Mining Web Hyperlink Data for Business Information: The Case of Telecommunications Equipment Companies

Liwen Vaughan
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, N6A 5B7, Canada
lvaughan@uwo.ca

Justin You
ApacBridge Consulting
8 Northgate Street
Ottawa, Ontario, K2G 6C7, Canada
justin.you@apacbridge.com

Abstract[1]

Few studies have explored the possibility of mining Web hyperlink data for e-commerce or business information. This study is an attempt to fill this gap. The project selected a group of telecommunications equipment companies and collected data on the number of links pointing to the company Websites (inlinks) and the company's revenue. A significant correlation between the two variables was found, suggesting that inlinks contain useful business information and can be objects for Web data mining. The project then explored the feasibility of using Web co-link data to map business competition positions. Co-link data (links pointing to a pair of company Websites) were collected and analyzed using multidimensional scaling (MDS). MDS maps correctly clustered these companies into sectors of the telecommunications equipment industry. Data collection was repeated after nine months and the results confirmed the reliability of the methodology developed in the study.

1. Introduction

Hypertext links on a Web page, designed for guiding users to related information and for navigational purposes, have been used in information retrieval algorithms to rank search results. However, few studies have explored the possibility of mining Web hyperlink data for E-commerce or business information. This study is an attempt to fill this gap.

It is necessary 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. If page X and Y are both linked to by page Z (i.e. page X and Y both have inlinks from page Z), then X and Y are co-inlinked. This study explores the potential of Web data mining using inlinks and co-inlinks (also called co-links later in the paper).

The following two hypotheses were tested in the study. First, the number of inlinks to a business Website correlates with the company's business performance. If this hypothesis is true, then Web hyperlinks contain business information and thus can be used for Web data mining. Second, the number of co-links to the Websites of a pair of companies can be used to cluster companies into a map of business competition. The reasoning behind this hypothesis is that the co-link count could be a measure of the similarity between the two companies. The more co-links the two companies have, the more closely related they are in the views of the sites that link to them. Since related businesses are competing businesses, Web co-link data can be used to cluster companies into a map of business competition. A group of telecommunications equipment companies were selected based on the following criteria to test the hypotheses: top companies in terms of revenue in the telecommunications equipment industry; top companies from different regions of the world; representation of companies from different sectors of the telecommunications equipment.

The remaining parts of the paper are organized as follows. First, different options for data collection and data analysis are described and the chosen options are justified. Next, the finding of a significant correlation between inlink count and the business performance measure of a company was reported, which paves the way for the co-link

[1] 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.

analysis that follows. Then results from the two rounds of co-link data were presented. Finally, the limitation of the study as well as the direction for future study is discussed.

2. Methods

2.1 Business performance data

There are many business performance measures such as revenue and profit. For public companies, these data are usually publicly available. For private companies, however, most financial data are kept secret. Since there were both public and private companies in the study, we had to restrict the choice of business performance measures to one that is most likely to be available for private companies, revenue. Other data, such as profit, would be useful but they are not available for private companies. Revenue data for public companies were collected through Yahoo! Finance for public companies (finance.yahoo.com) while that for private companies were collected from the companies' Websites.

2.2 Choice of Search Engine for Data Collection and Search Queries Used

There are two types of inlinks, total inlinks and external inlinks. Total inlinks include all links pointing to a particular 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. Our study only examines external inlinks because internal links are not indicators of online visibility or impact. Google cannot perform external link search as is indicated in the Google API documentation [1]. Yahoo was chosen for data collection.

The queries for external inlink search and co-link search are shown in Table 1. In Table 1, partial domain names were used in the "site" portion of the query (i.e. "www" was not include). The partial domain name search can do a more complete capture of all inlinks to the Website because it is conceivable that a related URL (e.g. mail.abc.com) could be used by the company in addition to the standard www.abc.com. This is consistent with earlier studies that collected inlink data through search engines [2, 3, 4].

Table 1. Yahoo! search queries

To search for	Query Used
inlinks to www.abc.com	link:http://www.abc.com - site:abc.com
co-links between www.abc.com and www.xyz.com	(link:http://www.abc.com – site:abc.com) AND (link:http://www.xyz.com – site:xyz.com)

It should be noted that Yahoo!'s "link" command only finds pages that link to a particular URL (in this study, links to a company homepage rather than all pages of that company's Website). Yahoo! has an undocumented command "linkdomain" which will search for links to all pages of a Website. Link count data were collected using both search commands. The comparison of the two data sets showed that the majority of links point to homepages. The result based on data collected with the "link" command correlates better with our knowledge of the telecommunications equipment industry. Thus the "link" command was considered to be a better choice for this study and all results reported below are based on data collected using this command. However, the relative advantages of the two commands for co-link studies in general are unknown and need to be explored in future studies.

As the Web is constantly changing, data collected from search engines in different time periods will also change. This provides us with an opportunity to collect different data sets to validate results and to determine if the market changes over time are reflected in the Web hyperlink data. Thus, two rounds of data were collected in this study. The first round of data was collected in July 2004 while the second round was collected in April 2005.

2.3 Data Processing

The co-link data collected were stored in the form of a symmetrical matrix with each row and column representing the URL of a company. The number in the cell row x and column y is the number of co-links between URL X and URL Y, i.e. number of pages with links that point to both URLs. The matrix (size 32 by 32, lower-triangle) is omitted here due to space limitations. This raw co-link matrix can be fed directly into the multidimensional scaling program for analysis. However, the raw co-link count may not be an accurate measure of the strength of the relationship between a pair of companies. For example, a co-link count of 5 is very high if the number of links

pointing to each company is 6. It will be low if the number of links pointing to each company is 100.

To measure the relative strength of the relationship between a pair of companies, we normalized the co-link counts by Jaccard Index as follows:

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

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

We fed both the raw co-link matrix and the normalized co-link matrix into the multidimensional scaling (MDS) program of SPSS version 12 and compared the mapping results from the two matrices. Judging from our knowledge of the telecommunications equipment industry, the map from the normalized co-link matrix depicts the company relationships better. So, the maps shown in the following sections are all from the normalized matrices. All MDS results reported below have a normalized raw stress value of less than 0.05, which indicates a very good fit between the input data and the out maps [5] p. 201.

3. Results

3.1 Significant correlation between inlink count and revenue

There is a highly significant correlation between revenue and inlink count data collected in July 2004 (Pearson correlation coefficient is 0.74,

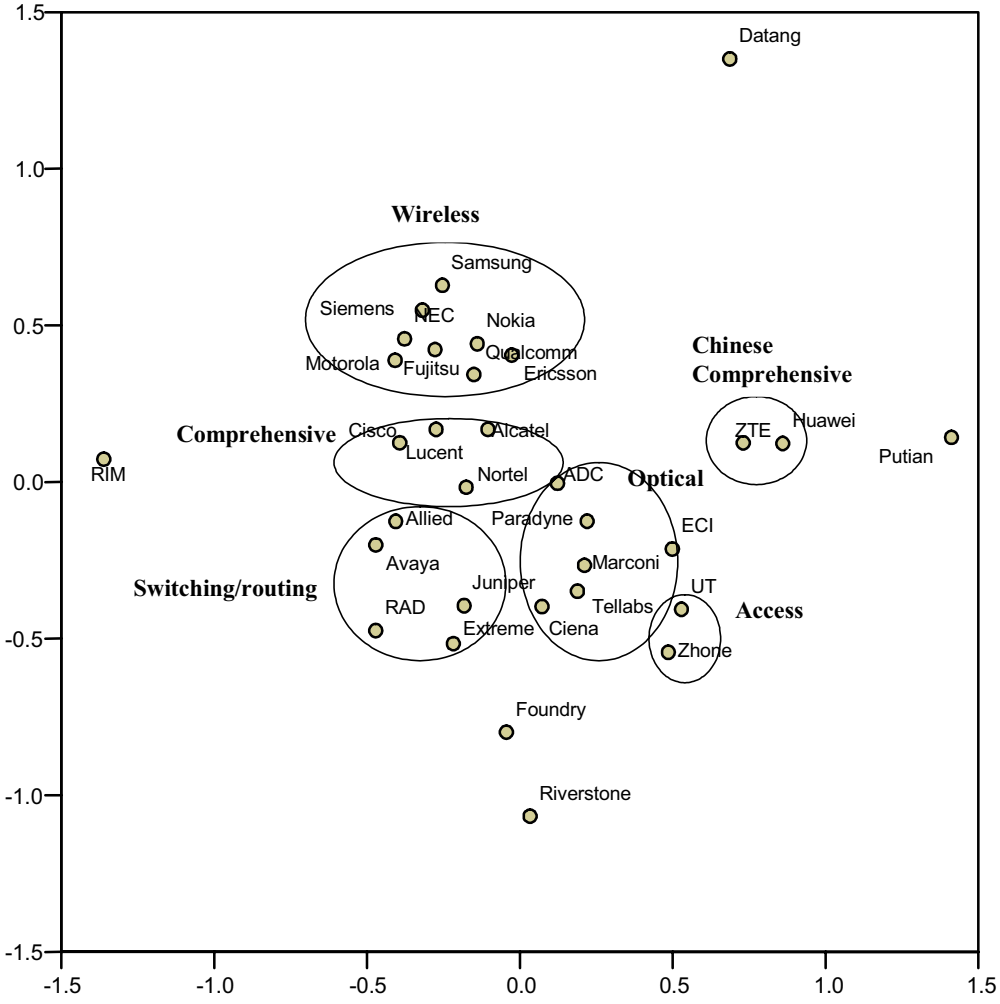

Fig. 1. MDS mapping result based on July 2004 Data

$p<0.01$). Data collection repeated in April 2004 confirms this correlation (Pearson correlation coefficient is 0.79, $p<0.01$). Thus it can be concluded that there is a correlation between a company's business performance and the number of inlinks to the company Website. Companies that are doing well in business also have higher Web profiles in that their Websites attracted more links and thus they are more visible on the Web. Web visibility is particularly important for E-commerce. This finding means that links to business Websites are not a random phenomenon but rather contain useful business information and can thus be objects for Web data mining. Based on this finding, we further explored the possibility of using co-link data (defined earlier in the *Methods* section of the paper) to map company market positions.

3.2 Co-link analysis of market positions

Fig. 1 is the MDS output map that shows the relative positions of the 32 companies based on the co-link data collected in July 2004. The companies are clearly clustered into sectors of the telecommunications equipment industry as labeled in Fig. 1, namely "wireless" companies, "comprehensive" companies, "optical" transport companies, "routing/switching" companies, "access" companies and "Chinese comprehensive" companies. The map shown in Fig. 1 is consistent with the competition landscape of the telecommunications equipment industry. For example, Canadian company RIM (Research in Motion) is not clustered with other companies. This shows the unique product and market position of RIM. RIM's main product, the BlackBerry, is a handheld wireless access device which has its own niche market and is not really competing with products from other wireless companies in this study.

The two other companies that are not grouped into clusters in Fig. 1 are Putian and Datang. These two Chinese companies are new comers in the global telecommunications equipment market. They are big telecommunications equipment companies in the Chinese market. However, they have little exposure so far in the international market. Their revenues are mainly from the domestic market. In contrast, the other Chinese powerhouses, Huawei, ZTE and UT Starcom, are in a closer position to their international competitors. This reflects the fact that these three companies are far more successful than Putian and Datang in the international market. For example, Huawei's revenue from the international market is around US $2 billion in 2004 which is 40% of its total revenue in that year. ZTE had 30% of its revenue from the international market in 2003. These two companies also pose more competitive threats to Western telecommunications equipment vendors.

The two Japanese companies that are not positioned according to the characteristics of their telecommunications products are NEC and Fujitsu. NEC should be treated as a comprehensive telecommunications equipment company while Fujitsu's telecommunications market focus is optical transport. The incorrect positioning could be attributed mainly to the following reason. These two companies have a wider product portfolio in IT and have other electronic products which do not belong to the telecommunications industry. Web links to those other products are included in the data we collected.

The success of this mapping from co-link data suggests that co-link data do contain information about the relationship among companies. Highly co-linked companies are highly related in their products and the market. Since related companies are competitors (they serve the same market needs), it follows that co-link data can be used to map the competitive position of companies. When combining the inlink data with the co-link mapping result of Fig. 1, we can get a better understanding of the competitive positions of these companies. For example:

- Cisco is the most competitive company in the "comprehensive" group and has the highest inlink count. Its main competitors are Lucent and Nortel which are positioned very close to it. Nortel, Alcatel and Lucent are located in the "comprehensive" group since they have wider product portfolios ranging from wireless infrastructure equipment and optical transport equipment to switching equipments and access equipment.
- Nokia, having the largest number of inlinks in the Wireless sector, is the most competitive player in this group. Compared with the companies in the comprehensive group, companies in the wireless group have product lines heavily focused in the wireless sector. Their product portfolios are not only in the wireless infrastructure equipment market but also in the wireless handset market.

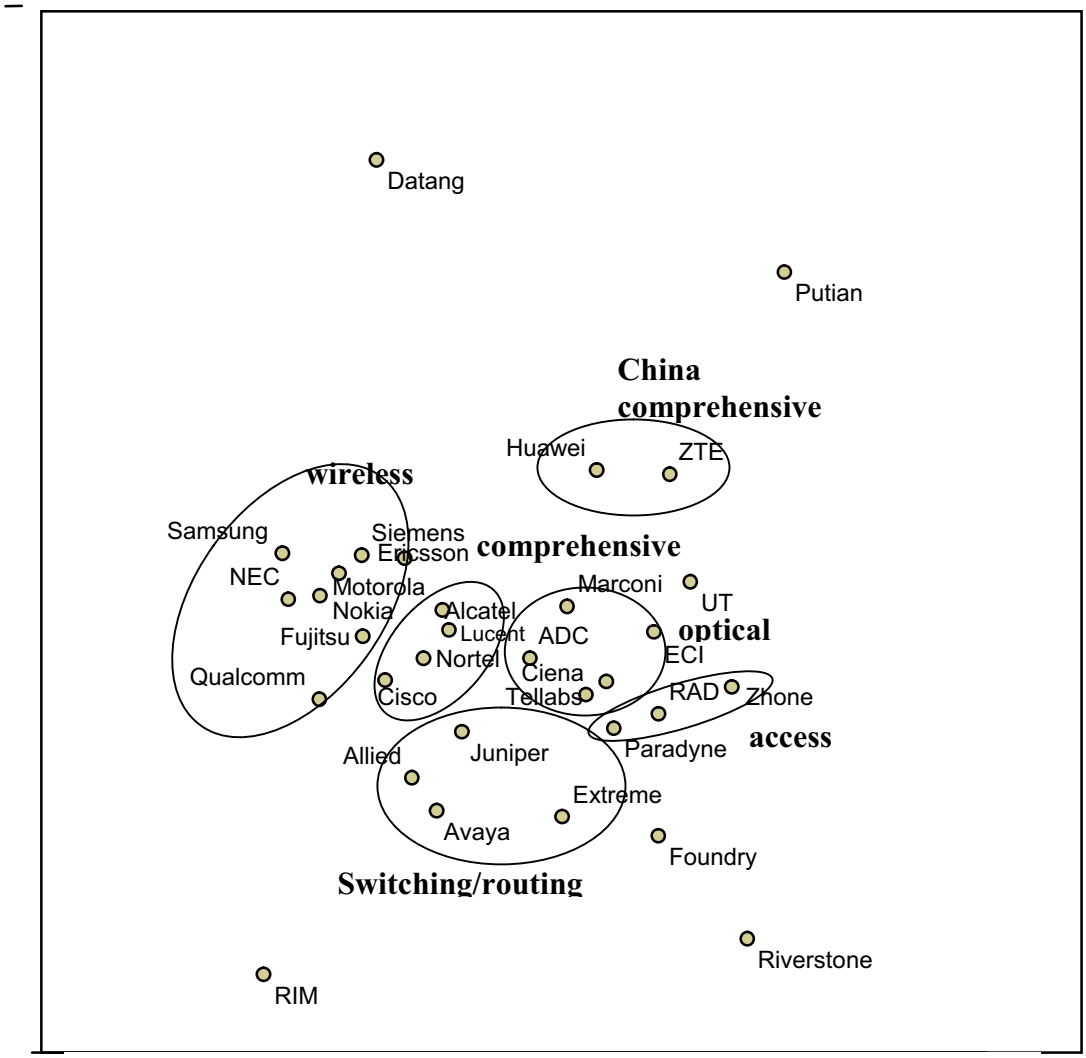

Fig. 2. MDS mapping result based on April 2005 Data

- The Routing/Switching group includes companies like Juniper, Extreme, Avaya, and Allied Telesyn. These vendors are mainly competing with each other in routing and switching equipment market. Extreme is very competitive in this group. It is the number one competitor for Juniper and their close positions in Fig. 1 shows this. Although Cisco competes heavily with Juniper in the router market, Extreme is positioned closer to Juniper in Fig. 1. Cisco is positioned in the comprehensive group in Fig. 1, which correctly reflects the fact that the overall competition between Juniper and Extreme in routing/switching is more significant than that between Juniper and Cisco.

- Huawei has not yet been Cisco's close competitor in the global market. It competes more with the optical transport companies in the router market. The co-link mapping result of Fig. 1 correctly puts Huawei in a closer position to the optical access group. The finding from Fig. 1, which is in line with the analysts' view, is that in the global market, the Chinese telecommunications equipment vendors will be competitive initially in the low margin optical access equipment market rather than in the high margin switching/routing equipment market.

Co-link data collected in April 2004 (9 months after the initial data collection of July 2004) resulted in the map as shown in Fig. 2. Fig. 2 shows similar clusters to that of Fig. 1 except the

slight changes as noted below. If you rotate Fig. 2 by 90 degrees clockwise, you will find a better match between the two maps. MDS maps can be rotated freely for interpretation. MDS maps are used to detect clusters or the relative positions of objects, not to show their absolute positions. Fig. 2 shows that the majority of the companies stayed at the same positions, especially those that have dominant positions in their market segments. This is to be expected as there is only a nine month gap between the two data collections. In fact, the relative stability of the two sets of results demonstrates the reliability of the methods developed in the study. The followings are changes from Fig. 1 to Fig. 2.

- RAD moved from the "Switching/Routing" group to the "Access" group. This correctly reflects RAD's market position. RAD has changed from a complimentary access company for switching/routing vendors to an "access solutions" company.
- Paradyne, an access company focusing on broadband access products, moved from the "Optical" group to the "Access" group. This is a more accurate market position for Paradyne.
- Juniper moved closer to Cisco, which reflects the fact that Juniper is a closer competitor of Cisco in the router product category.

4. Conclusions and Future Research

The study found that there is a significant correlation between inlink count and the company's business performance as measured by revenue. Companies that are doing well in business are more visible on the Web in that their Websites attracted more inlinks. The correlation coefficient of over 0.7 is higher than that of previous studies [4, 6], probably because the companies in the current study form a more homogenous group. This suggests that Web data mining based on inlinks needs to be applied to a particular industry.

The study applied MDS to co-link data to generate a map that clusters similar companies together. As similar companies are competitors, this map effectively shows the business competition landscape. Data collection repeated after nine months confirmed the reliability of the methodology and also reflected some market movements. Combining inlink data with co-link data provided a more complete depiction of the companies. Information from this kind of analysis can be used by industry analysts and decision makers who need an objective and global view of the competition landscape. The information can also complement the knowledge of business people who have direct experience in these sectors but would need information from a different angle to verify or expand their views. For example, the competitive position of the Chinese company Huawei is clearly illustrated in this study. However, Canadian companies were not as keenly aware of this competitor as they should be until Nortel's CEO pointed it out recently [7].

Further research is needed to test and verify the method in other sectors. Regular data collection is currently being conducted to further test and improve the methodology. It is hoped that data collected with longer time gaps will show more clearly the market movements of the industry.

References

[1] Google (2005). Google Web APIs Reference, available at http://www.google.com/apis/reference.html#2_2 (accessed June 8, 2005).

[2] Smith, A. & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363-380.

[3] Thelwall, M. (2001). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52 (13), 1157-1168.

[4] Vaughan, L. & Wu, G. (2004). Links to Commercial Web Sites as a Source of Business Information. *Scientometrics*, 60(3), 487-496.

[5] Meulman, J. & Heiser, W. (2001). *SPSS Categories® 11.0*. Chicago, USA: SPSS Inc.

[6] Vaughan, L. (2004). Exploring Website features for business information. *Scientometrics*, 61(3), 467-477.

[7] Maistre, R. L. (2004). Nortel Leaves All Doors Open. News published in June 02, 2004. Retrieved Dec. 1, 2004, from http://www.lightreading.com/document.asp?doc_id=53776&site=lightreading