Liwen Vaughan
Faculty of Information and Media Studies, UWO, London, Ontario
Rongbin Yang
Institute of Scientific and Technical Information of Shanghai, Shanghai, China
Chao Chen
Institute of Scientific and Technical Information of Shanghai, Shanghai, China
Weibo Liang
Shanghai Maritime University, Shanghai, China
Baoyi Li
Shanghai Maritime University, Shanghai, China

Extending Web Co-link Analysis to Web Co-word Analysis for Competitive Intelligence

Abstract: The study carried out Web co-link analysis and Web co-word analysis for a group of companies in the international shipping industry. Co-link and co-word data were collected and analyzed with MDS. Results from different data sets were compared and advantages and disadvantages of the two methods were examined.

Résumé : Cette étude a effectué une analyse des co-liens et des cooccurrence de termes sur le web pour un groupe de sociétés de l'industrie du transport international. Les données ont été recueillies et analysées par échelonnement multidimensionnel. Les résultats provenant de différents jeux de données ont été comparées et les avantages et désavantages des deux méthodes, examinées.

1. Background of the study

In the area of Web data mining for competitive intelligence, studies have found that the number of links pointing to a company's Website (inlinks) correlates significantly with the company's business performance measures such as revenue and profit (Vaughan, 2004). The correlation is significant even after factors such as the size of the company is accounted for. This demonstrates that inlinks contain useful business information.

Extending the concept of inlinks, the concept of co-links has been developed (Björneborn & Ingwersen (2004). If page X and page Y both received links from page Z, then page X and page Y are co-linked. For example, FIMS of University of Western Ontario and SLIS of University of Alberta are co-linked by ASIST. It has been found that companies that are co-linked are business competitors and co-link data have thus been successfully used to map business competition scenes (Vaughan & You, 2006).

The current study builds on the previous studies by extending the co-link concept to the co-word concept in order to extend data sources and to obtain richer information. Co-word refers to two words or terms that appear on the same Webpage. In this study, the words or terms are company names. Our hypothesis is that if the names of two companies co-occurred frequently on the same Webpages, then the two companies are likely to have similar or related business. Although companies can be related for non-competition reasons such as environmental concerns, on a macro level, related companies are likely to be business competitors (two unrelated companies such as a food company and software

company are not competitors). The co-word data can thus be used to map business competition scenes just as co-link data did.

The extension of the co-word concept is not just a technical one. It has the potential to greatly enrich data sources for competitive intelligence. Currently, some major search engines do not provide complete inlink search functions. For example, Google's inlink search will only retrieve a sample of inlinks that the database indexes. This limits sources for inlink data collection. In contrast, co-word search is based on keyword search and all major search engines provide strong keyword search functions. Another reason of exploring the co-word concept is to tap into the rich data source of Web 2.0. Currently, no inlink search is available in these sources. Our study specifically examined how co-word data retrieved from blogs can be used to map business competition scenes. A very important feature of blogs is that they all have explicit date and time stamps (ordinary Webpages do not have this feature). This feature is very useful in tracking business information over time. In addition to the value of practical application of Web data for competitive intelligence, our study also contributes to our knowledge of the Web. Studying co-word phenomenon helps us to understand keyword use and keyword distribution on the Web.

Our research questions are: what kind of business information can we get from the Web co-word data? Specifically, can the co-word data be used to map business competition scene? What are the advantages and disadvantages of the co-word data compared with co-link data? Which sources of co-word data (ordinary Webpages vs. blogs) is better?

2. Methodology

We took an empirical approach to the study. We selected a group of companies to study; located company Websites; collected both co-link and co-word data about these companies; analyzed the data using multidimensional scaling (MDS); compared MDS maps generated from different Web data sources to determine the type of business information that can be obtained from each data source and the advantages and disadvantages of each data source.

We selected the international shipping industry for the study. A main reason of this selection is that the international shipping industry is like a thermometer of the global economy in that it reflects the amount of trade activities (both import and export). A Web data mining method for this industry is useful in monitoring the global economy. The recent global economic recession had a major impact on the industry. There are plenty of Websites discussing the issue, which provided rich data source for the study. We chose the world's top 40 companies (measured by the operated fleets) listed on AXS-Alphaliner, an authoritative information platform. Three of these 40 companies had to be omitted from the study. One company had few co-links with other companies so it could not be used for co-link analysis. The names of two other companies coincided with other words so they could not be used for co-word analysis. As a result, a total of 37 companies were left for final analysis. It should be noted that the purpose of the study is not to find specific business information about these companies but rather to use our knowledge of the industry to judge if and how findings based on Web data match the industry reality. Two of the co-authors of the paper are information professionals specializing in this

industry. We also conducted extensive searches of information from various sources to extend our knowledge of the industry.

We collected co-link from Yahoo! because at the time of data collection (summer 2009) Google and MSN Bing could not do the co-link search that we needed. We collected co-word data from Google and Google Blogs. Google was chosen because it is the most commonly used search engine and it has a blog search engine at http://blogsearch.google.com (Yahoo! and MSN Bing did not have blog search engines at the time of the study). The three sets of data (co-link from Yahoo!, co-word from Google, and co-word from Google Blogs) were each collected within a one day period to mitigate the possible effect of Webpage changes on the data collected. The stress values of MDS analysis were 0.068, 0.063 and 0.057 for co-link, Google co-word, and Google Blog co-word respectively.

3. Findings and Directions of Future Research

The three MDS maps generated from the three sets of data have a similar pattern in that they positioned the companies by their global competitiveness rather than by country or language. Top competitive companies are positioned in the centre of the maps while lower ranked companies are positioned in the out skirt. So the MDS maps do reflect the business competition scene. This shows that the three Web data sources all contain useful business information and can be objects of Web data mining for competitive intelligence. A close comparison of the three MDS maps revealed the relative merits of each data source. The accuracy of the co-link data is contingent upon the URLs of the companies. If a company has multiple domain names for its branches, hyperlinks to the company Websites are split among different domain names. As a result, the co-link count will not correctly reflect its competitive position. This caused two top companies not being positioned in the center of the MDS map as they should be. The Web co-word data did not have this problem at all because the URLs have no bearing on the data. However, the co-word data are very sensitive to the company names. If the company name is not unique, e.g. it happened to coincide with a regular word or term, then the co-word data will be noisy (retrieving pages not about the company). This problem is more pronounced in data retrieved from Google than that from Google Blogs. The sheer size of the Google database and the diversity of pages it indexed increase the chance of noise.

Comparing the three MDS maps, we found that overall the MDS map generated from the blog data provides the most accurate visualization of the industry. It should be noted that the advantage of blog data over other types of data reported here may not be generalizable to other industries. We plan to test the co-word method developed in this study in other industries. We also plan to investigate how different types of Web data can be amalgamated into a single one that takes advantages of different data sources and offsets their disadvantages. We hope that the MDS map generated from this merged data set will more accurately reflect the competition scene than the MDS maps based on separate data sources.

Acknowledgments

This study is part of a larger project of Web data mining for business intelligence funded by the Social Sciences and Humanities Research Council of Canada (SSHRC). Research assistant Yijun Gao helped with data collection work.

4. References

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.

Vaughan, L. (2004). Web hyperlinks reflect business performance: A study of U.S. and Chinese IT companies. *Canadian Journal of Information and Library Science*, 28(1), 17-31.

Vaughan, L. & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3), 611-628.