Googling Companies - a Webometric Approach to Business Studies

Esteban Romero-Frías
University of Granada, Spain
erf@ugr.es

Abstract: So far Internet studies have focused mainly on using website content for gathering business information, however web hyperlinks have not been exploited enough for business purposes yet. Webometric techniques are based on the exploitation of information contained in the hyperlinks that connect the different documents contained on the Web. Webometrics could be considered as a new discipline that applies bibliometric techniques to the quantitative study of the Web, but also a discipline that progressively develops its own concepts and methodology. So far studies in this field have focused on academic and scholarly web spaces; however this methodology is equally applicable to commercial sites which are more predominant on the Web. This paper is intended to show how webometric techniques could be applied to business and management studies. Therefore, it describes a number of basic concepts and techniques and the way in which they have been applied to these fields so far. Firstly, some studies found that the number of links pointing to companies' websites correlates significantly with the business performance measures of the entity. This finding suggests that links to a website could be used as a timely indicator of business performance. Secondly, the examination of co-links, which refers to webpages that links two business sites simultaneously, have been used for competitive intelligence purposes. These studies are based on the idea that the number of co-links to the websites of a pair of companies is a measure of the similarity between them. For instance, this similarity measure between companies in the same industry can provide information about their competitive positions. Finally, motivations for the creation of hyperlinks to business sites could be analysed through a content analysis approach in order to get confirmation about the business relevance and nature of links. This view complements the quantitative perspective to link and co-link research, providing a brand new approach to business studies.

Keywords: web mining, webometrics, business intelligence, business management, internet studies

1. Researching the Web: why hyperlinks do matter

Since its creation in 1989, the World Wide Web (the Web) has revolutionised the Internet, facilitating the access to information to many potential users. Two decades later, the Web has become part of the daily lives of many people all over the world, causing deep social transformations that social scientists struggle to understand. Moreover, for the past five years, the Web has undergone significant changes by the popularisation of the so-called Web 2.0 (O'Reilly, 2005). This has provoked a democratisation of the information creation tools in such a way that millions of people have started to participate in a global conversation based on the use of no cost, user friendly, multimedia and Web-based software (Jenkins, 2006). The blossoming of publishing on the Web has made the media more difficult to understand, always in a process of constant change, described by many as chaotic, uncontrolled and of poor quality (Keen, 2007). However, the Web and the Internet as a whole are the largest repositories of information ever known in history.

In 2008, the official Google blog (2008) reported that the number of pages indexed had grown to more than one trillion (as in 1,000,000,000,000) unique URLs (Uniform Resource Locators). Technorati (2009), the main blog search engine, reported to be tracking on the order of 133 million blogs by the end of 2008. Most, if not all, of the human activities, whether they are political, social, educational or economical are reflected on the Internet, and some have developed native online phenomena that would not exist in the offline world.

The available collection of information makes it possible to define the Web as an enormous unstructured and heterogeneous database that, despite its appearance, is not randomly built. This implies that this could be exploited from different perspectives (based on content, structure and user behaviour) in order to study unique online phenomena or offline phenomena reflected in the Web. A basic component of the array of information available is the hyperlink.

A hyperlink is a reference or navigational element in a document to another section of the same document or to another document that may be on or part of a different domain. Hyperlinks represent the hidden structure of the Web connecting different sites and webpages that would stay isolated unless the specific URL is known (Berners-Lee, 1999). They can be regarded as an endorsement of a target page, especially if the creator has placed that link because it points to a useful or relevant

Reference this paper as:
Romero-Frías, E. "Googling Companies - a Webometric Approach to Business Studies." *The Electronic Journal of Business Research Methods Volume 7 Issue 1 2009, (pp.93 - 106), available online at www.ejbrm.com*

resource. The creation and exploitation of hyperlinks are not an irrelevant phenomena, but imply significant social repercussions (Turow and Lokman, 2008). It is not meaningless that currently most of the search engines use link analysis as part of their algorithms to crawl websites and to rank the pages (Batelle, 2005).

Google's market dominance derives originally from the technological lead established by the Pagerank system introduced in 1998 (Brin and Page, 1998). This Pagerank is based on hundreds of factors that have changed over time in order to avoid manipulation, but basically follows two principles:

- Webpages that have more links pointing to them are considered to be more relevant.
- Not all the hyperlinks have the same value. Those links established by relevant webpages are more valuable than others from less well known webpages.

However, hyperlinks are not a perfect source of evidence because many of them have not been created by a thoughtful process in order to endorse or discredit a webpage. Many links are created for navigational purposes within a site; others are just automatically created by content management systems or, in the worse case, they are just spam or lists of URLs created to perform better in the search engines rankings.

Hyperlinks constitute the raw basic material of quantitative research in the Web, as performed by Webometrics.

2. An introduction to Webometrics

The idea of the Web as a distributed database that is exponentially growing over time has been appealing for Data Mining research. There is a great potential for analysis by mining the website content, document structure, site relations, and user behaviour (Benoît, 2002). In this context, Webometrics has developed into a new discipline that studies Web-phenomena from a quantitative point of view.

2.1 Definition

The origin of Webometrics can be found in the field of Information Science. Thelwall, Vaughan and Björneborn (2005: 81) point out that the discipline "emerged from the realization that methods originally designed for bibliometric analysis of scientific journal article citation patterns could be applied to the Web, with commercial search engines providing the raw data". In fact, the idea that a link pointing to a webpage means a 'vote' to that webpage or document is based on bibliometric methods to rank scientific production (Garfield, 1979). The term Webometrics was first coined by Tomas Almind and Peter Ingwersen in 1997 and seems to be widely accepted by the research community together with the term Cybermetrics. Björneborn (2004) defined both terms by limiting their research areas. Webometrics is "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (in Björneborn and Ingwersen, 2004: 1217), while Cybermetrics does the same but on the whole Internet. Hence, Cybermetrics is more focused on the study of non web-based Internet phenomena, e.g. emails, chat, newsgroup studies, etc. Figure 1 shows the location and overlapping of these disciplines in the general context of Information Science.

Recent developments within the field suggest a move in the scope of the definition into a more general social science research approach instead of an approach that is mainly based on a informetric and bibliometric perspective. Thelwall (2009: 6) defines Webometrics as "the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study". Interdisciplinary research is getting more significant by enlarging the types of subjects of study and the techniques used. This evolution aligns with the definition of Internet research given by Hine (2008: 537): "Internet research itself is not a discipline but an interdiscipline, a field or a research network populated by heterogeneous perspectives."

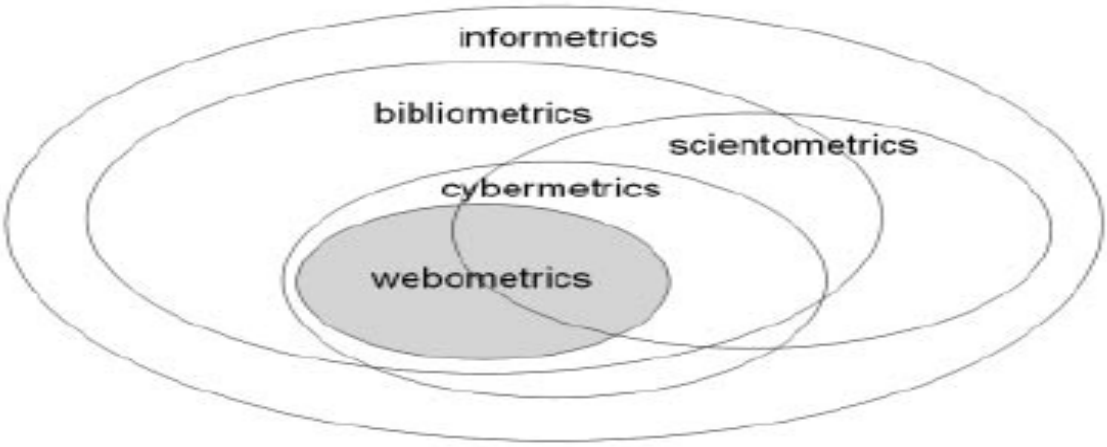

Figure 1: Webometrics and Cybermetrics in the context of Information Science (Björneborn and Ingwersen 2004: 1217)

2.2 Basic concepts

Björneborn and Ingwersen (2004) carried out the first attempt to develop a consistent terminology on the webometric field. Some years later, Thelwall and Wilkinson (2008) proposed a generic lexical framework that, based on the previous work, intended to unify and extend existing methods through abstract notions of link lists and URL lists.

Figure 2 shows a diagram of the Web where the circles represent different types of nodes (websites, webpages, etc.) and the arrows connections between them. The squares within the dashed line rectangle are nodes that are considered for analysis in a given research. They are specially useful to illustrate a specific type of analysis, the so called, co-link analysis.

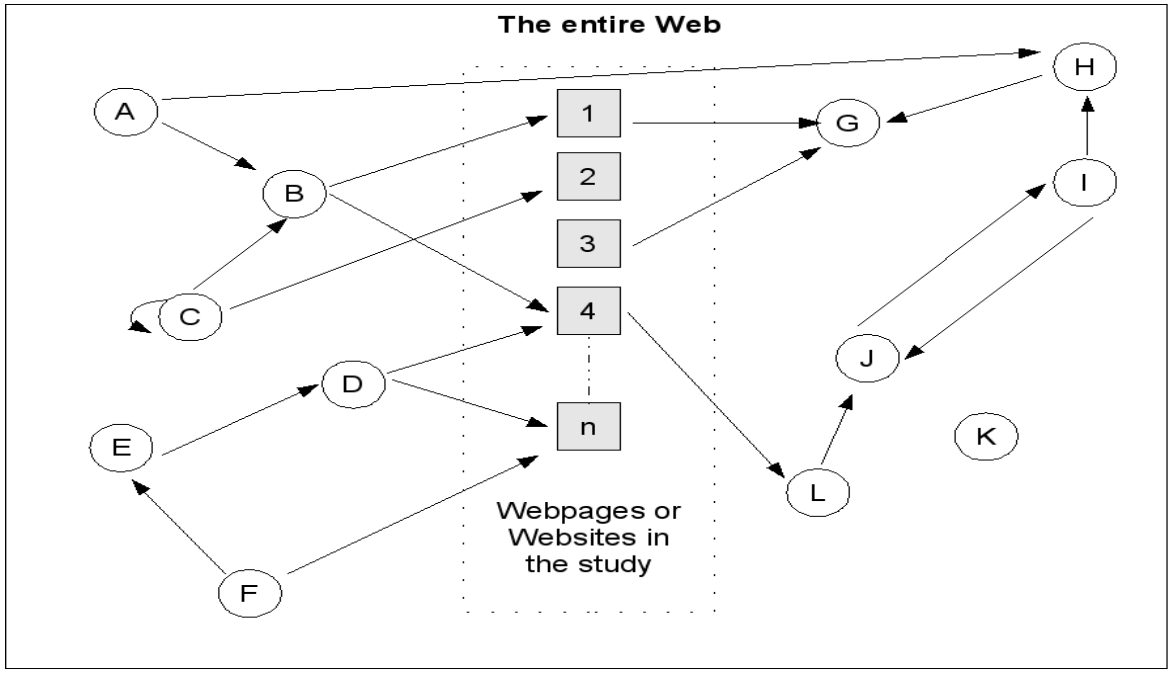

Figure 2: Diagram with different types of links existing in the Web.

Letters in the diagram represent any type of document in the Web, whether it is a webpage or a website, for instance. The following basic webometric terms (Björneborn and Ingwersen, 2004) can be explained by reference to Figure 2:

- **Inlink**: B has an inlink from A.
- **Outlink**: A has an outlink to B.

- **Self-link:** C has a self-link.
- **Page or site isolated**: K is isolated as it does not have any inlinks or outlinks.
- **Reciprocal links**: I and J have reciprocal links.
- **Transversal link**: A has a transversal outlink to H. This type refers to a link that joins to different areas of the Web that are not well interconnected.
- **Co-inlinks**: 1 and 4 have a co-inlink, as B links both of them simultaneously.
- **Co-outlinks**: G has a co-outlink, as 1 and 3 are linking to it.

This paper basically focused on inlink and co-inlink analysis, as most of the webometric research on business is based on these two concepts. Co-inlink analysis is referred simply as co-link analysis later in this paper.

2.3 Methodology

Before reviewing the application of webometric techniques to business research, this paper offers a general overview of the methodology. The application of bibliometric techniques to the Web had derived from the similarities between hyperlinks and academic citations, considering that both point from a source document to a target document. Nevertheless, important differences exist.

In this paper, we briefly describe three main types of techniques based in the use of commercial search engines to gather raw data. For a more systematic approach, a complete and didactic reference is Thelwall (2004; 2009).

2.3.1 Web impact analysis

Web impact analysis provides evidence for the impact or the spread of ideas, brands, organisations, etc. on the Web by measuring and analysing the URLs retrieved by commercial search engines in response to a specific query. This approach is especially useful as an exploratory approach for further research, although it has some significant drawbacks. For instance, one such problem is the extent to which the keyword used matches or does not the subject under research. This technique is the only one in this paper that is not based on the analysis of hyperlinks, however as it is the most intuitive one is appropriate to introduce the subject.

The main problems come from keywords that are too common or have different meanings, resulting in a wide variety of results that do not necessarily match the issue under study. In this case, it is necessary to filter out false matches. However, this is not always possible due to search engines limititations (see section 2.4).

Practical example: To carry out a basic impact analysis of two commercial banks in the UK, queries [“Royal Bank of Scotland”] and [Barclays] could be searched on Google, Yahoo! or Bing (MSN). Nevertheless, the number of matches is so high that the researcher might need to refine the search in order to focus on a specific issue concerning the companies. For instance, to explore the effects of the financial crisis on the banks, queries could be formulated such as [“Royal Bank of Scotland” AND “financial crisis”] and [Barclays AND “financial crisis”].

2.3.2 Link impact analysis

Link impact analysis is based on the comparison of the number of webpages or websites that link to a set of webpages or websites under research. The purpose of this type of research is, according to Thelwall (2009: 28), “to evaluate whether a given website has a high link-based web impact compared to its peers”. Also inlink counts can be an indirect gauge of other attributes of the organization represented by the website. For instance, this has been traditionally used, within the academic field, as a potential estimator of research performance (Smith and Thelwall, 2002). This method is more accurate than the previous one because false matches are surely avoided. However, other problems need to be taken into account, e.g. the search engines set restrictions on search of inlinks (see section 2.4), or the company is using more that one corporate URL.

Practical example: Following up the aforementioned example, to carry out a link impact evaluation of the two companies, the queries that should be used on Yahoo! (the search engine with the best array of functions for this purpose) would be as follows: [linkdomain:rbs.com -site:rbs.com] and

[linkdomain:barclays.com -site:barclays.com]. The results would be the estimated total number of links that point to the specific domain except the links that come from the same domain or self-links.

2.3.3 Co-link analysis

Co-link analysis could be classified as a type of link relationship mapping techniques (Thelwall, 2009). These are based on the link data that interconnect a set of websites in different ways in order to draw a diagram that illustrates the relationships between them. In particular, co-link analysis is based upon the number of webpages that link at the same time two webpages or sites belonging to the group of entities under study. This description fits the concept of co-inlink previously explained. As we mentioned before, co-link is often used as equivalent to the co-inlink concept. It is the case in this paper.

Co-links are analogous to the bibliometric concept of co-citation (Small, 1973). Co-link analysis has also been demonstrated to be a useful tool to reveal the cognitive or intellectual structure of a particular field of study (Zuccala, 2006). This method is particularly useful when websites interlink each other very rarely. It is the case of commercial websites that scarcely link the website of a competing company, especially when they are in the same industry (Vaughan, Gao and Kipp, 2006). The explanation to this could be that companies seem to avoid diverting web traffic to competitors (Shaw, 2001). Moreover, as Vaughan (2006) points out, co-link data are more robust than inlink data as the former are less easily manipulated. Multidimensional scaling and network diagrams are often used to show the data gathered and to interpret results.

Practical example: The query [linkdomain:rbs.com -site:rbs.com linkdomain:barclays.com -site:barclays.com] would retrieve the estimated total number of links that point simultaneously to both domains except the links that come from the same domain or self-links.

To conclude this section, some final considerations need to be done. Due to the origins of the discipline, so far the majority of webometric research has been carried out in the academic field. Thelwall, Vaughan and Björneborn (2005: 113) acknowledge this situation and point out that "This is ironic given that the Web is dominated by commercial sites". Webometrics is progressively enlarging its scope, focusing upon political websites (Foot and Schneider, 2006; Park and Thelwall, 2008), social networking (Thelwall, 2008b; 2008c; 2008e) and commercial websites (see section 3).

2.4 Collecting data using commercial search engines

Search engines are crucial for webometric research, because their databases are the source of information that cover most of the Web. Despite the fact that personal web crawlers can be used to automatically download pages and extract their links, commercial search engines have been used extensively for research especially when large areas or potentially the whole Web are the object of the study.

In order to perform a better research using commercial search engine data, it is fundamental to get a good understanding of the industry context, the advanced functions offered and the limitations.

A feature of the search-engine market is the oligopoly of three search-engine operators Google, Yahoo! and Live Search (Microsoft) which, from a global perspective, share the majority of the generalistic search-engine market. In specific areas of the Web the are other players such as Technorati for searching blogs. Search engine industry is under a constant process of change and innovation. This issue has been treated by many papers in recent years (Bar-Ilan, 2004; Vaughan and Thelwall, 2004; Lewandowski, Wahlig and Meyer-Bautor, 2006; Evans, 2007; Thelwall, 2008d).

As already mentioned, commercial search engines are the only source of data that covers the entire Web. However there are some significant limitations derived from the use of commercial search engines:

- Search engines do not index the entire Web (Sherman and Price, 2001; Bar-Ilan, 2004; Thelwall, Vaughan and Björneborn, 2005).
- Ranking systems eliminate similar or identical pages in their results, in order to avoid providing useless information (Gomes and Smith, 2003; Thelwall, 2008a).

- Crawling and reporting algorithms are commercial secrets and, therefore the exact criteria used to rank the information is unknown (Thelwall, Vaughan and Björneborn, 2005).
- The total number of results offered by search engines are estimates as they use algorithms that prioritise response time rather than exhaustiveness (Björneborn and Ingwersen, 2001).
- Results can be subject to national or language biases (Vaughan and Thelwall, 2004).
- The results can fluctuate and change over the time. In addition, only a few number of pages are accessible (usually just a maximum of 1000).

Commercial search engines are the best and unique source of information we have for certain types of webometric research, however they are not designed with this academic purpose and the results are not as exhaustive as we would desire. At this moment, Yahoo! is the search engine that is more useful for webometric research (Thelwall, 2008d). Yahoo! inlink data can be gathered in two different ways, through the general Yahoo! search engine and through Yahoo! Site Explorer. Despite the latter specializing in web structure information, complex queries can only be submited through the general Yahoo! search engine interface.

Nevertheless, collecting data can be a very time consuming process if using the web interface. This problem could be overcome by specialized software based on the application programming interfaces (API) developed by search engines and other services on the Web.

3. A webometric approach to business studies

Internet research applied to companies can provide new insights on Competitive Intelligence (CI) in order to help companies generate and maintain a competitive advantage (Porter, 1980). CI consists of a systematic plan to obtain and analyse information about competitors and general trends in the industry (Kahaner, 1996). The abundance of information available in the Internet generates new opportunities and challenges for businesses that need to monitor changes around them in order to compete in better conditions.

In the last decade, a few Web hyperlink analysis have been conducted for CI purposes with promising results (Reid, 2003; Tan *et al*., 2002). For example, Reid (2003) proposed a link analysis method which analysed a particular website, but not the global context of the Web. Over the last 5 years, Vaughan and colleagues have investigated the relationships between inlink counts and business performance measures as well as the application of co-link research to companies' websites.

3.1 Do inlinks and financial variables correlate?

Vaughan and colleagues' research on commercial websites has explored quantitative relationships between inlink counts and business performance variables. Vaughan and Wu (2004) made the first attempt to prove the hypothesis that the number of inlinks to commercial sites correlates with financial variables. This study tests the hypothesis in two groups of Chinese companies. Group 1 is made up of China's top 100 Information Technology (IT) companies and group 2 is comprised of top 100 privately owned companies.

The IT industry was selected because companies in this industry, by the nature of their activity, are likely to be leaders in utilizing the Web for commercial purposes. Experience has shown that different industries, have distinct patterns in the use of web technologies and therefore only homogeneous companies are likely to be comparable. Group 1 constitutes a homogeneous group of companies in terms of business activity, whereas group 2 is not homogeneous.

The financial variables available were: gross revenue, profit, export revenue and research and development expenses (for group 1); and, gross revenue (for group 2). Inlink data to each company website were collected using major commercial search engines at that time: Google, AltaVista, AllTheWeb and MSN Search. Also, the variable *website age* was used based on previous evidence (Vaughan and Thelwall, 2003) showing that inlinks to a website correlated with the age of the site, this is, older sites receives more inlinks. This data were retrieved from the Wayback machine in the Internet Archive (www.archive.org).

Spearman correlations in Table 1 show significant relationships between inlink counts and the three accounting variables. Results suggest that inlinks could be considered as a complementary performance indicator of a company. Vaughan and Wu (2004: 494) suggest that the strong correlation

found with research and development expense "could mean that companies that invest more in research and development have a better Web presence and that their sites are more visible and attract more links to them". No significant correlation was found between inlink count and export revenue. However, export revenue only represents a small fraction (around 8%) of the gross revenue and therefore it seems not to be of relevance as a global performance indicator.

Table 1: Spearman's correlations between business performance measures and inlinks (Vaughan and Wu, 2004)

	Gross revenue	Profit	Research and development expenses
Inlink count	.51	.30	.64
Inlink count / Website age	.50	.30	.63
All correlation coefficients in the table are statistically significant at .01 level.			

In relation to group 2 consisting of heterogeneous companies, there is no significant relationship between inlink count and gross revenue. As previously mentioned, this group is made up of companies belonging to different industries. Vaughan and Wu (2004: 494) conclude that "link count can be an indicator of business performance only when homogeneous group of companies are being compared".

In a second paper, Vaughan (2004b) studied the same relationships for the top 100 IT companies in China and the top 51 IT companies in the United States (US). Spearman's test in Table 2 shows significant correlations, lending support to the conclusion raised in the previous study. It is worth noting that the two sets of correlation coefficients are very similar in spite of the remarkable differences between both countries.

Table 2: Spearman's correlations between business performance measures and inlinks (Vaughan, 2004b)

	Inlink count & Gross revenue	Inlink count & Profit	Inlink count / Website age & Gross revenue	Inlink count / Website age & Profit
China	.51	.30	.50	.30
United States	.51	.35	.58	.37
All correlation coefficients in the table are statistically significant at .01 level.				

In a third study (Vaughan, 2004a), all Canadian and United States IT companies were examined. This time, in order to raise more general conclusions, the whole population of the companies in the industry was analysed. Also a new variable, the number of employees, is used to measure the size of the company. This variable is intended to oversee the effect that a larger company will tend to have larger revenue if the rest of the variables remain constant. Apart from confirming previous evidence, the results in Table 3 demonstrate that there is still a significant correlation, even after considering the company size.

Table 3: Correlation between business performance measures and inlinks (Vaughan, 2004a)

	Inlink count & Number of employees	Inlink count & Revenue	Inlink count & Revenue per employee	Inlink count / Website age & Revenue	Inlink count / Website age & Revenue per employee
Canada	.57	.55	.35	.55	.36
United States	.68	.71	.53	.67	.51
All correlation coefficients in the table are statistically significant at .01 level.					

More recently, new research has sought to confirm this evidence by exploring different types of industries.

Romero-Frías and Vaughan (2009a) analysed the top 50 international banks by using two different webometric techniques, inlink analysis and co-link analysis. The banking industry was selected due to

its economic significance in the financial crisis context and its high level of internationalization. Two sets of inlink counts (December 2008 and June 2009) and a set of financial variables for the year 2007 (including total assets, total liabilities, total revenue, net income, earnings before tax and return on assets) were taken into account in the study. Spearman's test showed that a majority of the correlation coefficients were significant at the .01 level. Only return on assets was found not significant.

Findings show that inlink counts could be used as a gauge of the banks' financial position and financial performance measures in absolute terms. However, there is no evidence when we refer to relative financial measures, such as return on assets. This could be explained by the absolute nature of the figure "inlink counts", as it accumulates all links pointing to a webpage since its creation. This is similar to the nature of financial position variables and, somehow, to the financial performance measures as revenue or total income tend to increase over time based on previous performance.

These results are also consistent with the results reported by Romero Frías, Vaughan and Rodríguez Ariza (2009). This study extended previous research by finding evidence of significant correlations between inlink counts and financial variables in several industries in the United States. The study analysed five different industries (commercial banks, construction of buildings, general merchandise store, utilities and mining), as well as the companies in the Dow Jones Industrials. The following economic variables for the years 2005-2007 were collected: number of employees, total assets, net income, total revenue, EBITDA, return on assets (ROA) and return on equity (ROE). Inlink data were retrieved from Yahoo! in January and May 2009. Due to the non normality of the inlink variables and other financial variables, Spearman's test was used. Correlations for the Dow Jones set of companies were not significant, confirming the evidence that only homogeneous companies in terms of activity are comparable (Vaughan and Wu, 2004). Significant correlations were found, to different extents, for all the industries, except for the Construction one. The Store industry had the positive and highest correlation coefficients with all the financial variables, including ROA and ROE. Utilities was the second industry regarding the level of correlation and the number of variables that were correlated, followed by the Banking and Mining industries.

In comparison to the results of Romero-Frías and Vaughan (2009a), correlation coefficients for the US banking industry are higher than for the global banking industry. This could be explained by the homogeneous competitive conditions that exist in the US market versus the heterogeneous markets where the top international banks operate. For instance, the extent to which the Internet is used for commercial purposes in the different countries could also be an explanatory variable.

It is worth remarking that correlation does not mean causation. The large number of inlinks that a company's website attracts does not cause the better financial performance. Although it is clear that a positive web image may constitute an intangible asset for a company and therefore can generate future incoming resources, it is more feasible that a bank that is doing well financially is able to maintain a high profile on the Web, maybe through the development of e-commerce practices. As the economy becomes more and more digital and information-based, companies are prompted to monitor if their web presence is in accordance to their financial importance. Web presence measured by inlink count could be an appropriate gauge to evaluate intangible assets related to the Internet.

3.2 Co-link analysis

Web co-link analysis for business information started by focusing on a single industry (Vaughan and You, 2006) or on a specific sector within an industry (Vaughan and You, 2008; Vaughan and You, 2009).

Based on the results of the past studies, Vaughan and You (2006) applied co-link analysis to map business competitive positions of 32 telecommunication companies. The hypothesis under examination is that the number of co-links to the websites of each pair of companies is a measure of the similarity between the two companies. This means that the more co-links the two companies have, the more related they are from the point of view of the sites that link to them. 32 companies were selected according to the following criteria: companies from different parts of the world, from different sectors within the telecommunications industry and top companies in terms of revenue. Co-link data were collected in order to reflect the business relationships in two markets, the global market and the Chinese market. With this purpose, global Yahoo! and Yahoo! China, the top search engine in China at the time of the research, were used respectively to collect the data. The symmetrical co-link

matrix obtained was analysed by using multidimensional scaling (MDS), which generated a map showing the relative positions of the companies in the industry. An MDS map depicting the relative positions of companies in the global market is shown as an example in Figure 3.

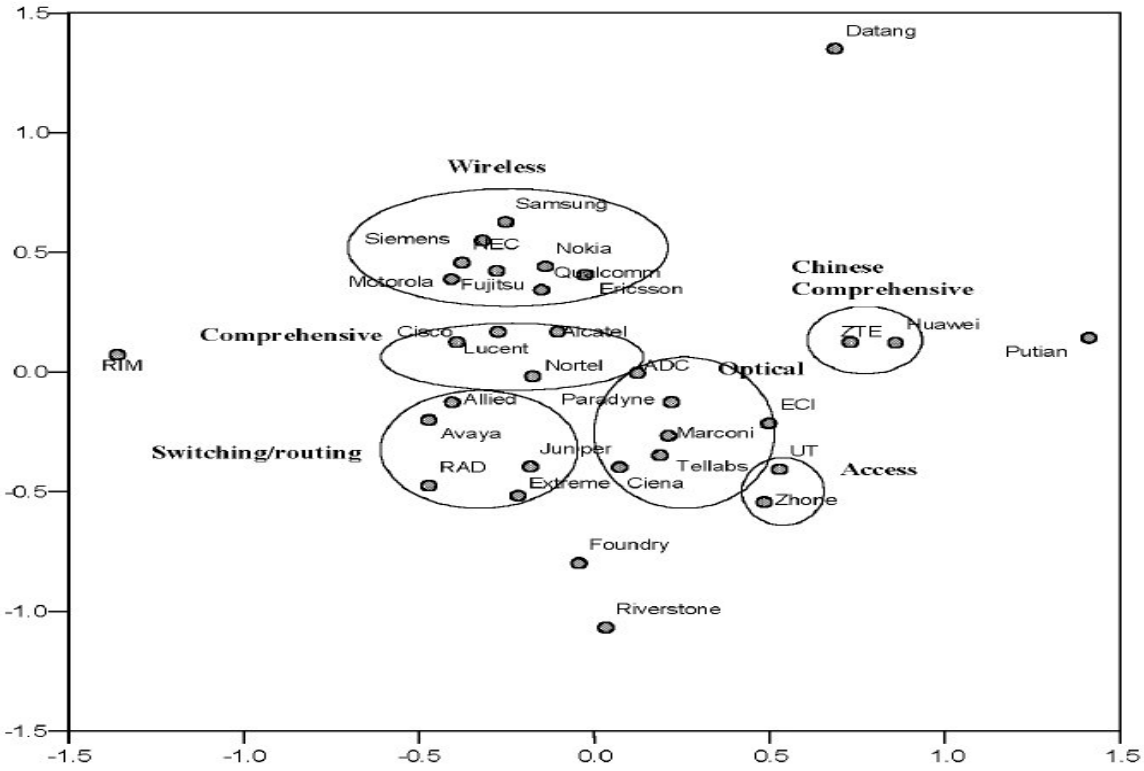

Figure 3: MDS mapping result based on global Yahoo! data (Vaughan and You, 2006: 619)

The companies are clearly clustered into the sectors of the telecommunications industry as labeled in Figure 2. The companies that are not grouped into clusters present specific features which explain their positions. The authors concluded that the maps obtained are consistent with the competition landscape of the industry. Findings suggest that co-link data do contain information about the relationship among companies. Vaughan and You (2006: 618) assert that: "Highly co-linked companies are highly related in their products and the market. Since related companies are competitors (they serve the same market needs), it follows that co-link data can be used to map the competitive position of companies."

A more recent paper (Vaughan and You, 2008) proposed a method that combines page content mining (keyword) and Web structure mining (co-link data) to achieve a more detailed picture of a particular sector within an industry. The WiMAX sector of the telecommunication industry was chosen specially because this acronym is used only to refer to this particular technology and therefore it avoids the problem of false matches. 39 companies were included in the study and two sets of co-link data were collected (with and without keyword) using the MSN search engine. By adding the keyword WiMAX to the search, only webpages that mention that word are retrieved. This implies removing pages that co-link the two companies for reasons other than the companies' activities in this particular sector. The maps obtained by applying Multidimensional Scaling are shown in Figure 4.

The first map, without the keyword (left), shows clusters of companies in terms of their overall competitive positions in the telecommunication industry. Three main groups are identified: "WiMAX chip", "WiMAX equipment" and "comprehensive" companies that offer a wide variety of products and services. The second map, with the keyword (right), is able to present companies in five main groups that reflect their competitive positions within the WiMAX sector. This map shows a more detailed analysis on the relationships between telecommunication companies but only from the point of view of the WiMAX services they provide. Authors compared their analysis of the sector with results obtained by an independent market research, finding their conclusions a great match. This study proves that by adding a keyword to the query in the search engine, the information obtained about the relative positions of the companies in the sector is more accurate and meaningful.

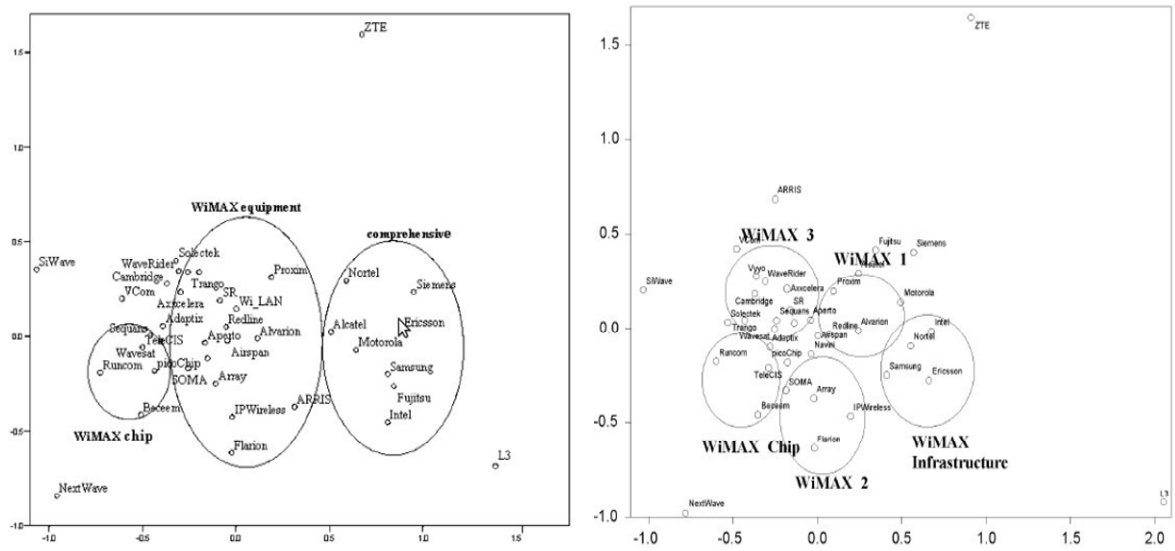

Figure 4: MDS map without the keyword (left) and MDS map with the keyword WiMAX (right) (Vaughan and You, 2008: 438 and 439)

The methodology developed in these papers has also been tested and verified in other countries and other industries. Romero-Frías and Vaughan (2009a) analysed the evolution of the top 50 international banks through co-link analysis, finding the existence of 3 main clusters of banks: Asian, European and the so called English-speaking banks that include banks from United States, United Kingdom, Australia and Canada. Vaughan, Tang and Du (2009) studied China's chemical industry and electronics industry. Finally, Romero-Frías and Vaughan (2009b) extended the use of co-link analysis into the banking industry in the US in order to test the feasibility of combining page content with co-link data to monitor financial crisis.

Parallel to these studies on homogeneous sets of commercial websites, research on heterogenous websites has been carried out to test the triple helix theory on the Web (Stuart and Thelwall, 2006; García-Santiago and de Moya-Anegón, 2009). In this line, Romero-Frías and Vaughan (2010) extended co-link analysis to websites of heterogeneous companies belonging to five stock exchange indexes. The companies selected are the biggest in their respective economies and therefore have a significant web presence. In addition, they are also likely to receive more attention from users, competitors, government and other economic agents. When applying co-link analysis to companies belonging to different industries, the interpretation of co-link as a similarity measure does not stand only for competitive relationships, but for a wide variety of interactions, such as alliances and other linkages between distinct industries. The main conclusion of the study indicates that the degree in which the industrial activity is information centered determines the position of the companies in the MDS maps of the different indexes. The so called information centered industries include mainly IT, media, and financial companies, among others. Co-link analysis could reveal the extent to which certain industries are involved in an ongoing process of business model transformation.

3.3 Content analysis

Qualitative research is a necessary complement to quantitative research because it provides confirming evidence about the relevant nature of the links being analysed. First study on commercial websites was carried out by Vaughan, Gao and Kipp (2006). This study examined three items from a set of webpages: country location of linking webpages, types of sites that created the links and motivations for linking. They analysed 418 links to US companies and 390 links to Canadian companies randomly taken from inlinks to companies studied in Vaughan (2004a). The results conclude that the vast majority of links to business sites are business related, supporting the relevance of link impact analysis for data mining purposes on commercial sites. Regarding motivations for linking, findings show that most types of links came from online directories (22,5%), list of companies (19,6%) and news articles (12,4%). The predominance of inlinks described as directories and list of companies, besides the fact that only 2 out of the 808 links studied pointed to competitors, indicate that co-link could be a fruitful direction to study business competition, as already explained in the previous section.

Content analysis has also been carried out to study the motivations for co-link creation. Vaughan, Kipp and Gao (2007) used content analysis of co-link pages to examine whether co-links had business purposes and why co-links were created. 495 co-link pages to 32 telecommunication companies (the same as used in Vaughan and You, 2006) were classified. 68.5% of co-links were created by commercial sites and 57.6% of co-links pointed to related products or companies. Also, these authors (2007: 447) conclude that "co-links to homepages are more likely to connect highly related businesses and to show a true business relationship between co-link companies."

4. Concluding ideas and future research

This paper has offered an overview of the webometric methodology to perform quantitative research of the web phenomena. The increasing importance of a hyperlinked society highlights the relevance of this approach to business studies. So far, most of the research on commercial websites has been carried out from an information science perspective, but there is a promising and wide field to explore different types of business issues by using information extracted from the Web. Additional work still needs to be done in analysing different industries and regions and in developing business applications to benefit from the findings up-to-date. It would be useful to monitor the Web periodically to analyse how web variables and economic variables correlate over time. Moreover, explanatory models based on multivariate regression need to be explored.

To conclude, the use of alternative sources to collect web data represent a new horizon for researching in social sciences. In this line, an improvement in the possibilities offered by search engines is to be expected, in particular based on a semantic analysis of internet contents and in the development of more powerful APIs.

References

Almind, T. C. and Ingwersen, P. (1997) 'Infometric analyses on the World Wide Web Methodological approaches to 'webometrics'', *Journal of Documentation*, 53(4), pp 404-426.

Bar-Ilan, J. (2004) 'The use of Web search engines in information science research', in Cronin, B. (Ed.), *Annual review of information science and technology*, pp. 231–288, Medford, NJ: Information Today.

Battelle, J. (2005) *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, London: Portfolio.

Benoît, G. (2002) 'Data mining', in Cronin B. (ed.), *Annual Review of Information Science and Technology*, Vol. 36 (1), pp 265-310, Medford, NJ: Information Today.

Berners-Lee, T. (1999) *Weaving the Web*, San Francisco: Harper.

Björneborn, L. (2004) *Small-world link structures across an academic Web space: A library and information science approach*, Doctoral dissertation, Royal School of Library and Information Science, Copenhagen, Denmark, [online], Available: http://vip.db.dk/lb/phd/phd-thesis.pdf, [consulted 21 Jul 2008].

Björneborn, L., and Ingwersen, P. (2001) 'Perspectives of webometrics', Scientometrics, 50(1), 65–82.

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

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual Web search engine', *Computer Networks and ISDN Systems*, 30, pp 1-7.

Evans, M. P. (2007) 'Analysing Google rankings through search engine optimization data', *Internet Research*, 17(1), pp 21-37.

Foot, K. and Schneider, S. (2006) *Web campaigning*, Cambrigde, MA: MIT Press.

García-Santiago, L. and de Moya-Anegón, F. (2009) 'Using co-outlinks to mine heterogeneous networks', *Scientometrics* , 79(3), pp. 681-702.

Garfield, E. (1979) *Citation indexing: Its theory and applications in science, technology and the humanities*, New York: Wiley.

Gomes, B. and Smith, B.T. (2003) Detecting query-specific duplicate documents. U.S. Patent 6,615,209, [online], Available: http://www.patents.com/Detecting-query-specific-duplicate-documents/US6615209/en-US/, [consulted 15 Nov 2009].

Google blog (2008) 'We knew the web was big...', [online], Available: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, [consulted 14 Jan 2009].

Hine, C. (2008) 'Internet Research as an Emergent Practice', in Hesse-Biber, S. N. and Leavy, P. (2008) *Handbook of Emergent Methods*, pp 525-541, New York: The Guilford Press.

Jenkins, H. (2006) *Convergence Culture: Where Old and New Media Collide*, Cambridge, MA: MIT Press.

Kahaner, L. (1996) *Competitive Intelligence – How to Gather, Analyze, and Use Information to Move Your Business to the Top*, 7th ed., New York, NY: Touchstone.

Keen, A. (2007) *The Cult of the Amateur: How the Democratization of the Digital World is Assaulting Our Economy, Our Culture, and Our Values*, New York: Doubleday Currency.

Lewandowski, D., Wahlig, H. and Meyer-Bautor, G. (2006) 'The freshness of web search engine databases', *Journal of Information Science*, 32(2), pp 131–148.

O'Reilly, T. (2005) 'What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software', [online], Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [consulted 17 Jan 2009].
Park, H. W. & Thelwall, M. (2008) 'Web linkage pattern and social structure using politicians' websites in South Korea', *Quality & Quantity*, 42(6), pp 687-697.
Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*, New York, NY: Free Press.
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*, pp 57-84, Hilliard, Ohio: Ideal Group Inc.
Romero Frías, E., Vaughan, L. and Rodríguez Ariza, L. (2009) 'El recuento de enlaces a sitios Web comerciales como indicador de las variables de desempeño y posición financiera de la empresa: estudio empírico de diversos sectores empresariales en Estados Unidos', XV Congreso de la Asociación Española de Contabilidad y Administración de Empresas, Valladolid, Spain, September 23-25, 2009.
Romero-Frías, E. and Vaughan, L. (2009a) 'A Webometric analysis of the global banking industry', 35th EIBA Annual Conference, Valencia, Spain, December 13-15, 2009.
Romero-Frías, E. and Vaughan, L. (2009b) 'Financial Distress of U.S. Banking Industry Viewed through Web Data', 12th International Conference on Scientometrics and Informetrics ISSI 2009, Rio de Janeiro, Brazil, July 14-17, 2009.
Romero-Frías, E. and Vaughan, L. (2010) 'Patterns of Web Linking to Heterogeneous Groups of Companies: The Case of Stock Exchange Indexes', to appear in Aslib Proceedings.
Shaw, D. (2001) 'Playing the links: interactivity and stickiness in .com and 'not.com' websites', *First Monday*, 6(3), [online], Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/837/746 [consulted 10 May 2009].
Sherman, C., and Price, G. (2001) *The invisible Web*, Medford, NJ: Information Today, Inc.
Small, H. (1973) 'Co-citation in the scientific literature: A new measure of the relationship between two documents', *Journal of the American Society for Information Science*, July-August: pp 265–269.
Smith, A. and Thelwall, M. (2002) 'Web Impact Factors for Australasian universities', *Scientometrics*, 54, pp 363-380.
Stuart, D. and Thelwall, M. (2006) 'Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry', *Research Evaluation*, 15(2), pp 97–106.
Tan, B., Foo, S. and Hui, S. C. (2002) 'Web information monitoring for competitive intelligence', *Cybernetics and Systems*, 33(3), pp 225-251.
Technorati (2009) 'State of Blogosphere', [online], Available at: http://technorati.com/blogging/article/state-of-the-blogosphere-introduction/ [consulted 15 Nov 2009].
Thelwall, M. (2004) *Link Analysis: An Information Science Approach*, San Diego: Academic Press.
Thelwall, M. (2008a) 'Extracting accurate and complete results from search engines: Case study Windows Live', *Journal of the American Society for Information Science and Technology*, 59(1), pp 38-50.
Thelwall, M. (2008b) 'How are social network sites embedded in the web? An exploratory link analysis', *Cybermetrics*, 12(1), paper 1.
Thelwall, M. (2008c) 'No place for news in social networking web sites?', *Online Information Review*, 32(6), pp 726-744.
Thelwall, M. (2008d) 'Quantitative comparisons of search engine results', *Journal of the American Society for Information Science and Technology*, 59(11), pp 1702-1710.
Thelwall, M. (2008e) 'Text in social network web sites: A word frequency analysis of Live Spaces', *First Monday*, 13(2).
Thelwall, M. (2009) *Introduction to Webometrics. Quantitative Web Research for the Social Sciences*, Morgan & Claypool.
Thelwall, M., Vaughan, L. and Björneborn, L. (2005) 'Webometrics', in Cronin B. (ed.), *Annual review of information science and technology*, 39, pp 81-135, Medford, NJ: Information today.
Thelwall, M. and Wilkinson, D. (2008) 'A generic lexical URL segmentation framework for counting links, colinks or URLs', *Library & Information Science Research*, 30, pp. 515–526.
Turow, J. and Tsui, L. (eds.) (2008) *The Hyperlinked Society: Questioning Connections in the Digital Age*, Ann Arbor: The University of Michigan Press.
Vaughan, L. (2006) 'Visualizing Linguistic and Cultural Differences Using Web Co-Link Data', *Journal of the American Society for Information Science and Technology*, 57(9), pp 1178-1193.
Vaughan, L. (2004a) 'Exploring website features for business information', *Scientometrics*, 61(3), pp 467-477.
Vaughan, L. (2004b) 'Web hyperlinks reflect business performance—A study of US and Chinese IT companies', *Canadian Journal of Information and Library Science*, 28(1), pp 17–31.
Vaughan, L. and Thelwall, M. (2003) 'Scholarly use of the Web: What are the key inducers of links to journal web sites?', *Journal of the American Society for Information Science and Technology*, 54, pp 29-38.
Vaughan, L. and Thelwall, M. (2004) 'Search engine coverage bias—Evidence and possible causes', *Information Processing and Management*, 40(4), pp 693–707.
Vaughan, L. and You, J. (2006) 'Comparing business competition positions based on Web co-link data—The global market vs. the Chinese market', *Scientometrics*, 68(3), pp 611–628.
Vaughan, L. and You, J. (2008) 'Content assisted web co-link analysis for competitive intelligence', *Scientometrics*, 77(3), pp 433-444.

Vaughan, L., and Wu, G. Z. (2004) 'Links to commercial websites as a source of business information', *Scientometrics*, 60(3), pp 487–496.
Vaughan, L., and You, J. (2009), 'Keyword enhanced Web structure mining for business intelligence', *Lecture Notes in Computer Science*, Vol. 4879, pp. 161–168.
Vaughan, L., Gao, Y. and Kipp, M. (2006) 'Why are hyperlinks to business Websites created? A content analysis', *Scientometrics*, 67(2), pp 291–300.
Vaughan, L., Kipp, M. and Gao, Y. (2007) 'Are co-linked business web sites really related? A link classification study', *Online Information Review*, 31(4), pp 440–450.
Vaughan, L., Tang, J. and Du, J. (2009) 'Examining the robustness of Web co-link analysis', *Online Information Review*, 33(5), pp. 956-972.
Zuccala, A. (2006) 'Author Cocitation Analysis Is to Intellectual Structure A Web Colink Analysis is to...?', *Journal of the American Society for Information Science and Technology*, 57(11), pp 1487-1502.

