Constructing Business Profiles based on Keyword Patterns on Websites

Liwen Vaughan[1], Jian Du, Juan Tang[2]

[1] *lvaughan@uwo.ca*
Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, N6A 5B7 (Canada)

[2] *jdu@libnet.sh.cn; jtang@libnet.sh.cn*
Institute of Scientific and Technical Information of Shanghai, Shanghai, 200031 (China)

Abstract

The study examined the possibility of constructing business profiles (specifically, product profiles) based on keyword patterns on various types of Websites including the company's own Website, blog sites, and Websites that have particular keywords and also have hyperlinks pointing the company's Websites. To test the proposed methods, we selected China's four major oil companies and two other companies that have related products. We collected data about these companies from these three Web sources and analyzed the numbers of retrieved pages to construct business profiles. The business profiles constructed were checked against business information collected from other sources such as company annual reports and company newsletters to determine the correctness of the profiles and thus the usefulness of the proposed methods. We found that the method of analyzing frequency distribution of product keywords on company's own Website worked very well. The method of searching for blogs with a company name and product keywords was useful in knowing recent changes and developments of the company. The method of searching for inlinks with keywords provided some information but the overall result is not as good.

Introduction

The increasing role that the Web plays in our society has prompted a whole new branch of academic research that focuses on various Web phenomena. The most commonly studied object in the information science community is Web hyperlinks (e.g. Björneborn & Ingwersen, 2004; Ingwersen, 1998). Recent studies began to analyze other Web objects such as Weblogs (Bar-Ilan, 2005; Smith, 2007). The type of Websites studied varied, ranging from academic sites (e.g. Aguillo, Granadino, Ortega & Prieto, 2006; Vaughan & Thelwall, 2005) to social sites (e.g. Thelwall, 2008) and political sites (e.g. Bar-Ilan & Echerman, 2005; Thelwall, 2007). Studies of business sites or studies that analyze Web related business phenomenon are relatively fewer. Still, there are plenty of such studies, as reviewed below, from which we drew experiences and developed the methodology for the current study.

Liu, Ma & Yu (2001) was among the first who studied business Websites. They developed a method of gathering business information by monitoring Websites of business competitors. Later studies analyzed how to obtain business information from Web hyperlinks (e.g. Reid, 2003; Vaughan & Wu, 2004) and a combination of hyperlinks and keywords (Vaughan & You, 2008). Web messages and Weblogs have also been studied. Das and Sisk (2005) analyzed messages posted to stock boards, mostly Yahoo! message board, using social network analysis methods. They found that network analysis of financial communities provides a way to base portfolio strategies. On the other hand, Tumarkin & Whitelaw (2001) studied the relationship between Internet message board activity and abnormal stock returns and trading volume. They found no causal link between message board activity and stock returns and volume. However, this study was carried out between 1999 and 2000 and the Web has changed substantially since then. In a recent study by Chen, Hu, & Liu (2007), the relationship between a company's blogging activity and its financial performance was studied for a list of Fortune 500 companies. It was found that corporate blogs are positively associated with future profitability and are negatively associated with future costs.

Proceedings of ISSI 2009, pages 636–647.
Edited by B. Larsen and J. Leta.

These earlier studies typically focused on a particular type of Web objects, e.g. Web hyperlinks or blogs, and examined how this Web object can provide useful information. The study reported in this paper attempts to analyze various Web objects to find out what kind of business information they contain and if we can construct business profiles of companies by combining information from various sources. The following three types of Web objects were analyzed: the company's own Website, blog sites in which the company is discussed, and hyperlinks that pointing to the company Website. A group of companies were selected for the study and data about these companies were collected from these three different sources. The research questions of the study are: (1) what kind of business information can we get from these sources? (2) how do these different information sources complement each other?

The specific business information that we are seeking is the relative weights a company places on its various products (i.e. a company's product profile) including new products that are under development. This kind of information is rarely available even for publically traded companies which release annual reports. For private companies, even annual reports are often not available, let alone these product information. Thus, it is very useful if we can develop methods that can be used to gather this kind of business information from the Web. The current study is an effort to develop such methods. It should be noted that information that is usually available in commercial databases such as Hoover's Company Records includes revenue and profit. For some companies, a simple list of products is available but no information on relative weights of products is provided. In addition, these commercial databases are not available for free as the Web does. An added advantage of Web is that it is constantly changing and thus information is potentially more up to date. The disadvantage of the Web information is that it is uncontrolled and some Webpages contain dated information. The challenge is to sift out useful information from potentially useless information. This is essentially what Web data mining is about.

Methodology

We approached our research questions by examining the frequency distribution of keywords over pages. We selected keywords that represent company products and conducted the following three types of searches: (1) searching for pages that contain these words on the company's own Website; (2) searching for pages on blog sites that contain the company name and these keywords; (3) searching for pages that contain these keywords and have a hyperlink pointing to the company's Website. We analyzed the numbers of the retrieved pages to find out if and how business information can be uncovered.

We gathered business information of the companies in the study from various sources such as company annual reports. We compared business information from these sources with the Web data we collected (i.e. the frequency distribution of keywords over pages) to find out if the information from the two sources match. In other words, the purpose of the study was not to find specific information of these companies but to use the information we had to find out if and how we could use Web data to uncover business information. We used this approach to develop methods of data collection and analysis that could be used for business intelligence.

Companies to Study

We chose China's oil industry to examine our research questions and to explore our data collection and data analysis techniques. The world oil market experienced a tremendous price hike in the first half of 2008. A major reason of the hike is China's growing demand on oil due to its rapidly growing economy. There was plenty of discussion of the oil industry on the Web and this provided a very good opportunity for our study. We selected four major Chinese

oil companies for our study: China Petroleum & Chemical Corporation (referred to as Sinopec below), China National Petroleum Corporation (CNPC), Sinochem Corporation (Sinochem), and China National Offshore Oil Corporation (CNOOC). These companies all have well developed Websites and they attract significant number of hyperlinks from other sites. These companies are also frequent subjects of blog discussion.

To test and explore our research method, we also selected two other Chinese companies as contrasts to these oil companies: China Shenhua Energy Company Limited (referred to as Shenhua below) and Shanghai Huayi (Group) Company (referred to as Huayi below). Shenhua is the leading integrated coal-based energy company in China. Although it is not an oil company, it is a hotly discussed company in the oil industry because it is developing a coal liquefaction technology to convert coal to oil products. This technology is mature enough to be at the verge of commercial production at the time of study (summer 2008). Huayi is a large-scale chemical enterprise group with many product lines. Some of its products are similar or related to that of the oil companies in the study. In addition, Huayi also has business in clean coal energy production which is related to Shenhua's business. The contrast between these two companies and the four oil companies helped us to test our method as will be reported in the *Results* section later in the paper.

The Websites of the companies in the study were manually searched for. For each site, possible alternative URLs in the form of an alias or a redirect were also considered and checked. Indeed two companies had multiple URLs. Shenhua has two URLs www.shhuayi.com and www.shhuayi.com.cn but there was no external inlinks to the latter URL so only the former URL was used in inlink search. When searching the company's own Website, both URLs were used. Sinopac has four URLs: www.sinopec.com, www.sinipec.com.cn, www.sinipec-ec.com, www.sinopec-ec.com.cn. It was not possible to combine all four URLs in a single search query. After extensive testing, we found that www.sinopec.com.cn was most appropriate for inlink search as it attracted more inlinks than other URLs. We also found that www.sinopec-ec.com.cn was Sinopec's e-commerce site where various products had comprehensive representation. Search results from this URL better reflected the company's product profile so this URL was used for searching the company's own Website.

Selection of Keywords

Three types of keywords were considered. The first type was about energy products. Six such keywords were chosen: crude oil, kerosene, gasoline, diesel, fuel oil, and coking. The second type relates to things produced from oil. The following three keywords in the study are of this type: ethylene, fertilizer, and chemical products. The word ethylene was chosen because it is an essential material of oil based chemical product. The total production of ethylene is a measure of the scale of a country's oil based chemical industry (China Investment Consulting Network, 2008). Fertilizer is a main product of China's oil industry because agriculture is very important for China's huge population. The broad keyword "chemical products" was chosen to include various products such as synthetic fibers and synthetic rubbers. The third type of keywords was related to a recent technology that attempts to relieve reliance on oil by generating oil products from coal. China has a huge coal reserve. The crisis of oil price hike in the first half of 2008 made this "coal-to-oil" technology very attractive. Six keywords were chosen for this type: coal-to-oil, coal gasification, coal liquefaction, direct liquefaction, indirect liquefaction, and coal chemicals. Altogether, 15 keywords were used in the study. The Chinese characters of these keywords were searched as we were studying Chinese oil

industry. For the same reason, when we searched for company names in blogs, we used Chinese names.

Choice of Search Engines and Query Syntax

All searches were all carried out through commercial search engines such as Google and Yahoo!. We used commercial search engines because no crawler from a single researcher or even a very large research team can cover the vast Web space as these search engines can. All three major commercial search engines, Google, Yahoo! and MSN, have a Chinese version. They are not just a version of the global search engine with a Chinese interface. The underlying databases are different so the search results are different. We need to decide whether to use the global search engines such as the global Google at www.google.com or the Chinese Google at www.google.cn. We conducted many experimental searches comparing global and the Chinese search engines (global Google vs. Chinese Google; global Yahoo! vs. Chinese Yahoo!) and found that the Chinese search engines consistently retrieved larger number of pages for queries in this study. This is not surprising given that the topic of the study is the Chinese oil industry. So we decided to use the Chinese search engines.

Once the Chinese search engines were chosen, decisions had to be made as to which search engines to use. For inlink search, we could only use Chinese Yahoo! because at the time of the study (summer 2008), only Chinese Yahoo! could do external inlink search while other Chinese search engines could not. For blog search, there were three candidates: Chinese Google, Chinese Yahoo!, and Baidu (a major Chinese search engine created in China). We compared search features and decided to choose Chinese Google. The main reason was that only the Chinese Google could limit search to blogs created in a particular time period while the two other search engines did not have this function. For keyword searching on the company's own Websites, all three major search engines plus Baidu could do. However, our tests showed that Chinese Yahoo! had more extensive coverage of these Websites (consistently retrieved more pages) so we decided to choose Chinese Yahoo!.

Example queries for the three types of searches are shown in Table 1 using the example of searching for keyword "crude oil" and for company Sinochem. In the case of keyword search and inlink search, we left truncated the company URL by taking the "www" out. This ensures that all subdomains of the company are included. We used "linkdomain" instead of the "link" command for inlink search to capture all inlink to the company Websites including its subdomain. In the example in Table 1, we used quotation marks around the keyword "crude oil" to show that phrase searches, which are more precise than word searches, were used. In fact, all keywords in the study have multiple Chinese characters, so quotation marks were used in all Chinese characters. The word Sinopec in the line of blog search of Table 1 represents the company name, not the company URL. The abbreviated form of the Chinese name of Sinopec has three Chinese characters. The three characters were searched as a phrase (using quotation marks) to make the search more precise. The abbreviated form of the company names instead of full names were used for all companies because the full names are all long and bloggers typically referred companies by their abbreviated names. All searches were carried out within a three-day time period (Aug. 25 to Aug. 27, 2008) to ensure that results are more comparable. It should be noted that results from the three sources are not mutually exclusive. For example, there could be an overlap between a blog search and an inlink search when a blog contains an inlink to the company Website and the blog is also indexed in the search engines' general index, not just the blog index. This potential overlap poses no threat to our methodology as the two types of searches are meant to provide two different perspectives and the blog item can be viewed from these two angles.

Table 1. Sample Queries

Type of search	Search engine used	Example query
Keyword search on company Website	www.yahoo.cn	"crude oil" site:sinochem.com
Blog search	blogsearch.google.cn	"crude oil" Sinochem
inlink with keyword	www.yahoo.cn	"crude oil" (linkdomain:sinochem.com – site:sinochem.com)

It should be noted that blog search was restricted to searching for blogs by using Chinese Google's blog search engine at http://blogsearch.google.cn. Some major Chinese companies have their own official cooperate blog sites. For example, Microsoft China Research & Development Group has its own blog site at http://www.microsoft.com/china/crd/blogcenter.mspx. However, we did not find official blog sites of the six companies included in the study. So we did not attempt to exclude the company's own blog site in the blog search.

Data Analysis Approach

We had access to the 2007 annual reports of all companies in the study except Huayi for which we had access to its internally circulated newsletters. We also gathered information on recent development of these companies and related technologies, e.g. the development of the "coal-to-oil" technology in response to the world oil price crisis. We used all these information to compare with our Web search results to find out if and how our Web search results revealed the company's product information.

Results

Keyword Distribution on Company Websites

The results of keyword searches on company's own Websites are shown in Table 2. In each cell of Table 2 is the number of pages that contained the keyword. This is not the number occurrences of the keyword because the same keyword can appear multiple times on a page. To get a clearer sense of the relative distribution of different keywords, the percentages of pages that contain the keywords were calculated and shown in Table 3. The percentage was calculated by dividing the number in each cell by the total number of pages in the last column of Table 2. It should be noted that the last column is the total number of pages retrieved from all these keyword searches, not the total number of pages of the company's Website. Dividing by the total number of pages retrieved ensures that the all percentages for a company in Table 3 add up to be 100%, thus making the comparison across companies meaningful. There were of course overlaps among pages retrieved, e.g. a page may contain two or more keywords in the study and thus retrieved two or more times. However, this did not invalidate the way we calculated the percentages as it was the relative distribution of the keywords over pages that was being examined here.

Table 2. Keyword Distribution on Company Websites – Raw Numbers

	crude oil	kerosene	gasoline	diesel	fuel oil	coking	ethylene	fertilizer
Sinopec	136	12	66	94	12	30	112	114
CNPC	23600	910	6560	8010	3310	900	5690	2890
Sinochem	872	3	6	8	855	2	1	953
CNOOC	166	1	10	16	74	3	106	118
Huayi	1	0	1	1	0	14	6	2
Shenhua	1	1	2	3	0	3	0	0

	chemical products	Coal-to-oil	coal gasifi-cation	coal lique-faction	direct lique-faction	indirect lique-faction	coal chemicals	Total
Sinopec	38	0	0	0	0	0	1	615
CNPC	9480	338	142	90	82	57	489	62548
Sinochem	16	0	1	0	0	0	2	2719
CNOOC	78	0	6	0	0	0	40	618
Huayi	4	1	2	1	1	0	2	36
Shenhua	4	11	0	7	4	1	11	48

Table 3. Keyword Distribution on Company Websites – Percentages

	crude oil	kerosene	gasoline	diesel	fuel oil	coking	ethylene	Fertilizer
Sinopec	22.1%	2.0%	10.7%	15.3 %	2.0%	4.9%	18.2%	18.5%
CNPC	37.7%	1.5%	10.5%	12.8 %	5.3%	1.4%	9.1%	4.6%
Sinochem	32.1%	0.1%	0.2%	0.3%	31.4 %	0.1%	0.0%	35.0%
CNOOC	26.9%	0.2%	1.6%	2.6%	12%	0.5%	17.2%	19.1%
Huayi	2.8%	0.0%	2.8%	2.8%	0.0%	38.9%	16.7%	5.6%
Shenhua	2.1%	2.1%	4.2%	6.3%	0.0%	6.3%	0.0%	0.0%

	chemical products	Coal-to-oil	coal gasification	coal lique-faction	direct lique-faction	indirect lique-faction	coal chemicals
Sinopec	6.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
CNPC	15.2%	0.5%	0.2%	0.1%	0.1%	0.1%	0.8%
Sinochem	0.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
CNOOC	12.6%	0.0%	1.0%	0.0%	0.0%	0.0%	6.5%
Huayi	11.1%	2.8%	5.6%	2.8%	2.8%	0.0%	5.6%
Shenhua	8.3%	22.9%	0.0%	14.6%	8.3%	2.1%	22.9%

The four oil companies all had significant percentages of keyword “crude oil” while the other two non-oil companies did not, which matches the nature of these companies. Sinopec’s annual report (Sinopec, 2008) shows its relative revenue from the three major oil products of gasoline, diesel, and kerosene to be about 7.9 to 14.3 to 1. This roughly matches the

percentage figures of these three products in Table 3. CNOOC has lower percentages of these keywords but higher on fertilizer and chemical products, which corresponds to the product description in its annual report (CNOOC, 2008). Table 3 shows that the most frequent keyword on Sinochem's site is "fertilizer" (35%). This keyword appeared more frequently on Sinochem site than on any other sites in the study. This reflects the fact that Sinochem is the largest fertilizer producer in China. Its fertilizer production increased by 21% from 2006 to 2007 (Sinochem Corporation, 2008a, p. 19). Sinopec has the highest percentages of ethylene pages followed by CNOOC, reflecting the fact that Sinopec is the largest producer of ethylene in China while CNOOC recently moved up to be a top producer (China Petroleum Network, 2007). We noted that Sinochem had much larger percent of "fuel oil" pages than pages of other oil products. A check of the company annual report revealed that the company placed more emphasis on fuel oil than other oil products. The Energy Business section of the annual report (Sinochem Corporation, 2008a, p. 16-17) discussed its market position and strategy of fuel oil but not other oil products. Sinochem's emphasis on fuel oil is further confirmed by the search options on the company Website. The dropdown menu of "searching by products and services" listed fuel oil as a search option while other oil products such as gasoline and diesel were not listed. The company's business overview of its petroleum business listed fuel oil as a major segment (Sinochem Corporation, 2008b).

The results of the two non-oil companies match very well with the company profiles. Huayi has the highest number in coking (almost 40%). This is because coking is a major business of Huayi and it owns a coking factory. Shenhua's largest portions of pages are in "coal chemicals" and "coal-to-oil" technology, about 23% each. Not only Shenhua's overall profile matches but also its specific coal-to-oil technology. Shenhua uses coal liquefaction instead of coal gasification technology and the two numbers in Table 3 (14.6% vs. 0%) clearly shows this. Shenhua uses direct liquefaction instead of indirect liquefaction and this again is reflected through the two numbers in Table 3 (8.3% vs. 2.1%). In contrast, CNOOC is developing the technology of coal gasification instead of coal liquefaction (CNOOC, 2008) and their two corresponding numbers in Table 3 show this (5.6% vs. 2.8%).

Company Profiles on Blog Sites

While the keyword distribution on company Website seems to match the profile of company products very well, the match is not so good in the results from the blog sites (search query see the line 3 of Table 1). Table 4 displays the raw data from the search. Table 5 was calculated from Table 4, which is in the same way that Table 3 was calculated from Table 2. Some numbers in Table 5 matches the company product line well, e.g. Sinochem is the largest producer of fertilizer and coking is the main business of Huayi, but others do not, e.g. the relative weight of various oil products.

Table 4. Company profiles on Blog Sites – Raw Numbers

	crude oil	kerosene	gasoline	diesel	fuel oil	coking	ethylene	fertilizer
Sinopec	33666	3201	23429	12637	1627	1883	2960	5795
CNPC	42220	3288	25750	12701	1574	2062	3821	7396
Sinochem	1666	130	287	252	116	285	244	1400
CNOOC	9745	359	3355	2241	343	1414	1225	2835
Huayi	339	25	133	112	8	579	231	295
Shenhua	12365	380	2745	2501	342	4968	1788	4739

	chemical products	Coal-to-oil	coal gasifi-cation	coal lique-faction	direct lique-faction	indirect lique-faction	coal chemicals	Total
Sinopec	2056	341	1304	130	130	96	2882	89473
CNPC	2167	344	1590	108	77	92	3078	106268
Sinochem	375	64	321	10	68	5	316	5539
CNOOC	408	280	389	79	34	35	434	23176
Huayi	201	55	72	42	47	42	264	2445
Shenhua	1698	2444	3831	408	408	403	4619	43639

Table 5. Company profiles on Blog Sites – Percentages

	crude oil	kerosene	gasoline	diesel	fuel oil	coking	ethylene	fertilizer
Sinopec	36.5%	3.5%	25.4%	13.7%	1.8%	2.0%	3.2%	6.3%
CNPC	39.7%	3.1%	24.2%	12.0%	1.5%	1.9%	3.6%	7.0%
Sinochem	30.1%	2.3%	5.2%	4.5%	2.1%	5.1%	4.4%	25.3%
CNOOC	42.0%	1.5%	14.5%	9.7%	1.5%	6.1%	5.3%	12.2%
Huayi	13.9%	1.0%	5.4%	4.6%	0.3%	23.7%	9.4%	12.1%
Shenhua	28.3%	0.9%	6.3%	5.7%	0.8%	11.4%	4.1%	10.9%

	chemical products	Coal-to-oil	coal gasification	coal lique-faction	direct lique-faction	indirect lique-faction	coal chemicals
Sinopec	2.2%	0.4%	1.4%	0.1%	0.1%	0.1%	3.1%
CNPC	2.0%	0.3%	1.5%	0.1%	0.1%	0.1%	2.9%
Sinochem	6.8%	1.2%	5.8%	0.2%	1.2%	0.1%	5.7%
CNOOC	1.8%	1.2%	1.7%	0.3%	0.1%	0.2%	1.9%
Huayi	8.2%	2.2%	2.9%	1.7%	1.9%	1.7%	10.8%
Shenhua	3.9%	5.6%	8.8%	0.9%	0.9%	0.9%	10.6%

A closer examination revealed that results from blog sites reflected more on the type of attention that a company attracted rather than the company's product lines. For example, there were more pages on gasoline than on other oil products and this is true for all companies. This is probably the result of great media attention on gasoline price hike in early 2008. There were overall more pages on "crude oil" than on any other topics, likely reflecting the crisis of world oil shortage in early 2008. A strong case in this regard is Shenhua which does not produce

crude oil. So there are only 2.1% of crude oil pages in Table 3 but it is 28% in Table 5. The company is developing the "coal-to-oil" technology in response to the oil crisis so it is natural that many blogs mentioned the company name when discussing the crude oil crisis. Another example is Sinochem's "coal chemicals" and "coal gasification" percentages. Sinochem is not working in this area itself so the company Website had few or no pages on this (0% and 0.1% in Table 3). However, Sinochem recently became a major share holder in Feng Feng Group, a company with strong coal reserve and coal chemical technology (Xu, 2008). This joint venture received much media attention and the blog page counts show this (about 6% each for coal chemicals and coal gasification for Sinochem in Table 5).

Inlinks with Keyword

Results of searching inlinks with the keyword (search query see the last line of Table 1) are shown in Table 6 and Table 7. Table 6 contains the raw data from the searches and Table 7 was calculated from Table 6, the same way that Table 3 was calculated from 2. Some data in Table 7 parallel company's product line, e.g., the relative weights of kerosene, gasoline, and diesel for Sinopec and the heavy weight of coking for Huayi. However, some cases do not match and the overall pattern is not as good as that of Table 3.

Table 6. Inlinks with Keyword – Raw Numbers

	crude oil	**kerosene**	**gasoline**	**diesel**	**fuel oil**	**coking**	**ethylene**	**fertilizer**
Sinopec	35200	2180	12600	16500	31100	679	2280	1510
CNPC	55600	2890	9770	16600	27100	1300	27300	2990
Sinochem	7930	1244	3700	3270	2600	790	1080	2090
CNOOC	30100	2020	8730	12300	20900	551	2350	2190
Huayi	10	3	11	11	6	254	26	99
Shenhua	5	0	2	0	0	8	0	2

	chemical products	**Coal-to-oil**	**coal gasifi-cation**	**coal lique-faction**	**direct lique-faction**	**indirect lique-faction**	**coal chemicals**	**Total**
Sinopec	2030	334	242	110	25	13	5430	110233
CNPC	2880	468	311	110	26	22	20300	167667
Sinochem	1560	24	189	10	7	8	693	25195
CNOOC	1570	440	279	19	16	12	687	82164
Huayi	239	0	3	0	0	0	16	678
Shenhua	10	9	6	0	1	0	9	52

Table 7. Inlinks with Keyword – Percentages

	crude oil	kerosene	gasoline	diesel	fuel oil	coking	ethylene	fertilizer
Sinopec	31.9%	2.0%	11.4%	15.0%	28.2 %	0.6%	2.1%	1.4%
CNPC	33.2%	1.7%	5.8%	9.9%	16.2 %	0.8%	16.3%	1.8%
Sinochem	31.5%	4.9%	14.7%	13.0%	10.3 %	3.1%	4.3%	8.3%
CNOOC	36.6%	2.5%	10.6%	15.0%	25.4 %	0.7%	2.9%	2.7%
Huayi	1.5%	0.4%	1.6%	1.6%	0.9%	37.5%	3.8%	14.6%
Shenhua	9.6%	0.0%	3.8%	0.0%	0.0%	15.4%	0.0%	3.8%

	chemical products	Coal-to-oil	coal gasification	coal lique-faction	direct lique-faction	indirect lique-faction	coal chemicals
Sinopec	1.8%	0.3%	0.2%	0.1%	0.0%	0.0%	4.9%
CNPC	1.7%	0.3%	0.2%	0.1%	0.0%	0.0%	12.1%
Sinochem	6.2%	0.1%	0.8%	0.0%	0.0%	0.0%	2.8%
CNOOC	1.9%	0.5%	0.3%	0.0%	0.0%	0.0%	0.8%
Huayi	35.3%	0.0%	0.4%	0.0%	0.0%	0.0%	2.4%
Shenhua	19.2%	17.3%	11.5%	0.0%	1.9%	0.0%	17.3%

Discussion and Conclusions

The study tested three methods of collecting and analyzing Web data for discovering company's product information. The method of analyzing distribution of product keywords on company's own Website worked very well. Comparing keyword distributions within and across companies can help to construct or discern product profiles. The method of searching for blogs with a company name and product keywords was found to be useful in finding recent changes and developments of the company. These two methods provide complementary information and can be used together to gain a fuller understanding of business situations. Companies can use this kind of information to gain a better understanding of not only their competitors but also their own relative market positions. The method of searching for inlinks with keywords provided some product information but the overall result is not as good as the method of keyword searching on company Websites. So this particular inlink searching method does not seem to be a promising direction to pursue.

Our future studies will further pursue the first two methods especially the blog search method. The significance of blogs for business communities was pointed out as early as 2005 in an article in Fortune magazine (Kirkpatrick, Roth, & Ryan, 2005). The finding from our current study confirms the importance of blogs. The Chinese blog activity has been growing rapidly in recent years together with the astonishing grow of the Chinese Internet. China overtook the United States as the world's largest Internet population in June, 2008 (China Daily, 2009). According to the 23rd statistics report released in 2009 by China Internet Network Information Center (CNNIC), the number of blog authors in China had reached 162 million by the end of

2008 (CNNIC, 2009, p3). Blogs related to the oil industry are written by various types of people such as stock brokers, journalists specializing in economics and finance, people who work in the oil industry and researchers. Most of these blogs are written from the business and economics perspective and very few from the environmental perspective. Since most blogs are business related, they are good materials for data mining for business intelligence.

A particularly useful feature of blogs is that they all have accurate date and time stamp, a feature that ordinary Web pages do not have. Our future study will limit blog searches to specific time periods to find out if we can detect and monitor particular events or developments during particular time periods. This is a promising direction to purpose because our current study found that the blog search method seemed to reflect new business developments or changes. We will also examine the content of sample blogs, as suggested by Smith (2007), to find out if they are discussing topics that we are searching for.

As the first exploration of the method, the current study used a set of predefined search terms. However, the method does not require that the terms be fixed. New search terms can be added to track new products if needed. The current study is limited to a particular country and industry. We plan to test these methods in other environments to find out if they can be applied there.

Acknowledgments

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

References

Aguillo, I. F., Granadino, B., Ortega, J. L. & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology*, 57(10), 1296-1302.

Bar-Ilan, J. (2005). Information hub blogs. *Journal of Information Science*, 31(4), 297-307.

Bar-Ilan, J. & Echerman, A. (2005). The anthrax scare and the Web: A content analysis of Web pages linking to resources on anthrax. *Scientometrics*, 63(3), 443-462.

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

Chen, D., Hu, N. & Liu, L. (2007). Corporate blogging and firm performance: An empirical study, International Conference on Wireless Communications, Networking and Mobile Computing, WiCom 2007. Sept. 21-23, Shanghai, China.

China Daily (2009). CNNIC: China had 298 million netizens by Dec 2008. Retrieved Mar. 24, 2009 from http://www.chinadaily.com.cn/bizchina/2009-01/13/content_7392547.htm.

China Investment Consulting Network. (2008). Investment consulting report of 2008 China's ethylene industry. Retrieved Dec. 9, 2008 from http://www.dragonraja.com.cn/20073/12007312146.html.

China Petroleum Network. (2007). CNOOC challenges Sinopec and CNPC: ethylene triads formed. Retrieved Oct. 3, 2008 from htpp://www.ljze.com/sygc/zuanjing/qyjs/200702/17327.html.

CNNIC. (2009). Statistical Survey Report on the Internet Development in China. Retrieved Mar. 24, 2009 from http://www.cnnic.net.cn/uploadfiles/pdf/2009/3/23/131303.pdf.

CNOOC. (2008). CNOOC starts its first coal to oil project. Retrieved Oct. 3, 2008 from http://www.cnooc.com.cn/zhyww/xwygg/2008_6_4/269691.shtml.

Das, S. R. & Sisk, J. (2005). Financial communities. *Journal of Portfolio Management*, 31(4), 112-123.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Kirkpatrick, D., Roth, D. & Ryan, O. (2005). Why there's no escaping the blog: freewheeling bloggers can boost your product – or destroy it. *Fortune*, Jan. 10, 2005, p. 44.

Liu, B., Ma, Y., & Yu, P. S. (2001). Discovering unexpected information from your competitors' Web sites. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 26-29, 2001, San Francisco, U.S.A.,

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

Sinochem Corporation. (2008a). 2007 annual report. Retrieved Aug. 22, 2008 from http://www.sinochem.com/Portals/0/nianbao/中化年报2007中文.pdf.

Sinochem Corporation. (2008b). Business overview–energy. Retrieve Nov. 25, 2008 from http://www.sinochem.com/tabid/638/Default.aspx.

Sinopec. (2008). 2007 Annual report. Retrieved Aug. 22, 2008 from http://www.sinopec.com/download/reports/2007/20080407/download/AnnualReport2007.pdf.

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

Thelwall, M. (2008). Text in social network web sites: A word frequency analysis of Live Spaces, *First Monday*, 13(2), available at http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2117/1939.

Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review*, 31(3), 277-289.

Tumarkin, R. & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3), 41-51.

Vaughan, L. & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing & Management*, 41(2), 347-359.

Vaughan, L. & Wu, G. (2004). Links to commercial Web sites as a source of business information. *Scientometrics*, 60(3), 487-496.

Vaughan, L. & You, J. (2008). Content assisted Web co-link analysis for competitive intelligence. *Scientometrics*, 77(3), 433 – 444.

Xu, W. (2008). Behind the scene of share holding of Feng Feng Group. Retried Oct. 3, 2008 from http://xcyszx.blog.163.com/blog/static/316139200822645044686.