Author: David Gunnarsson Lorentzen

Title: Webometrics Benefitting from Web Mining? An Investigation of Methods and Applications of Two Research Fields

Author affiliation: University of Borås, Swedish School of Library and Information Science

Author address: SE-501 90 Borås, Sweden

Corresponding author e-mail address: david.gunnarsson_lorentzen@hb.se

Corresponding author telephone number: 0046334354087

Corresponding author fax number: 0046334354005

Abstract

Webometrics and web mining are two fields where research is focused on quantitative analyses of the web. This literature review outlines definitions of the fields, and then focuses on their methods and applications. It also discusses the potential of closer contact and collaboration between them. A key difference between the fields is that webometrics has focused on exploratory studies, whereas web mining has been dominated by studies focusing on development of methods and algorithms. Differences in type of data can also be seen, with webometrics more focused on analyses of the structure of the web and web mining more focused on web content and usage, even though both fields have been embracing the possibilities of user generated content. It is concluded that research problems where big data is needed can benefit from collaboration between webometricians, with their tradition of exploratory studies, and web miners, with their tradition of developing methods and algorithms.

Keywords: webometrics; cybermetrics; web mining; web data mining; literature review.

1. Introduction

Webometrics, described as the "quantitative study of web-related phenomena" (Thelwall, Vaughan & Björneborn, 2005, p. 81) and *web mining* (also named "web data mining"), defined as "the discovery and analysis of useful information from the World Wide Web," (Cooley, Mobasher & Srivastava, 1997, p. 558) are research areas where mostly quantitative studies of web content, structure and usage are performed, but presumably from different perspectives, as they are sub-fields of information science and computer science respectively. This review paper aims to shed light on how they have evolved, what they have in common and what the differences are, by examining methods, research problems and citations. Building on an extensive literature survey, it is, arguably, the first paper to do so.

In 2005, Thelwall and Wouters suggested that the information scientist can be data evaluator and method developer, and also a broker of social science methods in a metadisciplinary context. The overall approach of the information sciences to new data sources have centered mainly on academic perspectives, such as scholarly communication, libraries and information dissemination. In contrast, the computer sciences have focused more on the development of algorithms and descriptive modeling (Thelwall & Wouters, 2005). The text by Thelwall and Wouters is a rare example of reflections on how research within these disciplines can benefit from strategic and systematic work on web-based big data. In this paper, such reflections are narrowed down to subfields of information science and computer science.

In the following attention is focused on historical dimensions, differences/similarities and any collaborative possibilities relating to these two fields. The potential of these two fields to benefit each other through closer contact and collaboration will be investigated. The main questions for this paper are:

- What characterizes and separates the fields of webometrics and web mining?
- To what extent and how have they embraced the possibilities of user generated web content?
- What kind of potential overlap is visible as resource for future collaboration?

To answer these questions I will work through the roots of the fields, review a substantial sample of research papers from each field, identify which communities are visible regarding specializations within the fields, and investigate how frequently researchers from both fields cite each other. In the literature review (section 5) it was not possible to identify any earlier discussion of both fields. Hence, I am using domain specific reviews as part of my data.

Research in both fields has been performed under slightly different labels. *Cybermetrics* is here included alongside *webometrics* as it has been the preferred label in Spain (Thelwall, 2009, p. 6), even though it is defined as being more general, and not just limited to the web (see Björneborn & Ingwersen, 2004). Papers dealing with cybermetric research that is not web related are not included in this review. *Web mining* and *web data mining* have been treated

synonymously by researchers (e.g. Zhang & Segall, 2008), and are treated the same way here. This is reflected in the queries for acquiring data (see section 4 and Appendix).

This review utilizes a taxonomy for classifying academic knowledge and disciplines as starting point (Becher and Trowler, p. 184), described in section 2. Section 3 includes other cross-disciplinary reviews and citation analyses. Section 4 outlines the data collection. Sections 5 and 6 include the results of the review, where the former is a literature comparison and the latter includes a co-word analysis and a cross-field citation analysis. Finally, the findings are discussed in section 7, and conclusions are drawn in section 8.

2. Theory

The data used to compare webometrics with web mining will be analyzed with the help of Becher and Trowler's (2001, p. 184) proposed taxonomy for classification of academic knowledge and disciplines. This is based on four basic sets of properties, two cognitive (i.e. type of knowledge) and two social (i.e. people and relationships, and how research is conducted). The cognitive properties are hard/soft and pure/applied; the social properties are convergent/divergent and urban/rural.

Becher and Trowler (2001, p. 39) acknowledge that "the boundaries between the hard/soft, pure/applied knowledge domains cannot be located with much precision" but their table of knowledge and disciplinary groupings (p. 36) gives us guidelines. Hard knowledge is more quantitatively oriented and value-free than soft knowledge, although qualitative approaches exist in the hard-applied disciplines. Pure knowledge is more self-regulating whereas applied knowledge is more open to influence from the outside. Applied knowledge is more oriented towards results in products, techniques, protocols, and procedures and pure knowledge is oriented towards results in discovery, explanation, understanding, and interpretation. Hence, pure knowledge is aiming to answer more basic questions and applied knowledge is aiming to solve more specific problems. Sometimes these specific problems originate from organizations and actors outside the academia. Regarding the social aspects, urban communities tend to have a higher people-to-problem ratio, a higher pace, a higher degree of collaboration and span a narrower area than their rural counterparts (Becher & Trowler, 2001).

As information science scholar, I became first interested in the collaborative opportunities between these two fields, having previous work experience from both. One reason for collaboration between these fields, at least from the webometric side, is that computer science methods could be very relevant for the big data on the web that is also unstructured or semi-structured. However, there seemed to be very few examples of collaboration. One problem could be language issues, also noted by Becher and Trowler (2001, p. 46): "the professional language and the literature of a disciplinary group play a key role in establishing its cultural identity. This is clearly so when they embody a particular symbolism of their own [...], or a significant number of specialized terms [...], placing them to a greater or lesser degree beyond the reach of an uninitiated audience." One example of this is the use of formulas, which is far more common within web mining.

3. Previous research

As mentioned, there are no cross-disciplinary reviews of webometrics and web mining. Other fields have been examined in this way, however. Ruller (1993) reviewed information science and computer science literature from an archivist's perspective, and Duane Ireland and Webb (2007) wrote a cross-disciplinary paper of entrepreneurship research, with the aim of finding common interests and bridging opportunities for scholars of the area. On a broader level, Fischer, Tobi and Ronteltap (2011) investigated collaboration between the natural and social sciences. The disciplines involved, approach used and objective of study in each paper were noted. However, none of these studies were comparisons of research fields. A comparison of research fields was provided by Glass, Ramesh and Vessey (2004) who studied computer science, software engineering, and information systems. They used a classification scheme comprised by topic, research approach, research method, reference discipline (the discipline of theory used), and level of analysis (technical or behavioral).

Both Ruller (1993) and Duane Ireland and Webb identified potential for collaboration. Fischer, Tobi and Ronteltap (2011) found that some disciplines were collaborating more frequently in their sample of articles (mainly economics with biology, physics, and environmental science). Glass, Ramesh and Vessey (2004) found little topical overlap between the fields, with computer science and software engineering more focused on the technical level of analysis of computer concepts and systems/software respectively, and information systems studying topics related to organizational concepts at a behavioral level of analysis. As a potential problem related to collaboration, they identified the differences in research approaches and methods among the fields.

Within the field of scientometrics, solutions to the investigation of cross-disciplinary comparisons of scientometric indicators have been proposed (e.g. Schubert & Braun, 1996; Ball, Mittermeier & Tunger, 2009). Kajikawa and Mori (2009) suggested a methodology for measuring inter-disciplinarity of articles based on network analysis of citations, and Williams and colleagues (2013) presented a bibliometric method for identifying areas that could benefit of cross-disciplinary research.

There are a number of examples of cross-field citation analyses. Biehl, Kim and Wade (2006) analyzed how 31 top journals in the management disciplines relate to each other by analyzing citation patterns, and Pratt, Hauser and Sugimoto (2012) analyzed citation patterns of journals from business disciplines. Kirby, Hoadley and Carr-Chellman (2005) did a journal based citation analysis of learning sciences and instructional systems design. Van Leeuwen and Tijssen (2000) studied to what extent disciplines are interrelated by using journal-to-journal cross-disciplinary citation analysis. Small (2010) showed that the share of interdisciplinary (co-citation) links increases with level of aggregation, where a high level corresponds to a global map of links. Following this, the citation analysis in this study is expected to yield lower cross-field citation counts than, for example, a cross-field citation analysis of information science and computer science.

4. Method

The brief review above highlights some different approaches to compare research within different fields, and to identify inter-disciplinary research. A review is always subjective in a way, even though the use of a classification scheme might reduce the subjectivity somewhat. The idea here is to balance this subjectivity with a quantitative co-word analysis. To identify traces of collaboration, a cross-field citation analysis was performed. In the following, the method for collecting data and selecting articles for qualitative analysis is outlined. Sections 4.3 and 4.4 describe the methods for analyzing keywords and cross-field citations.

4.1. Data collection

While the selection of research articles and papers in a literature review is always a trade-off between what is available to the researcher, the whole corpus of relevant and overviewable number of items, the aim was to select articles that have been regarded important for their field, as well as a set of items that can be regarded as a representative sample of both fields. The focus of the selection has been on coverage of sub-areas within both fields rather than being exhaustive. To guard against possible bias in the literature selection, a quantitative keyword analysis has been performed in an attempt to find the main topical themes within the fields. Finally, for this review only articles written in English have been selected.

The data has been collected in two stages. In the first stage, the aim was to find as many items as possible for further qualitative selection. Hence, some concepts not directly related to the research fields were included in the queries. The second stage involved identifying co-occurrences of keywords related to both research fields, and to perform a cross-field citation analysis between them, i.e. to investigate how frequently one field cites the other.

Answering the research questions required data from citation databases. Scopus and Web of Science where first considered, and as Scopus contained more field specific content it was used as a starting point and as a data source for the keyword and citation analyses. Although Web of Science contains roughly the same amount of items concerning webometrics, there was a substantially larger amount of web mining items in Scopus. Two queries were

run in Scopus, one using webometric terms and the other using web mining terms. Time limits were omitted to make it possible to retrieve the first items mentioning "webometrics" and "web mining". The terms used were a combination of the author's knowledge and terms found in lists of contents of books within both areas. Then the queries were refined and used with field search in Scopus and Web of Science. In Scopus, the fields *Title*, *Abstracts* and *Keywords* were used and in Web of Science the *Topic Search* was used. These were complemented by searches in DOAJ, the Encyclopedia of Library and Information Science, ACM Digital Library, Encyclopedia of Library and Information Sciences, Library Literature & Information Science, LISTA, IEEExplore, Inspec, Academic Search Premier and Business Source Premier. Where advanced search was not available, simple OR-searches were carried out using the terms *webometric**, *cybermetric**, "*web mining*" and "*web data mining*".

Two simple queries were used in Scopus for the keyword and citation analyses. These searches excluded items that were not categorized as article, article in press, conference paper, or review by Scopus. The analysis of citations was carried out not only across fields, but also within the fields ensuring that important works within both fields were included in the review. 307 items were retrieved by the webometric query and 2,518 were retrieved by the web mining query. 12 items intersected.

As can be seen in Appendix, some terms are used in both the complex queries but used in different contexts (with webometrics or its synonyms and with web mining or its synonyms). It must be noted that some topics within the areas do overlap and, as a consequence, some items have been returned by both queries. Furthermore, it could be argued that the terms *webmetrics* and "*web metrics*" should be included in the webometrics query, but the first is the name of a software tool (WebMetrics) and not the methodology. The latter is discussed by Thelwall (2010b) who compared it to webometrics, stating that the former is not a concept of the latter, and Aguillo (2009) emphasized not confusing one with the other. Examples of the use of "web metrics" can also be found Palmer (2002), where it is used as a term for the process of evaluating a web site based on its statistics.

Furthermore, these queries are by no means an attempt to define the areas. The purpose was to find a substantial body of research, thus the inclusion of terms like *informetric** and *scientometric**. Inevitably, *informetric** and *scientometric** retrieves items not related to webometrics.

4.2. Selection of items

From the sets of items retrieved an initial selection on the basis of the research questions was made. Early papers with substantial impact were identified as well as papers developing new definitions in order to uncover the roots and definitions of the research fields. A snowballing method was used as the most frequently cited works, according to the data sources, were chosen first, and reviews and other items that had a historical angle were chosen after searching in the abstracts of the remaining works. To find the items discussing the social web and using user generated data, a search within the set of items was conducted using social web related search terms, for example "social web", "participatory web", "web 2.0" and "new web", as well as narrower concepts such as "social media", "participatory media", "participatory culture", "social search", "social technologies" and "participatory technology". Names of various social media were also used[1]. Finally, items that seemed unique for both fields were selected, as it could be argued that these indicate the spread of problems and methods used. Items of a philosophical nature that attempt to formulate some kind of disciplinary identity are also included.

The second stage of the selection was the completion of the list of items with interesting references, including books, from the selected items, the most cited items from the domain specific citation analysis, and items not already in the lists written by researchers who have been the most productive within the areas. Some items from the sets were discarded as they were deemed as irrelevant to the topic, for example, a couple of articles and papers dealt with networks within physics. Items dealing with social network analysis combined with bibliometric methods were also discarded, since their main data sources were citation databases which is outside the scope of webometrics and web

[1] Source: http://www.go2web20.net/. See Appendix for a list of used social web terms.

mining. Some items were reclassified after reading their introductions and an analysis of their references. The first search was performed in June 2012, additional searches were performed in September 2012, and final searches were performed in September 2013.

All non-review items were categorized according to a classification scheme adapted from the one used by Glass, Ramesh and Vessey (2005). This is comprised by the following categories: 1) the topic, or research problem, 2) research approach (method development, exploration of a topic, or evaluation of a system/collection), 3) methods, 4) type of data and access to data, and 5) data type category (content, structure or usage).

It has to be mentioned that this review represents the tip of the iceberg as 156 papers and one book from a substantially larger body of published items reviewed, and some possibly important articles within the fields might have been excluded. Nevertheless, this review builds on a carefully selected sample of a large material that was seen to be, given a combination of qualitative and quantitative approaches, both adequate and sufficient for the research purposes.

4.3. Keyword analysis

To attain a more comprehensive view of what has been focused within the research fields, a co-keyword analysis was performed. This method facilitates identification of major themes within a scientific field, the relationships among them, and also minor areas (He, 1999). The analysis was based on the Scopus field *Index keywords*. For the items that were not assigned index keywords, the field *Author keywords* was used.

Co-keyword networks were created in Gephi (Bastian, Heymann, & Jacomy, 2009). Not all items were assigned keywords and these were excluded in this analysis, which is based on 259 and 2,455 items respectively. The words used in the queries were excluded from the networks, as were disconnected keywords. The top 100 keywords according to the number of papers they were assigned to were kept in both networks.

As part of this analysis, I was aiming to identify communities within the fields. For doing this, the community detection algorithm *Modularity* (Blondel et al., 2008; Lambiotte, Delvenne & Barahona, 2009) in Gephi was applied to both networks. The nodes in these networks are the keywords. These are sized according to the number of articles they have been assigned to. The edges between the nodes indicate co-occurrences, where the frequencies of the co-occurrences are reflected in the thickness of the edges.

4.4. Citation analysis

A cross-field citation analysis was also conducted for the purpose of getting an indication of how frequently authors from one field cite papers and authors from the other field. Journal-to-journal citations were not analyzed due to two main reasons. First, a problem with analyzing journal-to-journal citations is that journals can be interdisciplinary (Small, 2010), and second, there are differences in where research papers are published. In these sets, web mining has a heavy overweight in conference papers (68% of the retrieved papers), and webometrics is heavily dominated by articles (77%). Instead, the choice was to do cross-field citation counts, as this answers the question of how frequently papers from one field cite papers from the other field.

BibExcel was used alongside a semi-automatic method in order to investigate cross-field citations. Field specific lists of references with the number of times each item was cited were created by the program. Some manual editing was needed to create more accurate lists. Citations were analyzed in relation to two questions: 1) “How frequently do papers from both fields cite a paper from the other field?” and 2) “Which authors from one field are most frequently cited by a paper from the other field?” For question 1, the resulting lists from BibExcel were matched against the sets of records from Scopus and for question 2, the 20 most productive authors within each class were matched against the references from the papers from the other class. Items found in both retrieved sets were counted as belonging to the other field, i.e. those who have been retrieved by both queries and also been cited by both fields, were included in the citation count for both metrics and mining.

5. Literature comparison

Both fields came to life in the mid-90s as sub-fields of established fields. The earliest published items in this review are from 1996 (web mining) and 1997 (webometrics). Webometrics came from the information sciences as a subfield of informetrics and web mining came from the computer sciences as a subfield of data mining. In early works within the fields, the content, usage and structure of the web have been found to be interesting objects to study.

5.1. Webometrics and web mining – origins and definitions

In this section, the origins of the areas will be outlined including definitions made by the authors. Reviews, surveys and overviews of the fields will be reviewed here as well.

5.1.1. Webometrics

Webometrics was first mentioned by Almind and Ingwersen (1997, p. 404): "the approach taken here will be called webometrics, which covers research of all network-based communication using informetric or other quantitative measures". They used a case study of Denmark's use of the web comparing it with other Nordic countries to demonstrate a method which can be used for webometric analyses. They found informetric methods on the web to be useful for a diverse range of tasks, such as issue management, the gathering of business intelligence, and research evaluation. The article is credited as the birth of webometrics by Thelwall (2010b).

Björneborn and Ingwersen (2001) acknowledged similarities to informetric and scientometrics methods, comparing linking with citing but with the obvious difference that links can go either way. In their framework, Björneborn and Ingwersen included graph theory, path analysis, transversal links (short-cuts between heterogeneous web clusters), weak ties and small-world phenomena (see Milgram, 1967).

Björneborn and Ingwersen (2001) argue that webometric analyses of the nature, structures and content properties of web sites and pages, as well as of link structures are important for understanding the web and its connections. Interestingly, Björneborn and Ingwersen mention "web mining": "the distributed, diverse and dynamical nature of the web – combined with minimal use of metadata – makes it a difficult setting for knowledge discovery or 'web mining'" (2001, p. 72). It is also worth to note that the well-known webometrician Mike Thelwall the same year proposed a crawler for web link mining (2001a).

Björneborn and Ingwersen acknowledged the web as a directed graph[2] where its pages are the nodes and its hyperlinks are the edges (2001), and proposed a basic link terminology for describing linking relationships between the nodes (2004). This terminology was later supported by Thelwall, Vaughan & Björneborn (2005). A challenge was identified by Björneborn and Ingwersen (2001), which involves analyzing and synthesizing findings as well as the development of theories and methodologies, partly for understanding the complexity of the topology of the web but also for understanding its functionalities and potentials. In their paper from 2004, they also identified four main areas for webometrics: (1) web page content analysis, (2) web link structure analysis, (3) web usage analysis (including log files of users' searching and browsing behavior), and (4) web technology analysis (including search engine performance).

The unique characteristics of the web make webometrics something other than bibliometrics and in the paper from 2004, Björneborn and Ingwersen pointed out that webometrics draw on bibliometric and informetric approaches. This "denotes a heritage without limiting further methodological developments of web-specific approaches, including the incorporation of approaches of web studies in computer science, social network analysis, hypertext research, media studies, and so forth" (Björneborn & Ingwersen, 2004, p.1217). One reason for the name "webometrics" was to emphasize its roots in bibliometrics and informetrics, and thereby its information science perspective (Thelwall, Vaughan & Björneborn, 2005).

[2] A graph is considered directed if the edges imply a direction, e.g. page A links to page B.

Thelwall, Vaughan & Björneborn (2005) reviewed webometric research focusing on different types of link analysis as well as on basic concepts and methods. They argue that the semantic web renders the contents of web pages and their interconnections more explicit and facilitates automatic processes, and thereby more powerful webometric analyzes. They also predict that collaboration with other fields, for example cultural studies, computer-mediated communication, social network analysis and social informatics could be a future trend.

In 2009, Thelwall re-defined 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", in an attempt to free webometrics from informetrics and distinguish the field from field-specific methods such as "purely mathematical analyses of online language" (Thelwall, 2009, p. 6). Malinský and Jelínek (2010) stressed that webometrics is purely a quantitative research area, but which may be enhanced by sentiment analysis and opinion mining.

5.1.2. Web mining

Chen and Chau (2004, p. 290) credited Etzioni (1996) for coining the term web mining as "the use of data mining techniques to automatically discover web documents and services, extract information from web resources, and uncover general patterns on the web". Web mining is the area of data mining dealing with the extraction of interesting knowledge from the web (Etzioni, 1996). The classification into content, structure, and usage has been used by several authors, for example Facca and Lanzi (2005), and Sharma, Shrivastava and Kumar (2011). Kosala and Blockeel (2000) identified four subtasks for each area – resource finding, information selection and preprocessing, generalization and analysis. A similar finding was made by Zhang and Segall (2008) who in their survey of the field identified the subtasks resource finding and retrieving, information selection and preprocessing, patterns analysis and recognition, validation and interpretation, and visualization.

Srivastava, Cooley, Deshpande and Tan (2000) describe usage mining in three phases; preprocessing, pattern discovery and pattern analysis. They identified four main types of data that can be mined on the web; content, structure, usage and user profiles. A key issue identified was privacy, which they found "further complicated by the global and self-regulatory nature of the web", and revolving "around the fact that most users want to maintain strict anonymity on the web" (Srivastava et al., 2000, p. 20).

Eirinaki and Vazirgiannis (2003) focused their review on web mining for personalization, a subtask of usage mining. They discerned between content-based filtering and collaborative filtering as the former is based on the users' preferences and the latter on their actions (ratings, navigations etc.). They concluded that web personalization is gaining momentum, both in the research area and in the business area, and also acknowledged privacy as an important issue. Da Costa Jr and Gong (2005) surveyed web structure mining. They attempted to clarify the confusion concerning web mining, referring to the four subtasks defined by Kosala and Blockeel. They argue that the three categories sometimes can be reduced to two, by including web structure mining in one of the others.

Facca and Lanzi (2005) identified four main applications for usage mining: 1) personalization of the delivery of web content, 2) improving user navigation through pre-fetching and caching, 3) improving web design by analyzing usage and giving recommendations and 4) to improve customer satisfaction in e-commerce. They also discussed privacy and concluded that, so far, no usage mining solutions or approaches had taken privacy into account so far. Lappas (2007) surveyed web mining research relating to areas of societal benefit. The literature on applications and methods used within e-services, e-learning, e-government, e-politics and e-democracy was reviewed. Lappas found that there was a growing interest in applications of web mining that are of social interest. According to Lappas, web mining may benefit organizations that seek to utilize the web for supporting decision making.

In 2011, web mining was redefined as data mining techniques applied on the web, where the data used in the mining process belongs to at least one of the categories structure or usage (Kumar & Gosul, 2011). However, the authors concluded that there is no agreed definition on web mining to date.

5.1.3. Common denominators

Both fields have developed link analysis/link structure mining as sub-fields. The seminal works by Kleinberg (1999) and Brin and Page (1998) are cited by authors in both fields and it is perhaps the case that the greatest use of link analysis has been in the search engine algorithms HITS and PageRank.

There are differences in the use of link analysis or link mining. Link structure mining refers to computer science link analysis and "computer scientists tend to use links as the basis of new algorithms or to improve the functioning of existing ones" (Thelwall & Wouters, 2005, p.189). In an effort to distinguish the use of links in the social sciences from the use of links in computer science, they found that linking is analyzed for descriptive web mapping, as social phenomenon and as indicators for social relations in the social sciences, and for mapping scholarly communication in information science, whereas in computer science, link analysis had been used for algorithms to build web navigation and information retrieval tools, as well as descriptive modeling.

From link analysis, the step to social network analysis is not far as both concern connections between actors. Otte and Rousseau (2002) proposed social network analysis as a strategy that can be applied within information science. Björneborn, for example, noted that the social network analysis concept betweenness centrality can be used to identify and track gate-keepers and interdisciplinary crossings in an academic web space, knowledge that can be used to identify possible areas for interdisciplinary exploration (2006). Several studies within webometrics and web mining have used social network analysis as we will see in sections 5.2.1 and 5.2.2.

There are a few examples of studies where methods from both fields are used. Jonkers, De Moya Anegon and Aguillo (2012) both used webometric and web mining methods when they attempted to use a measurement of e-research usage as an indicator of research activity. They focused on research activities in research organizations in UK, Germany and Spain, and combined log analysis with an analysis of URL citations to the data source Expert Protein Analysis System, whose server logs were the subject of the usage mining. Polanco, Ivana and Dominique (2006) proposed a system based on co-usage for the statistical analysis of science and technical information. They described their work as "informetrics from the point of view of computer-based technologies" (p 171), as they combined log analysis with co-usage analysis and information about users' information needs. A third example is provided by Martínez-Torres, Toral, Palacios and Barrero (2012), who used factor analysis combined with link analysis to extract main web site profiles in terms of their internal structures based on 64 social network and correlation indicators.

5.2. Review of research in both areas

In this section, papers within these areas are briefly summarized. Results or conclusions from the papers are not included other than directions for future research, or identified problems. A striking difference is that webometric research tends toward exploratory studies with the aim of describing and analyzing a phenomenon, whereas in web mining research the focus is rather on developing methods and algorithms to deal with web data.

Out of the 61 webometrics-classed items, 57 were deemed exploratory, of which nine combined methodological with exploratory work, and five combined evaluative with exploratory work. 51 studies focused on structure, with 17 of them combining structure with content (16) or both usage and content (1), and six focused on content solely.

Of the 57 web mining-classed items, 52 were methodological, of which two combined methodology with evaluation, and one was exploratory and methodological. 23 items focused on content only, 17 on usage only, and three on structure only. Structure was combined with content in nine of the items, and four items focused on the combination of content and usage while two focused on structure and usage.

Fundamental differences relating to tradition are indicated. Although there are substantial similarities relating to the object of study and methodological instruments, in the dimensions of exploratory/methodological and structure/usage they appear to be inverse to each other. These basic differences are further illuminated through a closer reading of the material.

5.2.1. Webometrics

Most webometric research pertains to link analysis. This heavy emphasis on one cluster of tools is also marked by a focus on theoretical and conceptual development. This is evident in the intense discussion of the *web impact factor* (WIF) (see for example Bar-Ilan, 2008). In an early article on the subject, the WIF was defined by Ingwersen (1998) as the number of in-links to the domain divided by the number of pages of the domain accessed by the crawler. This measure has been debated ever since, and there seem to be a lack of consensus on how to use the measure as well as how to interpret the results. Thelwall (2001b) found that the WIF gave a higher correlation with research ratings if staff number were included in the measure. After an investigation of Iranian universities using the WIF (Noruzi, 2005), the measure was analyzed and its advantages and disadvantages were discussed (Noruzi, 2006). Its development and applications were reviewed, and it was concluded that the measure was useful for quantitative intra-country comparisons, but has little value beyond this.

Webometrics seems to be quite homogeneous in choices of case studies. Given its background in bibliometrics it is perhaps not surprising that scholarly web activities have been the subject of a wealth of research studies. Link analysis has been used for studying academic web site or page relations (Thelwall, 2001a; 2001b; 2002b), academic relations with geographical aspects (Thelwall, 2002a), the web impact of scientists' personal pages (Barjak, Li & Thelwall, 2007), whether linking can be used as an indicator of collaboration between universities and government and industry (Stuart, Thelwall & Harries, 2007), the intra-country WIFs of universities (Nwagwu & Agarin, 2008; Aminpour et al., 2009; Asadi & Shekofteh, 2009; Shekofteh et al., 2010; Erfanmanesh & Didegah, 2011; Islam & Alam, 2011; Islam, 2011), political communication (Park, 2010), and the web impact of Islamic top universities (Didegah & Goltaji, 2010).

Link analysis can be characterized as a cluster of tools that seem to be under continuous development. One interesting variation of link analysis is the analysis of co-links, which builds on the assumption that two web pages or sites that have inbound links from the same source are related. Co-links are especially useful when there is little interlinking between pages (Thelwall, 2009, p. 41). Holmberg and Thelwall (2009) did a study of the interlinking of government web sites with geographical aspects, and Holmberg (2010) then carried out a similar study, but using co-links as metrics. Lang, Gouveia and Leta (2010) also focused on co-links in their investigation of the relationships in small networks, using the Oswaldo Cruz Foundation institutes in Brazil as the object of study, and later used link analysis to map collaboration among health-research institutions around the world (Lang, Gouveia & Leta, 2013).

Another development of link analysis has been value-adding by combining it with other tools or perspectives. As web pages and sites are nodes and they are linked together, the combination of link analysis with social network analysis is a logical development. An excellent example is provided by Björneborn (2006). His investigation focused on the type of links that could function as small-world bridging structures between sites adhering to different topics in an academic web. Other examples using this combination on academic web include a network analysis of links between Spanish university departments and research groups (Ortega & Aguillo, 2007), an investigation of relationships between Danish, Finnish, and Swedish web sites and their relationships with the European web (Ortega & Aguillo, 2008), a mapping of the web presence of the European Higher Education Area (Ortega et al., 2008), and a visualization of the most important universities in the world (Ortega & Aguillo, 2009). Also related to this is Martínez-Torres and Díaz-Fernández (2013) comparison of global link visibility of academic web sites with local visibility.

These kinds of studies can also be found on political subjects, for example the political communication in South Korea (Park & Thelwall, 2008) and the web presence of Spanish media actors as well as if link analysis of media and political parties can provide insight into political orientation (Romero-Frías & Vaughan, 2012). Nam, Lee and Park (2013) studied the interrelationship among web sites during the 2010 local elections in South Korea, by utilizing co-link analysis and name mentions on Twitter, blogs and in news articles. Another topic studied is research and development of different organizations. Martínez-Ruiz and Thelwall (2010) examined the correlation between web visibility (in terms of link data) and research and development expenditures and technologies of firms. A related

example is the research and development support infrastructure associated with science parks as investigated by Minguillo and Thelwall (2012), who used a link analysis based method combined with social network analysis and manual content analysis of out-links.

Given the central position of the link within the field, it is not surprising to find many examples that focus on link motivation and linking theory. Vaughan and Thelwall (2003) studied the factors that attract in-links to journal web sites using link count analysis, Bar-Ilan (2004) searched for a better understanding of why links were created, using qualitative assessments of eleven different characteristics of university pages linked to universities in Israel, and Thelwall (2006) proposed a theoretical framework to find underlying reasons for link creation. Thelwall concluded that no single method for link interpretation is perfect and that method triangulation is required, including a direct method (i.e. interviewing link creators, or categorizing a random sample of links) and a correlation testing method. Thelwall warned however that due to fundamental problems such as "the rich get richer" phenomenon and the dynamic nature of the web, research conclusions should always be expressed cautiously. He found interpretative exercises inappropriate and considers that a theory of linking is unrealistic as a research goal. Another example of the non-triviality of interpreting links is provided by Zuccala's (2006) comparison of co-citation analysis with co-link analysis. Zuccala concluded that co-links not only add a new dimension compared to single links, they also significantly differ from co-citations as the web constitutes a much broader context, and that there can be broader spectra among the reasons to link compared to the reasons to cite.

Link analysis has typically been carried out by using a major search engine as their indexes are constantly growing as new data is gathered from the web. The heavy emphasis on link analysis through search engines is problematic as the availability of the empirical material cannot be taken for granted. Some problems related to link analysis have been identified, for example search engines often develop functions and change data accessibility without notice (Thelwall, 2001b). Other issues reported are content management systems that hide content in databases, universities that have index pages which automatically link to all staff and student pages, and universities that ban search robots tend to be less covered (Thelwall, 2002b).

One way of dealing with the problem of commercial search engines is to use non-commercial crawlers specifically created for link analysis purposes. Such crawlers have been utilized in webometrics. For example, Thelwall (2001a) proposed a distributed approach for web crawler design. The system was employed for analyzing the link structure of university web sites after finding that a single crawler could not work quickly enough to cover the necessary number of web sites. Non-commercial crawlers have been used alone or together with commercial search engines, for example by Thelwall (2001b), Park (2010), Didegah and Goltaji (2010), Minguillo and Thelwall (2012), and Lang, Gouveia and Leta (2013).

A difficulty concerning the use and value of link analysis is the withdrawal of the linkdomain[3] command in Yahoo!. To address this problem Thelwall (2011) compared link counting and URL citation counting, where a URL citation is the URL mentioned in the text but not necessarily accompanied by a link. After 15 case studies using Yahoo!'s linkdomain command, Thelwall concluded that if the linkdomain command in Yahoo! was withdrawn (as it later was), the power of link analysis was likely to be undermined, except in the case of academic spaces. Thelwall and Sud (2011) studied the same topic as they compared the in-link count method with URL citation and the organization title mentions. The results showed that there was quite a strong correlation between the metrics and that they can be used interchangeably for impact measurements, as well as for significant correlations between two chosen offline measures, the US News & World Report and the Research Assessment Exercise (RAE) 2008 statistics.

Another approach has been tested by Vaughan and You (2010), who used co-word analysis to address the relatedness of organizations using the blog sphere as data source. Their results showed that co-word analysis could potentially be as useful as co-link analysis. A later study by Vaughan and Romero-Frías (2012) came to similar conclusions, as

[3] The linkdomain command was useful for finding all pages linking to any page belonging to a given web site.

they correlated in-link counts and the number of web pages a company was mentioned on with the company's business measures. Co-words have also been tested on Chinese business environment (Vaughan, Yang & Tang, 2012), where the method was found to have some potential. A related study was performed by Vaughan and Yang (2012), who attempted to determine which source was the better for estimate business and academic quality. The sources were Alexa's in-link data, Google URL citation, and Yahoo! in-link data. The authors found high correlations between the three sources, but also some limitations of Alexa, which does not reveal pages providing in-links, and does not offer co-link data. Another example of evaluation of data sources is Thelwall and Sud (2012) evaluation of Bing for webometric tasks. They found that the search engine and its application programming interface (API)[4] was practical but with important limitations. As other search engines have changed their functionality or withdrawn or changed their APIs, some alternatives were investigated. Of these, the Russian search engine Yandex was judged as being the most promising.

Even though link analysis and its variations have dominated webometrics, other approaches apart from co-word analyses have been utilized. Kretschmer and Aguillo (2005) introduced gender co-operation, web visibility rates, and gender centrality in networks as new indicators for gender studies, combining social network analysis, web citation analysis, co-author analysis, and bibliographic and web co-authorship networks of the 64 COLLNET members. Thelwall, Vann and Fairclough (2006) introduced *web issue analysis* as the analysis of the spread or diffusion of particular issues of concern, and used integrated water resource management as an exemplifying issue. Aguillo, Granadino, Ortega and Prieto (2006) tested webometric indicators for describing and ranking university activities. On the same note, Aguillo (2009) introduced activity, impact and usage as a new set of web indicators to provide a complementary system for the evaluation of scholarly activities of academic organizations, and Shunbo Yuan and Weina Hua (2011) investigated the scholarly impact of open access journals within library and information science, focusing on the correlation between citation counts, links, pages, WIFs and PageRank. Finally, the presence of image files, multimedia files and blog content in a Spanish academic context was studied by Orduña-Malea (2012).

It has also been found fruitful to apply webometric methods to user generated content. Within the field, there are numerous studies of blogs, as Thelwall (2010a) points out, for example a longitudinal analysis of trends in blog linkages and online networking among Assembly members in South Korea (Park & Kluver, 2009), a study on cross-lingual linking in the blogosphere, where cross-lingual links are seen as bridges in information exchange online (Hale, 2012), and an investigation of South Koreans' protests against US beef imports with references to the slashdot effect, homophily theory, the Balkanization thesis and agenda-setting and issue-rippling (Woo-Young & Park, 2012). The combination of social web content and academic related research have resulted in the birth of altmetrics, and in one paper within the topic, Thelwall, Haustein, Larivière and Sugimoto (2013) evaluated eleven different social media altmetric measurements by comparing them to citation data from Web of Science.

Lately, Twitter has received some webometric attention. Thelwall, Buckley and Paltoglou (2011) studied sentiment in Twitter events, trying to assess whether popular events are typically associated with increases in sentiment strength. Cho and Park (2012) studied Korean government organization through the Ministry for Food, Agriculture, Forestry and Fisheries timeline on Twitter, using semantic network analysis combined with follower data and complementary interviews for the background data behind the tweet. From a media agenda-setting perspective, Wilkinson and Thelwall (2012) studied trending Twitter topics from tweets posted from six different English speaking countries/regions over a nine month period. The goal was to assess how trending topics vary by country. Hsu and Park (2011, 2012) used multiple sources as they studied the online networks of South Korean National Assembly members using social network analysis, focusing on Twitter, homepages and blog networks. An example of webometric YouTube research is provided by Van Zoonen, Vis and Mihelj (2011) who did a network analysis of videos that are reactions to the controversial short film Fitna, combined with qualitative discourse analysis. They looked at reactions, number and types of interactions, and the contents of networks of comments, subscriptions and

[4] An API is an interface set up by a service provider, which can be used to query the provider for certain data.

friends. Another paper on YouTube focused on attitudes towards wildlife conservation, by studying comments and data associated with commentators of a viral video (Nekaris et al., 2013).

Some other social web applications have been studied using webometrics. Angus, Thelwall and Stuart (2008) combined webometric data collection with classification and informetric analysis as they investigated tagging on Flickr; the function of tags, and whether they follow a power law distribution as well as the extent to which users tag their uploaded pictures. The statistics and semantic structure of the categories in Wikipedia have been studied (Holloway, Bozicevic & Börner, 2007), as well as co-authorship in the Simple English version (Biuk-Aghai et al., 2009), and co-authorship and collaboration patterns using social network analysis (Laniado & Tasso, 2011).

One study looked at the policy relevancy of webometric indicators within other scientific fields (Thelwall et al., 2010). In case studies of transdisciplinary fields it was concluded that the indicators could find patterns and relationships among organizations and sectors that traditional metrics could not find, as well as finding patterns and relationships in wider networks and in the intensity of interactions between actors in a field. However, webometrics is not suited to large or mature fields, but rather to new fields as it was concluded that large fields are difficult to analyze with these techniques. Furthermore, it was concluded that webometrics can provide evidence about research collaboration and identifying roles and impact of intermediary organizations as well as finding areas where collaboration needs to be enhanced.

5.2.2. Web mining

The common sub-fields of web mining are content, structure and usage. I found quite a few studies using a combination of two sub-fields. Most of the research focuses on either usage or content, and in some cases both of them, and there are a number of studies on personalization and recommender systems. These studies typically build on usage data often found in usage logs, but there are also examples of hybrid approaches.

An early architecture for personalization was proposed by Mobasher, Cooley, and Srivastava (2000). It was named the WebPersonalizer System and made use of a batch process and an online process. The former is comprised of data preparation and usage mining, including clustering of sessions and page views as well as mining of association rules, and the latter is the recommendation engine, which matches the active session with sessions in the database to make recommendations. Canny (2002) used log analysis on the data-sets of EachMovie, Jester and Clickthru, where the purpose was to develop a new collaborative filtering method that takes consideration of privacy using probabilistic factor analysis. Miller, Konstan and Riedl (2004) took privacy into consideration as they presented the recommender system PocketLens using peer-to-peer architecture, a system designed to run on disconnected palmtop computers. Deshpande and Karypis (2004) focused on the scalability problem as they used a two-step model based on the similarities between the items, and Huang, Chen, and Zeng (2004) focused on sparsity, which occurs when there are a small number of transactions in relation to the number of objects in the recommender system. To solve the sparsity problem they used associative retrieval techniques and a related spreading algorithm.

Four shortcomings of memory-based methods were identified by Hofmann (2004). These were the possible suboptimal accuracy of recommendations, the lack of an explicit statistical model which means that no general insight is gained from the data, scalability, and difficulties in systematical adaptation for achieving objectives for specific tasks. To deal with these shortcomings, a model-based method relying on a statistical latent class model for personalization was developed. The difference between memory-based and model-based algorithms is that the former uses the entire user database to make predictions, whereas the latter uses the database to learn a model, which in turn is used for making predictions (Breese, Heckerman & Kadie 1999).

Methods and models for recommender systems have been presented in a number of studies; for query recommendation in search engines based on recent history of search logs and similarity of queries (Zhang & Nasraoui, 2008), for learning platforms based on usage logs and personal profiles (Romero et al., 2009), and for personalized e-commerce based on users' product specific knowledge (Chou et al., 2010). Other examples include an

algorithm for recommending TV-programs using content-based and item-based hybrid approach (Barragáns-Martinez et al., 2010), a clustering-based profile matching system for a dating site (Alsaleh et al., 2011), and a log based cross-language model for recommending academic articles (Lai & Zeng, 2013). Mobasher, Dai, Luo, and Nakagawa (2002) focused on real time recommendations of web pages, and page recommender systems were also presented by Shyu, Haruechaiyasak, and Chen (2006), and Guerbas and associates (2013). Another example of a solution for web pages was presented by Bayir, Toroslu, Demirbas and Cosar (2012). In an attempt to solve the session problem, their model included link constraints, which entails that whenever a page is accessed without a link from the last accessed page it is excluded from the session construction process.

These three examples lead us to another important area within usage mining – the mining of user profiles and navigation paths. Nasraoui, Rojas and Cardona (2006) studied the task of tracking emerging topics and clusters in noisy and evolving text data sets and in mining evolving user profiles from clickstream data in a single pass and under different trend sequencing scenarios, where a trend sequencing scenario corresponds to a specific way of ordering web sessions or text documents. Ou, Lee and Chen (2008) introduced a flexible model of mining with dynamic thresholds for analyzing paths. They applied the Markov chain model and used dynamic thresholds to reduce the number of unnecessary rules produced. Das and Turkoglu (2009) proposed log analysis and path analysis for improving web design and web site administration. Other examples include a framework for evolving user profile mining with the assumption that the patterns are dynamic (Nasraoui et al., 2008), a sliding window-based algorithm for mining traversal patterns by Li (2009), the induction-based decision rule model for generating inferences and implicit hidden behavioral aspects that was presented by Poongothai and Sathiyabama (2012), and Kundu's (2012) proposed hybrid approach for web traffic analysis based on pattern discovery and pattern analysis, where the most recently accessed data was prioritized and used alongside clustering. Finally, Arbelaitz and associates developed a method for generating semantically enriched usage profiles for link prediction, web design, and marketing purposes (Arbelaitz et al., 2013).

It is common in web mining to use "single-site, multi-user, server-side usage data […] as input" (Srivastava et al., 2000, p. 17), and we have seen examples of this already. One exception to the single site norm was presented by Pierrakos and Paliouras (2010) who proposed a framework for the personalization of web directories that corresponded to user navigation throughout the web, not just one site. The usage data came from access log files of Internet service providers' cache proxy servers. Paliouras (2012) focused on the role of user communities in user modeling and personalization. A new research opportunity was identified: the discovery of communities of active users, and showed how this relates to recent efforts on analyzing social networks and social media. Paliouras suggested OpenID for interconnecting popular networks and social media sites, but pointed out privacy, trust and security as important and pressing issues.

The combination of usage and content is fruitful when studying social media applications. Compared to webometrics, web mining has had a slightly stronger and different focus on user-generated content. Some studies have focused on Twitter data for different purposes; to identify breakpoints in public opinion using tweets (Akcora et al., 2010), sentiments in tweets (Bifet & Frank, 2010), to predict the future revenues of movies (Asur & Huberman, 2010), and to find out if Twitter can work as an alternative to real-world sensor of hay fever (Takahashi, Abe & Igata, 2011). Efron (2011) wrote a review on microblogs information retrieval, on how to study microblog streams, and on aspects relevant to relevance ranking when searching in microblog streams. Problems in microblog retrieval, such as entity search and sentiment analysis, were identified. Other social media related studies have focused on hate groups in the blogosphere (Chau & Xu, 2007) and on identifying bloggers with marketing impact (Li, Lai & Chen, 2009). An example of how blogosphere discussions are related to scientific literature was presented by Gruzd, Black, Le and Amos (2012). They used crawling, social network analysis and qualitative content analysis to investigate the relationship between blogosphere discussions and biomedical literature in PubMed on diabetes.

As in webometrics, sentiment analysis has been studied using web mining, but is often referred to as opinion mining and review mining, particularly in the context of the social web. Ku and Chen (2007) mined opinions in blogs and

news in Chinese, and Shandilya and Jain (2009) demonstrated how an opinion mining framework can be used for extracting the opinions of consumers or customers. Fernández et al. (2011) evaluated the validity of the EmotiBlog corpus for sentiment analysis tasks, as they had identified a lack of resources, methods and tools for dealing with subjective data, and found that the resource compared well with the JRC corpus. There is also an example of a combination of methods using structure on the social web. Cheong and Lee (2011) presented a framework for extracting civilian sentiment and response on Twitter during terrorism scenarios, using sentiment analysis, social network analysis and demographic exploration. Kontopoulos, Berberidis, Dergiades and Bassiliades (2013) addressed the inefficiency of text-based sentiment classifiers for Twitter data, and aimed to solve this problem by utilizing an ontology based approach. Instead of assigning a sentiment score to the individual tweet, a score was assigned to each distinct notion in the tweet.

Another example of user-generated content is the review. A number of web mining studies have focused on the different aspects of reviews. Somprasertsri and Lalitrojwong (2010) presented an approach for summarizing reviews which included a review crawler, the parsing and dependency analysis of reviews, product feature extraction and related opinion extraction, and a relation extraction based on product ontology. Other examples include Qiu, Liu, Bu and Chen (2011) who studied context-dependent sentiment analysis on a test review collection, Moghaddam and Ester's (2010) Opinion Digger, which is a method for extracting important information about a product and determining the overall customer's satisfaction with it, and Algur, Patil, Hiremath and Shivashankar's (2010) proposed model using conceptual level similarity for spam detecting in reviews. Brejla and Gilbert (2012) developed a system for the tourist industry using holiday reviews of cruises that uses content mining, natural language processing and qualitative content analysis.

It can be seen that web mining is used for different commercial purposes. A method for acquiring market intelligence has been proposed by Ai, Zhang, Zuo and Wang (2006) who used a crawler with Bayes and inductive classifiers for collecting and grouping information fragments from web sites. Yeh, Lien, Ting and Liu (2009) attempted to identify the potential customers of online bookstores, and Popova, John and Stockton (2009) mined news items and press releases for the purpose of gaining sales intelligence. In a study with a related purpose, Richardson and Domingos (2002) used structure mining to identify customer influence using data from Epinions. Another system for marketing purposes was presented by Wang, Ting and Wu (2013). The proposed system used consumer social networks around influential blogs, and association rules based on the blog and response content.

Content mining can also be used to distinguish between true and false facts in conflicting information. Panchal, Pillai and Singh (2012) designed a general framework for this problem and invented the algorithm TruthFinder which utilizes the relationships between web sites and their content. A web site was considered trustworthy if it provided many pieces of true information, and a piece of information was likely to be true if provided by many trustworthy sites. They worked from the assumption that a true fact is likely to appear to be same or similar on different sites whereas a false fact is less likely to be same or similar.

As mentioned above, semantic web mining seems to be a promising sub-field of web mining, and some examples above are related to various semantic methods. On this topic, Wang, Lu and Zhang (2007) introduced the method key information mining, where key information can be distinctive menu items and navigation indicators that help classify the main contents of a web page. A survey showed that "well designed" web pages use information blocks, menus and navigation indicators, and use similar word lengths for items in menus and navigation indicators. The key information mining method is comprised of two steps; extracting a list of candidate information and then applying entropy measures to discover key information. They also developed a prototype for using the extracted key information for the development of ontology. Velásquez, Dujovne and L'Huillier (2011) defined the problem of extracting web site key objects, and presented a method where relations between objects were extracted using a site ontology. The method also included comparison of the objects using edit distance, the approximated time spent viewing the objects, and user session clustering. A different usage based method for identifying key objects has been presented by Velásquez (2013). This method combined eye-tracking technology with log usage data.

A new method for dealing with semantic web was presented by Rettinger and associates (2012). Scalability, the distributed and heterogeneous nature of web data, and the complexity of constructing ontological knowledge bases for advanced reasoning were identified as problems related to extracting knowledge from the semantic web. To deal with this problem, they presented a set of new methods based on statistical inference on the standard representations of semantic web knowledge bases. It was argued that machine learning research has to offer a wide variety of methods applicable to different expressivity levels of semantic web knowledge bases, for example, advanced machine learning techniques can automate relevant tasks for the semantic web by complementing and integrating logical inference with inductive procedures that exploit regularities in the data.

Finally, another semantic web related method was suggested by Wang, Sanin and Szczerbicki's (2011, 2012). The method combines the use of decisional DNA with data mining tasks. They described decisional DNA as a knowledge representation that can deal with noisy and incomplete data as well as being able to learn from experience. In their experiments on IMDB data information about 250 movies were converted into XML with the required decisional DNA-based structure. The authors conclude that "[t]he advantage of this promising approach is that after each crawl we formally define a new experience-based piece of knowledge that can be stored, reused, and shared between users and multiple applications" and that "[t]his new experimental structure can extract information from web sites and convert it into knowledge that can be reused or shared among different systems" (Wang, Sanin & Szczerbicki, 2012, p. 141).

Most of the research within the field has focused on usage and content, but there are also examples of structure mining, which could be said to be web mining's equivalent to link analysis. Structure mining is frequently used with other methods. Borges and Levene (2006) compared two methods for ranking web pages in a site, the Markov chain-based Site Rank (adaptation of PageRank to the granularity of a web site) and Popularity Rank (based on the frequencies of user clicks on the outlinks in a page that are captured by the navigation sessions of users through the web site). Zhang and Xu (2009) combined structure and usage mining with a co-citation-based algorithm used together with probabilistic latent semantic analysis, and Lin, Chu and Chiu (2011) developed a system based on the HITS algorithm for the automatic generation of hierarchical sitemaps for web sites. Yang, Liu and Feng (2012) explored the notion of network communities and their properties using a stochastic model (the next state is determined only by its previous state). They proposed a general framework for characterizing, analyzing and mining network communities, based on social network analysis and Markov chains. Finally, Yang and Sun (2013) developed a method for the automatic exploration of the topical structure of a given academic subject, based on the HITS algorithm, semantic clustering, co-link analysis and social network analysis.

5.3. Summary

As can be seen, there are both similarities and differences between these two fields. The most obvious difference between them is that webometrics mainly concerns exploratory studies of phenomena on the web, whereas in web mining there is a heavy overweight of methodological and experimental studies. Another difference is the focus on structure in webometrics whereas in web mining, the focus has been on content and usage.

Looking at methods, link analysis and social network analysis have dominated webometrics so far. Webometrics have mainly been performed on academic situations, even though there are examples of studies of political communication and relations and other areas or problems. Thelwall (2010b) argued the need for more applied webometrics stating that it may be at "a vulnerable growing stage with too few clear applications to make a strong case for its future value and vitality", and suggested health-related research as an interesting area for webometrics.

Web mining seems to have been dominated by applications and methods for usage studies. A common purpose has been to create better recommender systems. There has also been a large interest in mining opinions, and user-generated content has been studied more often using web mining than webometrics. The commercial benefits are of interest, but there are examples of recommender systems for aiding navigation in settings other than e-commerce.

Webometrics has a data collecting problem which has been identified by researchers within the field. The data needed for webometric research has been provided by search engines but the withdrawn linkdomain command makes global link analyses difficult to perform. Both fields are to some extent reliant on data providers, as various APIs have been used for data collecting. However, these APIs are subject to change and data providers may alter the conditions for data access and use. Single site studies based on log analysis, on the other hand, do not have this problem, as any organization needing to use these methods can simply apply them to their own data.

In webometrics, another kind of a problem was identified by Van Zoonen, Vis and Mihelj (2011, p. 1298), who suggested that "cybermetric search and analytic instrument runs the risk of suggesting 'completeness'", but this completeness is "by definition temporal, and biased towards well-established and maintained websites".

From an information and social science perspective, some interesting challenges have been outlined. For example, as new media types evolve, it becomes more difficult to compare election cycles due to differences regarding data access from one cycle to the other (Park & Kluver, 2009). Even if the same data source can be used, it is likely that the data provider has altered access conditions to the data. Another challenge is the need for further research into political and religious conflict, online interaction and gender (Van Zoonen, Vis & Mihelj, 2011).

In web mining, trust and privacy are important issues. They have been discussed by Cooley, Mobasher and Srivastava (1997), Srivastava, Cooley, Deshpande and Tan (2000), Canny (2002), Eirinaki and Vazirgiannis (2003), Miller, Konstan and Riedl (2004), Facca and Lanzi (2005), and Paliouras (2012). Other topics of discussion include performance related aspects, such as efficiency (e.g. Poongothai & Sathiyabama, 2012; Das & Turkoglu, 2009; Algur et al., 2010; Wang, Sanin & Szczerbicki, 2012), effectiveness (e.g. Wang, Lu & Zhang, 2007; Zhang & Xu, 2009; Somprasertsri & Lalitrojwong, 2010; Pierrakos & Paliouras, 2010), scalability (e.g. Yang, Liu & Feng, 2012; Li, 2009; Rettinger et al., 2012; Richardson & Domingos, 2002; Deshpande & Karypis, 2004), sparsity (e.g. Huang, Chen & Zeng, 2004; Zhang & Nasraoui, 2008), and accuracy (e.g. Hofmann, 2004).

6. Keywords and cross-field citations

The same sets of articles have been used for both the keyword and citation analyses. This section starts with the most prominent keywords in both sets. These are the keywords that have been assigned to most items, but excluding the search terms that retrieved the sets ("webometric*" and "cybermetric*" for webometrics, and "web mining" and "web data mining" for web mining). The next sub-section maps the top 100 keywords based on co-occurrences. The two sets of keywords formed different networks. There were 1,245 and 9,750 unique keywords in the webometrics and the web mining sets respectively. Finally, the section ends with a cross-field citation analysis.

6.1. Most prominent keywords

The top 30 keywords with their number of mentions are presented in tables 1 (webometrics) and 2 (web mining). The link analysis domination in webometrics is visible here through the keywords "link analysis", "hyperlinks", and "inlinks". The connection to library and information science is visible through terms containing "information" and the keywords "libraries" and "library and information science". Also among the top keywords are "bibliometrics", "informetrics", "scientometrics", and "citation analysis", and some scholarly concepts.

Keyword	#	Keyword	#	Keyword	#
world wide web	44	societies and institutions	15	scientometrics	10
websites	43	hypertext systems	13	libraries	10
internet	41	research	13	web presence	10
search engines	40	social network analysis	12	web visibility	9
bibliometrics	30	visibility	12	informetrics	9
information science	21	information management	12	article	9
link analysis	20	education	11	library and information science	9

information retrieval	20	user interfaces	11	inlinks	8
information analysis	18	citation analysis	11	ranking	8
hyperlinks	16	South Korea	10	information systems	7

Table 1 The top 30 keywords for webometrics (259 papers)

The focus on algorithms and methods is visible in the top 30 keywords for web mining. Information retrieval and related keywords such as "natural language processing" and "text processing" are visible here, indicating a focus on efficiency and effectiveness of search engines. Related to this is the semantic web, which a couple of reviewed papers have dealt with. Web usage has also been a focus, as the review section indicated. The keyword "electronic commerce" indicates an interest in e-commerce.

Keyword	**#**	**Keyword**	**#**	**Keyword**	**#**
data mining	1210	web services	184	information systems	116
world wide web	938	learning systems	182	natural language processing systems	111
websites	522	electronic commerce	179	clustering algorithms	111
information retrieval	415	artificial intelligence	159	ontology	110
algorithms	376	web usage mining	158	information management	106
mining	333	web page	151	information technology	104
search engines	299	database systems	147	information extraction	102
user interfaces	260	semantic web	144	data structures	101
internet	225	mathematical models	133	text processing	100
semantics	208	classification (of information)	119	text mining	97

Table 2 The top 30 keywords for web mining (2,455 papers)

In web mining, a strong connection with data mining can be seen, but not at the higher level with computer science. Webometrics, on the other hand, seems to connect stronger to the higher level with (library and) information science, but not as strong with informetrics, bibliometrics, and scientometrics, even though these connections exist. There are a few keywords in both sets. These are "world wide web", "websites", "internet", "search engines", "information retrieval", "information management", "user interfaces", and "information systems".

6.2. Mapping keywords

The webometric keywords (fig. 1) shared 1,211 edges. The average degree (i.e. the average number of keywords a keyword co-occurs with) was 24.22. The keywords were divided into four communities by the community detection algorithm. These are 1) *metrics and visualization* (bottom left; 29 keywords), 2) *retrieval systems and links* (bottom right; 29), 3) *investigation of scholarly activity* (top left; 28), and 4) *semantics and links* (top right; 14). Overall, scholarly activity and the academics have been an important topic for webometrics. Looking at methods and data, link analysis and the link have had central positions within the field, as has search engines.

Excluding the general keywords "world wide web", "websites", and "internet" reveals that among the strongest associations, keywords appear such as "information retrieval", "search engines", "hyperlinks", "societies and institutions", "education", "research", "bibliometrics", "scientometrics", "informetrics", "information analysis", and "information science". Finally "web mining" is among the strongest connections, here together with "search engines". Some of these keywords indicate an interest in scholarly activities, and others indicate the inheritance from information science and the different metrics fields. The "search engine" keyword might on one hand indicate an interest in retrieval systems, as the inclusion of "information retrieval" suggest, but it has also to do with the use of search engines as data source, for example in link analysis.

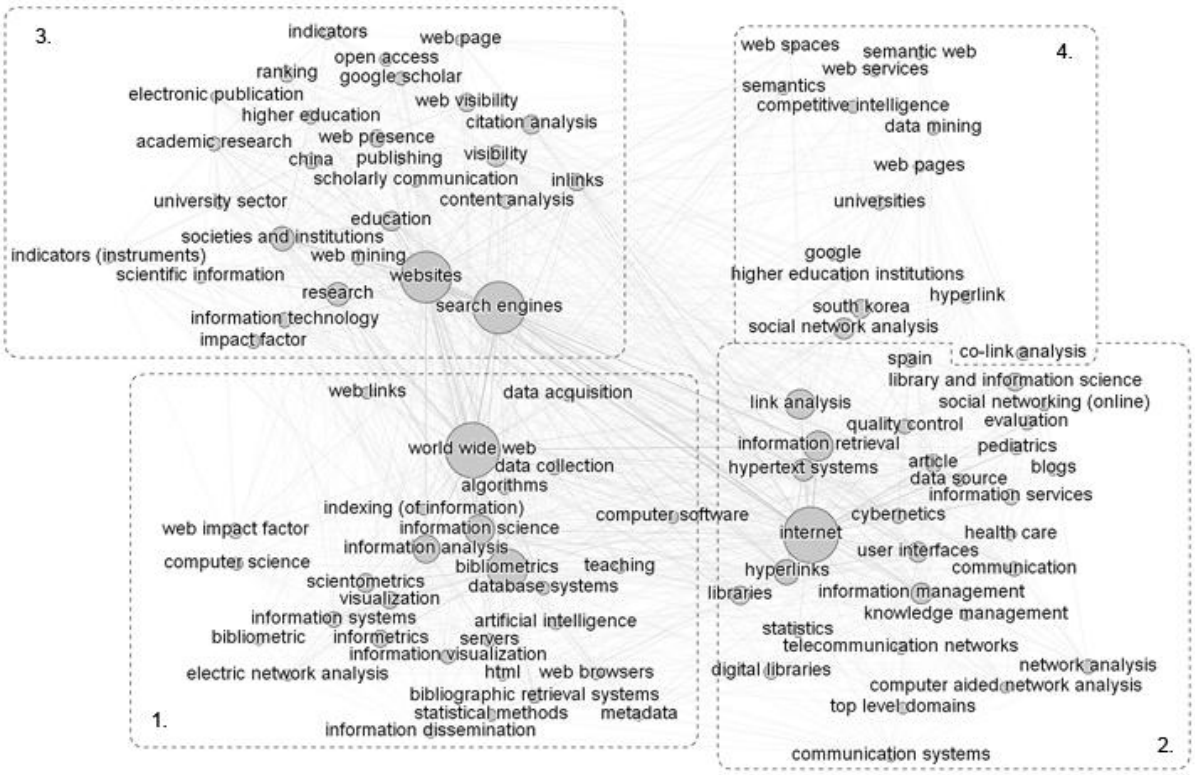

Fig. 1 Co-occurrences for webometric keywords

The web mining keywords (fig. 2) shared 3,812 edges, making the network much denser than the webometric ditto. The average degree was 76.24, which indicates that the average keyword was used with most of the remaining top 99 keywords. The keyword "data mining" was at first included in this network, but as it was connected to all the other keywords, the network became hard to interpret due to the number of connections. Excluding "data mining" resulted in a division into four communities. These are 1) *retrieval systems and machine learning* (top right; 47 keywords), 2) *electronic commerce and personalization* (top left; 33), 3) *information management and decision support* (bottom center; 11), and 4) *social and semantic web* (bottom right; 9). The largest community, number 1, seems to cluster around retrieval systems and machine learning, even though other topics are included here. This is the only community including keywords from other disciplines ("information science" and "bioinformatics").

When filtering out the general keywords "world wide web", "websites", "mining", and "internet", the keywords "information retrieval", "algorithms", and semantic web related concepts such as "natural language processing systems", "computational linguistics", "semantics" and "semantic web" are all involved in the strongest connections. This indicates a core interest in retrieval systems and search. Three of four communities are related to search in different ways; in community 1 and 4, the aim is to create more efficient and effective retrieval systems, with number 4 having a focus on semantics; and in community 2 for personalization purposes. The connectedness of the network could indicate a quite narrow focus, or be an effect of the larger activity within web mining. It could also be a result of the algorithmic focus within the field, and that most of the techniques can be applied for different types of problems. Even when filtering out the aforementioned general words, the average degree is still high (71.4). Web mining research spans over a wide area, ranging from recommender systems for searching and commercial purposes and information management for decision support, to linguistics and machine learning for various purposes.

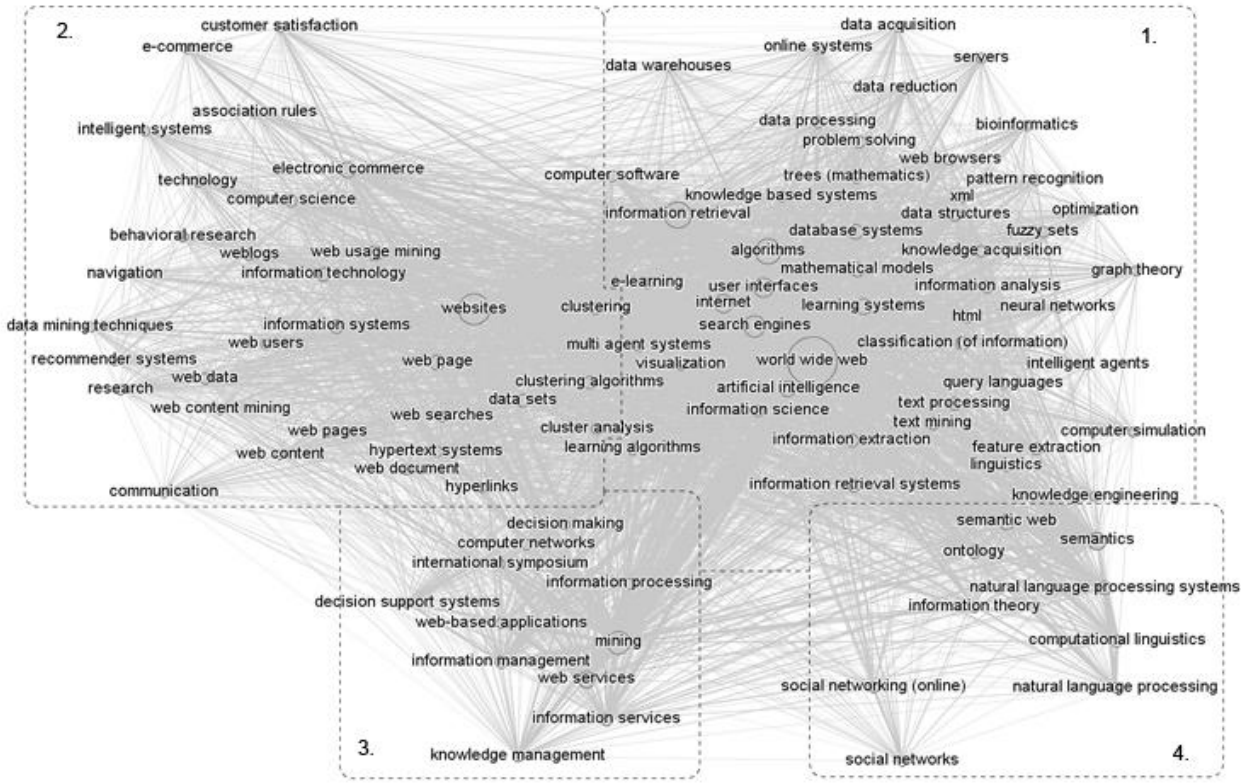

Fig. 2 Co-occurrences for web mining keywords

Taking both the lists of frequent keywords and the co-keywords maps into account, it seems that webometrics is more convergent with a few core topics. Most of the research seems to focus on the scholarly activity and the academics. Method-wise, the focus has been on link analysis. Both fields have traces of other disciplines. Most notably are bioinformatics in web mining, and in webometrics, pediatrics and health care. There are also some computer science related concepts within the webometric network. These are “algorithms”, “computer science”, and “artificial intelligence” in community 1, “web mining” and “data mining” in communities 3 and 4.

There are differences between the networks of the fields, for example the inclusion of commercial concepts in the web mining network which are not there in the webometrics network. Web mining also include more methodological keywords, especially those related to algorithms. One similarity is the core interest in retrieval systems. In both networks, “information retrieval” is among the most frequent combinations. This is perhaps not surprising, given that information retrieval is a sub-field of both information science and computer science. There is also evidence of semantic web research in both fields, but whereas the review section does not reveal webometric semantic web research other than it has been mentioned as a potential area for future research (e.g. Thelwall, Vaughan & Björneborn, 2005; Malínsky & Jelínek, 2010), the web mining field has had a stronger interest in the semantic web. There are only a few, not frequently used, social web related keywords in these networks, which suggest that social media research is a topic in the peripheries, when taking the whole histories of the fields into account.

6.3. Citation analysis

The results of the cross-field citation analysis revealed few signs of collaboration. In the web mining set of 2,518 items, only 55 references to the webometrics set were found, and in the webometrics set, 75 references to the web mining set were found (table 3). 0.13% of the web mining references pointed to the webometrics set which can be compared to the 0.8% pointing the other way around. However, of the references from webometric papers to web mining papers, 49 were to papers written by L. Vaughan, L. Björneborn, and M. Thelwall, authors considered being

webometricians who have written web mining papers. In total, 56 authors were found in both sets. Apart from the three webometricians mentioned, these include I. Aguillo, L. Björneborn, C.-C. Hsu, J.-L., Ortega, H.-W. Park, R. Baeza-Yates, and J. You.

With regard to webometric authors cited in web mining papers, M. Thelwall was found to be the far most cited author with 91citations followed by L. Vaughan with 67. R. Baeza-Yates had 29 citations in webometric papers, and was the only author with more than ten citations. Table 3 shows that web mining papers have cited 226 webometric authors while webometric authors cite only 57 web mining papers. It is difficult to compare these figures as this analysis takes all authors into account, and most papers have more than one author. In this set, webometric papers were more often authored by two authors (37%) followed by one author (24%), whereas in web mining two authors were most common (32%) followed by three authors (29%). All in all, web mining papers had an average of 2.86 authors and the corresponding average for webometric papers was 2.42.

Class		**Webometrics (n = 307)**	**Web mining (n = 2,518)**
No. of references		9,326	42,236
Avg. of references per item		30.4	16.8
References to prominent author from the other class		57	226
References to item from the other class	Total	75	55
	% of references	0.8%	0.13%
	References per paper	0.24	0.02
References to item from own class	Total	534	1,604
	% of references	5.7%	3.8%
	References per paper	1.74	0.64

Table 3 Citation analysis

Different disciplines have different standards and traditions regarding citing patterns and citing frequencies, which makes comparison between them difficult (e.g. Schubert & Braun, 1996; Ball, Mittermaier & Tunger, 2009). As there are large differences between these two fields regarding number of papers published, the distribution of number of papers per author, and the average number of references per paper, this analysis is only an indication of collaboration. For example, the 20th author in the webometric set had authored or co-authored three papers whereas the first author with three papers in the web mining set was found at the 307th place. Also note that there are an unknown number of cross-field references to items that were not retrieved in the Scopus database.

7. Discussion

The areas of webometrics that were identified by Björneborn and Ingwersen (2004) do match quite well with the three main areas of web mining. Similarities between both areas include the use of social network analysis as well as link analysis methods. Web mining, however, seems to focus on one site investigation whereas webometric research often tries to picture a part of the global web, for example, web mining is used for determining which site is the authority in a given context or how large the influence of a given site on the web is. Web mining applications have also been more focused on e-commerce where webometric research has tended to follow in the footsteps of bibliometrics, with a focus on scholarly activities and communication.

Looking at these two fields from the taxonomy of Becher and Trowler (2001, p. 184), how can they be categorized? Web mining appears to be more experimental and instrumental, and thus harder and more applied, but examples of applied knowledge are also seen within webometrics. Web mining is more focused on products and techniques, with webometrics more interested in discovery and interpretation. Webometrics is probably more hard than soft with its quantitative approach.

There are more researchers involved, and the pace of the research is higher in the web mining field, which indicates that it is more urban than webometrics. The people-to-problem ratio is however difficult to assess here. It would be

overly simplifying to assume that just because the fields are devoted to the web, they are also dealing with the same number of problems. When assessing the problems dealt with within a field, difficulties arise regarding distinguishing problems from each other, and a very detailed classification scheme would have been needed to identify and categorize the problems. The one indicator we have here is the number of authors per paper. There are 474 authors of 307 metrics papers, and 4,497 authors of 2,518 mining papers which gives one indication of web mining being slightly more urban as one thing that characterizes urban areas is that collaboration is more common, which entails co-authoring of papers. But urban researchers "tend to occupy a narrow area of intellectual territory and to cluster round a limited number of discrete topics that appear amenable to short-term solutions" (Becher & Trowler, 2001, p. 185), and that statement does not apply to the field of web mining, but perhaps to a more urban core of web mining, which is focused on search and retrieval, and the semantic web.

Web mining has been more focused on content and usage, but also used structure mining as part of algorithms. As both fields are evolving as specialized sub-fields within their well established mother fields, it is tempting to trace how they try to shape identities of their own. Both the review and the keyword analysis suggest that web mining is more divergent, probably as a consequence of the many papers written and researchers involved within the field. The number of unique keywords used within the fields should also indicate a greater diversity within web mining, even though this also is an effect of the number of papers. This might suggest that there is a clearer conceptualization, or an agreement of vocabulary, in webometrics, but a closer analysis of the keywords and their relations is needed to determine this.

In web mining, applications have been proposed for marketing and e-commerce, dating, cruises, identification of hate groups, response to terrorism, improving design and structure of web sites, medicine, movies, and news. Webometrics seemed to have difficulties moving beyond the link, but during the last couple of years, diversity has been increasing. The main topics found in the literature review were related to academics and politics, but other topics have been highlighted as well, for example, image tagging, discussions and networks related to YouTube videos, Wikipedia studies, news, and geography in relation to links. Webometric researchers show an increasing interest in conducting other types of studies, and social media research is gaining momentum. The studies focused on user generated data seem to be slightly more frequent in web mining, but webometrics have embraced the possibilities of the social web as well. Of the reviewed items, 19 webometric and 23 web mining papers dealt with the social web and user-provided data in some way. There should be a larger interest in web mining given that data mining has had some clear applications for business intelligence and marketing, and for these ends, social web data should be very interesting, especially reviews which have been used much more within web mining than webometrics.

Several studies have identified a need for webometrics to combine with other areas, or with qualitative research. Some areas of these two fields intersect and, although signs of collaboration between the areas were rarely found, Aguillo (2009) proposes using webometrics and web mining in combination, and Malinský and Jelínek's (2010) suggestion using mathematical and statistical methods and linguistic analysis in webometrics seems to be in line with this proposal. The citation analysis revealed that these two fields do not collaborate citation-wise. The results can be compared with Pratt, Hauser and Sugimoto's (2012) findings where one to three references per paper was to another discipline during the later investigated period, which is substantially larger figures than between webometrics and web mining. However, their study focused on specializations within the same main discipline (business). Perhaps Kirby, Hoadley and Carr-Chellman's study (2005) provide a better comparison, where 0.4% and 0.5% of the references from one field pointed to the other. Biehl, Kim and Wade (2006) also found lack of cross-field citations in their study, with the exception of human resources/organizational behavior and strategic management. On a broader level signs of collaboration were found by Fischer, Tobi and Rontelrap (2011), but their study focused on collaboration between natural and social sciences.

Considering Small's (2010) findings, it is perhaps not surprising to find that these two specialized sub-fields do not cite each-other frequently. As web mining has its focus on method and algorithm development, it was expected that

the webometric community would import web mining techniques, and thus cite web mining literature to a higher extent than vice versa. There is a significant larger amount of references from webometrics to web mining than the other way around, but a large portion of these were to webometricians. Even though there are issues with this citation analysis as mentioned in section 6.3, it does indicate that there is a lack of collaboration between these fields. It may be the case that both fields are holding on to research traditions inherited from their mother fields.

8. Conclusions

This article has compared the fields of webometrics and web mining, and investigated signs of collaboration between them. Roots have been outlined, as have similarities and differences. Webometrics evolved from informetrics and bibliometrics and web mining from data mining, so they both inherit different sets of methods and traditions. The latest definition of webometric taken up in this review re-focuses the field towards the social sciences (Thelwall, 2009 p. 6), and distances the field from the web mining approaches highlighted here. This can be contrasted with the definition of web mining as the application of data mining techniques on web structure or usage data (Kumar & Gosul, 2011).

I have shown examples of research of all of the data type categories content, structure and usage in both fields, with a webometric focus on structure and content, and a web mining focus on usage and content. A lack of collaboration was found in the cross-field citation analysis. 0.8% of webometric references were to web mining papers, and 0.13% of web mining references were to webometric papers. In relation to the taxonomy by Becher and Trowler, this review found both fields more oriented towards the hard and applied aspects of knowledge, but with webometrics purer in general, and in some areas softer than web mining. Web mining was found to be more divergent and more urban, both as a probable consequence of the larger body of papers, and the larger number of researchers involved.

One large difference between the two fields is that exploratory studies dominate within webometrics, whereas in web mining, methodological studies are most frequent. Common denominators for both fields is the use of link analysis/structure mining, sentiment analysis/opinion mining and the use of user generated data, but even when they use the same kind of data or data sources, for example, in studies of Twitter data, their approaches and foci are different. Topic-wise, there are more differences than similarities. There is an interest in e-commerce and marketing within web mining which is not there in webometrics. Webometrics is dominated by studies of the academia and politics, which are topics in the peripheries of web mining. Both fields have evaluated search and retrieval systems, but here too from different viewpoints. While web mining is more inclined to the efficiency and effectiveness of the systems, webometrics has evaluated their suitability for webometric purposes.

In the context of big data and the ever-changing web, it is interesting to consider what questions can be asked, what data can be combined and in what ways, and how we can formulate sustainable research problems that are not limited to current access to data. Within web mining, much focus has been on developing algorithms and methods for dealing with large quantities of unstructured, semi-structured or structured data. Within webometrics, the focus has been on exploratory studies. Webometrics can be said to have been influenced by the social sciences in this respect, which is in line with Thelwall's 2009 definition. And in the light of the social sciences, it can be concluded that big data needs the problem statements, the questions, and the methods from social science and information science, but also the algorithms from computer science. Access to data in the near future is more and more likely to be given through APIs and according to conditions set by data providers. This means that quantitative analyses of the web are limited to those with access to data, and access to data is limited to those with programming skills. Programming skills do exist within the information science community, but it is likely that the percentage of programmers within computer science is much greater. As programming skills are needed to take advantage of the web's vast data resources, webometricians are likely to need to collaborate with web miners if they do not already possess the skills needed. Hence, web mining will, in all likelihood, play an important part in future webometrics.

This review has shown that these two fields are very different in their research approaches, which could be an interesting ground for collaboration. For example, the potential need for web mining techniques in webometrics, and

conversely, a need for the webometric theoretical base in web mining, as webometrics as a field is less instrumental and more pure. The differences in approach can also explain the lack of collaboration, a conclusion also drawn by Glass, Ramesh and Vessey (2004). Another cause for the lack of collaboration is that these fields are subfields of disciplines that traditionally have found cross-discipline collaboration difficult, even though they share information retrieval as a joint sub-field.

As this review is the first, or one of the first, of its kind, there are limitations that need to be considered. A more exhaustive literature search could have been performed, and it would be interesting to conduct year-by-year co-keyword analyses to describe the topical development within the fields, however, the webometric body of research is probably too small for this. The citation analysis could be expanded by also including references to the other field's mother-discipline, and Web of Science data could also be used with or instead of Scopus data. Another interesting data source for studying collaboration would be the web.

References

Aguillo, I. (2009). Measuring the institution's footprint in the web. *Library Hi Tech*, Vol. 27 (4), 540-556.

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*, Vol. 57 (10), 1296-1302.

Ai, D., Zhang, Y., Zuo, H., & Wang, Q. (2006). Web Content Mining for Market Intelligence Acquiring from B2C Websites. In L. Feng et al. (Eds.), *WISE 2006 Workshops, LNCS 4256*, 159-170.

Akcora, C.G., Bayir, M.A., Demirbas, M. & Ferhatosmanoglu, H. (2010). Identifying breakpoints in public opinion. *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, 62-66.

Algur, S.P., Patil, A.P., Hiremath, P.S. & Shivashankar, S. (2010). Conceptual level similarity measure based review spam detection. *Proceedings of the 2010 International Conference on Signal and Image Processing, ICSIP 2010*, 416-423.

Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: methodological approaches to 'Webometrics'. *Journal of documentation*, Vol. 53 (4), 404-426.

Alsaleh, S., Nayak, R., Xu, Y. & Chen, L. (2011). Improving matching process in social network using implicit and explicit user information. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6612, 313-320.

Aminpour, F., Kabiri, P., Otroj, Z. & Keshtkar, A.A. (2009). Webometric analysis of Iranian universities of medical sciences. *Scientometrics*, Vol. 80 (1), 253-264.

Angus, E., Thelwall, M., & Stuart, D. (2008). General patterns of tag usage among university groups in Flickr. *Online Information Review*, Vol. 32 (1), 89-101.

Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J. M., & Perona, I. (2013). Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. *Expert Systems with Applications*, 40, 7478–7491.

Asadi, M. & Shekofteh, M. (2009). The relationship between the research activity of Iranian medical universities and their web impact factor. *Electronic Library*, Vol. 27 (6), 1026-1043.

Asur, S. & Huberman, B.A. (2010). Predicting the future with social media. *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, Vol. 1, 492-499.

Ball, R., Mittermaier, B., & Tunger, D. (2009). Creation of journal-based publication profiles of scientific institutions – A methodology for the interdisciplinary comparison of scientific research based on the J-factor. *Scientometrics*, Vol. 81 (2), 381–392.

Bar-Ilan, J. (2004). A Microscopic Link Analysis Of Academic Institutions Within A Country - The Case Of Israel. *Scientometrics*, 59 (3), 391-403

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century – A review. *Journal of Informetrics*, Vol. 2, 1-52.

Barjak, F., Li, X. & Thelwall, M. (2007). Which factors explain the Web impact of scientists' personal homepages? *Journal of the American Society for Information Science and Technology*, Vol. 58 (2), 200-211.

Barragáns-Martínez, A.B., Costa-Montenegro, E., Burguillo, J.C., Rey-López, M., Mikic-Fonte, F.A. & Peleteiro, A. (2010). A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. *Information Sciences*, Vol. 180 (22), 4290-4311.

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Bayir, M.A., Toroslu, I.H., Demirbas, M. & Cosar, A. (2012). Discovering better navigation sequences for the session construction problem. *Data and Knowledge Engineering*, Vol. 73, 58-72.

Becher, T. & Trowler, P.R. (2001). *Academic tribes and territories: intellectual enquiry and the culture of disciplines*. (2. ed.) Philadelphia, Pa.: Open University Press.

Biehl, M., Kim, H., & Wade, M. (2006). Relationships among the academic business disciplines: a multi-method citation analysis. *Omega*, 34 (4), 359-371.

Bifet, A. & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6332, 1-15.

Biuk-Aghai, R.P., Tang, L.V.-S., Fong, S. & Si, Y.-W. (2009). Wikis as digital ecosystems: An analysis based on authorship. *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, DEST '09*, 581-586.

Björneborn, L. (2006). 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space. *Scientometrics*, Vol. 68 (3), 395-414.

Björneborn, L. & Ingwersen, P. (2001). Perspectives of Webometrics. *Scientometrics*, Vol. 50(1), 65-82.

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

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, 1-12

Breese, J.S, Heckerman, D, & Kadie, C. (1999). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 43–52.

Brejla, P. & Gilbert, D. (2012). An Exploratory Use of Web Content Analysis to Understand Cruise Tourism Services. *International Journal of Tourism Research*.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, Vol. 30 (1-7), 107-117.

Borges, J. & Levene, M. (2006). Ranking pages by topology and popularity within web sites. *World Wide Web – Internet and Web Information Systems*, Vol. 9 (3), 301-316.

Canny, J. (2002). Collaborative filtering with privacy via factor analysis. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 238-245.

Chau, M. & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, Vol. 65 (1), 57-70.

Chen, H. & Chau, M. (2004). Web Mining: Machine Learning for Web Applications. *Annual Review of Information Science and Technology*, Vol. 38, 289-329+xvii-xviii.

Cheong, M. & Lee, V.C.S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, Vol. 13 (1), 45-59.

Cho, S.E. & Park, H.W. (2012). Government organizations' innovative use of the Internet: The case of the Twitter activity of South Korea's Ministry for Food, Agriculture, Forestry and Fisheries. *Scientometrics*, Vol. 90 (1), 9-23.

Chou, P.-H., Li, P.-H., Chen, K.-K. & Wu, M.-J. (2010). Integrating web mining and neural network for personalized e-commerce automatic service. *Expert Systems with Applications*, Vol. 37 (4), 2898-2910.

Cooley, R., Mobasher, B. & Srivastava, J. (1997). Web mining: Information and pattern discovery on the world wide web. In *International Conference on Tools with Artificial Intelligence*, 558-567.

Da Costa Jr, M.G. & Gong, Z. (2005). Web structure mining: An introduction. *ICIA 2005 - Proceedings of 2005 International Conference on Information Acquisition*, Vol. 2005, 590-595.

Das, R. & Turkoglu, I. (2009). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Systems with Applications*, Vol. 36 (3). 6635-6644.

Deshpande, M. & Karypis, G. (2004). Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, Vol. 22(1). 143-177.

Didegah, F. & Goltaji, M. (2010). Link analysis and impact of top universities of Islamic world on the world wide web. *Library Hi Tech News*, Vol. 27 (8). 12-16.

Duane Ireland, R. & Webb, J. W. (2007). A Cross-Disciplinary Exploration of Entrepreneurship Research. *Journal of Management*. Vol 33, (6), 891-927.

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, Vol. 62 (6). 996-1008.

Eirinaki, M. & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology*, Vol. 3(1), 1-27.

Erfanmanesh, M. & Didegah, F. (2011). Visibility and impact of Iranian research institutions on the web. *Library Hi Tech News*, Vol. 28 (1), 4-9.

Etzioni, O. (1996). The world-wide Web: quagmire or gold mine? *Communications of the ACM* 39 (11), 65–68.

Facca, F.M. & Lanzi, P.L. (2005). Mining interesting knowledge from weblogs: A survey. *Data and Knowledge Engineering*, Vol. 53(3), 225-241.

Fernández, J., Boldrini, E., Gómez, J.M. & Martínez-Barco, P. (2011). Evaluating EmotiBlog robustness for sentiment analysis tasks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6716, 290-294.

Fischer, A. R. H., Tobi, H., & Ronteltap, A. (2011). When Natural met Social: A Review of Collaboration between the Natural and Social Sciences. *Interdisciplinary Science Reviews*, Vol. 36 No. 4, pp 341-358.

Glass, R. L., Ramesh, V., & Vessey, I. (2004). An analysis of research in computing disciplines. *Communications of the ACM*, Vol. 47 (6), 89-94.

Gruzd, A., Black, F.A., Le, T.N.Y. & Amos, K. (2012). Investigating biomedical research literature in the blogosphere: A case study of diabetes and glycated hemoglobin (HbA1c). *Journal of the Medical Library Association*, Vol. 100 (1), 34-42.

Guerbas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., Ridley, M., & Alhajj, R. (2013). Effective web log mining and online navigational pattern prediction. *Knowledge-Based Systems*, 49, 50–62.

Hale, S.A. (2012). Net increase? Cross-lingual linking in the blogosphere. *Journal of Computer-Mediated Communication*, Vol. 17 (2), 135-151.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48 (1), 133-159.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, Vol. 22(1), 89-115.

Holloway, T., Bozicevic, M. & Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, Vol. 12 (3), 30-40.

Holmberg, K. (2010). Co-inlinking to a municipal Web space: A webometric and content analysis. *Scientometrics*, Vol. 83 (3), 851-862.

Holmberg, K. & Thelwall, M. (2009). Local government web sites in Finland: A geographic and webometric analysis. *Scientometrics*, Vol. 79 (1), 157-169.

Hsu, C.-L. & Park, W.H. (2011). Sociology of hyperlink networks of web 1.0, web 2.0, and twitter: A case study of South Korea. *Social Science Computer Review*, Vol. 29 (3), 354-368.

Hsu, C.-L. & Park, H.W. (2012). Mapping online social networks of Korean politicians. *Government Information Quarterly*, Vol. 29 (2), 169-181.

Huang, Z., Chen, H. & Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems*, Vol. 22(1), 116-142.

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, Vol. 54 (2), 236–243.

Islam, M. A. (2011). Webometrics study of universities in Bangladesh. *Annals of Library and Information Studies*, Vol. 58 (4), 307-318.

Islam, M. A. & Alam, M.S. (2011). Webometric study of private universities in Bangladesh. *Malaysian Journal of Library and Information Science*, Vol. 16 (2), 115-126.

Jonkers, K., De Moya Anegon, F. & Aguillo, I.-F. (2012). Measuring the usage of e-research infrastructure as an indicator of research activity. *Journal of the American Society for Information Science and Technology*, Vol. 63 (7), 1374-1382.

Kajikawa, Y. & Mori, J. (2009). Interdisciplinary Research Detection by Citation Indicators. *International Conference on Industrial Engineering and Engineering Management 2009 (IEEM2009) in Hong Kong*. (December 8-11, 2009).

Kirby, J. A., Hoadley, C. M., & Carr-Chellman, A. A. (2005). Instructional Systems Design and the Learning Sciences: A Citation Analysis. *ETR&D-Educational Technology Research and Development*, 53 (1), 37-48.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, Vol. 46(5), 604-632.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40, 4065–4074.

Kosala, R. & Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD Explorations*, 2 (11), 1-15.

Kretschmer, H. & Aguillo, I. F. (2005). New indicators for gender studies in Web networks. *Information Processing & Management*, Vol. 41 (6), 1481-1494.

Ku, L.-W. & Chen, H.-H. (2007). Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, Vol. 58 (12), 1838-1850.

Kumar, G. D. & Gosul, M. (2011). Web mining research and future directions. *Communications in Computer and Information Science*, Vol. 196, 489-496.

Kundu, S. (2012). An Intelligent Approach of Web Data Mining. *International Journal on Computer Science and Engineering*. Vol. 4 (5), 919-928.

Lai, Y. & Zeng, J. (2013). A cross-language personalized recommendation model in digital libraries. *The Electronic Library*, Vol. 31 (3), 264-277.

Lambiotte, R., Delvenne, J.-C. & Barahona, M. (2009). Laplacian dynamics and multiscale modular structure in networks. *arXiv*. http://arxiv.org/abs/0812.1770. Accessed 10 October 2013

Lang, P. B., Gouveia, F. C., & Leta, J. (2010). Site co-link analysis applied to small networks: a new methodological approach. *Scientometrics*, Vol. 83 (1), 157-166.

Lang, P. B., Gouveia, F. C., & Leta, J. (2013). Cooperation in Health: Mapping Collaborative Networks on the Web. *PLoS ONE*, Vol. 8 (8).

Laniado, D. & Tasso, R. (2011). Co-authorship 2.0 -Patterns of collaboration in Wikipedia. *HT 2011 - Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, 201-210.

Lappas, G. (2007). An overview of web mining in societal benefit areas. *Online Information Review* 32 (2), 179-195.

Li, H.-F. (2009). A sliding window method for finding top-k path traversal patterns over streaming Web click-sequences. *Expert Systems with Applications*, Vol. 36 (3), 4382-4386

Li, Y.-M., Lai, C.-Y. & Chen, C.-W. (2009). Identifying bloggers with marketing influence in the blogosphere. *ACM International Conference Proceeding Series*, 335-340.

Lin, S.-H., Chu, K.-P. & Chiu, C.-M. (2011). Automatic sitemaps generation: Exploring website structures using block extraction and hyperlink analysis. *Expert Systems with Applications*, Vol. 38 (4), 3944-3958.

Malinský, R. & Jelínek, I. (2010). Improvements of Webometrics by using sentiment analysis for better accessibility of the web. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6385, 581-586.

Martínez-Ruiz, A. & Thelwall, M. (2010). The Importance of Technology and R&D Expenditures in the Visibility of the Firms on the Web: An Exploratory Study. *Cybermetrics International Journal of Scientometrics, Informetrics and Bibliometrics*, 14 (1), 2.

Martínez-Torres, M.R., Toral, S.L., Palacios, B. & Barrero, F. (2012). An evolutionary factor analysis computation for mining website structures. *Expert Systems with Applications*, Vol. 39 (14), 11623-11633.

Martínez-Torres, M. R. & Díaz-Fernández, M. C. (2013). A study of global and local visibility as web indicators of research production. *Research Evaluation*, 22, 157–168.

Milgram, S. (1967). The small-world problem. *Psychology Today*, 1 (1), 60-67.

Miller, B.N., Konstan, J.A. & Riedl, J. (2004). PocketLens: Toward a personal recommender system. *ACM Transactions on Information Systems*, Vol. 22 (3), 437-476.

Minguillo, D. & Thelwall, M. (2012). Mapping the network structure of science parks: An exploratory study of cross-sectoral interactions reflected on the web. *Aslib Proceedings: New Information Perspectives*, Vol. 64 (4), 332-357.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic Personalization Based On Web Usage Mining. *Communications Of The ACM*, 43 (8), 142-151

Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2002). Discovery and evaluation of aggregate usage profiles for Web personalization. *Data Mining and Knowledge Discovery*, Vol. 6(1), 61-82.

Moghaddam, S. & Ester, M. (2010). Opinion digger: An unsupervised opinion miner from unstructured product reviews. *International Conference on Information and Knowledge Management, Proceedings*, 1825-1828.

Nam, Y., Lee, Y.-O., & Park, H. W. (2013). Can web ecology provide a clearer understanding of people's information behavior during election campaigns? *Social Science Information*, 52 (1), 91-109.

Nasraoui, O., Rojas, C. & Cardona, C. (2006). A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation. *Computer Networks*, Vol. 50 (10, SI), 1488-1512.

Nasraoui, O., Soliman, M., Saka, E., Badia, A. & Germain, R. (2008). A Web usage mining framework for mining evolving user profiles in dynamic Web sites. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20 (2), 202-215.

Nekaris, K. A.-I., Campbell, N., Coggins, T. G., Johanna Rode, E., & Nijman, V. (2013). Tickled to Death: Analysing Public Perceptions of 'Cute' Videos of Threatened Species (Slow Lorises – Nycticebus spp.) on Web 2.0 Sites. *PLoS ONE*, Vol. 8 (7).

Noruzi, A. (2005). Web Impact Factors for Iranian Universities. *Webology*, Vol. 2 (1).

Noruzi, A. (2006). The web impact factor: A critical review. *Electronic Library*, Vol. 24 (4), 490-500.

Nwagwu, W. E. & Agarin, O. (2008). Nigerian university websites: A webometric analysis. *Webology*, Vol. 5 (4).

Orduña-Malea, E. (2012). Graphic, multimedia, and blog content presence in the Spanish academic web-space. *Cybermetrics International Journal of Scientometrics, Informetrics and Bibliometrics*, 16 (1), 3.

Ortega, J. L. & Aguillo, I. F. (2007). Interdisciplinary relationships in the Spanish academic web space: A Webometric study through networks visualization. *Cybermetrics International Journal of Scientometrics, Informetrics and Bibliometrics*, 11 (1), 4.

Ortega, J. L. & Aguillo, I. F. (2008). Visualization of the Nordic academic web: Link analysis using social network tools. *Information Processing and Management*, Vol. 44 (4).

Ortega, J. L. & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, Vol. 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*, Vol. 74 (2), 295-308.

Otte, E. & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, Vol. 28 (6), 441-453.

Ou, J.-C., Lee, C.-H. & Chen, M.-S. (2008). Efficient algorithms for incremental Web log mining with dynamic thresholds. *VLDB journal*, Vol. 17 (4), 827-845.

Paliouras, G. (2012). Discovery of Web user communities and their role in personalization. *User Modelling and User-Adapted Interaction*, Vol. 22 (1-2), 151-175.

Palmer, J.W. (2002). Web site usability, design, and performance metrics. *Information Systems Research*, Vol. 13(2), 151-167.

Panchal, V., Pillai, S., & Singh, A. (2012). Truth Finder Algorithm for Multiple Conflicting Information Providers on the Web. *International Journal of Computer Applications*. Issue 5, 1-4.

Park, H.-W. (2010). Mapping the e-science landscape in South Korea using the webometrics method. *Journal of Computer-Mediated Communication*, Vol. 15 (2), 211-229.

Park, H.-W. & Kluver, R. (2009). Trends in online networking among South Korean politicians - A mixed-method approach. *Government Information Quarterly*, Vol. 26 (3), 505-515.

Park, H.-W. & Thelwall, M. (2008). Link analysis: Hyperlink patterns and social structure on politicians' Web sites in South Korea. *Quality & Quantity*, Vol. 42 (5), 687-697.

Pierrakos, D. & Paliouras, G. (2010). Personalizing Web Directories with the Aid of Web Usage Data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22 (9), 1331-1344.

Polanco, X., Roche, I. & Besagni, D. (2006). User science indicators in the Web context and co-usage analysis. *Scientometrics*, Vol. 66 (1), 171-182.

Poongothai, K. & Sathiyabama, S. (2012). Efficient web usage miner using decisive induction rules. *Journal of Computer Science*, Vol. 8 (6), 835-840.

Popova, V., John, R., & Stockton, D. (2009). Sales Intelligence Using Web Mining. In P. Perner (Ed.), *ICDM 2009, LNAI*, 5633, 131–145.

Pratt, J. A., Hauser, K., & Sugimoto, C. R. (2012). Cross-disciplinary communities or knowledge islands: examining business disciplines. *Journal of Computer Information Systems*, Vol. 53 No. 2, 9-21.

Qiu, G., Liu, B., Bu, J. & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, Vol. 37 (1), 9-27.

Rettinger, A., Loesch, U., Tresp, V., D'Amato, C. & Fanizzi, N. (2012). Mining the Semantic Web Statistical learning for next generation knowledge bases. *Data Mining and Knowledge Discovery*, Vol. 24 (3, SI), 613-662.

Richardson, M. & Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 61-70.

Romero, C., Ventura, S., Zafra, A. & De Bra, P. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers & Education*, Vol. 53 (3), 828-840.

Romero-Frías, E. & Vaughan, L. (2012). Exploring the relationships between media and political parties through web hyperlink analysis: The case of Spain. *Journal of the American Society for Information Science and Technology*, Vol. 63 (5), 967-976.

Ruller, T. J. (1993). A Review of Information Science and Computer Science Literature to Support Archival Work with Electronic Records. *American Archivist*. Vol 56.

Schubert, A. & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics*, Vol 36 (3), 311-324.

Shandilya, S.K. & Jain, D.S. (2009). Automatic opinion extraction from web documents. *Proceedings - 2009 International Conference on Computer and Automation Engineering, ICCAE 2009*, 351-355.

Sharma, K., Shrivastava, G. & Kumar, V. (2011). Web mining: Today and tomorrow. *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology*, Vol. 1, 399-403.

Shekofteh, M., Shahbodaghi, A., Sajjadi, S. & Jambarsang, S. (2010). Investigating Web impact factors of type 1, type 2 and type 3 medical universities in Iran. *Journal of Paramedical Sciences*, Vol. 1 (3), 34-41.

Shunbo Yuan & Weina Hua (2011). Scholarly impact measurements of LIS open access journals: based on citations and links. *The Electronic Library*, Vol. 29 (5), 682 - 697

Shyu, M.-L., Haruechaiyasak, C. & Chen, S.-C. (2006). Mining user access patterns with traversal constraint for predicting web page requests. *Knowledge and Information Systems*, Vol. 10 (4), 515-528.

Small, H. (2010). Maps of science as interdisciplinary discourse: co-citation contexts and the role of analogy. *Scientometrics*, Vol. 83 (3), 835–849.

Somprasertsri, G. & Lalitrojwong, P. (2010). Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal Computer Science*, Vol. 16 (6), 938-955.

Stuart, D., Thelwall, M., & Harries, G. (2007). UK academic web links and collaboration – an exploratory study. *Journal of Information Science*, Vol. 33 (2), 231–246.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.N. (2000). Web Usage Mining: Discovery And Applications Of Usage Patterns From Web Data. *Sigkdd Explorations*, 1 (2), 12-23

Takahashi, T., Abe, S. & Igata, N. (2011). Can Twitter be an alternative of real-world sensors? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6763, 240-249.

Thelwall, M. (2001a). A Web Crawler Design for Data Mining. *Journal Of Information Science*, Vol. 27 (5), 319-325.

Thelwall, M. (2001b). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, Vol. 52 (13), 1157-1168.

Thelwall, M. (2002a). A Research and Institutional Size Based Model for National University Web Site Interlinking. *Journal Of Documentation*, Vol. 58 (6), 683-694

Thelwall, M. (2002b). Evidence for the Existence of Geographic Trends in University Web Site Interlinking. *Journal Of Documentation*, Vol. 58 (5), 563-574

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology archive*, Vol. 57 (1), 60-68.

Thelwall, M. (2009). *Introduction to Webometrics: Quantitative Web research for the social sciences*. New York, NY: Morgan & Claypool.

Thelwall, M. (2010a). Webometrics. *Encyclopedia of Library and Information Sciences, Third Edition*, 5634-5643

Thelwall, M. (2010b). Webometrics: Emergent or doomed? *Information Research*, Vol. 15 (4).

Thelwall, M. (2011). A comparison of link and URL citation counting. *Aslib Proceedings: New Information Perspectives*, Vol. 63 (4), 419-425.

Thelwall, M., Buckley, K. & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, Vol. 62 (2), 406-418.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, Vol. 8 (5).

Thelwall, M., Klitkou, A., Verbeek, A., Stuart, D. & Vincent, C. (2010). Policy-Relevant Webometrics for individual scientific fields. *Journal of the American Society for Information Science and Technology*, Vol. 61 (7), 1464-1475.

Thelwall, M. & Sud, P. (2011). A Comparison of Methods for Collecting Web Citation Data for Academic Organizations. *Journal of the American Society for Information Science and Technology*, Vol. 62 (8), 1488-1497.

Thelwall, M. & Sud, P. (2012). Webometric research with the Bing Search API 2.0. *Journal of Informetrics*, Vol. 6 (1), 44-52.

Thelwall, M., Vann, K. & Fairclough, R. (2006). Web issue analysis: An integrated water resource management case study. *Journal of the American Society for Information Science and Technology*, Vol. 57 (10), 1303-1314

Thelwall, M., Vaughan, L. & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, Vol. 39, 81-135.

Thelwall, M. & Wouters, P. (2005). What's the deal with the web/blogs/the next big technology: a key role for information science in e-social science research? *CoLIS'05: Proceedings of the 5th international conference on Context: conceptions of Library and Information Sciences*.

Van Leeuwen, T. & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. *Research Evaluation*, Vol. 9 (3), 183–187.

Van Zoonen, L., Vis, F. & Mihelj, S. (2011). YouTube interactions between agonism, antagonism and dialogue: Video responses to the anti-Islam film Fitna. *New Media and Society*, Vol. 13 (8), 1283-1300.

Vaughan, L. & Romero-Frías, E. (2012). Exploring Web keyword analysis as an alternative to link analysis: a multi-industry case. *Scientometrics*, Vol. 93 (1), 217–232.

Vaughan, L. & 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*, Vol. 54(1), 29-38.

Vaughan, L. & Yang, R. (2012). Web Data as Academic and Business Quality Estimates: A Comparison of Three Data Sources. *Journal of the American Society for Information Science and Technology*, 63 (10), 1960–1972.

Vaughan, L., Yang, R., & Tang, J. (2012). Web co-word analysis for business intelligence in the Chinese environment. *Aslib Proceedings: New Information Perspectives*, Vol. 6, 653-666.

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*, Vol. 4 (4), 483-491.

Velásquez, J. D. (2013). Combining eye-tracking technologies with web usage mining for identifying Website Keyobjects. *Engineering Applications of Artificial Intelligence*, 26, 1469–1478.

Velásquez, J. D., Dujovne, L.E. & L'Huillier, G. (2011). Extracting significant Website Key Objects: A Semantic Web mining approach. *Engineering Applications of Artificial Intelligence*, Vol. 24 (8), 1532-1541.

Wang, C., Lu, J., & Zhang, G. (2007). Mining key information of web pages: A method and its application. *Expert Systems with Applications*, 33, 425–433.

Wang, K.-Y., Ting, I.-H., & Wu, H.-J. (2013). Discovering interest groups for marketing in virtual communities: An integrated approach. *Journal of Business Research*, 66, 1360–1366.

Wang, P., Sanin, C., & Szczerbicki, E. (2011). Application of Decisional DNA in Web Data Mining. *Knowlege-Based and Intelligent Information and Engineering Systems*. Vol. 6882, 631-639

Wang, P., Sanin, C., & Szczerbicki, E. (2012). Introducing the Concept of Decisional DNA–Based Web Content Mining. *Cybernetics and Systems: An International Journal*, 43, 136–142.

Wilkinson, D. & Thelwall, M. (2012). Trending Twitter Topics in English. *Journal of the American Society for Information Science and Technology*. Vol. 63 (8), 1631–1646.

Williams, C. J., O'Rourke, M., Eigenbrode, S. D., O'Loughlin, I, & Crowley, S. J. (2013). Using Bibliometrics to Support the Facilitation of Cross-Disciplinary Communication. *Journal of the American Society for Information Science and Technology*, Vol. 64 (9), p 1768-1779.

Woo-Young, C. & Park, H.W. (2012). The network structure of the Korean blogosphere. *Journal of Computer-Mediated Communication*, Vol. 17 (2), 216-230.

Yang, B., Liu, J. & Feng, J. (2012). On the spectral characterization and scalable mining of network communities. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24 (2), 326-337.

Yang, B. & Sun, Y. (2013). An exploration of link-based knowledge map in academic web space. *Scientometrics*, Vol. 96 (1), 239–253

Yeh, I.-C., Lien, C., Ting, T.-M., & Liu, C.-H. (2009). Applications of web mining for marketing of online bookstores. *Expert Systems with Applications*, 36, 11249–11256

Zhang, Q. & Segall, R.S. (2008). Web Mining: A survey of current research, techniques, and software. *International Journal of Information Technology and Decision Making*, Vol. 7 (4), 683-720.

Zhang, Y. & Xu, G. (2009). On web communities mining and recommendation. *Concurrency and Computation-Practice & Experience*, Vol. 21 (5), 561-582.

Zhang, Z. & Nasraoui, O. (2008). Mining search engine query logs for social filtering-based query recommendation. *Applied Soft Computing*, Vol. 8 (4), 1326-1334.

Zuccala, A. (2006). Author Cocitation Analysis Is to Intellectual Structure As Web Colink Analysis Is to...? *Journal of the American Society for Information Science and Technology*, 57(11), 1487-1502.

Appendix: The queries

General data collection

Initial Scopus queries

Webometrics
(webometric* OR "web metric*" OR cybermetric* OR scientometric* OR informetric*) AND ("web impact assessment" OR "web impact report*" OR "web impact analy*" OR "web citation analy*" OR "web content analy*" OR "link analy*" OR "webometric link analy*" OR "link relationship map*" OR "link relationship analy*" OR "link impact report*" OR "link impact analy*" OR "link network analy*" OR "colink relationship map*" OR "colink relationship analy*" OR "colink impact report*" OR "colink impact analy*" OR "colink network analy*" OR "co-link relationship map*" OR "co-link relationship analy*" OR "co-link impact report*" OR "co-link impact analy*" OR "co-link network analy*" OR "web analy*" OR "log analy*" OR "web memetic*" OR "social network analy*" OR "social network metric*")

Web mining
("web mining" OR "web data mining") AND ("social network mining" OR "social network metric*" OR "web personalization" OR "web recommend*" OR "web community analy*" OR "web linkage mining" OR "web usage mining" OR "web structure mining" OR "web content mining" OR "web knowledge discovery" OR "collaborative filtering" OR "opinion mining" OR "web community discovery" OR "web graph measur*" OR "web graph model*" OR "log analy*" OR "log mining" OR "web structural analy*" OR "web structure analy*" OR "web temporal analy*" OR "link analy*")

Refined queries for Scopus and Web of Science

Webometrics, Scopus
TITLE-ABS-KEY(webometric* OR cybermetric* OR scientometric* OR informetric*) AND TITLE-ABS-KEY("web impact" OR "web citation analy*" OR "web citing analy*" OR "web content analy*" OR "link analy*" OR "colink analy*" OR "co-link analy*" OR "link relationship*" OR "link impact*" OR "link network*" OR "colink relationship*" OR "colink*" OR "colink network*" OR "co-link relationship*" OR "co-link impact*" OR "co-link network*" OR "web analy*" OR "log analy*" OR "web content*" OR "web usage" OR "web memetic*" OR "virtual memetic*" OR "social network" OR "web knowledge")

142 items returned.

Webometrics, WoS
TS=(webometric* OR cybermetric* OR scientometric* OR informetric*) AND TS=("web impact" OR "web citation analy*" OR "web citing analy*" OR "web content analy*" OR "link analy*" OR "colink analy*" OR "co-link analy*" OR "link relationship*" OR "link impact*" OR "link network*" OR "colink relationship*" OR "colink*" OR "colink network*" OR "co-link relationship*" OR "co-link impact*" OR "co-link network*" OR "web analy*" OR "log analy*" OR "web content*" OR "web usage" OR "web memetic*" OR "virtual memetic*" OR "social network" OR "web knowledge")

133 items returned.

Web mining, Scopus
TITLE-ABS-KEY("web mining" OR "web data mining") AND TITLE-ABS-KEY("social network" OR "web personal*" OR "web recommend*" OR "web community" OR "web linkage mining" OR "web usage" OR "web structure" OR "web content" OR "web knowledge" OR "collaborative filtering" OR "opinion mining" OR "web

community" OR "web graph measur*" OR "web graph model*" OR "log analy*" OR "log mining" OR "web structural analy*" OR "web structure analy*" OR "web temporal analy*" OR "link analy*")

688 items returned.

Web mining, WoS
TS=("web mining" OR "web data mining") AND TS=("social network" OR "web personal*" OR "web recommend*" OR "web community" OR "web linkage mining" OR "web usage" OR "web structure" OR "web content" OR "web knowledge" OR "collaborative filtering" OR "opinion mining" OR "web community" OR "web graph measur*" OR "web graph model*" OR "log analy*" OR "log mining" OR "web structural analy*" OR "web structure analy*" OR "web temporal analy*" OR "link analy*")

338 items returned.

Data collection for citation and keyword analysis

Webometrics
TITLE-ABS-KEY(webometric* or cybermetric*) AND (LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "re") OR LIMIT-TO(DOCTYPE, "ip"))

307 items returned.

Web mining
TITLE-ABS-KEY("web mining" or "web data mining") AND (LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "re") OR LIMIT-TO(DOCTYPE, "ip"))

2,518 items returned.

Social web search terms
farmville, hulu, prezi, posterous, blipfm, boxee, friv, friendfeed, gliffy, kerpoof, mint, docstoc, animoto, fotoflexer, lijit, google docs, foxytunes, wufoo, twitter, openid, piczo, picnik, joost, footnote, digg, viddler, snap, wesabe, zamzar, linkedin, compete, weebly, typepad, ilike, slide, feedblitz, mybloglog, quantcast, blip.tv, songbird, widgetbox, panoramio, plazes, scrapblog, imagekind, zoho, metacafe, evernote, reddit, zyb, yelp, amie.st, finetune, pageflakes, feedburner, netvibes, zooomr, facebook, youtube, alexa, flickr, gmail, box, ebay, amazon, orkut, myspace, skype, meebo, delicious, del.icio.us, flock, stumbleupon, pandora, last.fm, smugmug, social, 2.0, new media, blog*, communit*, wiki, collabo*, participat*, new web