward [8].

Cluster analysis has been used to aid in similarity search on the web. Agglomerative and divisive hierarchical clustering algorithms; partitional clustering algorithms, like *k*-means; and density-based clustering algorithms have all been used to partition the web, in combination with link-based analysis [6].

Naturally, there are various link-based systems that determine hypertext documents' similarity to each other. Co-citation analysis and bibliographic coupling are two such means. Co-citation, as advanced by [4] and [7], occurs when two documents are cited by another document. Bibliographic coupling determines correspondence via the number of citations that two documents have in common [3].

Traditional link-based approaches like those do not take into account the internal structure of the web pages in question. The use of Citation Proximity Analysis (CPA) [1] would help in this regard. CPA adds another dimension to Co-citation by accounting for the joint appearance of citations with respect to their mutual proximity. The key presupposition is that citations have a greater probability of being related the closer they are to one another. CPA is a finer tool overall, and has the advantage of being able to determine the relatedness of subsections of documents.

We have not found CPA's underlying concept used in a link-based approach to similarity search in hypertext documents. It has previously been applied only to scientific articles, where it delivered good results [1].

3 Link Proximity Analysis

The following example is a screenshot from a Wikipedia article about the 25 largest daily newspapers in America.

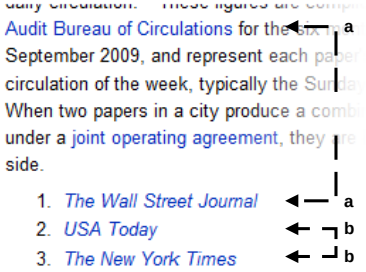

Figure 1: Example of website with links

Figure 1 illustrates that links on websites are usually the more related the closer they are listed to each other. Whereas the link to the "Audit Bureau of Circulation" is only to some extent related to the "Wall Street Journal" (see arrow *a*), the other entries listed before and after the New York Times (see arrow *b*) are closely related - all of them are newspapers.

In LPA, this fact is used to calculate the link proximity and so to identify related websites. If two links are given within one sentence they are probably addressing a similar topic. If, on the other hand, two links are separated by a whole paragraph, then they likely address less related topics. Unfortunately, most websites do not list information in as structured a way as Wikipedia. Nevertheless, by analyzing not one source (i.e., one website), but millions, and by only considering the most frequent link combinations, outliers such as the combination of "joint operation agreement" and "Wall Street Journal" (see Figure 1) are not significant enough to be considered.

So far, we have used a simple weighting approach that only considers the amount of words, sentences, paragraphs and section headings between links. We ignored factors like different fonts, font sizes etc. If links were ordered alphabetically, the position within the list was ignored. If more than one web page linked the URI, we used the average.

Algorithm: Calculate Link Proximity
Input: Crawled websites w_i containing links l_j, $i = 1, 2, \ldots n$, $j = 1, 2, \ldots m$
Output: Related websites

// to assign Link Proximity Index to pairs of links within a page for all web pages in w_i:
$\exists\, l_j \forall\, w_i$: // Exclude pairs of identical links
if distance d = same sentence, then LPI = 1,
if d = same paragraph, then LPI = ½,
if d = same section, then LPI = ¼ etc.
// to calculate the average LPI for a given pair of links:
$\forall\, (l_a, l_b) \in w_i$,

$$\frac{\sum_{\gamma=1}^{x} \mathrm{LPI}(l_{a\gamma}, l_{b\beta})}{x}$$

, where x is the number of pairs of $(l_a, l_b) \in w_i$

A first empirical study was conducted to evaluate the performance of this approach. The target websites were chosen from the most highly trafficked sites[1]. To determine related pages, the structures of 500,000 sites linking the targets were analyzed. Stimuli were culled from the top 50 targets where the outlined algorithm returned a suggestion different from Google's. In 552 cases, according to 20 volunteer test subjects, Google delivered better "related web page" results, whereas in 448 cases the described approach delivered superior results. Usually, the best recommendations are generated by hybrid approached such as combining text, user behavior and link analysis [2]. It seems likely that this is also true for LPA.

We plan to expand this study with the involvement of interested researchers to compare the performance of LPA with existing text-, link- and user behavior-based approaches. In order to facilitate this, we released the crawler and LPA-Software as Open-Source under the General Public License.

[1] According to Alexa.com

4 Conclusion

In this paper we proposed a new approach to identify related websites based on link analysis. In contrast to the traditional approaches, it additionally analyzes the proximity of links to each other within websites. A first study showed that this approach leads to good results despite its simplicity. However, a more comprehensive and comparative study needs to be done to evaluate the potential and fields of applications such as movie recommender systems etc.

References

[1] Bela Gipp and Jöran Beel. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In Birger Larsen and Jacqueline Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 2, pages 571–575, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935. Available on http://www.sciplore.org.

[2] Bela Gipp, Jöran Beel, and Christian Hentschel. Scienstein: A Research Paper Recommender System. In *Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09)*, pages 309–315, Virudhunagar (India), January 2009. Kamaraj College of Engineering and Technology India, IEEE. Available on http://www.sciplore.org.

[3] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[4] IV Marshakova. System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6(2):3–8, 1973.

[5] Dániel Fogarasand Balázs Rácz. Scaling link-based similarity search. In *International World Wide Web Conference. Proceedings of the 14th international conference on World Wide Web*, 2005.

[6] Anirban Kundu Ruma Dutta, Indranil Ghosh and Debajyoti Mukhopadhyay. An Advanced Partitioning Approach of Web Page Clustering utilizing Content & Link Structure. *Journal of Convergence Information Technology*, 4:65–71, 2009.

[7] H Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[8] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, 2000.

[9] Dan Klein Taher H. Haveliwala, Aristides Gionis and Piotr Indyk. Evaluating strategies for similarity search on the web. In *International World Wide Web Conference. Proceedings of the 11th international conference on World Wide Web*, 2002.

[10] Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *Proceedings of the eleventh international conference on Information and knowledge management*, page 506. ACM, 2002.