Vaughan, L. & Romero-Frías, E. (2012). "Exploring Web Keyword Analysis as an Alternative to Link Analysis: A Multi-industry Case". *Scientometrics*, 93 (1): 217-232. Postprint for research purposes.

Exploring Web Keyword Analysis as an Alternative to Link Analysis: A Multi-industry Case

[1]Liwen Vaughan and Esteban Romero-Frías[2]

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

[2] *erf@ugr.es*
Department of Accounting and Finance, University of Granada, Granada, Spain

Abstract

The study explored the feasibility of using Web keyword analysis as an alternative to link analysis and tested the feasibility in a multi-industry environment. The keyword is the organization's name, in this case the company name. American companies from five industries were included in the study. The study found that the Web visibility of a company as measured by the number of Webpages on which the company name appears correlates with the company's business measures (revenue, profits, and assets). The correlation coefficients are similar to that between the inlink counts and the business measures. This suggests that the keyword count (searched by the company name) could replace inlink count as an alternative indicator of some commonly used business measures. The co-word (the co-occurrence of the names of two companies on Webpages) count was used as a measure of the relatedness of the two companies. Multidimensional scaling (MDS) analysis was applied to the co-word matrices and generated MDS maps that showed relationships among companies in a multi-industry context. Keyword data were collected from three different types of Websites (general Websites, blog sites, and Web news sites) and results were compared. The study found blog sites to be the better source to collect data for this type of study. The comparison of MDS maps generated from co-link data and the blog co-word data showed that the co-word analysis is as effective as co-link analysis in mapping business relationships. The value of the study is not limited to the business sector as the co-word method could be applied to analysing relationships among other types of organizations.

Keywords

Web keyword analysis; Web visibility; business performance; business relationships; Webometrics.

Background of the Study

For well over a decade, Web hyperlink analysis has been a growing area and a main topic of Webometrics research. Starting from the early Web Impact Factor concept (Ingwersen 1998), many studies, both quantitatively (e.g. Thelwall 2001) and qualitatively (e.g. Bar-Ilan 2005), have been carried out that developed different concepts and techniques that use Web hyperlink data to find various types of information. Inlink analysis and co-link analysis have been two common types of Web hyperlink analysis. Parallel to the inlink and co-link concepts are the concepts of the number of keywords and the number of co-words. Thelwall and Sud (2011) proposed organisation title mentions as a measure of academic impact while Vaughan and You (2010) proposed the Web co-word analysis as a way to measure and visualize relationships among organizations. Extending these earlier studies, the current study examined the two concepts in a multi-industry environment to determine if the keyword count of the company name can be a measure of business performance and whether the co-word method can measure relationships among companies in a heterogeneous context. Further, the study were carried out in various types of Web environments including general Websites, blog sites and Web news sites to find out if and how results differ in these different environments and which type of Websites is more conducive for this type of keyword analysis.

Many studies have shown that the number of inlinks to an organization can be a measure of the organization' performance or position. For example, inlink counts have been found to correlate with university's teaching or research performance (Li, Thelwall, Musgrove and Wilkinson 2003; Smith and Thelwall 2002) and company business performance measures (Vaughan and Romero-Frías 2010). Co-link studies have been applied to various types of organizations and were found to be able to show relationships among the organizations studied, for example academic relationships (Ortega and Aguillo 2009; Thelwall and Wilkinson 2004), business relationships (Romero-Frías and Vaughan 2010b), political relationships (Kim, Barnett and Park 2010; Romero-Frías and Vaughan 2010a), and government relationships (Holmberg 2009). While these hyperlink studies have successfully contributed to our understanding of the Web link phenomenon and made significant contribution to Webometrics, the source of Web hyperlink data collection from commercial search engines has been decreasing over the years. MSN suspended its inlink search in 2007 (Seidman 2007). Although Google still provides inlink search, it only retrieves a sample of inlinks (Google 2011). Recent studies relied on Yahoo! for inlink and co-link data collection. However, Yahoo! stopped link search command from its Web interface (www.yahoo.com) in the summer of 2010 which made co-link search impossible there. Then in Apr. 2011 Yahoo! terminated its API service (Yahoo! 2011) so no inlink nor co-link search could be done through API. At the time of writing (Oct 2011), Yahoo! Site Explorer still provides inlink search but it is not clear how long this service will be available.

Given the diminishing data source for link analysis, researchers have tried Web keywords as an alternative object to supplement Web hyperlinks. Vaughan and You (2010) proposed the Web co-word analysis concept and tested the method in the telecommunications industry. They found that the co-word method could generate business competition maps as the co-link method did. Later, Thelwall and Sud (2011) proposed the organisation title mentions (the number of hits of keyword searching of the organization title) as a Web impact measurement. Building on these earlier studies, the current study attempts to further our knowledge of Web keyword analysis in the following ways. First, we will determine if the number of occurrences of a company name on Websites can be used as an indicator of some commonly used business measures. While Thelwall and Sud (2011) showed that organization title mention can be academic impact measure, no study has examined if the company name mention can be a measure of business performance and our study attempted to determine this. Second, the current study extends the co-word analysis method to multi-industry environment to find out its feasibility there (Vaughan and You (2010) study was in a single industry environment). Third, the current study was carried out in different types of Web environments including general Websites, blog sites, and Web news sites to find out if and how results differ in these different environments and which environment is more conducive to the keyword analysis. While Thelwall and Sud (2011) study was carried out only on general Websites, Vaughan and You (2010) compared results from general Websites with that from blog sites and found that the latter is a better source to collect Web co-word data. The current study also includes the news Websites to find out how it compares with other types of sites that have been studied before. These purposes of the study lead naturally to our research questions as follows.

Research questions:

1. Does the Web visibility of a company as measured by the number of Webpages on which the company name appears correlate with the company's business measures.
2. Can co-word analysis show business relationships among companies when applied to a multi-industry context.

3. Which type of Websites (general Websites, blog sites, or Web news sites) is more conducive for data collection.

Methodology

To address our research questions, we selected a group of American industries as well as companies within each industry to study, collected company financial data, selected search engines and collected various types of keyword data using the search engines. In addition, we also collected Web hyperlink data (both inlink and co-link data) because we want to contrast results from keyword analysis with that from inlink and co-link analysis to find out if keyword analysis can supplement inlink and co-link analysis in light of the shortage of commercial search engines from which to collect hyperlink data.

Industries and companies in the Study

Five diverse U.S. industries were selected for the study: information technology, media, heavy construction and engineering, mining, and banking. These industries cover a broad range of economic features and various degrees of exposure on the Internet. They range from traditional industries (mining and construction) to more information-centred industries (IT and media). To make an objective selection of companies within each industry, we consulted industry reports produced by Mergent (http://www.mergentonline.com), a reputable business database. Mergent reports list top companies (usually nine to ten) for each industry. All companies listed were included in the study. All the reports are dated 2010 (Mergent 2010a, 2010b, 2010c, 2010d) except the one for the heavy construction which is dated 2009 (Mergent 2009) and which was the most recent report for that industry at the time the reports were consulted (September 14, 2010). The IT industry report listed nine companies and the other four industry reports each listed ten companies. All these 49 companies were included in the study. The complete list of all companies together with all data about the company that were used in the study is shown in Appendix 1.

Collecting Company Financial Data

For the purpose of the study, we decided to use financial variables of revenue, profit, and assets because they are the most commonly used variables of financial performance (revenue and profit) and financial position (assets). We collected financial data from Yahoo! Finance (http://finance.yahoo.com/) as it contained these three types of data. Yahoo! Finance data were provided by Capital IQ (a Standard and Poor’s business). Specifically, we entered the company ticker (see Appendix 2) into the search box of “GET QUOTES” and then retrieved the financial data that we wanted. Company Massey Energy Co. was not available at Yahoo! Finance because the company was acquired by Alpha Natural Resources Inc. in June 2011. We obtained this company’s financial data directly from the 2010 Annual Report as registered in the SEC's Edgar System (http://www.sec.gov/edgar.shtml). All financial data used in the study were for year 2010, the year that we collected all Web data, and they are shown in Appendix 2.

Collecting Web Keyword Data

Two types of Web keyword data were collected: the number of occurrences of company names (keyword count) and the number of co-occurrences of names of a pair of companies (co-word count). In both scenarios, the acronym, rather than the full name of the company, was used. For example, Intel is used instead of Intel Corp. while Cisco is used for Cisco Systems Inc. The decision to use the acronym rather than the full name was based on the fact that the former is more likely to be used on Webpages. This is also consistent with the co-word data collection method in earlier studies (e.g. Vaughan and You 2010). The proper acronym for each company was determined based on common use as shown on Webpages. Appendix 1 shows acronyms used in the study.

If an acronym consists of more than one word, it was searched as a phrase by using quotation marks around the acronym. For example, the acronym of Time Warner was searched as “Time Warner”. Keyword counts were collected by entering the company name as the query term and then recording the number of hits of the query. The co-word counts were determined by entering the pair of company names as the query and then recording the number of hits of the query. For example, the co-word count of companies Time Warner and Intel was searched as *“Time Warner” Intel*. Boolean operator AND was not used to connect the two acronyms because AND was the default search operator in Google which was used to collect Web keyword data.

Web keyword data were collected from three types of Web sources: the general Web, blogs, and Web news. Three Google search engines (www.google.com, www.google.com/blogsearch, and news.google.com) were used to collect the three types of Web data respectively. Google was chosen because it is the most popular search engine on the Web and had the largest coverage of Websites. Another reason that we used Google was that at the time of the study, fall 2010, Bing and Yahoo! did not have blog search engines. Data from the general Web were collected on Oct. 6, 2010 while data from blogs and news sites were collected on Oct.11, 2010.

Collecting Web hyperlink Data

The Website address of each of companies in the study was searched using Google and then manually checked to ensure that it was correct. The vast majority of companies in the study have only one URL for their Websites. When a company had more than one valid URL, we checked each URL to find out which one had more inlinks and used that one for collecting inlink data. Ideally, we should use all URLs of a company in collecting inlink data. However, the search engine used for collecting inlink data, Yahoo!, could not handle the complex queries need for collecting co-link data with two or more URLs.

As discussed earlier in the "Background of the Study" section of the paper, only Yahoo! could be used for inlink data collection at the time of the study (fall 2010). Further, co-link data could only be collected from Yahoo! API while inlink data were still available through Yahoo!'s Site Explorer. So we collected all inlink and co-link data through Yahoo! API. Yahoo! had two inlink search operators: link and linkdomain. The "link" operator retrieved links to a particular page while the linkdomain operator retrieved all links to all pages of a particular Website or domain. We used the linkdomain operator because all links to the Website or the domain of a company are relevant to the company's Web visibility and connectivity.

The query syntax for inlink data was: linkdomain:website1.com –site:website1.com; whereas the query syntax to collect co-link data was: (linkdomain:website1.com –site:website1.com) (linkdomain:website2.com –site:website2.com). We truncated the www portion of the URLs in the queries in order to capture links to all subdomains (e.g. mail.website1.com). The "-site:website1.com" part of the query let us filter out internal links coming from within the domain of the company itself. All inlink and co-link data were collected on Oct. 5, 2010.

Methods of Data Analysis

Descriptive statistics were generated (1) for each industry individually and for all industry as a whole to provide an overall view of the industries; (2) for each type of Web data to for a comparison of different types of Web data. Correlation coefficient tests were carried out to address research question 1. Spearman correlation coefficient tests rather than the Pearson correlation coefficient tests were used because the frequency distributions of Web data were very skewed. Correlation coefficients for different types of Web data were compared to determine if the keyword count data can replace inlink data and which type of Websites (general Websites, blog sites, and news Websites) is better for data collection (research questions 3).

To address research question 2, co-link and co-word matrices were analyzed using multidimensional scaling (MDS) to generate MDS maps. The raw co-link and co-word counts were normalized by Jaccard index to obtain a relative measure of the relatedness and then fed into SPSS version 17 for MDS analysis. We compared MDS maps of co-link with co-word data to find out if co-word data can replace co-link data. We also compared co-word data from different Web data sources (general Websites, blog sites, and Web news sites) to find out which data source is better (research question 3). Vaughan and You (2010) showed that MDS analysis of co-word data can position companies in a particular industry according to their business relationships. The current study extended that study to a multi-industry context and attempted to find out if the co-word analysis would map companies in the way that reflects a multi-industry business scenario: (1) companies are clustered according to their industry membership; (2) similar industries would be positioned closer. These are the criteria that we used to compare different MDS maps.

Results

Descriptive Statistics

Descriptive statistics of inlink and keyword count data (keyword search of company names) are shown in Table 1. The overall pattern is that the IT industry was the most visible on the Web (having the highest inlink and keyword counts) while the mining and the construction industries had the lowest inlink and keyword counts. This pattern echoes the Web profile of these industries as we know them: IT industry is the leader in Web use while mining and construction had lower use of the Web for business purposes. When all industries are combined, there are more blog counts than inlink counts, which suggests that there will be no shortage of blogs from which to collect keyword data if we are going to replace inlink data with keyword data.

Table 1. Descriptive statistics of inlink and keyword (keyword search of company names) data

Industry		Inlink count	Google count	Google Blog count	Google News count
All industries (n=49)	Mean	4,236,745.9	68,784,616.33	4,742,513.9	4,714.18
	Median	27,900	778,000	55,359	434
	Std. Deviation	1.727E7	2.360E8	1.381E7	9,105.39
Banking (n=10)	Mean	159,600	4,625,550	4,245,574.9	6,091.3
	Median	81,550	2,620,000	354,673.5	3,826
	Std. Deviation	154,879.04	5,513,953.93	1.157E7	6,544.55
IT (n=9)	Mean	20,740,788.89	3.51E8	20,424,257.11	15,833
	Median	3,780,000	1.29E8	6,743,219	13,810
	Std. Deviation	3.739E7	4.732E8	2.527E7	15,570.85
Media (n=10)	Mean	1,923,758	15,680,220	556,630.3	2,458.4
	Median	459,500	3,715,000	305,259	1,561
	Std. Deviation	3,418,389.89	2.219E7	635,183.39	2,466.3
Mining (n=10)	Mean	2,640.3	423,590	29,095.2	229.9
	Median	1,945	106,800	14,525	131
	Std. Deviation	2,684.45	894,511.36	35,901.45	228.83
Construction (n=10)	Mean	7,346.6	647,670	25,186.3	70.2
	Median	5,405	255,000	6,030	48.5
	Std. Deviation	7,231.33	1,092,584.89	54,915.8	55.68

Correlation between Web data and Financial Data

Spearman correlation coefficients between Web data and financial data are shown in Table 2. All correlation coefficients are statistically significant (p<0.01). Relating to research question 1, data here show that the Web visibility of a company as measured by the number of Webpages on which the company name appears correlates with the company's business performance measures of revenue, profits, and assets. Comparing the three types of keyword data sources, correlations are higher for data retrieved from blog and news sites than that from the general Websites. This suggests that blog and news sites are better than the general Websites for this type of keyword analysis, a conclusion that is also reached in our co-word analysis that will be reported below.

Are keyword count data as good as inlink data as Web visibility or impact measures? We suggest that they are comparable. This is based on the comparison of correlation coefficients in Table 2 (inlink vs. Google Blogs and Google News) where the numbers are very close. So we conclude that the keyword counts of company names could potentially replace inlink counts to company Websites especially when the latter are not available. A further evidence that supports our conclusion is that the correlation between inlink counts and Google blog counts is 0.81 while that between inlink counts and Google news counts is 0.82; both are very high and significant (p<0.01).

Table 2. Correlation between Web data and financial data

	Assets	Revenue	Profit
Inlink data (Yahoo)	0.650	0.720	0.651
Google count	0.460	0.582	0.519
Google Blog count	0.632	0.660	0.598
Google News count	0.668	0.670	0.615

Co-link and Co-word Analysis

Four MDS analyses were carried out, one for the co-links and the other for the three sets of co-word data collected from Google, Google Blogs, and Google News. The stress values are all under 0.05 (0.049, 0.034, 0.029, and 0.025 respectively) which indicate that the MDS maps fit the data well. In the MDS maps reported below (Fig. 1 to Fig. 4), companies are labelled in a way that will easily identify its industry membership. The first two letters in the labels identify the industry, e.g. "Ba" for banking and "Mi" for mining. The number following the two letters is the order that the company shows up in Appendix 1, e.g. Ba1 is the first bank in Appendix 1. For the convenience of reading the maps, circles that represent the companies in the maps are shown in different shades (from solid black to transparent) for different industries.

Fig. 1 is the MDS map generated from the co-link data. Companies are clustered by the industries except those of the media industry. All IT companies are clustered close together except companies Ingram Micro (IT8) and Tech Data (IT9). These two are computer wholesalers, much smaller and different from the giants such as Apple, IBM and Microsoft. Industries that rely more on information technology (IT, banking, and media) are positioned on one side, contrasting with mining and construction industries that are located on the other side (the dotted line in Fig. 1 shows the division). A contrast between traditional industries and the information centred industries was also seen in an earlier co-link study of multi-industry companies (Romero-Frías and Vaughan 2010b).

Figure 1. MDS map based on co-link data

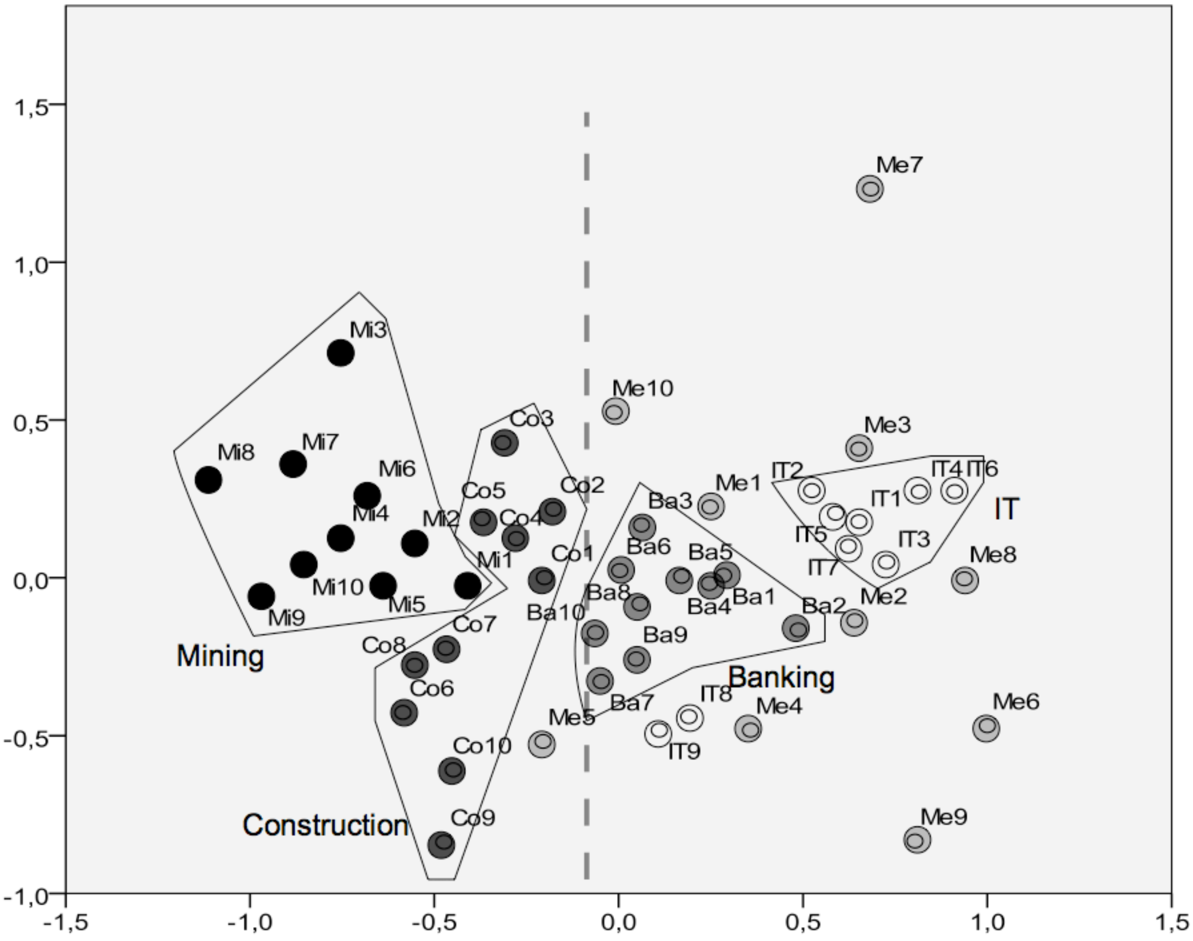


Fig. 2 is the MDS map generated from the co-word data collected from Google Blog. Clustering by industries is clear here than in Fig. 1 where media companies are not clustered together. The contrast between traditional industry (mining and construction) and the three more IT oriented industries seen in Fig. 1 is also shown Fig. 2. Overall, co-word data collected from Google Blogs is as good as or even better than co-link data in showing business relationships among companies.

Figure 2. MDS map based on co-word data collected from blogs

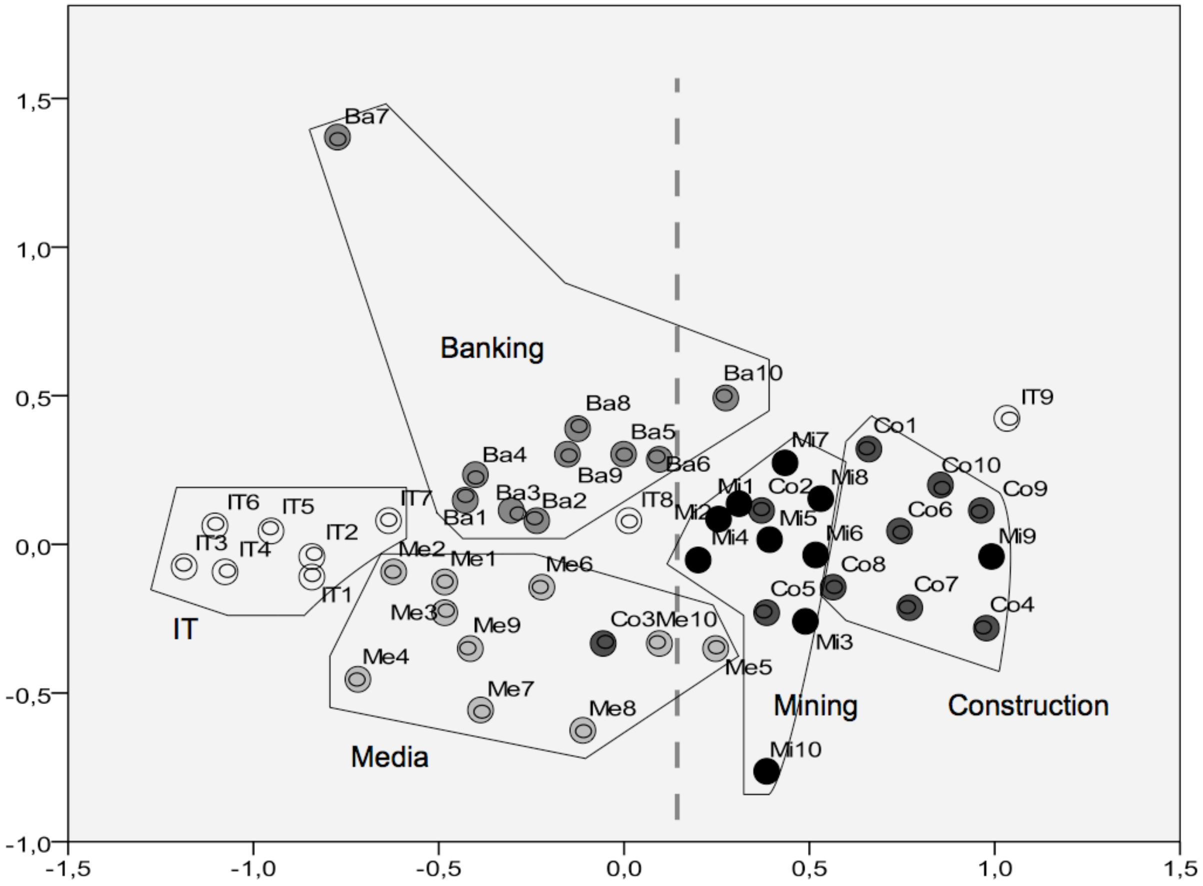
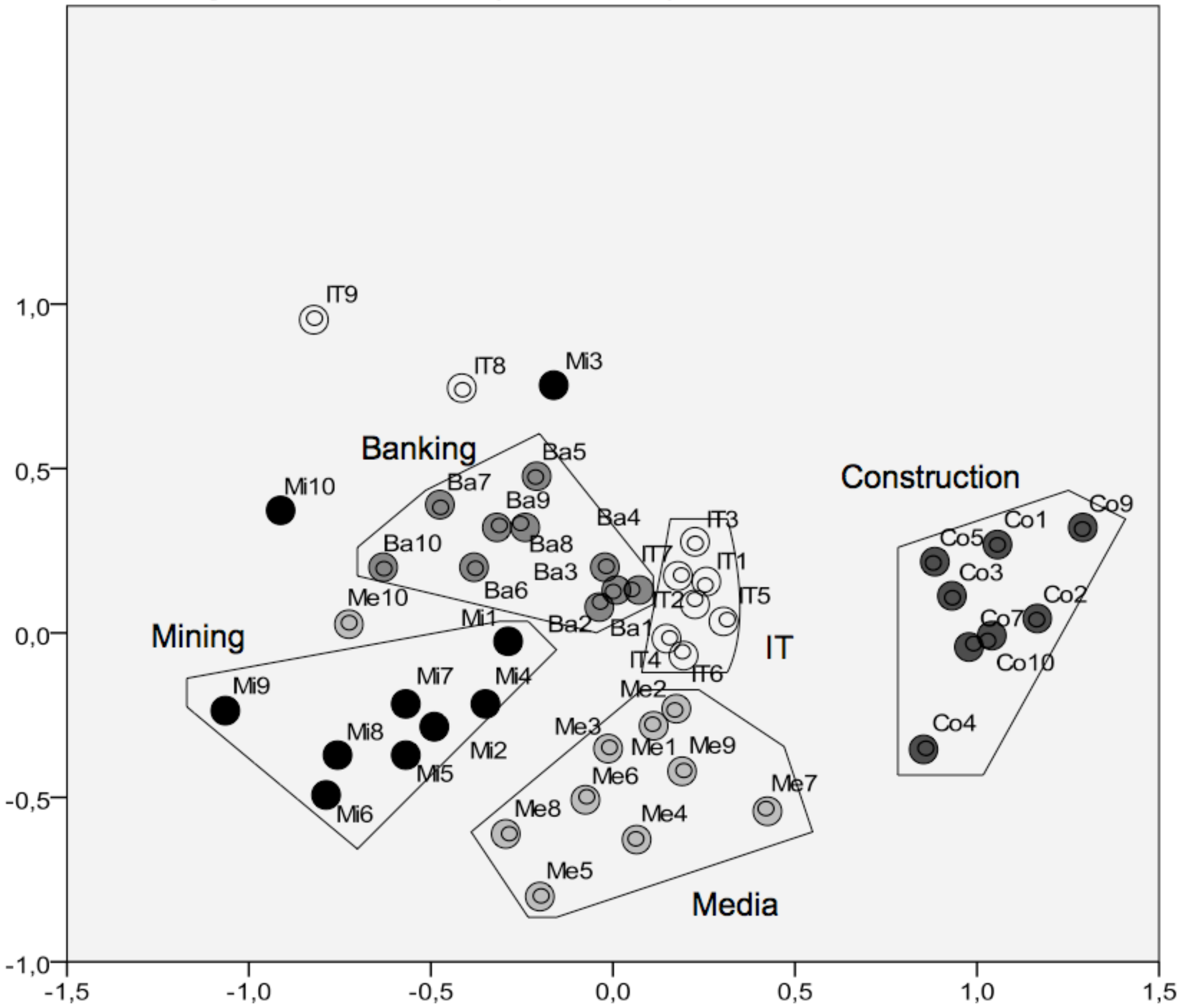


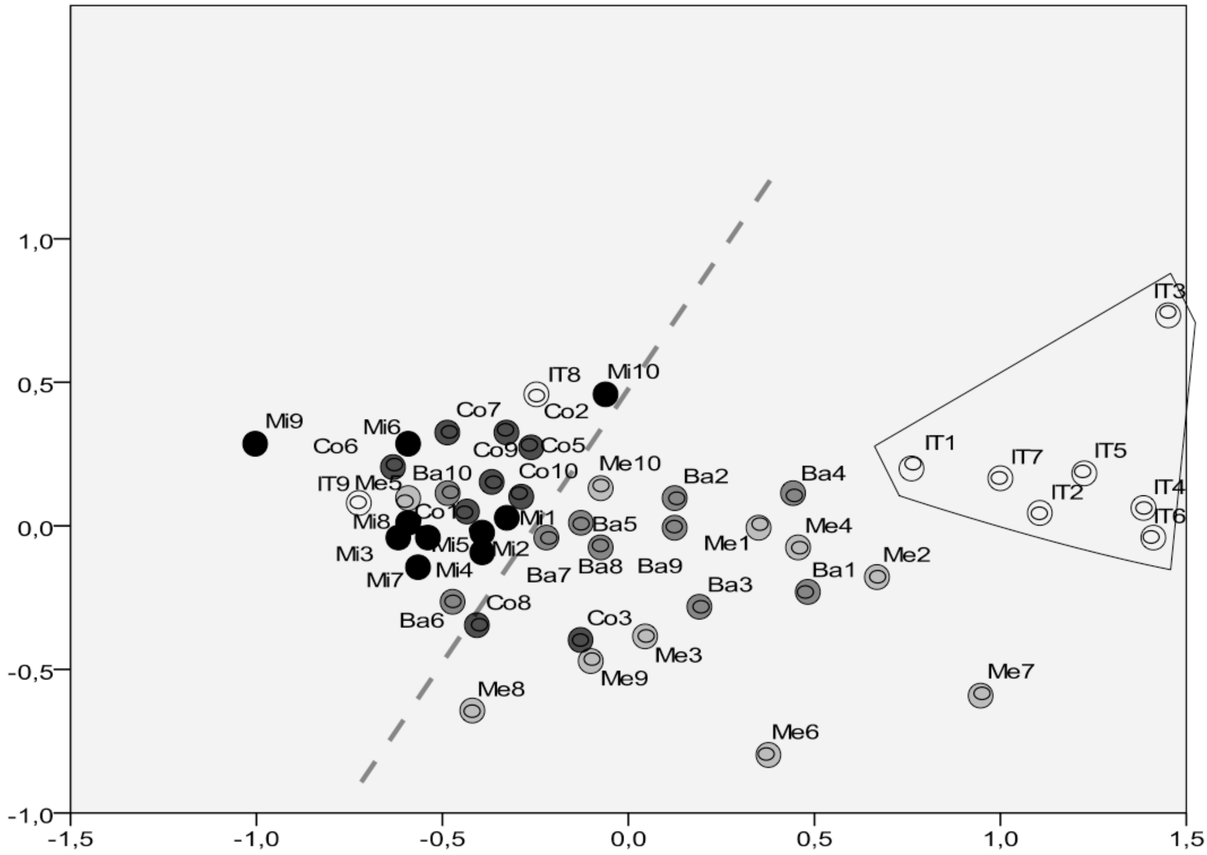
Fig. 3 is the MDS map of co-word collected from Google News. There are 47 instead of 49 companies in this map. Two companies, MDU Resources group (Co6) and Martin Marietta Materials In (Co8), had to be omitted from the MDS analysis because the co-word counts between these two companies and other companies are too few to have proper MDS analysis. Like in Fig. 2, companies are clearly clustered into the five industries with the exception of the two smaller IT companies as explained earlier and a few other companies. However, the pattern of division between the traditional industries (mining and construction) vs. the other industries is not shown here.

Figure 3. MDS map based on co-word data collected from Web News



The MDS map generated from co-word data collected from general Google search engine is shown in Fig. 4. There is a rough division between traditional industries (mining and construction) vs. other industries. However, companies are not clustered by industries except the IT industry. Overall, this map does not show relationship among companies, which suggests that the general Web is not an appropriate source from which to collect co-word data.

Figure 4. MDS map based on co-word data collected from general Websites



Discussion and Conclusions

The study found that the Web visibility of a company as measured by the number of Webpages on which the company name appears correlates with the company’s business measures. This finding parallels findings from earlier research which showed that the number of inlinks pointing to a company’s Website correlates with the company’s business performance measures (Vaughan and Romero-Frías 2010). The current study also found significant correlations between inlink counts and the keyword counts (the number of pages on which the company name appears), suggesting that the keyword count could substitute inlink count as an alternative indicator of business measures. Thelwall and Sud (2011) found that the organisation title mentions could be a measure of academic impact. Tying all these findings together, we conclude that the keyword count (the number of mentions of the organization name) could be a measure of Web visibility or Web impact for academic and business organizations, replacing the role that the inlink count has played in this regard. In terms of sources for data collection, the study found blog sites and Web news sites to be better than general Websites.

The study also found that the co-word analysis could show business relationships among companies even in a multi-industry context. This extends earlier studies that tested the co-word method in a single industry (Vaughan and You 2010; Vaughan, Yang and Tang, in press). When different data sources are compared, the study found that blogs to be a better source than general Websites. Vaughan and You (2010) reached the same conclusion so the advantage of blog pages over general Webpages seems to be clear, at least for studies of business Websites. The study also tested data collection on news Website; no previous study used this data source. It found that Web news sites is a better data source than the general Web but may not be as good as blog sites. Comparing results of co-link data and that of co-word data collected from blog sites, the latter is as good or even slightly better. So co-word analysis could potentially replace co-link analysis if an appropriate co-word data source is used.

A limitation of the study is that it was focused on one particular environment (business related Websites), so the conclusions on the potential of co-word analysis replacing co-link analysis and the relative advantage of blog data over general Web data may not be applicable to other studies (e.g. mapping academic relationships). This is the first study that tried collecting data from news Websites and the study is limited in scale, so the conclusion on the usefulness of news Websites for data collection may not be generalizable.

It is very important to note that the value of the study is not limited to business related Websites. Earlier studies have shown that the co-link analysis can be used to map relationships among various types of organizations such as academic (Ortega, Aguillo, Cothey and Scharnhorst 2008; Thelwall and Wilkinson 2004), business (Vaughan and Romero-Frías 2010), political (Romero-Frías and Vaughan 2010a) and government (Holmberg 2009). The co-word analysis parallels the co-link analysis in logic and method, so it is conceivable that the co-word analysis could be useful in mapping other types of relationships as well. It is important that we develop new Webometrics method such as co-word analysis in light of the declining data source for hyperlink analysis. This will not only keep the healthy development of Webometrics but also let us take advantage of the rich information available on the Web.

Acknowledgments

The first author is supported by a research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC) for the research program of Web data mining for business intelligence. Research assistant Qiang Ning helped with data collection. Part of the findings of the study were reported at the 7th International Conference on Webometrics, Informetrics and Scientometrics (WIS) & 12th COLLNET Meeting.

References

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing and Management*, 41(4), 973-986.

Google (2011). Links to your site. http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=55281. Accessed 14 July 2011.

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

Kim, J.H., Barnett, G.A., & Park, H.W. (2010). A Hyperlink and Issue Network Analysis of the United States Senate: A Rediscovery of the Web as a Relational and Topical Medium. *Journal of the American Society for Information Science and Technology*, 61(8), 1598–1611.

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

Li, X., Thelwall, M., Musgrove, P. & Wilkinson, D. (2003). The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001, *Scientometrics*, 57(2), 239-255.

Mergent (2009). North America – Heavy Construction Sectors. March 2009, available at http://webreports.mergent.com (accessed 14 September 2010).

Mergent (2010a, b, c, d). North America – (Banking Sectors/ IT & Technology Sectors/ Media Sectors/ Mining Sectors). April 2010. http://webreports.mergent.com. Accessed 14 September 2010.

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

Ortega, J.L., Aguillo, I., Cothey, V., & Scharnhorst, A. (2008), Maps of the academic web in the European Higher Education Area – an exploration of visual web indicators. *Scientometrics*, 74(2), 295–308.

Romero-Frías, E., & Vaughan, L. (2010a). European Political Trends Viewed Through Patterns of Web Linking. *Journal of the American Society for Information Science and Technology*, 61(10), 2109-2121.

Romero-Frías, E., & Vaughan, L. (2010b). Patterns of Web Linking to Heterogeneous Groups of Companies: The Case of Stock Exchange Indexes. *Aslib Proceedings: New Information Perspectives*, 62(2), 144-164.

Seidman, E. (2007). We are flattered, but... Retrieved Oct. 24, 2011 from http://web.archive.org/web/20081219045957/http://blogs.msdn.com/livesearch/archive/2007/03/28/we-are-flattered-but.aspx. (Could not retrieve the original page. This is the archived page on Internet Archive)

Smith, A., & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics*, 54(1-2), 363-380.

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., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organisations. *Journal of the American Society for Information Science and Technology*, 62(8), 1488–1497.

Thelwall, M., & Wilkinson, D. (2004). Finding similar academic Web sites with links, bibliometric couplings and colinks. *Information Processing & Management*, 40(3), 515-526.

Vaughan, L., & Romero-Frías, E. (2010). Web hyperlink patterns and the financial variables of the global banking industry. *Journal of Information Science*, 36(4), 530–541.

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

Vaughan, L., Yang, R., & Tang, J. (in press). Web co-word analysis for business intelligence in the Chinese environment. *Aslib Proceedings: New Information Perspectives*.

Yahoo! (2011). Web Search APIs from Yahoo! Search. http://developer.yahoo.com/search/web/webSearch.html. Accessed 24 October 2011.

Vaughan, L. & Romero-Frías, E. (2012). “Exploring Web Keyword Analysis as an Alternative to Link Analysis: A Multi-industry Case”. *Scientometrics*, 93 (1): 217-232. Postprint for research purposes.

Apenddix 1. Web data of Companies in the Study

label in MDS map	Company	Industry	Domain	Company acronym	Inlink count	Google serch count	Google blog count	Google news count
Ba1	Bank of America Corp	Banking	https://www.bankofamerica.com/	“Bank of America”	442,000	16,200,000	1,840,007	15,918
Ba2	JPMorgan Chase & Co	Banking	https://www.chase.com/	jpmorgan	348,000	4,200,000	559,427	10,314
Ba3	Citigroup Inc	Banking	http://www.citigroup.com/	citigroup	61,600	6,270,000	1,248,905	9,060
Ba4	Wells Fargo & Co	Banking	https://www.wellsfargo.com/	"wells fargo"	339,000	12,300,000	1,326,229	6,888
Ba5	US Bancorp (DE)	Banking	http://www.usbank.com/	us bancorp	130,000	716,000	51,216	304
Ba6	PNC Financial Services Group	Banking	https://www.pnc.com/	"pnc financial services"	86,200	325,000	31,463	685
Ba7	BB&T Corp	Banking	http://www.bbt.com/	bb&t	52,100	906,000	37,127,851	16,301
Ba8	Suntrust Banks Inc	Banking	https://www.suntrust.com/	suntrust	48,200	1,690,000	149,920	764
Ba9	Fifth Third Bancorp	Banking	https://www.53.com/	“Fifth Third”	76,900	3,550,000	107,845	596
Ba10	State Street Corp	Banking	http://www.statestreet.com/	"State Street Corp"	12,000	98,500	12,886	83
IT1	Hewlett-Packard Co	IT Tech	http://www.hp.com/	"Hewlett Packard"	4,250,000	34,800,000	1,426,990	8,236
IT2	International Business Machines Corp	IT Tech	http://www.ibm.com/	ibm	3,780,000	129,000,000	6,743,219	13,810
IT3	Dell Inc	IT Tech	http://www.dell.com/	dell	4,220,000	1,430,000,000	37,027,949	50,563
IT4	Microsoft Corporation	IT Tech	http://www.microsoft.com/	Microsoft	66,200,000	599,000,000	50,939,791	16,617
IT5	Intel Corp	IT Tech	http://www.intel.com/	intel	2,380,000	219,000,000	16,611,752	17,943
IT6	Apple Inc	IT Tech	http://www.apple.com/	apple	103,000,000	648,000,000	67,526,194	26,694
IT7	Cisco Systems Inc	IT Tech	http://www.cisco.com/	cisco	2,790,000	96,400,000	3,497,152	8,360
IT8	Ingram Micro Inc	IT Tech	http://www.ingrammicro.com/	"ingram micro"	27,900	438,000	43,724	254
IT9	Tech Data Corp	IT Tech	http://www.techdata.com/	"tech data corp"	19,200	37,900	1,543	20
Me1	Time Warner Inc	Media	http://www.timewarner.com/	"time warner"	98,100	6,960,000	1,094,944	3,351
Me2	Comcast Corp	Media	http://www.comcast.com/	Comcast	2,430,000	38,000,000	2,008,371	6,184
Me3	News Corp	Media	http://www.newscorp.com/	"news corp"	774,000	1,750,000	513,680	1,912
Me4	Dish Network Corp	Media	http://www.dishnetwork.com/	"dish network"	105,000	4,690,000	994,722	963

	Communications Inc			communications"				
Me9	McClatchy Co	Media	http://www.mcclatchy.com/	McClatchy	1,880,000	2,740,000	381,518	6,977
Me10	IAC/InterActiveCorp	Media	http://www.iac.com/	InterActiveCorp	22,200	951,000	36,816	274
Mi1	Freeport-McMoRan Copper & Gold	Mining	http://www.fcx.com/	"Freeport McMoRan"	9,290	333,000	22,518	647
Mi2	Peabody Energy Corp	Mining	http://www.peabodyenergy.com/	"Peabody Energy"	5,220	203,000	29,363	185
Mi3	Southern Copper Corp	Mining	http://www.southernperu.com/	"Southern Copper"	1,930	50,100	13,085	113
Mi4	Massey Energy Co	Mining	http://www.masseyenergyco.com/	"Massey Energy"	2,340	323,000	55,359	557
Mi5	Arch Coal Inc	Mining	http://www.archcoal.com/	"Arch Coal"	2,070	115,000	15,965	68
Mi6	Alpha Natural Resources Inc	Mining	http://www.alphanr.com/	"Alpha Natural"	1,300	89,700	10,059	48
Mi7	Cliffs Natural Resources Inc	Mining	http://www.cliffsnaturalresources.com/	"Cliffs Natural"	1,960	98,600	11,714	149
Mi8	Patriot Coal Corp	Mining	http://www.patriotcoal.com/	"patriot coal"	360	52,700	8,722	78
Mi9	Alliance Resource Partners	Mining	http://www.arlp.com/	"Alliance Resource Partners"	985	20,800	2,157	20
Mi10	Headwaters Inc	Mining	http://www.headwaters.com/	Headwaters	948	2,950,000	122,010	434
Co1	Fluor Corp	Construction	http://www.fluor.com/	"Fluor Corp"	15,600	146,000	5,326	23
Co2	Jacobs Engineering Group Inc	Construction	http://www.jacobs.com/	"Jacobs Engineering"	8,560	421,000	12,748	106
Co3	KBR Inc	Construction	http://www.kbr.com/	KBR	9,760	3,690,000	180,781	197
Co4	URS Corp	Construction	http://www.urscorp.com/	"urs corp"	23,000	81,700	4,407	45
Co5	Foster Wheeler Ltd	Construction	http://www.fwc.com/	"Foster Wheeler"	6,010	579,000	19,604	119
Co6	MDU Resources Group Inc	Construction	http://www.mdu.com/	"MDU Resources"	2,490	133,000	2,551	35
Co7	Granite Construction Inc	Construction	http://www.graniteconstruction.com/	"Granite Construction"	4,800	185,000	6,435	69
Co8	Martin Marietta Materials In	Construction	http://www.martinmarietta.com/	"Martin Marietta"	1,570	778,000	10,783	52
Co9	Dycom Industries Inc	Construction	http://www.dycomind.com/	dycom	496	325,000	3,603	19
Co10	MasTec Inc	Construction	http://www.mastec.com/	mastec	1,180	138,000	5,625	37

Apenddix 2. Financial Data of Companies in the Study

label in MDS map	Company	Tickers	Assets (in thousands)	Net Income (in thousands)	Revenue (in thousands)
Ba1	Bank of America Corp	BAC	2,264,909,000	-2,238,000	134,194,000
Ba2	JPMorgan Chase & Co	JPM	2,117,605,000	17,370,000	102,694,000
Ba3	Citigroup Inc	C	1,913,902,000	10,602,000	60,559,000
Ba4	Wells Fargo & Co	WFC	1,258,128,000	12,362,000	93,249,000
Ba5	US Bancorp (DE)	USB	278,267,000	3,317,000	20,518,000
Ba6	PNC Financial Services Group	PNC	264,284,000	3,011,000	17,096,000
Ba7	BB&T Corp	BBT	157,081,000	816,000	11,072,000
Ba8	Suntrust Banks Inc	STI	172,874,000	189,000	10,072,000
Ba9	Fifth Third Bancorp	FITB	111,007,000	753,000	7,218,000
Ba10	State Street Corp	STT	160,505,000	1,556,000	9,716,000
IT1	Hewlett-Packard Co	HPQ	124,503,000	8,761,000	126,033,000
IT2	International Business Machines Corp	IBM	113,452,000	14,833,000	99,870,000
IT3	Dell Inc	DELL	38,599,000	2,635,000	61,494,000
IT4	Microsoft Corporation	MSFT	108,704,000	23,150,000	69,943,000
IT5	Intel Corp	INTC	63,186,000	11,464,000	43,623,000
IT6	Apple Inc	AAPL	75,183,000	14,013,000	65,225,000
IT7	Cisco Systems Inc	CSCO	87,095,000	6,490,000	43,218,000
IT8	Ingram Micro Inc	IM	9,084,032	318,060	34,588,984
IT9	Tech Data Corp	TECD	6,488,292	214,243	24,375,973
Me1	Time Warner Inc	TWX	66,524,000	2,578,000	26,888,000
Me2	Comcast Corp	CMCSA	118,534,000	3,635,000	37,937,000
Me3	News Corp	NWSA	61,980,000	2,739,000	33,405,000
Me4	Dish Network Corp	DISH	9,632,153	984,729	12,640,744
Me5	Liberty Global Inc	LBTYA	33,328,800	388,200	9,016,900
Me6	Gannett Co Inc	GCI	6,816,844	588,201	5,438,679
Me7	McGraw-Hill Cos Inc	MHP	7,046,561	828,063	6,168,331
Me8	Discovery Communications Inc	DISCA	11,019,000	653,000	3,773,000
Me9	McClatchy Co	MNI	3,136,359	36,273	1,375,232
Me10	IAC/InterActiveCorp	IACI	3,439,554	99,359	1,636,815
Mi1	Freeport-McMoRan Copper & Gold	FCX	29,386,000	4,273,000	18,982,000
Mi2	Peabody Energy Corp	BTU	11,363,100	774,000	6,860,000
Mi3	Southern Copper Corp	SCCO	8,128,019	1,554,051	5,149,500
Mi4	Massey Energy Co	---	4,611,000	-166,600	3,039,000
Mi5	Arch Coal Inc	ACI	4,880,769	158,857	3,186,268
Mi6	Alpha Natural Resources Inc	ANR	5,179,283	95,551	3,917,156
Mi7	Cliffs Natural Resources Inc	CLF	7,778,200	1,019,900	4,682,200
Mi8	Patriot Coal Corp	PCX	3,810,036	-48,026	2,035,111
Mi9	Alliance Resource Partners	ARLP	1,501,278	321,017	1,610,065
Mi10	Headwaters Inc	HW	888,974	-49,482	654,699
Co1	Fluor Corp	FLR	7,614,923	357,496	20,849,349
Co2	Jacobs Engineering Group Inc	JEC	4,683,917	245,974	9,915,517
Co3	KBR Inc	KBR	5,417,000	327,000	10,099,000
Co4	URS Corp	URS	7,351,355	287,889	9,177,051
Co5	Foster Wheeler Ltd	FWLT	3,083,539	215,407	4,067,719
Co6	MDU Resources Group Inc	MDU	6,303,549	240,659	3,909,695

Co7	Granite Construction Inc	GVA	1,535,533	-58,983	1,762,965
Co8	Martin Marietta Materials In	MLM	3,074,743	97,012	1,782,857
Co9	Dycom Industries Inc	DY	724,755	16,107	1,035,868
Co10	MasTec Inc	MTZ	1,655,828	90,528	2,308,031