Online Information Review

Emerald Article: Examining the robustness of web co-link analysis

Liwen Vaughan, Juan Tang, Jian Du

Refereed article received
20 October 2008
Approved for publication
30 January 2009

Examining the robustness of web co-link analysis

Liwen Vaughan
Faculty of Information and Media Studies, University of Western Ontario, London, Canada, and
Juan Tang and Jian Du
Institute of Scientific and Technical Information of Shanghai, Shanghai, China

Abstract

Purpose – The purpose of this paper is to examine the robustness of web co-link analysis for business intelligence.

Design/methodology/approach – The method is tested in two different Chinese industries, the electronics/IT industry and the chemical industry. Web co-link data are collected in two different time periods from a different search engine in each period. Multidimensional scaling (MDS) is used to map the co-link data into business competition positions.

Findings – Web co-link analysis is fairly robust in that the mapping results reflect fairly well the business competition landscape for both industries and in both time periods. The mapping results are better when the data collection is restricted to Chinese language webpages only. The study also finds that the Chinese webpages are very consumer-oriented, a phenomenon that is not seen in previous studies of international companies.

Originality/value – This paper contributes to the understanding of the robustness and applicability of the co-link analysis method. The method is useful for business intelligence and can also be applied to the non-business environment. The paper also contributes to the understanding of a specific Chinese web phenomenon.

Keywords Worldwide web, Intelligence, China

Paper type Research paper

Introduction

It has been well established that web hyperlinks contain useful information that can be explored for various purposes. For example, Google (Page *et al.*, 1998) and other major search engines (Sullivan, 2007) all use sophisticated algorithms to rank webpages based on the ways that webpages link to each other. Patterns of webpage linking have been used to identify web communities (Flake *et al.*, 2002), to obtain scientific research information (Thelwall, 2005) and business information (Reid, 2003; Thuraisingham, 2003), and to carry out counter-terrorism research (Mooney *et al.*, 2004).

In the business world, the websites of business competitors tend not to link to each other (Vaughan *et al.*, 2006), but the website of a third party, such as a customer or a

This study was part of a larger project funded by the Initiative on the New Economy (INE) Research Grants programme of the Social Sciences and Humanities Research Council of Canada (SSHRC). Research assistants Karl Fast and Gloria Liu helped with data collection. Senior chemical engineer Lei Xu of the Science and Technology Information Institute of Shanghai Chemical Industry helped with the interpretation of the results of the chemical industry.

Online Information Review
Vol. 33 No. 5, 2009
pp. 956-972

1468-4527
DOI 10.1108/14684520911001936

retailer, may link to a pair of business competitors, that is the websites of two business competitors may be co-linked by a third party. In analysing the co-link patterns of international telecommunications companies, Vaughan and You (2005a) found business information in the form of a business competition map. The current study applied the method developed in that earlier study to the Chinese environment to determine if the method is applicable to an environment that has a different language, culture and business practice. The testing was carried out in two different time periods, Winter (February and March) 2007 and Fall (September) 2008. The results of the Winter 2007 testing were reported in an earlier paper (Vaughan *et al.*, 2008).

The overall approach of the study was to select a group of companies, locate company websites and for each possible pair of companies use a search engine to find all webpages that had hyperlinks that pointed to the pair of company websites (i.e. find the number of co-links of these two companies), then construct a co-link matrix of all the companies in the group and analyse the co-link matrix using a statistical method called multidimensional scaling (MDS). The result of MDS is a map that attempts to position all companies according to their similarities, with similar companies positioned closer to each other. The similarity of a pair of companies is measured by the number of co-links they have. The more co-links pointing to the two companies, the more similar they are. This is based on the idea that the more two companies are related, the more likely it is that they will be co-linked. For example, the website of a computer retail store is likely to link to two computer companies that are their suppliers, but it is unlikely that the website will have links to a computer company and a food company. As similar or related companies are likely to be competitors (two computer companies are competitors but a computer company and a food company are not competitors), the MDS map will effectively place competing companies together. Thus the MDS map will in effective show the business competition landscape.

China's two major industries, the electronics/information technology industry and the chemical industry, were chosen for the study. These two Chinese industries are major international players in their respective markets in that they are both major exporters and importers. They are both a main focal point of their respective international competitors. These two industries are also very different so they can be used as contrasts to test the co-link analysis method. Each industry has thousands of companies – too many to be included in the study, therefore it was decided to study the top companies because reliable business information on these companies is more readily available. Details of the company selection are reported in the Methodology section below.

China was chosen for several reasons. First, there is a great interest in business intelligence in China, as shown by the existence of a national organisation, the Society of Competitive Intelligence of China (www.scic.org.cn), and several commercial portals on competitive intelligence such as China's Network of Competitive Intelligence (www.chinaci.com).

Second, an examination of research literature showed a lack of research on business intelligence methods in the Chinese environment. In China, there is plenty of discussion of business intelligence but relatively little has been done to develop original research methods for business intelligence. Although the methods used to produce business intelligence reports are very rigorous, they tend to be traditional methods without the use of web resources (China's Network of Competitive Intelligence, 2007).

Third, we cannot assume that methods of analysing web hyperlinks for business information that are developed in the Western environment will fit the Chinese environment by default. Not only does China have a different language, culture and business practice, its history of web development is also different from the rest of the world. Web development started later in China but the pace of development in recent years has been astounding. By mid-2008, China surpassed the United States for total web users and thus became the largest online nation (Mokey, 2008). However, a recent study (Vaughan and Zhang, 2007) comparing the coverage of websites of different countries by major search engines found that Chinese sites received a lower rate of coverage relative to their US counterparts. The discrepancy was more pronounced in the areas of commercial sites, which are the focus of the study presented here.

Methodology

Terminology

Some terms need to be defined before discussing the details of the study. Inlinks (also called back links) are links coming into (or pointing to) a webpage. 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 homepage" type of navigational links within the site. 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-linked.

Companies in the study

China's Ministry of Industry and Information Technology conducts an annual ranking of the top 100 Chinese electronics/IT companies. The ranking is based on revenues and other financial measures, and the ranking results are published on the website www.ittop100.gov.cn. The website also contains other information such as company profiles, industry trends and business research reports. As this is an official government website, the site content is considered authoritative and reliable, therefore it was used as the main information source for this study. At the time of the first round of data collection (Winter 2007), the most current ranking was 2006 (China's Ministry of Industry and Information Technology, 2006) so this ranking was used. The chemical companies in the study were from the ranked list of China's top 50 chemical enterprises compiled by SRI Consulting, Beijing Office (Alibaba.com, 2006). SRI Consulting is a business research service for the global chemical industry. It publishes research reports and conducts client-sponsored research (SRI Consulting, 2008). It has almost 60 years of history and their reports are considered fairly authoritative and reliable. The ranking of the companies was based on 2004 revenue data. When we started the project in late 2006, this was the most current list publicly available.

The list of the electronics/IT companies did not contain company URLs. The URLs of these companies were manually searched for and carefully verified. The list of chemical companies contained the URLs of these companies. Each URL was manually checked.

Not all 100 electronics/IT companies were included in the study as some companies had few co-links with other companies in the study. The similarity values (the proximate values in MDS terms) between these companies and the other companies

were too low to generate a meaningful MDS map. For example, if one data point had a proximate value that was 100 times less than most other proximate values, this data point had to be placed about 100 times away from other data points. This would effectively squeeze the rest of the data points so tightly together that their relative positions would not be visible. This mapping problem is analogous to trying to map 50 locations within New York City with one location in California. The map would only show the vast distance between New York and California and would not be able to show the relative positions of the 50 locations.

There is no hard and fast rule on how high the proximate values have to be for a data point to be included in MDS mapping. We took an empirical approach to the problem. We ranked the 100 electronics/IT companies by inlink counts and then took the top 60 for the study. As was expected, the websites of chemical companies had on average fewer inlinks than those of the electronics/IT companies, so only the top 30 chemical companies were used for the study. Because inlink counts tend to correlate significantly with business performance measures (Vaughan and Wu, 2004; Vaughan and You, 2005b), top ranked companies by inlink counts tend to be top performing companies. So our method effectively selected top companies – companies that are of interest for business intelligence analysis (see Tables I and II for the lists of companies in the study).

Search engines and query syntax for data collection

Vaughan and Zhang (2007) compared Chinese search engines (e.g. Yahoo China at www.yahoo.cn) with global search engines (e.g. Yahoo! at www.yahoo.com) and found that the former provide better coverage of websites from China than the latter. Since the companies in this study were all Chinese and the corresponding websites were Chinese sites, we preferred to use Chinese search engines for data collection. However, at the time of the study, we could not find a Chinese search engine that provided the inlink search function that was needed for the study. Baidu (www.baidu.com), a major Chinese search engine, never had an inlink search function. Google China (www.google.cn) had the same problem as the global version of Google in that it could not filter out internal inlinks (details below). Yahoo China and MSN China (http://cn.msn.com) had occasionally provided inlink search functions, but unfortunately they did not work at the time of our data collection (the first round of data were collected in the Winter of 2007 while the second round were in the Fall of 2008). Therefore, we had to use global search engines. However, we did examine the effect of limiting searches to Chinese pages when using the global search engines (see details below).

MSN Live Search (www.msn.com) was used for the first round of data collection because two other major search engines in the market, Google and Yahoo!, could not serve the purpose of the study at that time. Google could only search for total inlinks, that is, it could not filter out internal links in the search results. The Google (2006) documentation on back-link search query stated, "No other query terms can be specified when using this special query term". The "link" query had to be combined with the "site" query in order to filter out internal links. Yahoo! has been used as a search engine for inlink data collection in recent years (e.g. Ortega *et al.*, 2006; Vaughan and You, 2005a) after it acquired AltaVista and AllTheWeb. However, at the time of the first round of data collection, we found that Yahoo! did not support co-link searches although inlink searching still seemed to work. A co-link query such as (link:www.uwo.ca – site:uwo.ca) AND (link:www.ubc.ca – site:ubc.ca) was interpreted

Rank	Company name	URL	Label in Figures 1 and 3
1	Lenovo Group	www.lenovo.com.cn	Lenovo
2	Haier Group	www.haier.com	haier
3	BOE Technology Group Co.	www.boe.com.cn	boe
4	TCL Corporation	www.tcl.com	tcl
5	Huawei	www.huawei.com.cn	huawei
6	Midea Group	www.midea.com.cn	midea
7	Hisense	www.hisense.com.cn	hisense
8	Sva Group	www.sva.com.cn	sva
9	Panda Electronics Group Co., Ltd	www.chinapanda.com.cn	panda
10	Founder Group	www.founderpku.com	founder
11	ZET Corporation	www.zte.com.cn	zte
12	Changhong	www.changhong.com.cn	changhong
13	Huaqiang Holdings Ltd	www.szhq.com	szhq
15	Galanz Group Co. Ltd.	www.galanz.com.cn	galanz
16	Skyworth Group Co., Ltd	www.skyworth.com	skyworth
17	Inspur Group	www.langchao.com.cn	langchao
18	Alcatel Shanghai Bell	www.alcatel-sbell.com.cn	alcatel_sbell
19	Desay Corporation	www.desay.com	desay
20	Konka Group Co., Ltd	www.konka.com	konka
22	Foryou Group	www.foryougroup.com	foryou
23	Tsinghua Tongfang	www.thtf.com.cn	thtf
24	Ningbo Bird Company	www.chinabird.com	bird
26	Shenzhen Electronics Group CO., Ltd	www.seg.com.cn	seg
27	Xiamen Overseas Chinese Electronic Co., Ltd	www.xoceco.com.cn	xoceco
28	Shanghai Feilo Co., Ltd	www.feilo.com.cn	feilo
30	Qindao Aucma Co., Ltd	www.aucma.com.cn	aucma
31	BYD Company Limited	www.byd.com.cn	byd
32	Cosun Group	www.qiaoxing.net	qiaoxing
33	Hengtong Group	www.hengtonggroup.com	hengtong
34	Wanlida Group Co., Ltd	www.malata.com	malata
36	XJ Group Corporation	www.xjgc.com	xigc
37	East China Electronics Group Co., Ltd	www.hdeg.com	eastchina
39	ETERN Group Ltd	www.chinayongding.com	yongding
40	IRICO Group Corporation	www.ch.com.cn	ch
41	Frestech	www.xinfei.com	xinfei
43	Insigma Technology	www.insigma.com.cn	insigma
44	Huahong Group	www.huahong.com.cn	huahong
47	China Hualu	www.hualu.com.cn	hualu
49	Zhongtian Technologies Co., Ltd	www.jszt.com.cn	jszt
53	Tsinghua Unisplendour Corporation Limited	www.thunis.com	thunis
58	Neusoft	www.neusoft.com	neusoft
59	Hengdian Group	www.hengdian.com	hengdian
60	FiberHome Technologies Group	www.wri.com.cn	wri
61	HEDY Holding	www.hedy.com.cn	hedy
62	Jiuzhou Electric Group Co., Ltd	www.jiuzhou.com.cn	jiuzhou
63	China Silian Instrument Group Co. Ltd	www.sicc.com.cn	sicc
64	China Zhenhua Electronics Group	www.czelec.com.cn	czelec
65	Aisino Corporation	www.aero-info.com.cn	aero
66	Yangtze Optical Fibre and Cable Company Ltd	www.changfei.com.cn	changfei

Table I.
Electronics/IT companies in the study

(*continued*)

Rank	Company name	URL	Label in Figures 1 and 3
67	Shanghai Jinling Co., Ltd	www.jin-ling.com	jin-ling
70	China Lucky Film Corporation	www.luckyfilm.com	luckyfilm
71	Pianzhuan Group	www.pianzhuan.com.cn	pianzhuan
73	China Resources Microelectronics (Holdings) Ltd	www.crmh.com.cn	crmh
82	Hasee Computer Co., Ltd	www.hasee.com	hasee
83	Jinpeng Group Co., Ltd	www.gzjpg.com	gzjpg
84	Huaqi Information Digital Technology Co., Ltd	www.huaqi.com	hauqi
86	Nantong Fujitsu Microelectronics., Ltd	www.fujitsu-nt.com	fujitsu
90	Nantian Electronics Information Corp, Ltd	www.nantian.com.cn	nantian
92	Tellhow Sci-tech Co., Ltd	www.tellhow.com	tellhow
94	Shanghai Automation Instrumentation Co., Ltd	www.saic.sh.cn	saic

Table I.

Rank	Company name	URL	Label in Figures 2 and 4
1	China Petroleum and Chemical Corporation (Sinopec)	www.sinopec.com.cn	sinopec
2	China National Petroleum Corporation (CNPC)	www.cnpc.com.cn	cnpc
2	PetroChina Company Ltd	www.petrochina.com.cn	petrochina
3	Sinochem Corporation	www.sinochem.com	sinochem
4	China National Offshore Oil Corporation	www.cnooc.com.cn	cnooc
5	Shanghai Huayi (Group) Company	www.shhuayi.com	huayi
7	Shandong Haihua Group Co.Ltd.(SHG)	www.haihua.com.cn	haihua
9	GITI Tire	www.gititire.com	gititire
11	Shandong Chengshan Group Co., Ltd	www.chengshan.com	shengshan
12	Liaoning Huajin Chemical Industry Group Co., Ltd	www.huajinchem.com	haujinchem
13	Hangzhou Zhongce Rubber Company Limited	www.chaoyang.com	chaoyang
16	Jiangyin Chengxing Industrial Group Co., Ltd phosphatechina	www.phosphatechina.com	
19	Yuntianhua Group	www.yth.com.cn	yth
21	Triangle Group	www.triangle.com.cn	triangle
22	Juhua Group Corporation	www.juhua.com.cn	juhua
23	Baochem Group	www.baochem.com	baochem
25	Tianjin Dagu Chemical Corporation Ltd	www.daguchem.com	daguchem
28	Shandong Linglong Rubber Co., Ltd	www.linglong.cn	linglong
31	Red Sun Group Corporation	www.china-redsun.com	china_redsun
32	Lihuayi Group Co. Ltd	www.lihuayi.com	lihuayi
34	Yantai Wanhua Group	www.wanhua.com.cn	wanhua
36	Lutianhua Group Inc.	www.chinalth.com	chianlth
37	Aeolus Tyre Co., Ltd	www.aeolustyre.com	aerolus
38	Qingdao Huanghai Rubber Group	www.yellowsea.com.cn	yellowsea
42	Guizhou Wengfu Chemi-Phos Imp. and Exp. Corp.	www.wengfu.com	wenfu
43	Yabang chemical industry Corporation (Group)	www.yabang.com	yabang
44	ChemChina Group Corporation	www.chemchina.com.cn	chemchina
46	Hubei Yihua Chemical Industry Co., Ltd	www.hbyh.cn	hbyh
47	Guangzhou Kingfa Sci. and Tech. Co., Ltd	www.kingfa.com.cn	kingfa
50	Xiamen Cheng Shin Rubber Ind., Ltd (XCS)	www.xcs.com.cn	xcs

Table II.
Chemical companies in the study

by Yahoo! as a keyword search – the word "link" and the URL in the query were bolded in the search result screen as they would be in a keyword search. Many retrieved pages did not have the actual links that the query was meant to search for but had the word "link" in them. So Yahoo! could not be used at that time.

Assume that we were searching for co-links between websites www.abc.com and www.xyz.com. The query that we used in MSN Live Search was:

(link:www.abc.com –site:abc.com) (link:www.xyz.com –site:xyz.com).

Note that MSN Live Search, like other major search engines, adds the Boolean operator AND by default in between query terms so the AND operator was omitted in our data collection. In other words, the query that we used was effectively:

(link:www.abc.com –site:abc.com) and (link:www.xyz.com –site:xyz.com).

Data collection was carried out using MSN's API (Application Programming Interface), which usually returns the same results as its web interface (Thelwall, 2008). Although data collected from commercial search engines have limitations (Bar-Ilan, 2004; Mayr and Tosques, 2005; Thelwall, 2008), in practice there is no way that an individual researcher or even a large group of researchers can crawl the whole web to build up a search engine that is large enough to rival commercial search engines. For this reason, we used a commercial search engine for data collection.

The "link" command of MSN Live Search retrieved webpages that linked to a particular URL (in this study, links to a company homepage rather than all pages of the company website). The "linkdomain" command of MSN Live Search retrieved webpages that linked to all pages of a site including the homepage. We collected data using both the "link" and the "linkdomain" commands and compared the results. Since all companies in the study were Chinese companies, we wanted to find out what difference, if any, it would make to restrict data collection to webpages in the Chinese language. So we collected data in two scenarios, one to include all webpages regardless of the language and the other to include only pages in the Chinese language. Altogether, there were four sets of data:

(1) "link" command and include all pages;

(2) "link" command and Chinese language pages only;

(3) "linkdomain" command and include all pages; and

(4) "linkdomain" command and Chinese language pages only.

The outcomes from these four data sets are reported in the Results section below.

The second round of data was collected in the Fall of 2008. This time, the co-link search in MSN Live Search did not work, but it did work in Yahoo!. So Yahoo! was used for the second round of data collection. The query syntax of Yahoo! is similar to that of MSN Live Search, so the same co-link query as for MSN Live Search was used. Although we were forced to use two different search engines for the two rounds of data collection because of the availability of search engines at the time of data collection, we do not think that this has prevented us from making a valid comparison of the two rounds of results. In our comparison, we kept in mind that the differences between the two search engines may have caused different mapping results. We were examining the mapping results based on our knowledge of business competition in the two

industries. In other words, we wanted to know if the mapping results reflected the real business situations. If they did, then the co-link analysis method is successful in these industries. The second and the third authors of this paper are information professionals in the chemical industry and the electronics/IT industry respectively. Our knowledge allowed us to judge the accuracy of the mapping results. In this sense, it is actually useful that we used two search engines, which allowed us to test the robustness of the co-link method through two different search engines.

We collected the second round of data in two scenarios as with the first round of data collection – one to include all webpages regardless of the language, and one to include only pages in the Chinese language. Because the results from the first round of the study showed that data collected with the "linkdomain" command generated better MDS mapping than that with the "link" command (details reported in the Results section below), we collected the second round of data initially using only the "linkdomain" command. Later when we were revising this paper in January of 2009, we collected data using the "link" command, also using Yahoo!, to check if the "linkdomain" command was better.

Data processing

The co-link data collected needed to be normalised to obtain a relative measure of the number of co-links because a co-link count of 5 is very high if the number of links pointing to each website is 6, while it is low if the number of links pointing to each website is 100. The normalisation was done through the Jaccard Index as follows:

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

where:

- A is the set of webpages that link to Website X;
- B is the set of webpages that link to Website Y;
- $n(A \cap B)$ is the number of pages that link to both Website X and Website Y (i.e. the raw co-link count); and
- $n(A \cup B)$ is the number of pages that link to either Website X or Website Y.

MDS was then applied to the normalised co-link matrices using SPSS. The MDS output is a map that attempts to position companies according to their normalised co-link counts. The higher the normalised co-link count between a pair of companies, the closer they tend to be placed in the MDS map. Essentially the map has the potential to cluster competing companies together as competing companies are likely to have higher co-link counts.

Results

Link command versus linkdomain command

For both industries and both rounds of data, the map from the "linkdomain" command had clearer clusters so the results reported below are based on the data collected using the "linkdomain" command. An earlier study (Vaughan and You, 2005a) that examined major global telecommunications companies also compared the results of these two commands and found that the results from the "link" command reflected the industry

reality better. A follow up study that qualitatively examined the reasons for co-linking found that links to homepages were more likely to be business related than links to non-homepages, which explained why the "link" command had better results (Vaughan *et al.*, 2007). The fact that the "linkdomain" command worked better for the Chinese companies in this study suggests that the Chinese pages may have a different reason for co-linking. Further qualitative study is needed to gain a better understanding of the Chinese co-link patterns.

Global webpages versus Chinese language webpages

The MDS maps from the two scenarios were similar but the maps generated from the data of Chinese language pages were better. This was true for both industries in the study. To examine this phenomenon further, the two sets of data were compared in terms of inlink counts. For the first round of electronics/IT company data, the global links (links from all pages without language restriction) were only 12.3 per cent more than the links restricted to Chinese language pages only. For the first round of chemical company data, this ratio was 11.7 per cent. The ratios for the second round of data were similar, 15.1 per cent and 10.6 per cent for the electronics/IT industry and chemical industry respectively. Given that the two search engines used for data collection (MSN Live Search and Yahoo!) are global search engines that index webpages from around the world, the small number of non-Chinese language pages retrieved could not be attributed to a possibility of under-representation of non-Chinese language pages in the search engines. In other words, the only explanation for the small number of non-Chinese language pages is that these company websites did not attract many inlinks from international websites. We considered the possibility of collecting data excluding Chinese language pages and then mapping this data set to find out how the companies were positioned in the international market. However, this idea was not feasible because excluding Chinese language pages would result in very low co-link counts, that is, a very sparse co-link matrix where many cells have zeros, which is not appropriate for an MDS analysis.

Business competition maps from the first round of data

All the MDS maps presented below are based on the "Chinese language pages only" data collected with the "linkdomain" command because this data set generated MDS maps that better reflected the industry reality as reported above. In all the figures in this paper, the data points are labelled with abbreviated company names. Please refer to Tables I and II for the full names and other information of the companies.

Figure 1 is the MDS mapping result of the electronics/IT industry. The stress value of this MDS mapping was 0.05.

Figure 1 can be divided into two parts: the right side are companies that produce consumer products such as televisions, refrigerators and home computers, while the left side are companies whose products or services (e.g. cables, software outsourcing and integrated circuits) are not geared directly towards the consumer market. The right side can be further divided into two groups: companies in the upper circle produce home appliances including televisions, home air conditioning machines and refrigerators. The lower circle encompasses companies that produce personal computers. Unlike the right side, the left side of Figure 1 does not have clear clusters of companies by products or services. For example, cable companies (e.g changfei at the

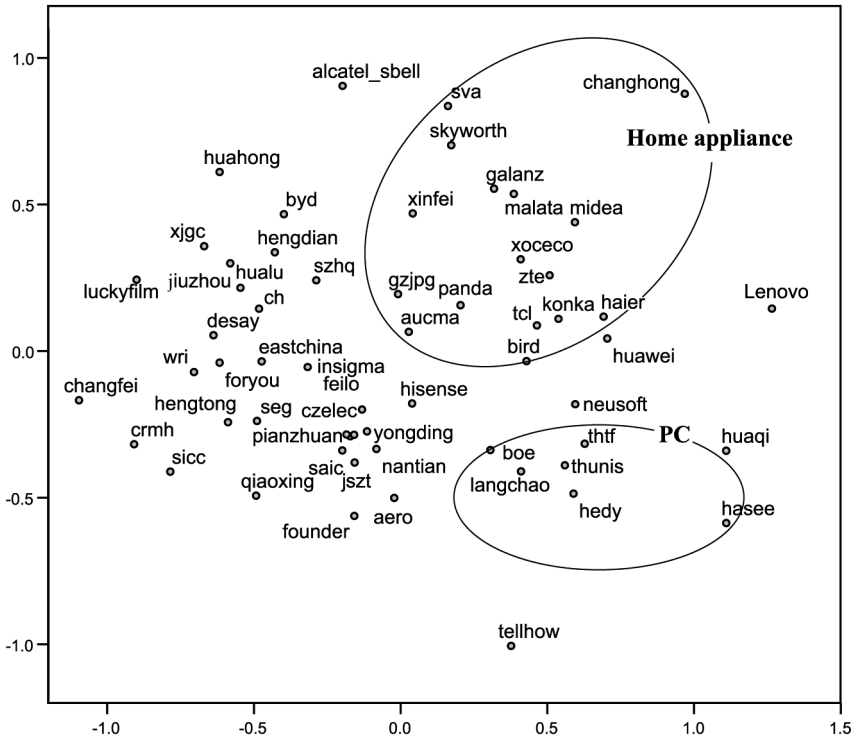

Figure 1.
MDS mapping of the electronics/IT industry (first round of data)

far left, jszt at the lower middle and hengtong at the left center) are mixed with other types of companies rather than being grouped together.

We hypothesised that the lack of clear clusters on the left side of Figure 1 was caused by the lack of a sufficient number of inlinks to show a statistical pattern of clustering. A large portion of the Chinese webpages were consumer-related (i.e. targeting the consumer market). Companies on the left do not attract links from consumer webpages because their products and services are not consumer-oriented. To test this hypothesis, we compared the number of inlinks to the companies on the left side with that on the right side. A t-test showed that the inlink counts on the two sides were indeed significantly different ($p < 0.05$). The average number of inlinks to the companies on the right side was 8579, while that on the left side was only 983. It is clear that companies whose products and services are consumer-oriented attract far more inlinks. For example, Huaqi Information Digital Technology Co. Ltd attracted 10,974 inlinks while Alcatel Shanghai Bell received about half of that number of inlinks (5609). The former is a much smaller company than the latter by various measures. For example, Huaqi was ranked 84th among the top 100 companies while Alcatel Shanghai Bell was ranked 18th. The revenue of the former was RMB 1.8 billion compared with RMB 12 billion of the latter (China's Ministry of Industry and Information Technology, 2006). The strong inlink contrast between the two can be explained by their products: the main products of Huaqi are MP3 players, digital cameras, USB flash drives, etc.,

while the main products of Alcatel Shanghai Bell are telecommunications networks such as voice networks, mobile networks and broadband networks.

A few outliers on the right side of Figure 1 are companies that are not competing directly with the companies in the two clusters. The most noticeable one is Lenovo, positioned on the far right. Lenovo was formed as a result of the acquisition by China's Lenovo Group of the IBM's Personal Computing Division in 2004. Although Lenovo's main products are microcomputers (both desktop and notebook), it is not grouped into the PC cluster of Figure 1. This might be explained by the fact that other companies in the PC cluster are not really at the same level of competitiveness as Lenovo, a leader in the PC market not only in China but around the world.

Figure 2 shows the MDS mapping of the chemical companies based on the first round of data. The stress value of this MDS mapping was 0.04.

The MDS analysis clustered the chemical companies into four distinct sectors. The five major oil companies are located in a very small area, reflecting a very competitive oil market. Tire companies and chemical fertiliser companies are well positioned within the respective circles. If it were not for Huayi, the new chemical materials circle would be much tighter. The reason that Huayi is somewhat distant from the other companies of this group is that it is a chemical enterprise that consists of several companies with a very broad spectrum of products, ranging from new chemical materials to biomedicines. Another company that has a broad range of products is ChemChina. It is not included in any circle but is located close to the oil and the new material circles.

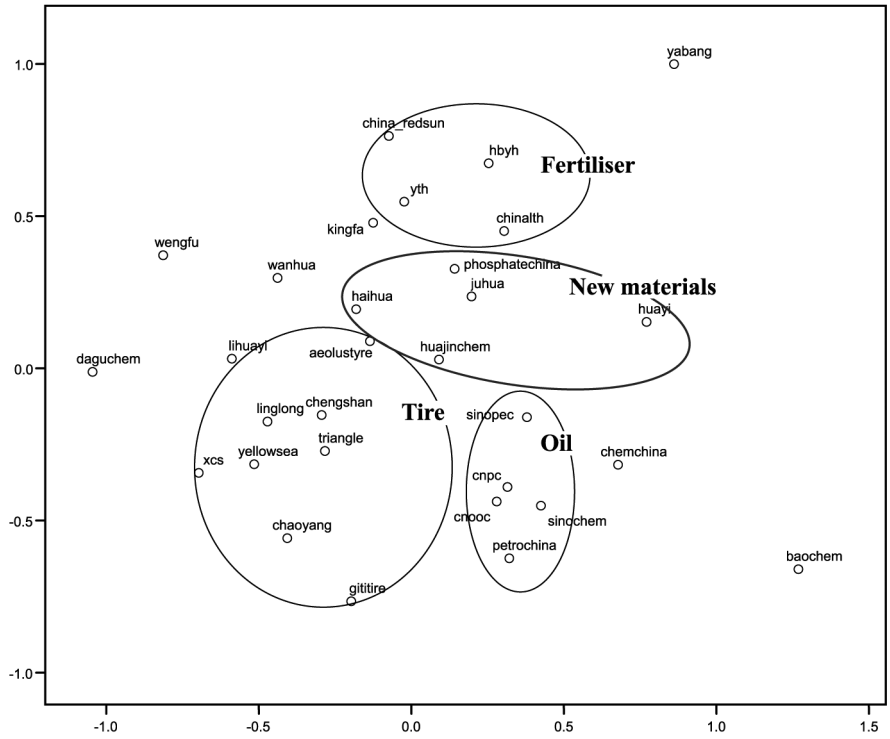

Figure 2.
MDS mapping of the chemical industry (first round of data)

The main products of ChemChina are new chemical materials but ChemChina also has oil and other chemical products. The few companies that are positioned on the outer skirts of Figure 2 are those that are not main competitors of the companies within the five circles and have different products. For example, the main products of Yabang Goup, located on the top right corner of Figure 2, are dyestuff and paints. The main business of BaoChem, located at the bottom right corner of Figure 2, is coal tar processing.

Business competition maps from the second round of data

Figure 3 is the MDS map for the electronics/IT industry based on the second round of data. The stress value of this MDS mapping was 0.05.

This map also has the two main clusters of PC and home appliance firms. As with Figure 1 (from the first round of data), companies on the left side of Figure 3 are those whose products (e.g. cables) are not directly consumer-oriented. Overall, Figure 1 and Figure 3 have similar positioning of companies, which suggests that the success of the co-link method is not affected seriously by the choice of search engine – a sign of the robustness of the method.

A main difference between Figure 1 and Figure 3 is that three telecommunication companies form a clear cluster in Figure 3, while they are scattered in Figure 1. This is

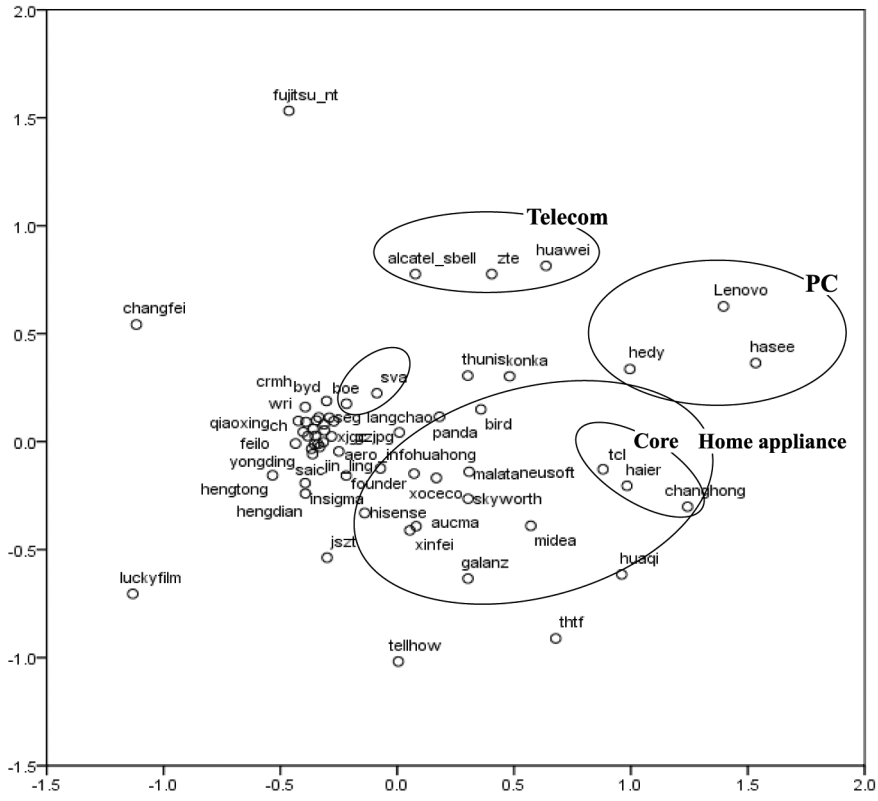

Figure 3.
MDS mapping of the electronics/IT industry (second round of data)

a sign that the data collected from Yahoo! reflected the market position better. Supporting this assertion is the cluster labelled "Core" inside the home appliance cluster of Figure 3. The three companies in this core group are the top three home appliance makers in China and they are main competitors of each other. They are placed very close to each other in Figure 3 but not in Figure 1.

The PC circle has fewer companies in Figure 3 compared with Figure 1. BOE Technology Group Co ("boe") in the PC circle of Figure 1 has moved to be very close to Sva Group ("sva") in Figure 3. This reflects very well the changes in the two companies between 2007 and 2008. These two companies are the largest LCD panel producers in China. Talks of merging these two companies with another related company attracted much attention in 2007. So it is not surprising that they had higher co-link counts and thus are placed closer to each other in Figure 3. Two other companies that moved out of the PC circle of Figure 3 are thtf and thunis. They are PC producers but they also have several other product lines. So the PC circle of Figure 3 contains only core PC companies while the PC circle of Figure 1 includes peripheral companies. Interestingly, Lenovo has moved into the PC circle in Figure 3. This might be partly explained by a change in the company in January 2008 when it sold its cell-phone division, giving the company a purer focus on PC production.

Figure 4 is the MDS map for the chemical industry based on the second round of data. The stress value of this MDS mapping was 0.056.

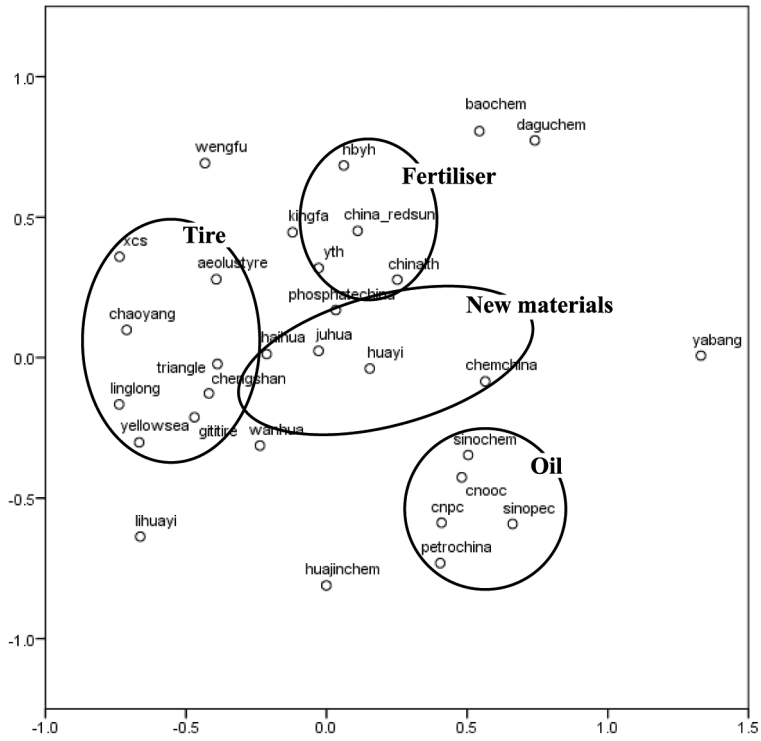

Figure 4.
MDS mapping of the chemical industry (second round of data)

There are no major changes in the four clusters between Figures 2 and 4. Most companies maintained their relative positions within the clusters. A more noticeable change is with ChemChina. It is included in the new materials cluster in Figure 4 but is not inside this cluster, although very close to it, in Figure 2. As explained earlier, this company's main products are new materials but it also has other products such as oil products. So the change between Figure 2 and Figure 4 is not inconsistent with the company profile. In fact, it is consistent with our observation that the MDS maps from the second round of data (collected from Yahoo!) have tighter and more accurate clusters overall. This is true for both the chemical industry and the electronics/IT industry. Unlike the electronics/IT industry, the chemical industry had little change between the two rounds of data collection. This is consistent with the nature of the chemical industry. Production in the chemical industry needs large infrastructure and equipment with large investment. This requires that major products be relatively stable and do not change much from year to year.

Discussion and conclusions

The method of web co-link analysis for business intelligence that was developed and applied successfully to international companies in previous studies (Vaughan and You, 2005a, 2005b) was applied to two Chinese industries and tested in two different time periods. Data were collected from two different search engines. The MDS maps generated from the data are consistent with the competitive situations of each industry. The stress values of the MDS mapping were below or very close to 0.05, which suggests a very good fit between the data and maps. For the electronics/IT industry, the changes in the MDS maps from the two rounds of data reflect fairly well the changes in the industry. For the chemical industry, there was little change between the rounds of data collected, reflecting the relatively stable nature of the chemical industry. These results suggest that the co-link analysis method is fairly robust. The results also suggest that the web co-link analysis method is probably fairly widely applicable to major industries of major countries.

The experience of data collection is worth noting as it could be useful for future studies. The study found that it was better to restrict data collection to Chinese language webpages. Excluding webpages in other languages in the data collection resulted in a more homogenous collection of pages, which is probably the reason for the better mapping result. The study also found that data collected from Yahoo! generated more accurate MDS maps compared with data collected from MSN Live Search. However, it should be noted that the data were collected from Yahoo! and MSN Live Search in two different time periods so the comparison between the two search engines was not a strictly direct one. An ideal test would be to collect data using both search engines in the same time period. However, as the co-link search functions of the two search engines were not available during the same period prevented us from such a comparison.

The significance of the study is two-fold. On a practical level, the co-link analysis of business websites allows us to gain information on business competition. Although business people usually know which companies are their competitors, they may not necessarily have the big picture of their entire industry. The fact that a market research report of an industry often costs thousands of dollars to buy attests the value of this kind of information. This is particularly true when dealing with a foreign country such as China. Companies in the West that are interested in investing in China have a dire

need of Chinese business information but they face many difficulties, among them the language barrier. The co-link analysis method we developed does not require one to know the language of the country. One can also carry out the analysis over time to see changes in the industry, as is shown in this study.

On a theoretical level, our study contributes to knowledge of the web linking phenomenon. This knowledge is useful not just for business intelligence but for information science in general.

An interesting and somewhat unexpected finding of the study was that the Chinese webpages are very consumer-oriented, a phenomenon that was not seen in our previous studies of international companies. Websites of companies whose products (e.g. home appliances) were geared directly towards consumers attracted significantly more inlinks than the websites of companies whose products and services (e.g. telecommunications networks) were not directly consumer-oriented. This also explains why the websites of chemical companies attracted far fewer inlinks than those of the electronics/IT companies. Chemical industry products are generally not sold directly to consumers and these companies' websites were not consumer-oriented.

The consumer orientation of Chinese websites is probably the result of the huge consumer market of China, the consequence of a huge population. This consumer market is developing rapidly along with the rapid development of China's economy. Meanwhile, e-commerce is gaining momentum in China as well. Thus competitive intelligence using web data will be a fruitful area to explore further. In our future studies, we plan to qualitatively examine the nature and characteristics of the Chinese web space to complement the current quantitative study. Another direction for future projects is to apply more advanced web data-mining techniques, such as combining web co-link analysis with web keyword analysis in order to gain more in-depth business intelligence.

References

Alibaba.com (2006), "China's top 50 chemical enterprises among Asian's top chemical enterprises ranked by revenue", Alibaba Group, Hangzhou, available at: http://info.china.alibaba.com/news/detail/v4-d5871708-p2.html#newsdetail-content (accessed: 30 September 2008).

Bar-Ilan, J. (2004), "Search engine ability to cope with the changing web", in Levene, M. and Poulovasilis, A. (Eds), *Web Dynamics*, Springer-Verlag, Berlin.

China's Ministry of Industry and Information Technology (2006), "The 20th annual ranking of the top 100 electronic and information technology companies", China's Ministry of Industry and Information Technology, Beijing, available at: www.ittop100.gov.cn/detail?record=1&channelid=1270&presearchword=ID=188766 (accessed 6 October 2008).

China's Network of Competitive Intelligence (2007) Vol. 30, "Our work procedure", Gungho Group, Hong Kong, available at: www.chinaci.com/about/flow.htm (accessed 30 September 2008).

Flake, G.W., Lawrence, S., Giles, C.L. and Coetzee, F.M. (2002), "Self-organization and identification of web communities", *IEEE Computer*, Vol. 35 No. 3, pp. 66-71.

Google (2006) Vol. 30, "Google SOAP search API reference", Google, Mountain View, available at: www.google.com/apis/reference.html#2_2 (accessed 30 September 2008).

Mayr, P. and Tosques, F. (2005), "Google web APIs – an instrument for webometric analyses?", available at: www.ib.hu-berlin.de/ ~ mayr/arbeiten/ISSI2005_Mayr_Toques.pdf (accessed 30 September 2008).

Mokey, N. (2008), "China passes US in web population", available at: http://news.digitaltrends.com/news-article/17385/china-passes-u-s-in-web-population (accessed 15 October 2008).

Mooney, R.J., Melville, P., Tang, L.R., Shavlik, J., de Castro Dutra, I., Page, D. and Costa, V.S. (2004), "Relational data mining with inductive logic programming for link discovery", in Kargupta, H., Joshi, A., Sivakumar, K. and Yesha, Y. (Eds), *Data Mining: Next Generation Challenges and Future Directions*, MIT Press, Cambridge, MA, pp. 239-54.

Ortega, J.L., Aguillo, I., Cothey, V. and Scharnhorst, A. (2006), "Maps of the academic web in the European higher education area – an exploration of visual web indicators", *Book of Abstracts, 9th International Conference on Science and Technology Indicators, Leuven, 7-9 September*, Vol. 2006, pp. 107-10.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), "The PageRank citation ranking: bringing order to the web", available at: http://dbpubs.stanford.edu:8090/pub/1999-66 (accessed 30 September 2008).

Reid, E. (2003), "Using web link analysis to detect and analyze hidden web communities", in Vriens, D. (Ed.), *Information and Communications Technology for Competitive Intelligence*, Ideal Group, Hilliard, OH, pp. 57-84.

SRI Consulting (2008), "About us", SRI Consulting, Menlo Park, CA, available at: www.sriconsulting.com/SRIC/Public/Aboutus.html (accessed 30 September 2008).

Sullivan, D. (2007), "How search engines rank web pages", available at: http://searchenginewatch.com/showPage.html?page=2167961 (accessed 30 September 2008).

Thelwall, M. (2005), "Scientific web intelligence: finding relationships in university webs", *Communications of the ACM*, Vol. 48 No. 7, pp. 93-6.

Thelwall, M. (2008), "Extracting accurate and complete results from search engines: case study Windows Live", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 1, pp. 38-50.

Thuraisingham, B. (2003), *Web Data Mining and Applications in Business Intelligence and Counter-terrorism*, CRC Press, Boca Raton, FL.

Vaughan, L. and Wu, G. (2004), "Links to commercial web sites as a source of business information", *Scientometrics*, Vol. 60 No. 3, pp. 487-96.

Vaughan, L. and You, J. (2005a), "Mapping business competitive positions using web co-link analysis", in Ingwersen, P. and Larsen, B. (Eds), *Proceedings of ISSI 2005 – the 10th International Conference of the International Society for Scientometrics and Informetrics in Stockholm, 24 -28 July*, Vol. 2005, pp. 534-43.

Vaughan, L. and You, J. (2005b), "Mining web hyperlink data for business information: the case of telecommunications equipment companies", *Proceedings of the First IEEE International Conference on Signal-Image Technology and Internet-Based System in Yaoundé, 27 November-1 December*, Vol. 2005, pp. 190-5.

Vaughan, L. and Zhang, Y. (2007), "Equal representation by search engines? A comparison of websites across countries and domains", *Journal of Computer-Mediated Communication*, Vol. 12 No. 3, available at: http://jcmc.indiana.edu:80/vol12/issue3/vaughan.html (accessed 30 September 2008).

Vaughan, L., Gao, Y. and Kipp, M. (2006), "Why are hyperlinks to business websites created? A content analysis", *Scientometrics*, Vol. 67 No. 2, pp. 291-300.

Vaughan, L., Kipp, M. and Gao, Y. (2007), "Are co-linked business websites really related? A link classification study", *Online Information Review*, Vol. 31 No. 4, pp. 440-50.

Vaughan, L., Tang, J. and Du, J. (2008), "Exploring web co-link patterns for business intelligence: the case of two Chinese industries", *Proceedings of the 36th Annual Conference of the Canadian Association for Information Science in Vancouver, 5-7 June*, available at: www.cais-acsi.ca/2008proceedings.htm (accessed 30 September 2008).

About the authors

Liwen Vaughan is a Professor of the Faculty of Information and Media Studies at the University of Western Ontario in Canada. The focus of her current research is web data mining for business intelligence. Her research interests also include web search engines, webometrics and informetrics. Liwen Vaughan is the corresponding author and can be contacted at: lvaughan@uwo.ca

Juan Tang is an Assistant Professor at the Shanghai Library and the Institute of Scientific and Technical Information of Shanghai, China. Her research interests include business intelligence and web data mining.

Jian Du is a Research Analyst in the InfoLib Consulting Division of the Institute of Scientific and Technical Information of Shanghai, China. His research is focused on content analysis and webometrics.