1-1-2009

# Keyword enhanced web structure mining for business intelligence

L. Vaughan

*Vaughan, L.; Faculty of Information and Media Studies, University of Western Ontario, London, ON N6A 5B7, Canada; email: lvaughan@uwo.ca*, lvaughan@uwo.ca

J. You

# Keyword Enhanced Web Structure Mining for Business Intelligence

**Liwen Vaughan*** **and Justin You****
* Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, N6A 5B7, Canada
lvaughan@uwo.ca
** ApacBridge Consulting
8 Northgate Street
Ottawa, Ontario, K2G 6C7, Canada
justin.you@apacbridge.com

**Abstract** The study proposed the method of keyword enhanced Web structure mining which combines the ideas of Web content mining with Web structure mining. The method was used to mine data on business competition among a group of DSLAM companies. Specifically, the keyword DSLAM was incorporated into queries that searched for co-links between pairs of company Websites. The resulting co-link matrix was analyzed using multidimensional scaling (MDS) to map business competition positions. The study shows that the proposed method improves upon the previous method of Web structure mining alone by producing a more accurate map of business competition in the DSLAM sector.

**Keywords**: Web content mining, Web structure mining, E-commerce, business intelligence

## 1 Introduction

Web data mining can be classified into the following three sub-areas based on the type of data used [1]: Web content mining, Web structure mining, and Web usage mining. Web content mining uses Web page content, the most common of which is Web page texts. Web structure mining tries to discover the model underlying the Web hyperlink structure. Web usage mining tries to discover patterns from Web usage data [2]. Web data mining has been applied to various areas; one of which is E-commerce or business information [3]. Earlier studies used Web content mining [4] and Web structure mining [5] to gather information on business competition, an important topic of business intelligence. Building on to earlier studies, the current study proposes a method that combines both Web content mining and Web structure mining. The method was tested in the Websites of a group of companies in DSLAM (Digital Subscriber Line Access Multiplexer) sector of the telecommunication industry. The result shows that the proposed method of keyword enhanced Web structure mining improves upon the method that is based on Web structure mining alone.

We need to define some terms before discussing the details of the study. Inlinks (also called back links) are links coming into (or pointing to) a Web page while outlinks are links going out from a Web page, i.e the hyperlinks embedded in the Web page. Two different types of inlinks need to be distinguished, total inlinks and external inlinks. Total inlinks include all links pointing to a particular page or site while external inlinks include only links coming from Websites outside the site in question. In other words, external inlinks do not include links within the site itself, such as the "back to home" type of navigational links within the site. Björneborn & Ingwersen [6] further distinguished between co-inlinks and co-outlinks. If page X and page Y are both linked to by page Z (i.e. page X and page Y both have inlinks from page Z), then X and Y are co-inlinked. The method proposed in this paper analyzed patterns of co-inlinks (also called co-links later in the paper). Our study only examines co-links in the form of external inlinks because they are objective measures of relatedness between the co-linked sites while internal in-links do not indicate such relationships.

## 2 Problem Statement

In an earlier study [5], co-link method was used to analyse business competition among a group of companies in the overall telecommunication area. The study showed that this method was able to successfully cluster the competitors based on their overall business strength, market segments and regional market focus. The value of this method is to reveal competitive positions at a macro level.

In a particular product or market segment, for example DSLAM products which are used for DSL broadband Internet access, there are various types of companies ranging from big telecommunication equipment companies such as Ericsson and Alcatel which have many product lines in addition to DSLAM products to small companies which solely focus on one particular product line. Using Web structure mining alone will result in an unbalanced comparison between big companies, which have rich portfolio of products and much wider market presence, and small companies, which specialize in certain products. The Websites of these large companies attract huge number of inlinks and co-links, many of which are related to product lines other than the one being analyzed. In other words, inlink or co-link analysis alone will not result in an accurate comparison among these companies in a particular sector.

To solve this problem, we propose a "keyword enhanced Web structure mining" method in this paper. DSLAM sector is chosen as a case study of a particular market segment. The purpose of the study is to find out whether the proposed method will provide a more accurate competitive analysis for a particular product or a market segment.

## 3 Proposed Method

The proposed method is based on the idea that co-links to a pair of Websites is an indicator that the two sites are similar or related. The more such co-links, the stronger

the relationship. In business world, co-links are particularly useful as business competitors tend not to link to each other so simple in-links do not contain much useful information on business competition. However, two related companies will be co-linked by a third party such as a customer or a reseller [7]. The more co-links the two companies have, the more closely related they are. Since related companies are competing companies, Web co-link data can be used to cluster companies into a map of business competition. Using this method, an earlier study [5] successfully generated a map of business competition for a group of companies in the telecommunication equipment industry. The current method improves upon the previous method by incorporating Website content, specifically keywords on the sites, to achieve a more accurate mapping.

The method was tested in a group companies in DSLAM sector of the telecommunication industry. DSLAM was chosen for this study as the acronym DSLAM is usually used in the industry and on Websites instead of the complete spelling of the term. So the unique acronym is ideal to test the proposed method as incorporating this keyword into co-link data collection will filter out Websites that co-linked two companies for reasons other than DSLAM, e.g. two sites were co-linked because of their charity activities. We selected 35 DSLAM companies that are included in a reliable research report on DSLAM market [8]. These 35 companies are major DSLAM product companies worldwide.

We located Websites of these companies and then searched for co-links to a pair of these companies using search engine Yahoo! (search details below). The keyword DSLAM is added in search queries. The query result is a matrix of 35 by 35 symmetrical by the diagonal. Each row or column represents a company. Each cell of the matrix records the number of Web pages retrieved by the co-link query. The matrix is not sparse as there was at least one co-link between the majority pairs of companies. This raw co-link matrix needs to be normalized to obtain a relative measure of the strength of the relationship because a co-link count of 5 is very high if the number of links pointing to each company is 6 while it will be low if the number of links pointing to each company is 100. The normalization is done through Jaccard Index as follows:

$$NormalizedColinkCount = n(A \cap B) / n(A \cup B)$$

Where A is the set of Web pages which links to company X
B is the set of Web pages which links to company Y
$n(A \cap B)$ is the number of pages which link to both company X and company Y, i.e. the raw co-link count
$n(A \cup B)$ is the number of pages which link to either company X or company Y.

Multidimensional Scaling (MDS), a statistical analysis method, was then applied to the normalized co-link matrix using version 12 of SPSS software. The MDS output includes a map that positions each company according to their similarity to other companies as measured by co-link counts. The higher the co-link count, the closer the

two companies will be placed. Essentially the map will cluster competing companies together so the map will show the competition landscape of the DSLAM sector.

## 4 Data Collection Details

Yahoo! was used for data collection as it is more suitable for the study than two other major search engines in the market, Google and MSN. As explained earlier, the study needs to search for external inlinks but Google can only search for total inlinks, i.e. it cannot filter out internal links in search results. In order to filter out internal links, the link query needs to be combined with the "site" query. However, Google's link query term cannot be combined with other query terms as is stated in Google search API reference [9]. MSN can search for external inlinks. However, at the time of data collection (summer 2006), MSN indexed a much smaller number of inlinks than Yahoo! did as MSN usually retrieved a much smaller number of pages for the same inlink query. So Yahoo! was preferred over MSN for data collection.

Two sets of data were collected. One set is the co-link count alone and the other set added the keyword DSLAM in search queries. MDS mapping results from the two data sets were compared to determine if the one with the keyword (i.e. combining link structure data with keyword data) is better. The syntax of queries used to collect the two sets of data is shown in Table 1 in a hypothetical scenario of searching for co-links between www.abc.com and www.xyz.com. Note that Yahoo!, like other major search engines, adds Boolean operator AND by default in between query terms so the AND operator is omitted. In other words, the query syntax for the data without the keyword is effectively "(link:http://www.abc.com –site:abc.com) AND (link:http://www.xyz.com –site:xyz.com)".

**Table 1. Yahoo! Query Syntax**

| Type of Data Collected | Yahoo! Query |
|---|---|
| Without keyword | (link:http://www.abc.com –site:abc.com)<br>(link:http://www.xyz.com –site:xyz.com) |
| With keyword DSLAM | ((link:http://www.abc.com –site:abc.com)<br>(link:http://www.xyz.com –site:xyz.com)) DSLAM |

The "link" command of Yahoo! finds Web pages that link to a particular URL (in this study, links to a company homepage rather than all pages of the company Website). The "linkdoamin" command of Yahoo! will search for Web pages that link to all pages of a site. We decided to use the link command as earlier testing [5] showed that data collected using this command generated better mapping result than that using the "linkdomain" command. A content analysis study [10] that examined reasons of co-linking found that links to homepage were more likely to be business related than

links to non-homepage. This confirms that the "link" command is better than the "linkdomain" command for business Websites.

## 5 Results

Fig. 1 is the MDS mapping result for the set of data that were collected using co-link queries alone (i.e. the first search query in Table 1) while Fig. 2 is the result for the set of data that combined the keyword DSLAM with co-link search queries (i.e. the second search query in Table 1). One needs to be knowledgeable about the industry to interpret the two Figures. The second author has over a decade of experience working in the telecommunication industry. In the interpretation below, the notation of "business competition" is used based on the author's knowledge of the competitive environment of DSLAM industry rather than a strict definition of what constitutes a business competition.

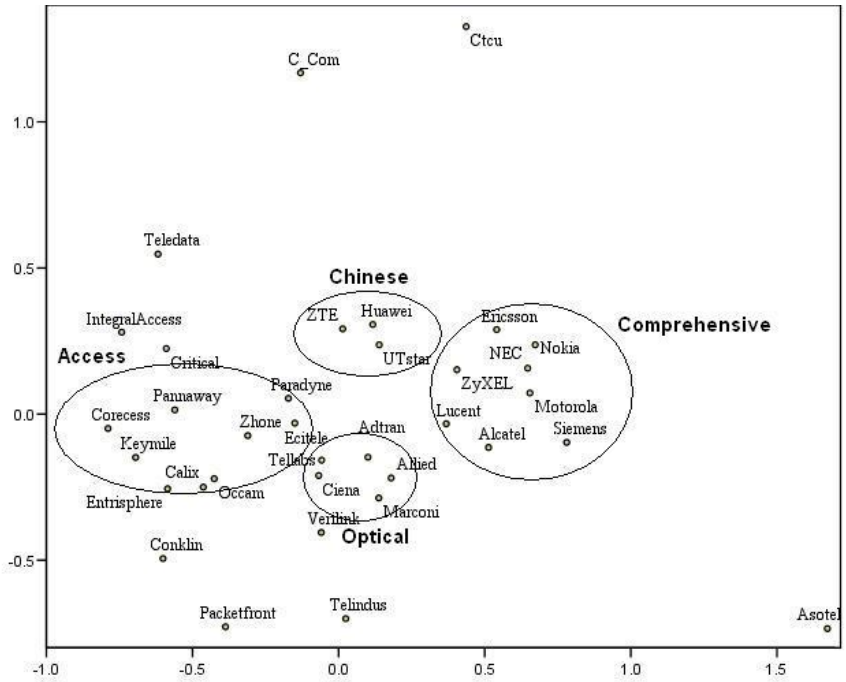


**Fig. 1** MDS map without keyword

In Fig. 1, company positions are determined by their overall business and Internet marketing strength in the general telecommunication equipment industry, not specifically in DSLAM marketplace. In this map, tier one telecommunication

equipment companies are clustered together in the "comprehensive" category. This group includes Alcatel, Ericsson, and Nokia who are major players in various sectors of the telecommunication industry. The only exception in this group is ZyXEL. ZyXEL is a relatively smaller Taiwanese company with US $325 million revenue a year. However, its strong Internet presence as shown in the Internet Archive (www.archive.org) makes it appear in tier one group. Tier two companies are grouped into several categories including "optical" and "access" companies. The former category includes Ciena, Tellabs, and Marconi which have strong market presence in optical networking while the latter includes Zhone, Paradyne, and Calix, companies that specialize in broadband access products including DSLAM. The Chinese companies are positioned together as a separate group. This reflects the fact that in the general telecommunication equipment market, these companies' main market focus and main competition are mainly in China rather than in Europe and North American. Other small companies are scattered all over the map.

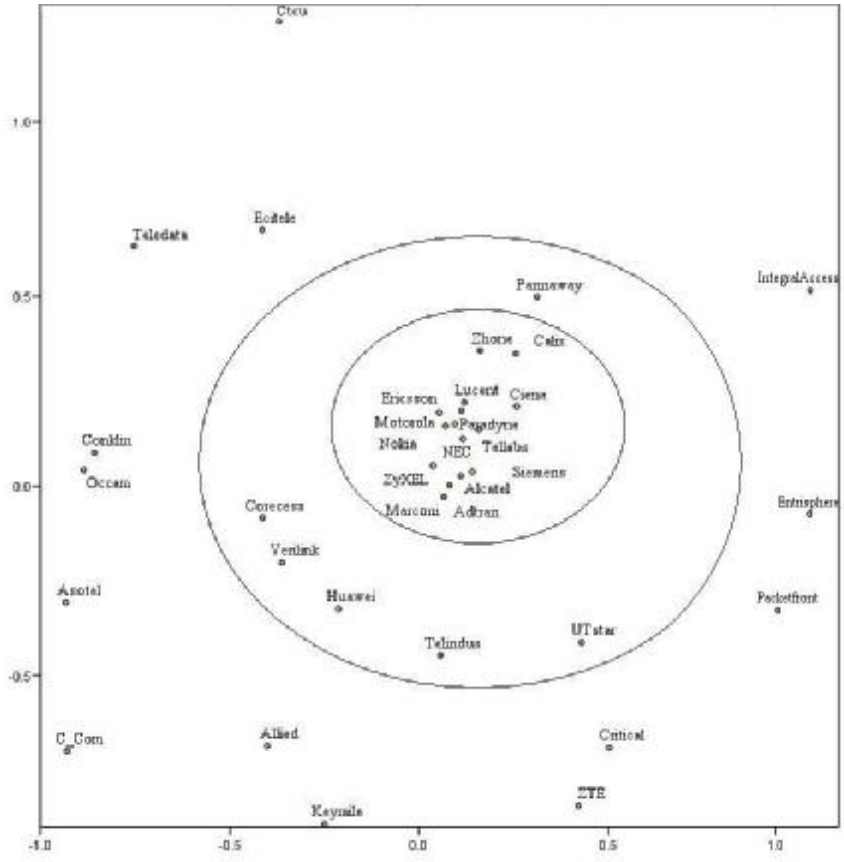


**Fig. 2**. MDS map with keyword DSLAM

By introducing the keyword DSLAM in search queries (see Table 1), we added a constraint to the relationships among these companies. The effect of this constraint can be seen from the strong contrast between Fig. 1 and Fig. 2. The two Figures are significantly different in that Fig. 1 shows company positions in the general telecommunication market while Fig. 2 shows their specific market positions in DSLAM sector. In Fig. 2, key DSLAM vendors, such as Paradyne, Zhone, and Calix who are tier two companies in Fig. 1, joined tier one companies to form the center of DSLAM sector (see the inner circle in the middle of Fig. 2). This cluster is denser than any cluster in Fig. 1. The close proximities of these companies in Fig. 2 reflect a very competitive DSLAM market. Other companies took different positions in Fig. 2 as well to form the second tier of DSLAM market as shown by the outer circle in Fig. 2. For example, the Chinese companies do not exist as a separate group in Fig. 2 anymore. Huawei and UTStarcom, the two leading DSLAM equipment companies in the Chinese market, are positioned fairly close to the center of Fig. 2. This reflects the competitive positions of these two companies in the world DSLAM market. In summary, adding the keyword constraint in data collection does serve the purpose of providing a more accurate competitive analysis for a particular market segment (in this case the DSLAM sector).

## 6 Conclusions and Discussion

The study proposed the method of "keyword enhanced Web structure mining" which combines the ideas of Web content mining (keyword) with Web structure mining (hyperlink). The method was used to mine data on business competition among a group of DSLAM companies. Specifically, the keyword DSLAM was incorporated into queries that searched for co-links between pairs of company Websites. Two sets of data were collected: one with the proposed method and one with co-link search alone. The resulting two data matrices were analyzed using multidimensional scaling (MDS) to generate maps of business competition. The comparison between the two maps shows that the proposed method produced a more accurate map of the business competition in the DSLAM sector. However, the proposed method does not refute the previous method of Web structure mining alone. The two methods are suited for different purposes of business intelligence. The method of structure mining alone is suitable for macro-level analysis of business competition as it generates a map of overall competition of an industry. The proposed method is better for micro-level analysis because it produces a map of business competition of a specific sector or segment. This new method represents a level of business intelligence that is higher than what was achieved before.

This is an exploratory study. The method was tested successfully only in one business sector. More testing in other environments is needed to determine if the method can be generalized. The scalability of the method needs to be examined as well. The study is a very first step in combining Web content mining with Web structure mining in that only one keyword was used. Future study will make more sophisticated use of keywords to further improve the method.

**Acknowledgements.** This study is part of a larger project funded by the Initiative on the New Economy (INE) Research Grants program of the Social Sciences and Humanities Research Council of Canada (SSHRC). Research assistant Karl Fast helped with the programming work for data collection.

## References

1. Madria, S., Bhowmick, S.S., Ng, W.K. & Lim, E.P. (1999). Research Issues in Web Data Mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, Florence, Italy, Aug. 30 – Sept. 1, 1999. Lecture Notes in Computer Science, Vol. 1676, 303-312.
2. Lu, Z., Yao, Y. & Zhong, N. (2003). Web Log Mining. In Zhong, N., Liu, J., & Yao, Y. (Eds) *Web Intelligence*. Berlin: Springer, 173-194.
3. Thuraisingham, B. (2003). *Web Data Mining and Applications in Business Intelligence and Counter-Terrorism*. Boca Raton, Florida: CRC Press.
4. Liu, B., Ma, Y., & Yu, P. S. (2001). Discovering Unexpected Information from Your Competitors' Web Sites, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 26-29, 2001, San Francisco, U.S.A., available at www.cs.buffalo.edu/~sbraynov/seminar/unexpected_information.pdf.
5. Vaughan, L. & You, J. (2005). Mining Web hyperlink data for business information: The case of telecommunications equipment companies. In Proceedings of the First IEEE International Conference on Signal-Image Technology and Internet-Based Systems, pp. 190–195, Yaoundé, Cameroon, Nov. 27–Dec. 1, 2005.
6. Björneborn, L., & Ingwersen, P. (2004). Towards a basic framework of webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.
7. Vaughan, L, Gao, Y. & Kipp, M. (2006). Why are hyperlinks to business Websites created? A content analysis. *Scientometrics*, 67(2), 291-300.
8. Beniston, G. (2005). IP DSLAMs, A Heavy Reading Competitive Analysis. Heavy Reading report series, Vol. 3, No. 15, August 2005. URL: http://www.heavyreading.com/details.asp?sku_id=836&skuitem_itemid=793&promo_code=&aff_code=&next_url=%2Flist%2Easp%3Fpage%5Ftype%3Dall%5Freports.
9. Google (2006). Google SOAP Search API Reference. Retrieved Aug. 18, 2006 from http://www.google.com/apis/reference.html#2_2.
10. Vaughan, L., Kipp, M.E.I., Gao, Y (2006). Are colinked business Web Sites really related? A qualitative study. Paper under review in *Online Information Review*.