Web Co-word Analysis for Business Intelligence: Applicability to the Chinese Language

Liwen Vaughan[1], Rongbin Yang[2], and Juan Tang[3]

[1]*lvaughan@uwo.ca*
Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, (Canada)

[2] *rbyang@libnet.sh.cn;* [3]*jtang@libnet.sh.cn*
Institute of Scientific and Technical Information of Shanghai, Shanghai, (China)

Introduction

Webometrics started from Web hyperlink analysis but recent studies covered a broader spectrum of the Web such as blogs (Bar-Ilan, 2007; Smith, 2007), RSS feeds (Thelwall & Prabowo, 2007), and social network space (Wilkinson & Thelwall, 2010). Cumulatively, a large, if not the largest, body of the literature focused on link analysis that examined various types of Websites ranging from university (Ortega & Aguillo, 2009) and government sites (Holmberg, 2009) to business (Vaughan, Tang & Du, 2009) and political sites (Romero-Frías & Vaughan, 2010). However, inlink data collection from major commercial search engines has been limited to Yahoo! alone in recent year. Since the summer of 2010, Yahoo! Web interface at www.yahoo.com no longer supports co-link search, although this search is still available at Yahoo! API.

In order to taping into more Web data sources and to develop new Webometrics methods, Web co-word analysis has been proposed (Vaughan & You, 2010). The test in the English environment showed the potential of the method but more studies are needed to find out if the method can be applied to other contexts. In this study, we tested the applicability in the Chinese environment. This test is needed because the Chinese Web environment is different from that of the West. More importantly, the nature of the Chinese language is also very different from that of English. Unlike the co-link method which is language independent as URLs are language independent, the co-word method is language dependent as it uses words for data collection. A very important characteristic of the Chinese language in terms of Web searching is that there is no space in between words. In English as well as in many other languages, words are separated by spaces or punctuation marks so computer programs can recognize words because of this. For the Chinese language, an individual character is usually not specific enough for searching. Typically, two or more characters are needed to form a meaningful search string. However, the lack of space between characters poses a challenge for computer searching and this could affect co-word search results.

Methodology

We took an empirical approach to the study. We selected a group of companies in two Chinese industries (IT and chemical); collected co-link and co-word data for the companies; analyzed the data with multidimensional scaling (MDS); and compared maps generated from the co-link data with that from the co-word data to find out if the co-word method works. An earlier study (Vaughan & You, 2010) compared co-word data collected from the general Web with that from blogs and found that the latter is a better data source. The current study collected data from both sources. However, due to the space limitation, this paper will not report co-

word analysis results from the general Web.

Co-link data were collected from Yahoo! API. The query of searching for co-links between company with URL www.xyz.com and company with URL www.abc.com was (linkdomain:xyz.com –site:xyz.com) (linkdomain:abc.com –site:abc.com). To find a search engine to collect co-word blog data, we examined all major Chinese search engines to find out which one was the most appropriate. At the time, Sept. 2010, Yahoo! China www.yahoo.cn did not have a blog search engine. Baidu www.baidu.com had a blog search engine at http://blogsearch.baidu.com but it was in Beta and did not have advanced search features. In contrast, Google China (located in Hong Kong) had a blog search engine at http://blogsearch.google.com.hk and it was not in Beta. It also had advanced search features including searching for specific time periods, specific languages, authors and titles. So we used this blog search engine for data collection.

Web co-word refers to the co-occurrence of two words or terms on a Webpage. In this study, the unit of the analysis is individual companies, so the words/terms are company names. The mentioning (co-occurrence) of two company names on the same Webpage suggests that two companies are related. As related companies are likely to be business competitors, the word co-occurrence data could be used to mine business competition information. This parallels the method of co-link analysis for business intelligence which has been proven successful in numerous studies (e.g. Vaughan, Tang & Du, 2009). We used company acronyms instead of full names for co-word data collection because acronyms are more likely to be used on Webpages, especially blog pages.

We examined various sources to determine the appropriate Chinese acronyms for each company. If there were two or more variations of acronyms for a company, we used the most common one for data collection. All acronyms involved two or more Chinese characters and we searched acronyms with quotation marks around them to make the search more accurate. Both co-link and co-word data were normalized by Jaccard Index and then analyzed with multidimensional scaling (MDS). A total of 36 chemical and 44 IT companies were included in final analysis.

Findings and Discussion

Summarizing the results from the two industries, we found that the co-word method worked to some extent but it was not as good as the co-link method. An early study (Vaughan & You, 2010) comparing the two methods in the English language environment found that the co-word method was as good as co-link method. This suggests that the co-word method could potentially be applied to the Chinese language but currently it is not as effective as in the English environment. Could the co-word method be improved so that better results would be achieved in the Chinese environment or the method just would not work as well as in English due to the nature of the Chinese language? More research is needed to find out.

Comparing the two methods, we found that the co-link and co-word methods both have advantages and disadvantages. Co-link method is very sensitive to the domain name problem while the co-word method is not. On the other hand, the co-word method is sensitive to the acronym problem (a company has more than one acronym) but the co-link method is not. This is, however, not an inherent problem of the co-word method but rather a limitation of the search engine used for data collection. The blog engine of Google China that we used did not support full Boolean operations. When a company had two or more variations of acronyms, we were force to use the most common one and ignore others. Theoretically, we should use Boolean operator OR to connect all variations of acronyms of one company

and then use AND to connect it with variations of acronyms of the other company. This nested Boolean operation was not supported in the blog search engine at the time of data collection. If this operation is possible in future, the result of co-word analysis is likely to be better. Furthermore, if we could find a way to combine the co-link and co-word data to take advantages and avoid problems of both methods, we would have a more robust tool. This is a direction that is worth pursuing in further research.

Acknowledgments

This study is part of a large 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.

References

Bar-Ilan, J. (2007). The use of Weblogs (blogs) by librarians and libraries to disseminate information. *Information Research*, 12(4), Available at: http://informationr.net/ir/12-4/paper323.html.

Holmberg, K. J. (2009). Webometric network analysis: Mapping cooperation and geopolitical connections between local government administration on the Web. PhD dissertation, Åbo Akademi University, Finland.

Ortega, J.L. & Aguillo, I. (2009). Mapping world-class universities on the Web, Information *Processing & Management*, 45(2), 272-279.

Romero-Frías, E. & Vaughan, L. (2010). European political trends viewed through patterns of Web linking. *Journal of the American Society for Information Science and Technology*, 61(10), 2109-2121.

Smith, A.G. (2007). Issues in "blogmetrics": case studies using BlogPulse to observe trends in weblogs. In D. Torres-Salinas & H. Moed (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*, Madrid, 25-27 June 2007, p. 726-730.

Thelwall, M. & Prabowo, R. (2007). Identifying and characterising public science-related fears from RSS feeds. *Journal of the American Society for Information Science and Technology*, 58(3), 379-390.

Vaughan, L., Tang, J. & Du, J. (2009). Examining the robustness of Web co-link analysis. *Online Information Review*, 33, (5), 956-972.

Vaughan, L. & You, J. (2010). Word co-occurrences on Webpages as a measure of the relatedness of organizations: a new Webometrics concept. *Journal of Informetrics*, 4(4), 483–491.

Wilkinson, D. & Thelwall, M. (2010). Social network site changes over time: The case of MySpace. *Journal of the American Society for Information Science and Technology*, 61(11), 2311–2323.