Links to Commercial Web Sites as a Source of Business Information

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

Guozhu Wu
Shanghai Library/ISTIS
Shanghai, China, 200031
E-mail: wgz3@263.net

Abstract

A study of hyperlinks on commercial Web sites was carried out to fill a gap in Webometrics research in this area. Web sites of China's top 100 information technology (IT) companies were examined. The link count to a company's Web site was found to correlate with the company's revenue, profit, and research and development expenses. This suggests that Web hyperlinks to commercial sites can be a business performance indicator and thus a source of business information. This information is useful for Web business intelligence and Web data mining. As a comparison to IT companies, China's top 100 privately owned companies were also studied. No relationship between link count and the business performance measure was found for these companies due probably to the heterogeneous nature of this group. Data collection issues for Webometrics research were also explored in the study.

1 Background and Purpose of the Study

Numerous Webometric studies have been carried out to examine the nature and structure of Web hypertext links. Useful information has been uncovered from the link count and link topology. For example, link counts to university Web sites were found to correlate with the universities' research ratings (Thelwall, 2001; Smith & Thelwall, 2002); links to faculty Web sites were found to correlate with the rankings of these faculties (Chu et al., 2002). Link counts to journal Web sites were found to correlate with the quality of the journals as measured by the Journal Impact Factor (Vaughan and Hysen, 2002; Vaughan and Thelwall, 2003). However, most link analysis studies so far have focused on academic or scholarly Web sites.

There have been studies on commercial Web sites, ranging from general design issues, e.g. site structure and appeal (Nel et al, 1999; White and Manning, 1998) to technical issues, e.g. meta tags for indexing (Thelwall, 2000). The value of the Web for business intelligence purposes has been discussed (Graef, 1997; McGonagle and Vella, 1999) and specific techniques have been proposed (Nordstrom and Pinkerton, 1999; Bauer and Scharl, 2000; Zanasi, 1998; Burwell, 1999; Fleisher and Blenkhorn, 2001). A sophisticated program has been developed to gather information from competitors' Web sites (Liu et al, 2001). Various computer programs are available for Web data mining for business purposes (Madnick and Siegel, 2002). However, all

Liwen Vaughan and Guozhu Wu, *Links to commercial web sites as a source of business information*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 321-329.

these studies used Web site *content* for gathering business information. Web hyperlinks, a rich information source, have yet to be exploited directly for business purposes.

In short, we know much more about the nature and patterns of links to academic and scholarly Web sites than we do about those to commercial Web sites. This is an ironic situation given that the Web is dominated by commercial sites numerically, with 83% of Web servers containing commercial content (Lawrence and Giles, 1999). Obviously, more research into commercial Web sites is needed if we want to gain a full understanding of the Web. Toward this end, we carried out a Webometric study that examined the inlinks to commercial Web sites (inlinks means links coming to a Web site rather than going out from the site). Our hypothesis is that the number of inlinks to commercial sites correlates with some business measures. If this hypothesis is true, then inlink counts can be used as a business performance indicator or a source of business information. This information is useful for gathering business intelligence from the Web and Web data mining.

2 Methodology

The methodology we used to test the hypothesis was to select coherent groups of companies; locate their Web sites; collect data on the number of inlinks to each Web site; collect data on each company's business measures; and carry out statistical tests to determine if the link count correlates with any of the business measures.

2.1 Companies in the Study

China's top 100 Information Technology (IT) companies, as ranked by China's Ministry of Information Technology Industry (Ministry of Information Technology Industry, 2002), were selected for the initial investigation. The IT industry was selected because companies in this industry are likely to be leaders in using the Web for business purposes. The top 100 companies were selected rather than a random sample of all companies because the top companies are more likely to have stable Web sites to be studied and reliable data on their business measures are more likely to be publicly available. The pattern of Web use by these companies, although more sophisticated than average and thus atypical, could be the trend that will be followed by other companies. As a comparison, China's top 100 privately owned companies, as ranked by All-China Federation of Industry & Commerce (2002), were also selected for the study. This is a less homogenous group of companies in terms of coming from a particular industry so it forms a useful contrast with the 100 IT companies (most of the IT companies in the study are publicly traded companies). Results from the two groups were compared to determine similarities and differences.

2.2 Company Information

Business performance data for the top 100 IT companies were taken from statistics compiled by China's Ministry of Information Industry (Ministry of Information Technology Industry, 2002). The following business measures were available and used in the study: gross revenue, profit, export revenue, and R&D expenses (research and development expenses). Business performance data for the top 100 privately owned companies were obtained from All-China Federation of Industry & Commerce (2002). Only gross revenue data were available for this group of companies.

Web sites of the companies in the study were located by searching Google and major Chinese search engines such as NetEase (http://search.163.com) and SOHU (http://www.sohu.com/). Each Web site found was manually checked to make sure that it did exclusively belong to the company in question. Not all companies had an exclusive Web site at the time of the study and they had to be excluded from the study. Of the 100 IT companies, 91 had an independent Web site. In contrast, 71 of the 100 private companies had an independent Web site.

2.3 Link Data

Inlinks to each company's Web site were searched using major commercial search engines. At the time of data collection (summer 2002), the following five search engines could do link searches: Google, AltaVista, AllTheWeb, Lycos, and MSN Search (Notess, 2002a). However, Lycos was not used in the study because it used the FAST database, which is the same database that is used by AllTheWeb (Notess, 2002b). The remaining four search engines were all used in data collection.

There are two types of inlinks: the external inlinks and total inlinks. The former are those that originate from Web sites outside the site in question while the latter include all links that point to the Web site in question regardless of the origin of the links. Because the total inlinks include navigational links (e.g. "back to the home page" links) which do not represent quality or impact of the site being linked to, it has been argued that external inlink count is a better measure than total inlink count (Ingwersen, 1998; Smith, 1999). A recent study (Vaughan and Hysen, 2002), where both types of data were collected and compared, showed only a very slight advantage for external inlinks. The study reported here collected both types of inlink data to ensure that the hypothesis was tested in a correct way and to obtain further evidence regarding the relative advantages of the two types of inlinks.

Of the four search engines used in the study, AltaVista and AllTheWeb can search both external inlinks and total inlinks while Google and MSN can only search for total inlinks. The search queries are illustrated in Table 1 using an imaginary URL of www.abc.com.

Table 1 Examples of Inlink Search Queries

Search Engine	Total Inlink Counts	External Inlink Counts
AltaVista (advanced search)	link:www.abc.com	link:www.oclc.com AND NOT host:www.abc.com
AllTheWeb (advanced search)	In Word Filters. Entered "Must include" "**www.abc.com**" "in the link to URL".	In Word Filters. Entered "Must include" "**www.abc.com**" "in the link to URL"; in Domain Filter, enter "www.abc.com" in the "Exclude" window.
Google (basic search)	link:www.abc.com	N/A
MSN Search (advanced search)	In "search the web for" window, enter **http://www.abc.com/**; in the "find" drop down menu, select "links to URL".	N/A

Some earlier studies have used partial domain name (abc.com) instead of the full domain name (www.abc.com) in the inlink search (e.g. Smith & Thelwall, 2002; Thelwall, 2001; Vaughan & Thelwall, 2003). It was reasoned that the partial domain name search would do a more complete capture of *all* inlinks to the Web site because it is conceivable that a related URL (e.g. mail.abc.com) exists other than the standard www.abc.com. AltaVista and AllTheWeb can accommodate this kind of partial domain name search while Google and MSN Search cannot. To test the relative advantages of searching for partial domain names, data collection in AltaVista and AllTheWeb were carried out with both partial domain name and full domain name. This means that there were ten different inlink searches for each URL (the six full domain searches listed in Table 1 plus four partial domain name searches in AltaVista and AllTheWeb).

2.4 Web Site Age Data

An earlier study found that inlinks to a Web site correlated with the age of the site in that older sites received more inlinks (Vaughan and Thelwall, 2003). To test whether this is true for commercial Web sites and to control the age variable if this is true, data on the ages of the Web sites were collected using the Internet Archive (www.archive.org). The earliest date that the Web site appeared in the Archive was used as the measure of the site's age. This date was then converted to the number of months from June 30, 2002 and used for data analysis.

2.5 Measures to Improve Data Quality

There are two known problems with using commercial search engines for data collection. First, search results from different search engines will be different because different search engines index different Web sites. Second, search results from the same Web search engine can vary by day because the Web is constantly changing. Furthermore, searches performed at different times on a particular day may also differ (Rousseau, 1998/99; Snyder and Rosenbaum 1999). Three measures were taken to improve the reliability of the data collected. First, four search engines, as listed above, were used for data collection in order to avoid the possible bias of a particular search engine. Second, two rounds of data were collected by executing the same search at the same search engine on different days about four weeks apart. Data from the two rounds were combined. This was shown to be beneficial in a previous study (Vaughan and Hysen, 2002). Third, all searches were carried out during the low traffic time of the Web because some search engines may "truncate" search results to improve response time when the traffic on the engine is high.

3 Data Analysis and Results

Spearman correlation coefficient tests were conducted to determine if inlink counts correlate with any of the business measures. The Spearman rather than the Pearson correlation test was used because the frequency distributions of both the business measures and the inlink counts were very skewed. The total inlink count was used in this investigation because Google and MSN cannot perform external inlink search. Total inlink counts collected from the four search engines and from the two rounds of data collection were merged by taking a simple arithmetic average. This average was calculated for each Web site and used to correlate with the business measures of that company. To gain knowledge on data collection issues for Webometrics research, the consistency of data collected through different search engines and in different rounds was investigated. In addition, external inlink data collected in AltaVista and AllTheWeb were compared with total

inlink data collected in those two engines to find out if there is any advantage to using external inlink data.

3.1 IT Companies

Spearman correlation tests show significant relationships ($p<0.01$) between inlink count to a company's Website and the following three business measures of the company: gross revenue, profit, R&D expense (correlation coefficients are 0.51, 0.3 and 0.64 respectively). This suggests that link count could be a good performance indicator of a company, particularly its research and development capability. However, there is no significant relationship between inlink count and export revenue (correlation coefficient is 0.08, $p=0.47$). This is not surprising given that export revenue is a very small part, typically around 8%, of the gross revenue for these companies.

There is a statistically significant correlation between the age of a Web site and its inlink count (Spearman correlation coefficient is 0.72, $p<0.01$). The older the Web site, the more inlinks it receives. This is consistent with findings from another study on commercial Web sites (Thelwall & Vaughan, 2003). However, it raises the question of whether Web site age is a confounding variable in the relationship between inlink count and the business measures reported above. To investigate this possibility, a new variable called inlink-age ratio was created. Inlink-age ratio was calculated as the inlink count divided by the age. This ratio can be interpreted as the number of inlinks received per month of the site's existence. Spearman correlation tests were carried out and significant relationships ($p<0.01$) were found between inlink-age ratio and the business measures of gross revenue, profit, and R&D expense (correlation coefficients are 0.5, 0.3, and 0.63 respectively). Note that this set of correlation coefficients is almost identical to that reported above, confirming that the relationships between inlink count and the three business measures are genuine. Again, the relationship between inlink-age ratio and export revenue is not statistically significant.

3.2 Privately Owned Companies

Gross revenue is the only business performance measure available for this group. The analysis of the privately owned companies is parallel to that of the IT companies. First, the Spearman correlation test was performed but no significant relationship was found between inlink count and gross revenue. Second, Web site age was found to correlate significantly with inlink count (correlation coefficient is 0.57, $p<0.01$). Third, the variable inlink-age ratio was calculated in the same way as for the IT companies. No significant relationship was found between inlink-age ratio and gross revenue. This confirms the result from the IT companies: age is not a confounding variable so that the relationship between inlink count and gross revenue remains the same with or without the presence of the age variable.

It is not surprising that inlink count is not a useful indicator of business performance for this group given the heterogeneous nature of the group. Unlike the IT group, companies in this group did not come from a particular industry. The industry type ranged from high tech (e.g. computer and telecommunication) to traditional (e.g. food processing) with everything else in between. Having a strong Web presence is important for some of these companies but not others. Their use of the Web for business purposes also varies. Thus, there is no correlation between Web link metrics and business performance.

3.3 Comparison between IT and Privately Owned Companies

Direct comparisons between the IT companies and the privately owned companies were made in terms of Web site age and inlink counts. One would expect that the Web sites of the IT companies would be more visible, i.e. receive more inlink counts, because these companies would be leaders in using the Web for business purposes. However, this is not the case. A Mann-Whitney test shows that there is no significant difference ($p=0.26$) between the two groups in inlink counts. The median inlink counts for the IT companies and the privately owned companies are very close, 42.6 and 44.4 respectively. There is no significant difference in the Web site age between these two groups either ($p=0.21$ for an independent T test). The average Web site age for the IT companies and the private companies are 27.5 and 25.2 months respectively.

3.4 Data Collection Issues for Webometrics Research

In addition to testing the main hypothesis of the study, data collected from different search engines and using different queries were analyzed in various ways to explore issues concerning the methodology of Webometric research. Specifically, the following questions were addressed (1) are the conclusions reached from the study dependent on the search engines used; (2) will two rounds of data give the same conclusions, i.e. how stable are search engine results; (3) is external link count a better measure than the total link count; (4) is using partial domain names better than using full domain names in link searches.

The inlink counts returned from different search engines are all highly correlated ($p<0.01$). Table 2 and Table 3 show the Spearman correlation coefficients among different search engines for IT companies and privately owned companies. All correlation tests examining the relationship between inlink count and business measures reported above were re-run using data from individual search engines and the results remained the same. There is no definitive answer as to which search engine is better in terms of achieving a higher correlation coefficient (coefficients for different engines are only slightly different and no engine consistently scored higher). Combining data from different search engines shows no noticeable advantage over using data from a single search engine.

Table 2 Correlation Coefficients of Inlink Count among Different Search Engines – Data from IT companies

	Google	MSN	AltaVista	AllTheWeb
Google	1.000			
MSN	0.738	1.000		
AltaVista	0.768	0.946	1.000	
AllTheWeb	0.743	0.929	0.945	1.000

Table 3 Correlation Coefficients of Inlink Count among Different Search Engines –Data from Privately Owned Companies

	Google	MSN	AltaVista	AllTheWeb
Google	1.000			
MSN	0.603	1.000		
AltaVista	0.597	0.935	1.000	
AllTheWeb	0.588	0.876	0.915	1.000

The two rounds of data collected four weeks apart were remarkable similar. The Spearman correlation coefficient was calculated between the two rounds of data for each search engine. The coefficients were all highly significant ($p<0.001$) and ranged from 0.828 for Google to 0.999 for AltaVista. This shows that search engine results are very stable, which provides assurance of the validity of the conclusions reached from the study. In fact, all correlation tests examining the relationships between the inlink count and business measures were run in various ways (with the two rounds of data combined and with each round of data separately) and the results remained the same.

The two types of inlink counts (external link vs. total link) collected from AltaVista and AllTheWeb were compared directly. The numbers are identical or very close in most cases. The correlation coefficients between inlink counts and business measures are essentially the same regardless which type of link count is used. There is no evidence from this study that the external links are a better measure than the total links.

AltaVista and AllTheWeb can accommodate link searches using both full domain names and partial domain names. Data collected using these two methods were compared directly and very little difference was found. Conclusions reached in the study all remain the same regardless of which type of domain name was used in the inlink search.

4 Discussion and Conclusions

For the IT companies in the study, the inlink count to the company Web site was found to correlate positively with the company's business performance measures of revenue, profit, and R&D expenses. It is intriguing to see that the correlation with R&D expenses is the strongest while the correlation with profit is much lower. This could mean that companies that invest more in research and development have a better Web presence and that their sites are more visible and attract more links to them. As a comparison to the IT companies, China's top privately owned companies were also studied. However, no relationship between the inlink count and the business measure (revenue) was found for this group of companies. The most probable reason for the lack of the correlation is that the companies in this group are very heterogeneous in that they came from various industries. Their business performance measures are not directly comparable and their needs for a strong Web presence differ too. This suggests that link count can be an indicator of business performance only when homogenous groups of companies are being compared.

For both groups of companies studied, the age of the Web site correlates with the inlink count. Generally speaking, older Web sites have more links to them. This raises the question of whether the Web site age is a confounding variable in the relationship between inlink counts and business performance measures. This question was investigated by creating the variable inlink-age ratio and then correlating this variable with business performance measures. The conclusions reached in this new analysis were the same as those without the age variable. Therefore, it is safe to conclude that the relationships between inlink counts and business performance measures (revenue, profit, R&D expenses for IT companies) are genuine rather than spurious.

The findings from this study contribute to our understanding of the nature of the Web links to commercial Web sites. Although links could be created for various reasons without any authority

or quality control, they contain useful information collectively. Because of this, links to commercial Web sites can be a source of business information. This knowledge is useful for collecting business intelligence information on the Web. It is foreseeable that a competitor's business performance data are not readily available while its Web site information can be easily collected. This knowledge is also useful for Web data mining.

Our knowledge on commercial Web sites will become increasingly important due to the rapid growth of e-commerce. This study fills a gap in Webometrics research in this important sector. The results from the study confirm conclusions reached in studies involving academic and scholarly Web sites and thus further our understanding of the nature of Web hyperlinks. Links to a Web site indicate the calibre of the hosting organization regardless of whether it is academic or commercial in nature. The link count is a sign of academic performance for university Web sites and business performance for commercial sites.

The study also investigated data collection issues for Webometrics studies by employing multiple search engines, using various queries, and collecting multiple rounds of data. It found that data from different search engines were highly correlated; conclusions of the study remained the same regardless of which was search engine was used or whether the data from different engines were merged. There is no significant advantage for any particular search engine included in the study. External link count showed no significant advantage over the total link count and the use of partial domain names in link searches made no noticeable improvement over the use of full domain names. Based on these results, it is recommended that future Webometrics research not be limited to using external links. The belief that external link count is better than total link count has prevented many previous studies from using Google because it cannot perform external link search. Google has much larger database (Notess, 2002b) than AltaVista, which has, to date, been the most commonly used search engine for Webometrics data collection. It should be noted, however, that Web sites in this study are relatively small compared to university Web sites. External link counts might have advantages for university Web sites because these sites are more likely to have internal links. Further empirical studies are needed to determine the merits of external links in these situations.

The study found the two rounds of data collected to be highly correlated, demonstrating that the search results are very stable over time. Conclusions of the study did not change regardless of which round of data was used or whether the two rounds of data were merged. The robustness of the results regarding search engine used and the time data were collected is very encouraging not only for this study but for Webometrics research in general.

Acknowledgement: We thank Linzhong Wang, research assistant for this project, for all his work in data collection.

References

All-China Federation of Industry & Commerce. (2002) Top 100 Privately Owned Companies. Retrieved June 1, 2002 from **http://www.chinese-voices.net/minqibaiqiang/index.htm**.

Bauer, C. and Scharl, A. (2000). Quantitative Evaluation of Web Site Content and Structure, Internet Research: Electronic Networking Applications and Policy, 10(1), 31-43.

Burwell, H. P. (1999). Online Competitive Intelligence: Increase Your Profits Using Cyber-Intelligence, Tempe, AZ, U.S.A.: Facts on Demand Press.

Chu, H., He, S. & Thelwall, M. (2002, in press). Library and Information Science Schools in Canada and USA: A Webometric Perspective. Accepted by Journal of Education for Library and Information Science.

Fleisher, C. S. and Blenkhorn, ed. (2001). Managing Frontiers in Competitive Intelligence, Westport, Connecticut, U.S.A.: Quorum Books.

Graef, J. L. (1997). Using the Internet for Competitive Intelligence: A Survey Report, Competitive Intelligence Review, 8(4), 41-47.

Ingwersen, P. (1998). The Calculation of Web Impact Factors, Journal of Documentation, 54(2): 236-43.

Lawrence, S., and Giles, C. L. (1999). Accessibility of Information on the Web, Nature, 400, 107-109.

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. Retrieved Sept. 20, 2002 from www.cs.buffalo.edu/~sbraynov/seminar/unexpected_information.pdf.

Madnick, S. and Siegel, M. (2002). Seizing the Opportunity: Exploiting Web Aggregation, MIS Quarterly Executive, 1(1), 35-46.

McGonagle J. J. and Vella, C. M. (1999). The Internet Age of Competitive Intelligence, Westport, Connecticut, U.S.A.: Quorum Books.

Ministry of Information Technology Industry. (2002). Top 100 Information Technology Companies. Retrieved June 1, 2002 from **www.ittop100.gov.cn**.

Nel, D., van Niekerk, R. Berthon, J. & Davies, T. (1999). Going with the Flow: Web Sites and Customer Involvement, Internet Research: Electronic Applications and Policy, 9(2), 109-116.

Nordstrom, R. D. and Pinkerton, R. L. (1999). Taking Advantage of Internet Sources to Build a Competitive Intelligence System, Competitive Intelligence Review, 10(1), 54-61.

Notess, G. R. (2002a). Search Engines by Search Features. Retrieved July 11, 2002 from http://www.searchengineshowdown.com/features/byfeature.shtml.

Notess, G. R. (2002b). Search Engine Statistics: Relative Size Showdown. Retrieved July 11, 2002 from http://www.searchengineshowdown.com/stats/size.shtml.

Rousseau, R. (1998/99). Daily Time Series of Common Single Word Searches in Alta Vista and Northern Light, Cybermetrics, 2/3(1), Retrieved Sept. 9, 2002 from www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html.

Smith, A. G. (1999), A tale of two web spaces: comparing sites using Web Impact Factors, Journal of Documentation, 55(5), 577-592.

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian Universities, Scientometrics, 54(3), 363-380.

Snyder H. and Rosenbaum, H. (1999), Can Search Engines be Used as Tools for Web-link Analysis? A Critical View, Journal of Documentation, 55(4), 375-84.

Thelwall, M. (2001). Extracting Macroscopic Information from Web Links, Journal of the American Society for Information Science and Technology, 52 (13), 1157-1168.

Thelwall, M. (2000). Commercial Web sites: Lost in Cyberspace?, Internet Research: Electronic Networking Applications and Policy, 10(2), 150-159.

Thelwall, M. and Vaughan, L. (2003). A Fair History of the Web? Examining Country Balance in the Internet Archive. Manuscript under review by Library & Information Science Research.

Vaughan, L. and Thelwall, M. (2003). Scholarly Use of the Web: What are the Key Inducers of Links to Journal Web Sites? Journal of the American Society for Information Science and Technology, 54(1), 29-38.

Vaughan, L. and Hysen, K. (2002). Relationship between Links to Journal Web Sites and Impact Factors, Aslib Proceedings: New Information Perspectives, 54(6), 356-361.

White, G. K. and Manning, B. J. (1998). Commercial WWW Site Appeal: How does it Affect Online Food and Drink Consumers' Purchasing Behavior?, Internet Research: Electronic Networking Applications and Policy, 8(1), 32-38.

Zanasi, A. (1998). Competitive Intelligence through Data Minding Public Sources, Competitive Intelligence Review, 9(1), 44-54.

A revised version has been published as:

Liwen Vaughan, Guozhu Wu (2004). Links to commercial websites as a source of business information. Scientometrics, 60(3), 487-496.