Bibliometrics to webometrics

Mike Thelwall

Journal of Information Science 2008; 34; 605 originally published online Jun 13, 2008;
DOI: 10.1177/0165551507087238

JIS celebrating 5 years

Bibliometrics to webometrics

Mike Thelwall

University of Wolverhampton

Abstract.

Bibliometrics has changed out of all recognition since 1958; becoming established as a field, being taught widely in library and information science schools, and being at the core of a number of science evaluation research groups around the world. This was all made possible by the work of Eugene Garfield and his Science Citation Index. This article reviews the distance that bibliometrics has travelled since 1958 by comparing early bibliometrics with current practice, and by giving an overview of a range of recent developments, such as patent analysis, national research evaluation exercises, visualization techniques, new applications, online citation indexes, and the creation of digital libraries. Webometrics, a modern, fast-growing offshoot of bibliometrics, is reviewed in detail. Finally, future prospects are discussed with regard to both bibliometrics and webometrics.

Keywords: bibliometrics; scholarly publishing; webometrics

1. Introduction

The last 50 years have seen two major technological changes in scholarly publishing and two major changes in the way research can be quantitatively analysed, alongside numerous less significant developments. The two publishing changes are the computerization of the printing process, reducing costs significantly and allowing more journals and books to appear in print; and the conversion of the entire publishing cycle (submission of articles, refereeing and publication) to the internet, allowing faster and possibly cheaper communication throughout. Historically, the first major change for the development of quantitative analysis of academic publishing (bibliometrics) was the creation of the Institute for Scientific Information (ISI, now Thomson Scientific) citation database, which began functioning in 1962 [1, 2] together with associated post-war sociological theory allowing it to be used to assess the impact of scientific work [3]. Since then there has been a continuous increase in the computing power available in universities, which has helped to make increasing numbers of bibliometric analyses possible. The second major development for bibliometrics was the web publishing of an increasingly broad range of research-related documents, from articles to email discussion lists, allowing the creation of a range of new metrics relating to their access and use.

In this article, the focus is on the measurement of science. Conveniently for this special issue, the two significant changes fall just after the beginning and just before the end of the period in question, although in between bibliometrics has arisen as a recognized scientific specialism: taught in

Correspondence to: Mike Thelwall, School of Computing and IT, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. Email: m.thelwall@wlv.ac.uk

universities as part of information science courses, with a substantial body of techniques, some theories, and an international group of specialist science evaluators. This review article has a dual focus: general bibliometric issues and the field of webometrics, a new research area that has grown out of bibliometrics. The next section discusses bibliometrics originating within the first half of 1958–2008 and the following section discusses a selection of more recent developments. A further section then focuses exclusively on webometrics.

2. Bibliometrics

Bibliometrics encompasses the measurement of 'properties of documents, and of document-related processes' [4]. The range of bibliometric techniques includes word frequency analysis [5], citation analysis [6], co-word analysis [7] and simple document counting, such as the number of publications by an author, research group or country. In practice, however, bibliometrics is primarily applied to science-related documents and hence has considerable overlap with scientometrics, the science measurement field.

Although recognizably bibliometric techniques have been applied for at least a century, the emergence of bibliometrics as a scientific field was triggered (in the 1960s) by the development of the Institute for Scientific Information (ISI) Science Citation Index (SCI) by Eugene Garfield [2], as a logical continuation of his drive to support scientific literature searching. The SCI was created as a database of the references made by authors, to earlier articles, in their articles published in the top scientific journals, originally focusing on general science and genetics. The underlying idea, still highly relevant today, is that if a scientist reads an article, then s/he would benefit from knowing which articles cited it, since they may cover a similar topic and might update or correct the original article. The importance of the SCI is also consistent with Bradford's [8] law of scattering: although a scientist may keep up-to-date with a research specialism by reading all relevant journals when they appear, a minority of relevant articles will be scattered throughout other journals. Hence citation searching protects researchers from missing relevant articles in non-core journals.

Almost a by-product of the SCI, and later also the Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI), was the ability to generate easily a range of new statistics: not just the number of citations to any given article but also, using other fields in the SCI database, aggregated publication and citation counts. These aggregated statistics include the number of citations to all articles in a journal or all articles by an author, research group, or country. Some were further developed into named indicators with supporting theories and reasonably well accepted standard interpretations. Perhaps the most well known is the journal impact factor (JIF), defined below.

Since the publication of the SCI, two types of bibliometric application have arisen: evaluative and relational [4]. Evaluative bibliometrics seeks to assess the impact of scholarly work, usually to compare the relative scientific contributions of two or more individuals or groups. These evaluations are sometimes used to inform research policy and to help direct research funding [6]. In contrast, relational bibliometrics seeks to illuminate relationships within research, such as the cognitive structure of research fields, the emergence of new research fronts, or national and international co-authorship patterns.

2.1. Evaluative bibliometrics

Most evaluative bibliometric techniques use citations as their raw data. The theory for this stems from Robert Merton's [3] sociology of science, which postulates that citations are the way in which scholars acknowledge influential prior work. On this basis, citation counting could be used as an indicator of scientific value because more influential work would tend to be more frequently cited. In fact the term 'impact' is now accepted as appropriate for that which citations measure or indicate. Subsequent research has shown that Merton's perspective is a simplification of reality: there are many different reasons to cite articles as well as many influences on which articles to select, when multiple options are available [9, 10]. From an alternative perspective, de Solla Price [11] showed that a cumulative

advantage process could be at work for highly cited papers, where papers that are initially well cited then tend to continue to be cited partly because they have been cited rather than for their intrinsic worth. This is similar to Merton's [12] 'Matthew effect' in science, whereby recognized scholars tend to be awarded a disproportionate credit for their research. Despite complications such as these, indicators based upon citation counts have been widely adopted.

The journal impact factor, introduced in the early 1960s [13–15] is the number of citations from ISI-indexed articles published in the year X to articles in the journal published in years X – 1 and X – 2, divided by the number of (citable) items published in the journal in the years X – 1 and X – 2. On the basis of Merton [3], journals with higher JIFs tend to publish higher impact research and hence tend to be better regarded. Nevertheless, there seems to be general agreement that, even within discrete subject fields, ranking journals based upon JIFs is problematic [4, 6]. Moreover, as the JIF has gained in popularity, there seem to have been attempts by journal editors to recommend authors to cite other articles in the same journal to improve its JIF.

A second common application is tenure and promotion decisions which may take into account the JIFs of the journals in which an academic has published, or the citation counts of their publications. This is not recommended by many bibliometricians, however, since citation counts at the level of individual authors are unreliable and those making the decisions may be unaware of field differences [6].

A third application, usually conducted by expert bibliometricians, is comparing academic departments through citations to their publications. Even carefully constructed bibliometric indicators, which are reasonably robust because of aggregation over the publications of entire departments, need to be combined with other sources of evidence (e.g. funding, sources of esteem, peer review, narrative) in order to give solid evidence for major decisions, such as those involving funding.

2.2. Relational bibliometrics

There were several early attempts to develop bibliometric methods to examine relations within science through ISI data, although the growth of relational analysis methods was probably constrained by the lack of sufficient computing power in the early days, especially for visualizations. Nevertheless, early relational analyses produced interesting insights into the structure of science through simple means, such as network diagrams of the flow of citations between key sets of articles [16]. This idea was apparently invented by the geneticist Dr Gordon Allen in 1960, who sent his citation diagram to an enthusiastic Garfield [16]. Journal citation diagrams were another early invention: these can illustrate the connections between journals within a field, detect journals that cross field boundaries and identify central or peripheral journals.

One important relational method, sometimes attributed to Garfield, is co-citation as a measure of similarity [17, 18]. The basis of this is that pairs of documents that often appear together in reference lists (i.e. are co-cited) are likely to be similar in some way. This means that if collections of documents are arranged according to their co-citation counts then this should produce a pattern reflecting cognitive scientific relationships. Author co-citation analysis (ACA) is a similar technique in that it measures the similarity of pairs of authors through the frequency with which their work is co-cited [19]. ACA operates at a high enough level of aggregation to be a practical tool for mapping the structures of fields [20].

3. Bibliometrics today

Mainstream bibliometrics has evolved rather than undergone revolutionary change in response to the web and web-related developments. The core citation-based impact measures are still in place, but are now supplemented by a range of complementary techniques. In addition, there is now a body of theory and case studies to draw upon so that an experienced bibliometrician can be reasonably sure of finding good ways to generate indicators from citations for any common task and also of how to interpret the results. In particular there has been an ongoing debate about the validity of using citations to measure impact, in parallel with the development of theories of citer motivations, which have recently been extensively reviewed [21].

Aside from the core citation analysis methods, the biggest change in bibliometrics stems from the availability of new significant sources of information about scholarly communication, such as patents, web pages, and digital library usage statistics. Of course, the wider field of scientometrics has never been exclusively interested in academic papers and has also used other data such as funding as well as qualitative indicators, such as peer review judgments.

There are perhaps three main trends in the recent history of bibliometrics, and citation analysis in particular. These are to improve the quality of results through improved metrics and careful data cleaning, to develop metrics for new tasks, and to apply bibliometrics to an increasing range of problems, particularly in descriptive relational contexts (see the knowledge domain visualization section below for examples of the latter).

3.1. The h-index and current bibliometric indicators

Perhaps the most significant new evaluative metric is the *h*-index [22] which, for a scientist, is the largest number *h* such that s/he has at least *h* publications cited at least *h* times. A high *h* index indicates that a scientist has published a considerable body of highly cited work. It is a metric that is easily calculated and intuitive to understand, and hence its appeal. There have been a number of studies of the *h*-index, evaluating it, proposing new versions or applying it to sets of scholars. For example, ranked lists of the *h*-indexes of UK and US LIS professors [23, 24] may be of interest to researchers in the field.

Apart from the *h*-index, the most important evaluative bibliometric indicators seem to have evolved gradually. For example, because there are widely differing field-based citation norms, comparing citation counts is inappropriate across multiple fields. Hence, it is now best practice to field-normalize citation indicators [6] when using them to evaluate an academic department. Even if a set of departments in the same discipline are to be compared, raw citation counts, or average citations per researcher, would not be an accurate reflection of their citation impact because the departments might specialize in fields with different average citation rates. Hence, departments engaging in research in fields with high average citation counts would have an unfair advantage unless the indicators were normalized, for example through dividing each department's citations by the field average. Hence the evaluative citation analysis goal has shifted from evaluating the impact of research to evaluating its impact relative to a field.

3.2. National research evaluation exercises

Systematic research assessment exercises seem set to become an important but still controversial application area for bibliometrics. Four countries now have periodic national research evaluation exercises that help to direct a significant proportion of their research funding. The UK's Research Assessment Exercise (RAE, see www.rae.ac.uk) was the first and has taken place in 1986, 1989, 1992, 1996, 2001 and in 2008. It is primarily based upon peer review, with small panels of subject experts awarding ratings to relevant groups from each university. In addition to peer review, normally based on the top four publications of each researcher, the panels take into account other factors, such as funding, PhD completions and a narrative. Although bibliometrics have not yet played a formal role, they can be included in the narrative part of the submission and the panels may also use them as part of their decision making process. Anecdotal evidence suggests that many disciplines have also developed informal publication guidelines and lists of target journals that are influenced by JIFs. The 2008 RAE was set to be conducted in parallel with bibliometrics, however, and subsequent RAEs may have a much heavier bibliometric component [25], something that has been argued for on the basis that academic publications are already peer reviewed, e.g. [26]. The rationale behind a shift towards bibliometrics is that bibliometric indicators such as citations and JIFs are transparent and cheap compared to peer review. Nevertheless, bibliometrics is inadequate on its own and so RAEs are likely to always contain a range of other inputs, probably including peer review as a final criterion.

New Zealand's Performance Based Research Fund,[1] which started in 2003, ran again partially in 2006 and is set to run again in 2012. It is quite similar to the RAE except that it has always given grades to individual academics rather than whole submissions. The UK RAE converted to individual grades rather than group grades in 2008, but in the New Zealand system every academic submits an evidence

portfolio, rather than a fixed number of publications, although for most academics the heart of their portfolio would probably have been the component of up to four 'nominated research outputs' [27].

In Australia, the Institutional Grants Scheme (IGS), which predominantly replaced the earlier Research Quantum, is effectively a national evaluation exercise. Australian assessment has never included a significant element of peer review but has always been based primarily upon external funding for research. Starting in 2002, the funding was based upon 'success in attracting research students (30% of funding), in attracting research income (60%) and in the quality and output of its research publications (10%)' [28].

The world's largest science system, that of the USA, does not have a national science evaluation exercise. Instead, US research funding is allocated competitively on a grant-by-grant basis, with ex-ante evaluations being of secondary importance and carried out by the funding agencies [29]. The Netherlands' research evaluation process is a different type again, comprising a 'patchwork' of evaluations by different funding agencies and stakeholders [30]. Different again is the evaluation system of the former Foundation for Research Development (FRD), now superseded by the National Research Foundation (NRF) in South Africa, which combined the New Zealand style retrospective evaluation of individual researchers with a second stage, the evaluation of the researchers' future plans, before awarding funding grants [31]. The current NRF system is similar, with a preliminary stage of rating individual researchers by peer review,[2] with rated researchers allowed to apply for grant funding. Finally, and despite the above descriptions, probably the majority of countries conduct ad-hoc evaluations rather than a systematic exercise, with Italy being an example of this [32].

How effective and useful are these research evaluation systems? This is clearly a controversial issue and one that does not have a simple answer. In Australia, however, there is evidence of the problems of a simplistic scheme: a negative impact of counting publications but not evaluating their quality seems to be an increased volume of lower quality journal articles [33]. The UK's RAE has been congratulated for its success, based upon bibliometric evidence of UK performance relative to the world [34], but it is not clear whether this assessment is widely believed.

3.3. New bibliometric databases: Google Scholar and Scopus

In 1992 the ISI was sold by Garfield and other shareholders to a company that later became Thomson Scientific, and which continued the citation indexes. Recent years have seen the emergence of significant challengers to the ISI indexes in the form of alternative large-scale online scholarly article databases like Google Scholar and Scopus (Elsevier), which contain embedded citation information. In addition, there are smaller-scale field-specific digital libraries and archives that contain citation indexes, such as CiteSeer for Computer Science, and the CiteBase initiative[3] to index the citations of free online scholarly publications, including those in archives like arXiv[4] (physics, mathematics, computer science and quantitative biology).

One article has compared the Web of Science (with ISI data), Google Scholar and Scopus with the explicit objective of assessing the extent to which the results of a citation analysis depend upon the data source used, using the task of ranking the faculty in a single library and information science school. The findings showed that Google Scholar was probably too difficult to use for a large-scale citation analysis and that the other two gave similar results overall [35]. Nevertheless, weaknesses in the coverage of certain fields resulted in significant disadvantages to some faculty members, depending upon the database used. Hence the use of both in conjunction with each other would give the fairest results. In addition, the poor coverage of conferences by both in comparison to Google Scholar illustrates that neither gives fair results to academics who publish in fields which emphasize conferences, such as computer science and computational linguistics [35]. Another investigation compared different databases for coverage of social sciences research, finding Scopus to offer particularly good coverage [36]. Other studies with wider disciplinary coverage have also shown that the coverage of Google Scholar is variable and can be low or particularly unreliable for some disciplines [37, 38].

Despite the limitations of all current citation sources, hopefully the existence of challengers to the ISI will make it easier than ever before to critically assess the extent of ISI coverage and to identify national and other biases.

3.4. Knowledge domain visualization

The increasing use of sophisticated visualizations is probably the most significant development in relational bibliometrics and has led to the creation of a new field: knowledge domain visualization (KDViz), within the information visualization research area [39]. This involves significant computing resources and is part of a wider 'eResearch' trend to harness computers for social science research goals. In addition to Chen's [39] three dimensional information-rich visualizations of individual research fields, others have implemented ambitious plans to map large areas of science via citations in the ISI database [40, 41].

Whilst early relational bibliometric analyses might have produced simple hand-drawn diagrams of citations between authors, journals or articles, later researchers developed software to automate this process. For example, Olle Persson's Bibexcel[5] can be fed citation data from the ISI and used to produce a range of two-dimensional diagrams, such as an ego network (a term borrowed from social network analysis) of researchers with the strongest citation relationship with a given author. Similarly, Loet Leydesdorff has a set of programs[6] that can convert ISI data into a format that can produce diagrams, especially to illustrate the citation relationships between individual journals [42]. Eugene Garfield's HistCite[7] produces timeline visualisations for citations within a collection of articles. It is unusual that it is aimed at supporting literature searches rather than bibliometrics.

There are several sets of visualization software with significant inputs from computer scientists that are free to use and can easily process ISI data to produce three dimensional visualizations. Katy Borner's InfoViz Cyberinfrastructure[8] is a general purpose suite of open source software with many algorithms to process and display the data. A particular advantage is its ability to process massive amounts of data. For example Boyack [41] produced a combined science and technology map using bibliometric coupling on the references from over a million papers in the Science Citation Index. Chaomei Chen's CiteSpace[9] software focuses exclusively on bibliometric analysis and can produce beautiful three dimensional visualizations of citation networks. One of the interesting features of some of Chen's networks is the ability to include not just the basic structure but also many layers of additional information chosen by the researcher through the use of colour and other features. For example a circle representing an article in a network of articles may reveal the article's annual citation counts by representing them with differently coloured layers [43].

3.5. Patents

A patent is a set of time-limited exclusive rights to an invention, normally granted by a government patent office. The term patent is also used for the officially registered invention descriptions. These documents are similar in some ways to academic papers, for example through the inclusion of a set of references. The adoption of patents as an indicator of scientific value stems from a recognition that academic researchers can be, and perhaps sometimes should be, directly involved in the development of useful technologies [44].

Patent indicators can be direct, in the sense of counting researchers' patent awards to reward those with novel research with potential commercial value. Patent indicators can also be indirect by using patent references to identify the cited academic work which is thereby endorsed as having applicable value [45]. Patent analyses have also been used to evaluate the performance of a country's technology and to identify flows of knowledge transfer between science and technology [46]. For example, one empirical study of patents relative to the Netherlands concluded that the results did not fit the existing theoretical models of university–industry relationships, which therefore needed to be reconsidered [47].

3.6. Usage data from digital libraries

Perhaps the most significant challenge for bibliometrics in the long run is that the new digital libraries are producing large-scale evidence of the usage patterns of academic articles for the first time [48]. Editors are already receiving usage statistics in addition to impact factors from publishers in some cases and it seems likely that the two give useful complementary information [49, 50].

Research into log files may also be able to connect usage patterns to user demographics in some cases, which may give additional insights into users [51] and information seeking patterns [52].

Two important questions concern the impact of open access publishing upon the visibility of articles and publishers' revenues. One study of mathematics articles in ArXiv addressed both questions and suggested that open access articles tend to be cited more often but that more citable articles tend to be deposited in ArXiv, rather than necessarily attracting more citations because of being deposited there. In addition, there was some evidence that open access articles that were subsequently published were less frequently downloaded (a reduction of 23%) from publishers' web sites [53].

Digital library usage data can correlate with citation counts. For instance, early readership data seems to be moderately good at predicting future citation counts for an article [54]. Nevertheless, there will be many articles for which citations and usage statistics differ greatly. This raises the possibility that usage data may be used as evidence of a different kind of impact. For example, it may be seen as valuable in some subjects that research is used in undergraduate teaching and usage statistics would be more valuable than citations to assess this. Perhaps in the future we will have 'usage classics' as well as 'citation classics'. Publishers already often distribute lists of the most downloaded articles to editorial boards of journals but the lack of standardization of usage statistics seems to have prevented the creation of universal lists, say for physics or information science. Moreover, the currently available usage statistics are not unproblematic: for example publishers have already noticed that individual articles can pick up high usage rates through being set as compulsory reading by the instructor of a large class.

4. Webometrics

Webometrics is the quantitative analysis of web phenomena, drawing upon informetric methods [55], and typically addressing problems related to bibliometrics. Webometrics was triggered by the realization that the web is an enormous document repository with many of these documents being academic-related [56]. Moreover, the web has its own citation indexes in the form of commercial search engines, and so it is ready for researchers to exploit. In fact, several major search engines can also deliver their results automatically to investigators' computer programs, allowing large-scale investigations [57]. One of the most visible outputs of webometrics is the ranking of world universities based upon their web sites and online impact [58].

Webometrics includes link analysis, web citation analysis, search engine evaluation and purely descriptive studies of the web. These are reviewed below, in addition to one recent application: the analysis of Web 2.0 phenomena. Note that there is also some research into developing web-based metrics for web sites to evaluate various aspects of their construction, such as usability and information content, but this will not be reviewed here.

4.1. Link analysis

Link analysis is the quantitative study of hyperlinks between web pages. The use of links in bibliometrics was triggered by Ingwersen's [59] web impact factor (WIF), created through analogy to the JIF, and the potential that hyperlinks might be usable by bibliometricians in ways analogous to citations, e.g. [60]. The standard WIF measures the average number of links per page to a web space (e.g. a web site or a whole country) from external pages. The hypothesis underlying early link analysis was that the number of links targeting an academic web site might be proportional to the research productivity of the owning organization, at the level of universities [61], departments [62], research groups [63], or individual scientists [64]. Essentially the two are related because more productive researchers seem to produce more web content, on average, although this content does not attract more links per page [65]. Nevertheless, the pattern is likely to be obscured in all except large-scale studies because of the often indirect relationship between research productivity and web visibility. For example, some researchers produce highly visible web resources as the main output of their research, whilst others with equally high quality offline research attract less online attention.

Subsequent hyperlink research has introduced new metrics and applications as well as improved counting methods, such as the alternative document models [66]. In most cases this research has focused on method development or case studies. The wide variety of reasons why links are created and the fact that, unlike citing, linking is not central to any areas of science, has led to hyperlinks rarely being used in an evaluative role. Nevertheless, they can be useful in describing the evolution or connectivity of research groups within a field, especially in comparison with other sources of similar information, such as citations or patents [67]. Links are also valuable to gain insights into web use in a variety of contexts, such as by departments in different fields [68, 69].

A generic problem with link analysis is that the web is continually changing and seems to be constantly expanding so that webometric findings might become rapidly obsolete. A series of longitudinal investigations into university web sites in Australia, New Zealand and the UK have addressed this issue. These university web sites seem to have stabilized in size from 2001, after several years of rapid growth [70]. A comparison of links between the web sites from year to year found that this site size stabilization concealed changes in the individual links, but concluded that typical quantitative studies could nevertheless have a shelf life of many years [71].

4.2. Web citation analysis

A number of webometric investigations have focused not on web sites but on academic publications; using the web to count how often journal articles are cited. The rationale behind this is partly to give a second opinion for the traditional ISI data, and partly to see if the web can produce evidence of wider use of research, including informal scholarly communication and for commercial applications. A number of studies have shown that the results of web-based citation counting correlates significantly with ISI citation counts across a range of disciplines, with web citations being typically more numerous [37, 72–74]. Nevertheless, many of the online citations are relatively trivial, for example appearing in journal contents lists rather than in the reference sections of academic articles. If this can be automated then it would give an interesting alternative to the ISI citation indexes.

4.3. Search engines

A significant amount of webometrics research has evaluated commercial search engines [75]. The two main investigation topics have been the extent of the coverage of the web and the accuracy of the reported results. Research into developing search engine algorithms (information retrieval), and into how search engines are used (information seeking) are not part of webometrics. The two audiences for webometrics search engine research are researchers who use the engines for data gathering (e.g. the link counts above) and web searchers wanting to understand their results.

Search engines have been a main portal to the web for most users since the early years. Hence, it has been logical to assess how much of the web they cover. In 1999, a survey of the main search engines estimated that none covered more than 17.5% of the 'indexable' web and that the overlap between search engines was surprisingly low [76]. Here the 'indexable' web is roughly the set of pages that a perfect search engine could be expected to find if it found all web site home pages and followed links to find the remainder of pages in the sites. The absence of comparable figures after 1999 is due to three factors: first, an obscure Hypertext Transfer Protocol technology, the virtual server, has rendered the sampling method of Lawrence and Giles ineffective; second, the rise of dynamic pages means that it is no longer reasonable to talk in terms of the 'total number of web pages'; finally, given that search engine coverage of the web is only partial, the exact percentage is not particularly relevant, unless it has substantially changed. One outcome of this research, however, was clear evidence that meta-search engines could give more results through combining multiple engines. Nevertheless, these have lost out to Google, presumably because the key task of a search engine is to deliver relevant results in the first results page, rather than a comprehensive list of pages.

Given that web coverage is partial, is it biased in any important ways? This is important because the key role of search engines as intermediaries between web users and content gives them considerable economic power in the new online economy [77, 78]. In fact, coverage is biased internationally

in favour of countries that were early adopters of the web [79]. This is a side effect of the way search engines find pages rather than a policy decision.

The issue of accuracy of search engine results is multifaceted, relating to the extent to which a search engine correctly reports its own knowledge of the web. Bar-Ilan and Peritz [80] have shown that search engines are not internally consistent in the way they report results to users. Through a longitudinal analysis of the results of the query 'Informetric OR Informetrics' in Google they showed that search engines reported only a fraction of the pages in their database. Although some of the omitted pages duplicated other returned results, this was not always the case and so some information would be lost to the user. A related analysis with Microsoft Live Search [81] suggested that one reason for lost information could be the search engine policy of returning a maximum of two pages per site.

Many webometric studies have used the hit count estimates provided by search engines on their results pages (e.g. the '50,000' in 'Results 1–10 of about 50,000') rather than the list of matching URLs. For example, Ingwersen [59] used these to estimate the number of hyperlinks between pairs of countries. The problem with using these estimates is that they can be unreliable and can even lead to inconsistencies [82–84], such as expanded queries giving fewer results. In the infancy of webometrics these estimates could be highly variable and so techniques were proposed to smooth out the inconsistencies [85], although the estimates subsequently became much more stable.

A recent analysis of the accuracy of hit count estimates for Live Search found a surprising pattern. The estimates tended to be stable for large numbers (>8000) and small numbers (<300) but unstable for mid-range numbers. The reason seems to be that the high estimates were of the total number of matches known to the search engine, whereas the low estimates were of the number of matches after the elimination of duplicate pages, near-duplicate matches and multiple pages from the same site. The instability in the middle of the results was due to the transition between these two types of estimate [81]. The different nature of the estimates is a problem for webometrics investigations that use queries with both high and low hit counts.

4.4. Describing the web

Given the importance of the web, some webometrics research has been purely descriptive. A wide variety of statistics have been reported using various survey methods. These include: the average web page size; average number and type of meta-tags used and the average use of technologies like Java and JavaScript [86, 87]. In addition, many commercial web intelligence companies have reported basic statistics such as the number of users, pages and web servers, broken down by country. Here only two types of descriptive analysis are reported, however: link structure characterizations and longitudinal studies.

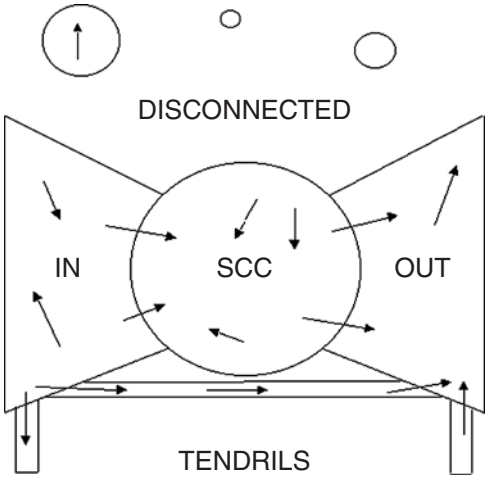

Fig. 1. The bow tie model of the web.

There are two separate key findings about web links, concerning the overall web structure and how the links evolve. Researchers at AltaVista used a copy of a web crawl to construct a holistic picture of the link structure of the web [88]. They found a bow tie model (Figure 1), with a core 'strongly connected component' (SCC) of 28% of pages that could all reach each other by clicking on one or more links. This seems to be the heart of the web, being relatively easy to navigate and containing the well-linked portal sites like Yahoo! Directory and the Dmoz Open Source Directory. In addition 21% of pages could be reached by clicking on one or more links, starting at any SCC page, but could not 'get back' to the SCC by following chains of links. This 'OUT' component includes many web sites that are linked to by Yahoo! or other SCC pages but do not contain any links to pages outside of the site. In contrast, the 'IN' component is the 21% of pages that link to the SCC directly or by following links. These seem to be pages or sites that are unknown to the core of the web. Finally, some groups of pages do not link to the rest of the web in any way (8%, DISCONNECTED) and a substantial proportion has more exotic connections (22%, TENDRILS).

The bow tie model was later recast as a corona model by Björneborn [89], in order to emphasize the centrality of the SCC and the often close connection between IN and SCC, and between SCC and OUT. For example, the core of a site could be in the SCC but with many pages deeper inside the site being in OUT. Björneborn [90] also investigated shortest link paths between subsites of UK university web sites, finding that computer science often connects otherwise disparate research fields.

Web dynamics research, in contrast to the structural analysis described above, is concerned with measuring, characterizing and modelling changes in the web. A key finding is that the cumulative advantage/Matthew effect phenomenon of bibliometrics (and elsewhere) applies to web links [91]. On the web, a few pages attract millions of links whereas hundreds of millions of pages attract one or none. This imbalance can be accounted for on the basis that when a new link is created it is more likely to target pages that already have many links. Of course nobody counts links to pages before deciding where their link should target, but the mediating factor is search engines. People are most likely to know about pages that have many links to them because search engines use links to find pages and to rank them [92]. Hence pages with many links to them are more visible online.

A study of the distribution of links to pages found that this cumulative advantage phenomenon fitted some types of pages, such as company home pages, but not others, like university home pages. In the latter case a second factor must have been at work, such as pre-knowledge of the existence of the page [93]. A consequence of either case is that a page with no links to it is unlikely to attract many more. Hence web site owners should initiate the creation of a few links to their site to help attract more. An implication for webometrics is that counts of links to a page are not reliable as indicators of the quality of the page's contents: pages may have many links because they became highly visible at some time in the past.

Finally, a different type of web dynamics is the analysis of changes in online information. Koehler [94] tracked a set of web pages from 1996, finding that they initially disappeared regularly, but there was then a period of stabilization, during which the surviving pages stopped disappearing. Koehler also claims that the survival rate of web pages will differ according to specialism. A differently organized study analysed 738 web sites in 1997 and 2004, finding a sevenfold growth in size during this period [95]. There were significant increases in some types of web content, such as dynamic pages and internal web site links, although the number of site outlinks (pointing to pages outside a site) did not grow as fast as the number of pages inside the site. The study found that only 25% of the site outlinks from 1997 still existed in the same place in 2004.

In summary, the web is clearly a complex, evolving entity that, despite its unregulated nature, exhibits strong patterns when analysed on a large scale.

4.5. Measuring Web 2.0

Web 2.0 is a term coined by the publisher Tim O'Reilly mainly to refer to web sites that are driven by consumer content, such as blogs, Wikipedia and social network sites. The growth in volume of web content created by ordinary users has spawned a market intelligence industry and much measurement research. The idea behind these is data mining: since so many people have recorded informal

thoughts online in various formats, such as blogs, chatrooms, bulletin boards and social network sites, it should be possible to extract patterns such as consumer reactions to products or world events. In order to address issues like these, new software has been developed by large companies like IBM's Web Fountain [96] and Microsoft's Pulse [97]. In addition, specialist web intelligence companies like Nielsen BuzzMetrics and Market Sentinel have been created or adapted.

A good example of a research initiative to harness consumer generated media (CGM) is an attempt to predict sales patterns for books based upon the volume of blog discussions of them [98]. The predictions had only limited success, however, perhaps because people often blogged about books after reading them, when it would be too late to predict a purchase. Other similar research has had less commercial goals. Gruhl et al. [99] analysed the volume of discussion for a selection of topics in blogspace, finding several different patterns. For example, some topics were discussed for one short period of time only, whereas others were discussed continuously, with or without occasional bursts of extra debate. A social sciences-oriented study sought to build retrospective timelines for major events from blog and news discussions, finding this to be possible to a limited extent [100, 101]. Problems occurred, for example, when a long running series of similar relatively minor events received little discussion but omitting them all from a timeline would omit an important aspect of the overall event.

In addition to the data mining style of research, there have been many studies of Web 2.0 sites in order to describe their contents and explain user behaviour in them. Here, research into social network sites is reviewed. A large-scale study of the early years of Facebook provides the most comprehensive overview of user activities. The data came from February 2004 to March 2006, when Facebook was a social network site exclusively for US college students [102]. Users seemed to fit their Facebooking with their normal pattern of computer use whilst studying, rather than allocating separate times. In terms of the geography of friendship, members mainly used Facebook to communicate with other students at the same college rather than school friends at distant universities. This suggests that social networking is an extension of offline communication rather than promoting radically new geographies of communication, although the latter is enabled by the technology of Facebook. This conclusion is supported by qualitative research into another popular site, MySpace [103, 104].

A webometric study of MySpace has indirectly investigated activity levels but focused on member profiles [105]. Amongst other findings, this showed that about a third of registered members accessed the site weekly and the average reported age was 21. Although other research found that MySpace close friends tended to reflect offline friendships [103], both male and female users preferred to have a majority of female friends [105]. Another study looked at the geography of friendship, finding that the majority of friends tended to live within a hundred miles, although a minority lived in the same town or city [106].

Finally, many statistics about Web 2.0 have been published by market research companies. Despite the uncertain provenance of this data, the results sometimes seem reasonable and also, because of the cost of obtaining the data, seem unlikely to be duplicated by academic researchers. An example is the announcement by HitWise that MySpace had supplanted Google as the most visited web site by US users by December 2007 [107]. The data for this was reported to come from two million US web users via an agreement between HitWise and the users' internet service providers. Making the results of overview analyses public gives useful publicity to HitWise and valuable insights to web researchers.

5. Conclusions and future prospects

5.1. Bibliometrics

Bibliometrics has changed out of all recognition since 1958, when it did not exist as a field or even as a coordinated group of researchers. Today it is taught widely in library and information science schools, and is at the core of a number of science evaluation research groups around the world, such as the Centre for Science and Technology Studies in the Netherlands. A number of bibliometric indicators are now internationally well known, principally the JIF, and bibliometrics are at least taken into account in a number of countries when making important policy decisions about the future of government funded

research. At the same time the state of the art for bibliometrics indicators has moved on so that most of the indicators that are well known and easy to calculate also have significant flaws in which practitioners will be well versed, but casual users may overlook. Hence one important task for bibliometric practitioners seems to be to convince policy makers of the importance of commissioning high quality robust indicators, as well as ensuring that no indicator is taken at face value.

Bibliometrics has also changed in the sense of expanding the number of data sources that can be drawn upon. Currently, Scopus and Google Scholar are the most important international bibliometric databases to challenge those of Thomson Scientific. More importantly, large-scale patent analysis is now much easier than before with the digitization and indexing of patent databases. This opens up an aspect of the commercial value of scientific research for bibliometric study.

Finally, bibliometrics has also changed by expanding the range of tasks investigated. In particular, the current wide range of relational bibliometric studies opens up new ways of understanding the scholarly communication process and the structure of science through citation relationships between journals, between scholars and between papers. Moreover, citation analysis in conjunction with visualization also helps to understand the structure of individual fields, and is particularly useful for emerging and rapidly developing important research areas, such as nanotechnology and biotechnology [42].

5.2. Webometrics

Webometrics research has been conducted by both information scientists and computer scientists, with different motivations. Within information science, webometrics has expanded from its initial focus on bibliometric-style investigations to more descriptive and social science-oriented research. It seems likely that webometric techniques will continue to evolve in response to new web developments, seeking to provide valuable descriptive results and perhaps also commercially applicable data mining techniques.

There are three main appeals of webometrics in contrast to traditional bibliometrics. First, the web can be timelier than the ISI databases. A typical research project might get funded, conduct research, report findings and then submit articles to journals. The time lag between the start of the project and the publication of the results in a journal is likely to be at least two years. Hence ISI-based bibliometrics is inevitably always retrospective, describing the research of years ago. In contrast, a research project might start by publishing a web site and could therefore be analysed with webometrics long before its research is published. The second advantage of the web is that it contains a wide range of scholarly-related artefacts, including presentations, patents, data, software and general web sites. Hence webometrics is potentially able to gather a wide range of evidence of research impact or connections. Finally, the web is free to access for all web users and so it potentially opens bibliometric-style analyses to those who could not access or afford ISI data.

Research into webometrics has also revealed many shortcomings, some of which are related to its advantages. First, the web is not quality controlled, unlike the ISI publication lists. Hence web data tends to be of lower quality, which means that webometric results are normally indicative rather than providing robust evidence. Second, web data is not standardized and so it is difficult to extract all except the simplest data (e.g. link counts). In particular, it is difficult to separate out the different types of publication. For example, there does not seem to be a simple way to separate out web citations in online journal articles from those in online course reading lists. Hence webometric results (e.g. link counts, web citation counts) tend to be the total of a mix of sources with variable value [e.g. 68, 108]. Third, although web data can be very timely, it can be impossible to find the publication date of a web page and so webometric results typically combine new and old web pages into one data set. Finally, web data is incomplete in several senses and in arbitrary ways. Although some academic articles are freely available online, the majority probably are not. Similarly, some researchers and research groups maintain extensive and comprehensive web sites but others do not. Hence the results reflect the web, which in turn is a very partial reflection of the activities of research.

Comparing the advantages and disadvantages of webometrics, it seems that it is unlikely to replace traditional bibliometrics but can be useful for several other purposes. First, it can be used for fast pilot studies to identify areas for follow-up systematic bibliometric analyses, e.g. [109].

Second, it can be used to assess the extent to which researchers are successful in publicizing their work online, given that this is an important activity. Third, it can be used for relational analyses of communication in disciplinary or geographic areas of science. Finally, its methods can help the analysis of Web 2.0 and online repositories for social sciences and humanities research goals.

Endnotes

1 PBRF, see www.tec.govt.nz
2 results published at http://evaluation.nrf.ac.za/Content/Facts/ratings.aspx
3 www.citebase.org
4 www.arXiv.org
5 www.umu.se/inforsk/Bibexcel/
6 http://users.fmg.uva.nl/lleydesdorff/software.htm
7 http://www.histcite.com
8 http://iv.slis.indiana.edu/sw/
9 http://cluster.cis.drexel.edu/~cchen/citespace/

References

[1] B. Thackray and H.B. Brock, Eugene Garfield: history, scientific information and chemical endeavour. In: B. Cronin and H.B. Atkins (eds) *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield* (Information Today, Medford, NJ, 2000) 11–23. [ASIS Monograph Series]
[2] E. Garfield, *Citation Indexing: Its Theory and Applications in Science, Technology and the Humanities* (Wiley Interscience, New York, 1979).
[3] R.K. Merton, *The Sociology of Science: Theoretical and Empirical Investigations* (University of Chicago Press, Chicago, 1973).
[4] C.L. Borgman and J. Furner, Scholarly communication and bibliometrics, *Annual Review of Information Science and Technology* 36 (2002) 3–72.
[5] G.K. Zipf, *Human Behavior and the Principle of Least Effort: an Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA, 1949)
[6] H.F. Moed, *Citation Analysis in Research Evaluation (Information Science and Knowledge Management)* (Springer, New York, 2005).
[7] L. Leydesdorff, Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science* 48(5) (1997) 418–27.
[8] S.C. Bradford, Sources of information on specific subjects. *Engineering: an Illustrated Weekly Journal* 137(26 January) (1934) 85–6.
[9] C. Oppenheim and S. Renn, Highly cited old papers and the reasons why they continue to be cited, *Journal of the American Society for Information Science* 29(5) (1978) 225–31.
[10] B. Cronin, The *Citation Process: the Role and Significance of Citations in Scientific Communication* (Taylor Graham, London, 1984).
[11] D. de Solla Price, A general theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science* 27(4) (1976) 292–306.
[12] R.K. Merton, The Matthew effect in science, *Science* 159(3810) (1968) 56–63.
[13] E. Garfield, Citation analysis as a tool in journal evaluation, *Science* 178(4060) (1972) 471–9.
[14] E. Garfield, *The Agony and the Ecstasy: the History and the Meaning of the Journal Impact Factor* (2005). Paper presented at the Fifth International Congress on Peer Review in Biomedical Publication, in Chicago, USA, 2005. Available at: http://garfield.library.upenn.edu/papers/jifchicago2005.pdf (accessed 27 September 2007)
[15] S.J. Bensman, Garfield and the impact factor, *Annual Review of Information Science and Technology* 41 (2007) 93–155.
[16] A. Cawkell, Visualizing citation connections. In: B. Cronin and H.B. Atkins (eds) *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield* (Information Today, Medford, NJ, 2000) 177–94. [ASIS Monograph Series]
[17] H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents, *Journal of the American Society for Information Science* 24(4) (1973) 265–9.

[18] I.V. Marshakova, System of document connections based on references, *Nauchno-Teknicheskaia Informatsiia* 2(1) (1973) 3–8.
[19] H.D. White and B.C. Griffith, Author co-citation: a literature measure of intellectual structure, *Journal of the American Society for Information Science* 32(3) (1982) 163–72.
[20] H.D. White, Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists, *Journal of the American Society for Information Science* 54(5) (2003) 423–34.
[21] J. Nicolaisen, Citation analysis, *Annual Review of Information Science and Technology* 41 (2007) 609–41.
[22] J.E. Hirsch, An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences* 102(46) (2005) 16569–72.
[23] C. Oppenheim, Using the *h*-index to rank influential British researchers in information science and librarianship, *Journal of the American Society for Information Science and Technology* 58(2) (2007) 297–301.
[24] B. Cronin and L.I. Meho, Using the *h*-index to rank influential information scientists, *Journal of the American Society for Information Science and Technology* 57(9) (2006) 1275–8.
[25] S. Harnad, Open access scientometrics and the UK Research Assessment Exercise. In: D. Torres-Salinas and H.F. Moed (eds) *Proceedings of 11th Annual Meeting of the International Society for Scientometrics and Informetrics* (CINDOC, Madrid, Spain, 2007) 27–33.
[26] V. Bence and C. Oppenheim, The influence of peer review on the Research Assessment Exercise, *Journal of Information Science* 30(4) (2004) 347–68.
[27] Tertiary Education Commission, *Performance-Based Research Fund – a Guideline for 2003* (2003). Available at: www.tec.govt.nz/upload/downloads/pbrffinal-july03.pdf (accessed 7 September 2007).
[28] DEST, *Institutional Grants Scheme* (n.d.) Available at: www.dest.gov.au/sectors/higher_education/programmes_funding/general_funding/operating_grants/institutional_grants_scheme.htm (accessed 12 September 2007).
[29] S.E. Cozzens, Assessing federally-supported academic research in the United States, *Research Evaluation* 9(1) (2000) 5–10.
[30] B. van der Meulen and A. Rip, Evaluation of societal quality of public sector research in the Netherlands, *Research Evaluation* 9(1) (2000) 11–25.
[31] M. Pienaar et al., The South African system of evaluating and rating individual researchers: its merits, shortcomings, impact and future, *Research Evaluation* 9(1) (2000) 27–36.
[32] A. Silvani, G. Sirilli and F. Tuzi, R&D evaluation in Italy: more needs to be done, *Research Evaluation* 14(3) (2005) 207–15.
[33] L. Butler, Explaining Australia's increased share of ISI publications – the effects of a funding formula based on publication counts, *Research Policy* 32(1) (2003) 143–55.
[34] J. Adams, Research assessment in the UK, *Science* 296(5569) (2002) 805.
[35] L.I. Meho and K. Yang, Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar, *Journal of the American Society for Information Science and Technology* 58(13) (2007) 2105–25.
[36] M. Norris and C. Oppenheim, Comparing alternatives to the Web of Science for coverage of the social sciences literature, *Journal of Informetrics* 1(1) (2007) 161–9.
[37] K. Kousha and M. Thelwall, Google Scholar citations and Google Web/URL citations: a multi-discipline exploratory analysis, *Journal of the American Society for Information Science and Technology* 58(7) (2007) 1055–65.
[38] P. Jacsó, Google Scholar: the pros and the cons, *Online Information Review* 29(2) (2005) 208–14.
[39] C. Chen, *Information Visualization: Beyond the Horizon*, 2nd Edition (Springer, New York, 2004).
[40] H. Small, Visualising science through citation mapping, *Journal of American Society for Information Science* 50(9) (1999) 799–813.
[41] K. Boyack, Using detailed maps of science to identify potential collaborations. In: D. Torres-Salinas and H.F. Moed (eds), *Proceedings of ISSI 2007 Volume 1* (CSIC, Madrid, 2007) 124–35.
[42] L. Leydesdorff, Betweenness centrality as an indicator of the interdisciplinarity of scientific journals, *Journal of the American Society for Information Science & Technology* 58(9) (2007) 1303–19.
[43] C. Chen, CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of the American Society for Information Science and Technology* 57(3) (2006) 359–77.
[44] M. Gibbons et al., *The New Production of Knowledge* (Sage, London, 1994).
[45] M. Meyer, Academic patents as an indicator of useful research? A new approach to measure academic inventiveness, *Research Evaluation* 12(1) (2003) 17–27.

[46] C. Oppenheim, Do patent citations count? In: B. Cronin and H.B. Atkins (eds) *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield* (Information Today, Medford, NJ, 2000) 405–32. [ASIS Monograph Series]
[47] L. Leydesdorff, The university-industry knowledge relationship: analyzing patents and the science base of technologies, *Journal of the American Society for Information Science and Technology* 54(11) (2004) 991–1001.
[48] H.-R. Ke et al., Exploring behavior of e-journal users in science and technology: transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan, *Library & Information Science Research* 24(3) (2002) 265–91.
[49] K. Marek and E.J. Valauskas, Web logs as indices of electronic journal use: tools for identifying a 'classic' article, *Libri* 52(4) (2002) 220–30.
[50] M.J. Kurtz et al., The bibliometric properties of article readership information, *Journal of the American Society for Information Science & Technology* 56(2) (2005) 111–28.
[51] S. Jones et al., A transaction log analysis of a digital library, *International Journal on Digital Libraries* 3(2) (2000) 152–69.
[52] P. Huntington, D. Nicholas and H.R. Jamali, Site navigation and its impact on content viewed by the virtual scholar: a deep log analysis, *Journal of Information Science* 33(5) (2007) 598–610.
[53] P.M. Davis and M.J. Fromerth, Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics* 71(2) (2007) 203–15.
[54] T.S. Brody, S. Harnad and L. Carr, Earlier web usage statistics as predictors of later citation impact, *Journal of the American Society for Information Science and Technology* 57(8) (2006) 1060–72.
[55] L. Björneborn and P. Ingwersen, Toward a basic framework for webometrics, *Journal of the American Society for Information Science and Technology* 55(14) (2004) 1216–27.
[56] T.C. Almind and P. Ingwersen, Informetric analyses on the World Wide Web: methodological approaches to 'Webometrics', Journal of Documentation 53(4) (1997) 404–26.
[57] P. Mayr and F. Tosques, *Google Web APIs: an Instrument for Webometric Analyses?* (2005) Available at: www.ib.hu-berlin.de/%7Emayr/arbeiten/ISSI2005_Mayr_Toques.pdf (accessed 7 January 2008).
[58] I.F. Aguillo et al., Scientific research activity and communication measured with cybermetrics indicators, *Journal of the American Society for Information Science and Technology* 57(10) (2006) 1296–1302.
[59] P. Ingwersen, The calculation of Web Impact Factors, *Journal of Documentation* 54(2) (1998) 236–43.
[60] B. Cronin, Bibliometrics and beyond: some thoughts on web-based citation analysis, *Journal of Information Science* 27(1) (2001) 1–7.
[61] M. Thelwall, Extracting macroscopic information from web links, *Journal of the American Society for Information Science and Technology* 52(13) (2001) 1157–68.
[62] O. Thomas and P. Willet, Webometric analysis of departments of librarianship and information science, *Journal of Information Science* 26(6) (2000) 421–8.
[63] F. Barjak and M. Thelwall, A statistical analysis of the web presences of European life sciences research teams, *Journal of the American Society for Information Science and Technology* 59(4) (2008) 628–43.
[64] F. Barjak, X. Li and M. Thelwall, Which factors explain the web impact of scientists' personal home pages? *Journal of the American Society for Information Science and Technology* 58(2) (2007) 200–211.
[65] M. Thelwall and G. Harries, Do better scholars' Web publications have significantly higher online impact? *Journal of American Society for Information Science and Technology* 55(2) (2004) 149–59.
[66] M. Thelwall, *Link Analysis: an Information Science Approach* (Academic Press, San Diego, 2004).
[67] G. Heimeriks, M. Hörlesberger and P. van den Besselaar, Mapping communication and collaboration in heterogeneous research networks, *Scientometrics* 58(2) (2003) 391–413.
[68] G. Harries et al., Hyperlinks as a data source for science mapping, *Journal of Information Science* 30(5) (2004) 436–47.
[69] X. Li et al., National and international university departmental web site interlinking, part 2: link patterns, *Scientometrics* 64(2) (2005) 187–208.
[70] N. Payne and M. Thelwall, A longitudinal study of academic webs: growth and stabilisation, *Scientometrics* 71(3) (2007) 523–39.
[71] N. Payne, *A Longitudinal Study of Academic Web Links: Identifying and Explaining Change* (University of Wolverhampton, Wolverhampton, 2007)
[72] L. Vaughan and D. Shaw, Bibliographic and web citations: what is the difference? *Journal of the American Society for Information Science and Technology* 54(14) (2003) 1313–22.
[73] L. Vaughan and D. Shaw, Web citation data for impact assessment: a comparison of four science disciplines, *Journal of the American Society for Information Science & Technology* 56(10) (2005) 1075–87.

[74] K. Kousha and M. Thelwall, Motivations for URL citations to open access library and information science articles, *Scientometrics* 68(3) (2006) 501–17.
[75] J.Bar-Ilan, The use of Web search engines in information science research. *Annual Review of Information Science and Technology* 38 (2004) 231–88.
[76] S. Lawrence and C.L. Giles, Accessibility of information on the web, *Nature* 400(6740) (1999) 107–9.
[77] L. Introna and H. Nissenbaum, Shaping the web: why the politics of search engines matters, *The Information Society* 16(3) (2000) 1–17.
[78] E. Van Couvering, New media? The political economy of Internet search engines. In: *Annual Conference of the International Association of Media & Communications Researchers* (Porto Alegre, Brazil, 2004). Available at: http://personal.lse.ac.uk/vancouve/IAMCR-CTP_SearchEnginePoliticalEconomy_EVC_2004-07-14.pdf (accessed 7 January 2008).
[79] L. Vaughan and M. Thelwall, Search engine coverage bias: evidence and possible causes, *Information Processing and Management* 40(4) (2004) 693–707.
[80] J. Bar-Ilan and B.C. Peritz, Evolution, continuity, and disappearance of documents on a specific topic on the Web: a longitudinal study of 'informetrics', *Journal of the American Society for Information Science and Technology* 55(11) (2004) 980 –90.
[81] M. Thelwall, Extracting accurate and complete results from search engines: case study Windows Live, *Journal of the American Society for Information Science and Technology* 59(1) (2008) 38–50. Available at: www.scit.wlv.ac.uk/%7Ecm1993/papers/2007_Accurate_Complete_preprint.doc (accessed 22 May 2007).
[82] H.W. Snyder and H. Rosenbaum, Can search engines be used for Web-link analysis? A critical review, *Journal of Documentation* 55(4) (1999) 375–84.
[83] J. Bar-Ilan, Search engine results over time – a case study on search engine stability, *Cybermetrics* (1999). Available at: www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html (accessed 7 January 2008).
[84] W. Mettrop and P. Nieuwenhuysen, Internet search engines – fluctuations in document accessibility, *Journal of Documentation* 57(5) (2001) 623–51.
[85] R. Rousseau, Daily time series of common single word searches in AltaVista and NorthernLight, *Cybermetrics* 2/3 (1999). Available at: www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html (accessed 25 July 2006).
[86] A.G. Smith, Does metadata count? A Webometric investigation. In: M. Tegelaars (ed.), *Proceedings of DC-2002, Florence, 14–17 October 2002* (Firenze University Press, Firenze, 2002) 133–8.
[87] T. Craven, Variations in use of meta tag keywords by web pages in different languages, *Journal of Information Science* 30(3) (2004) 268–79.
[88] A. Broder et al., Graph structure in the web, *Journal of Computer Networks* 33(1/6) (2000) 309–20.
[89] L. Björneborn, *Small-world link structures across an academic web space – a library and information science approach*. PhD thesis, Department of Information Studies (Royal School of Library and Information Science, Copenhagen, Denmark, 2004).
[90] L. Björneborn, 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space, *Scientometrics* 68(3) (2006) 395–414.
[91] A.L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* 286(5439) (1999) 509–12.
[92] S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine, *Computer Networks and ISDN Systems* 30(1/7) (1998) 107–17.
[93] D. Pennock et al., Winners don't take all: characterizing the competition for links on the web, *Proceedings of the National Academy of Sciences* 99(8) (2002) 5207–11.
[94] W. Koehler, A longitudinal study of Web pages continued: a report after six years, *Information Research* 9(2). Available at: http://informationr.net/ir/9–2/paper174.html (accessed 20 September 2007).
[95] J.L. Ortega, I. Aguillo and J.A. Prieto, Longitudinal study of content and elements in the scientific web environment, *Journal of Information Science* 32(4) (2006) 344–51.
[96] D. Gruhl et al., How to build a WebFountain: an architecture for very large-scale text analytics, *IBM Systems Journal* 43(1) (2004) 64–77.
[97] M. Gamon et al., Pulse: mining customer opinions from free text (IDA 2005), *Lecture Notes in Computer Science* 3646 (2005) 121–32.
[98] D. Gruhl et al., The predictive power of online chatter. In: R.L. Grossman et al. (eds), *KDD '05: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (ACM Press, New York, 2005) 78–87.

[99] D. Gruhl et al. Information diffusion through Blogspace. In: *WWW2004, New York*. (2004). Available at: www2004.org/proceedings/docs/1p491.pdf (accessed 10 July 2006).
[100] M. Thelwall, R. Prabowo and R. Fairclough, Are raw RSS feeds suitable for broad issue scanning? A science concern case study, *Journal of the American Society for Information Science and Technology* 57(12) (2006) 1644–54.
[101] M. Thelwall and R. Prabowo, Identifying and characterizing public science-related concerns from RSS feeds, *Journal of the American Society for Information Science & Technology* 58(3) (2007) 379–90.
[102] S.A. Golder, D. Wilkinson and B.A. Huberman, Rhythms of social interaction: messaging within a massive online network. In: *3rd International Conference on Communities and Technologies (CT2007), East Lansing, MI*, (2007) Available at: www.hpl.hp.com/research/idl/papers/facebook/facebook.pd (accessed 7 January 2008).
[103] d. boyd, Friendster and publicly articulated social networks. In: *Conference on Human Factors and Computing Systems (CHI 2004, Vienna: April 24–29)* (ACM Press, New York, 2004). Available at: www.danah.org/papers/CHI2004Friendster.pdf (accessed 3 July 2007).
[104] d. boyd, Friends, Friendsters, and MySpace Top 8: writing community into being on social network sites, *First Monday* 11(2) (2006). Available at: www.firstmonday.org/issues/issue11_12/boyd/index.html (accessed 23 June 2007).
[105] M. Thelwall, Social networks, gender and friending: an analysis of MySpace member profiles. *Journal of the American Society for Information Science and Technology* (forthcoming). Available at: www.scit.wlv.ac.uk/~cm1993/papers/MySpace_d.doc (accessed 23 August 2007).
[106] T. Escher, *The Geography of (Online) Social Networks (Web 2.0, York University)* (2007). Available at: http://people.oii.ox.ac.uk/escher/wp-content/uploads/2007/09/Escher_York_presentation.pdf (accessed 18 September 2007).
[107] L. Prescott, *Hitwise US Consumer Generated Media Report* (2007). Available at: www.hitwise.com/ (accessed 19 March 2007).
[108] D. Wilkinson et al., Motivations for academic Web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication, *Journal of Information Science* 29(1) (2003) 49–56.
[109] S. Robinson et al., *The Role of Networking in Research Activities (NetReAct D4.1)* (Empirica Gesellschaft für Kommunikations- und Technologieforschung mbH, Bonn, Germany, 2006).