Jointly published by Kluwer Academic Publishers, Dordrecht and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 50, No. 1 (2001) 65–82

Perspectives of webometrics

LENNART BJÖRNEBORN, PETER INGWERSEN

Center for Informetric Studies (CIS), Department of Information Studies, Royal School of Library and Information Science, Copenhagen (Denmark)

Since the mid-1990s has emerged a new research field, *webometrics*, investigating the nature and properties of the Web drawing on modern informetric methodologies. The article attempts to point to selected areas of webometric research that demonstrate interesting progress and space for development as well as to some currently less promising areas. Recent investigations of search engine coverage and performance are reviewed as a frame for selected quality and content analyses. Problems with measuring Web Impact Factors (Web-IF) are discussed. Concluding the article, new directions of webometrics are outlined for performing knowledge discovery and issue tracking on the Web, partly based on bibliometric methodologies used in bibliographic and citation databases. In this framework graph theoretic approaches, including path analysis, transversal links, "weak ties" and "small-world" phenomena are integrated.

Introduction

Since the mid-1990s increasing efforts have been made to investigate the nature and properties of the World Wide Web, named the Web in this article, by applying modern informetric methodologies to its space of contents, link structures, and search engines. Studies of the Web have been named "webometrics" by *Almind* and *Ingwersen* (1997) or "cybermetrics" as in the electronic journal of that name (1997). This article attempts to point to selected areas of webometric research that demonstrate interesting progress and space for development as well as to some currently less promising areas. The contribution is not an exhaustive review, but rather a view on the specialty.

Webometrics displays several similarities to informetric and scientometric studies and the application of common bibliometric methods. For instance, simplistic counts and content analysis of web pages are like traditional publication analysis; counts and analyses of outgoing links from web pages, here named outlinks, and of links pointing to web pages, called inlinks, can be seen as reference and citation analyses, respectively. Outlinks and inlinks are then similar to references and citations, respectively, in scientific articles. However, due to its dynamic and distributed nature, the Web often demonstrates web pages simultaneously linking to each other – a case not possible in the

0138–9130/2001/US $ 15.00

traditional paper-based citation world. The coverage of search engines of the total Web can be investigated in the same way as the coverage of domain and citation databases in the total document landscape and possible overlaps between engines detected. Since the Web consists of contributions from anyone who wishes to contribute, the quality of information or knowledge value is opaque due to the lack of kinds of peer reviewing; but citation-like link analyses may reveal clusters of sites to be reviewed. Patterns of Web search behaviour can be investigated as in traditional information seeking studies. Issue tracking on the Web is carried out and knowledge discovery attempts are made, similar to common data or text mining in administrative or textual (bibliographic) databases.

Since the Web is an information space quite different from the common scientific or professional databases, the similarities mentioned above may sometimes be superficial. For example, we do not know for sure why people on the Web link up to other pages. There exists no convention of citation as in the scientific world. Further, time plays a different role on the Web. On the other hand, because the Web is a highly complex conglomerate of all types of information carriers produced by all kinds of people and searched by all kinds of users, it is tempting to investigate; and informetrics indeed offers some methodologies to start from. However, one must be aware that like for the online application of the ISI controlled citation databases, for instance by means of the Dialog command language, *data collection* on the Web depends on the retrieval features of the various search engines and web robots. Prior to the appearance of the "set postings on" command feature in Dialog during the 1990s, online citation counts were not possible; one would have to download all the citing documents to be analysed locally for the actual number of citations – within the ISI-defined information space. At present this is exactly the case in most Web engines, as demonstrated by *Rousseau* (1997; 1999). The engines do not index the entire Web, their overlaps are not substantial (*Lawrence* and *Giles*, 1998), and their retrieval features too simplistic for extensive webometric analyses online. Sampling is thus crucial but difficult to perform, and filtering becomes necessary. Hence, there is a strong element of re-engineering and clean-up in webometric analyses.

The article points to some recent investigations of Web engine coverage and performance as a frame for selected quality and content analyses. We then look into sitation analyses, i.e., link-page analysis in terms of *Rousseau* (1997) and Web Impact Factor (Web-IF) studies. Attempts to provide new exciting directions of webometrics by performing knowledge discovery and issue tracking, for instance by means of transversal link structures and "weak ties", are discussed, concluding the article.

Web engine coverage and quality studies

Lawrence and *Giles* (1998) provide a substantial contribution with respect to the commercial search engine coverage of the Web space by introducing the concept of "indexable Web". The concept signifies the portion of the Web that can be indexed by the engines, excluding documents from Web-databases like Dialog. A comprehensive test run on December, 1997, between 6 major commercial search engines including AltaVista, HotBot, NorthernLight, Excite, Lycos, and Infoseek, showed a lower bound of the indexable Web to be 320 million pages. The study also demonstrates that the coverage of any one engine is significantly limited by indexing only up to a third of the indexable Web. There may be many reasons for this result. For instance, the depth (exhaustiveness) of indexing at the local servers visited by the engines, which depends on the site structure and organisation, may influence the retrieval output, as may the method of indexing query words, e.g., by imposed truncation of long web documents. Other attempts to evaluate Web engines have been carried out, for instance, to observe the quality of the ranked lists of web documents retrieved by major engines (*Courtois* and *Berry*, 1999). Aside from the findings* the paper discusses the more or less publicly available knowledge about the different indexing/retrieval features used by any one engine. The methodology of evaluation of Web engines is addressed by *Clarke* and *Willett* (1997). They compared AltaVista, Excite and Lycos. In addition, that paper offers a critical evaluation of earlier research and provides a realistic methodology, including relative recall measures. It was found that AltaVista performed significantly better than Lycos and Excite. *Oppenheim* et al. (2000) provides a detailed up-to-date review on the evaluation of Web search engines, including a discussion of test methodologies.

While many coverage and evaluation studies look into the relevance and amount of web pages at a given point in time, other critical analyses cover link-page retrieval (*Snyder* and *Rosenbaum*, 1999) or carry out Web coverage or structure investigations based on time series. Like *Ingwersen* (1998), *Snyder* and *Rosenbaum* observed large variations and inconsistency, in particular concerning the AltaVista engine's link-page recovery. The irregularities of that engine has also been reported by *Bar-Ilan* (1999) in a longitudinal study as well as by *Rousseau* (1999) who compared AltaVista and NorthernLight on a day-to-day basis over 21 weeks during 1999. The latter study used the same three common single words as queries during the evaluation period.

* Relevant pages among the top-20 or top-100 ranked documents were such that formally contain *all* query words. No expert assessments were used.

Where NorthernLight, as expected, showed a steady increase of hits, in line with the Web growth, AltaVista demonstrated large variations over time until the particular date (October, 25, 1999) when it became re-launched in a new and more stable form. At that date the number of retrieved web pages increased dramatically – with this *nova*-like effect depending on the query (*Rousseau*, 1999, p. 5) – later to drop a bit probably due to removal of dead link pages. *Rousseau* proposes still to apply a median filter to reduce the effect of result variation in that particular engine. Another result of the study is that the Web Impact Factor (Web-IF) results published by *Ingwersen* (1998) most probably are highly doubtful, since his data collection results for both web pages and inlink pages derive from the "old" and unstable AltaVista version. The reason why focus is put on AltaVista is that the engine has a large Web coverage and (thus far) provides search features suitable for informetric studies of the Web. Time series seem thus very useful as a tool when monitoring Web engine performance.

Web page properties and quality

According to *Cronin* and *McKim* "the Web is reshaping the ways in which scholars communicate with one another. New kinds of scholarly and proto-scholarly publishing are emerging. Thanks to the Web, work-in-progress, broadsides, early drafts and refereed articles are now almost immediately sharable … with authors able to choose between narrowcasting and broadcasting. And peer review has emerged from the closet to reveal a spectrum of possibilities…" (1996, p. 170). This belief and vision is already reality. Hence, webometric analyses of the nature, structures and content properties of web sites and pages, as well as link structures are important in order to understand the virtual highway and their interconnections. *Larson* was one of the first information scientists to perform an exploratory analysis of the intellectual structure of cyberspace (1996). Shortly after, *Almind* and *Ingwersen* (1997) applied a variety of bibliometric-like methods to the Nordic portion of the Web in order to observe the kinds of page connections and define the typology of web pages actually found at national Nordic level. The methodology involved stratified sampling of web pages and download for local analysis purposes. Among the interesting findings the analyses revealed that each web page, capable of outlinking, on average provides 9 outlinks – a proportion which nowadays still holds in the exponentially growing Web space. The contribution also attempted a comparison between the estimated share of scientific web pages and the distribution found in the citation indexes between the Nordic countries. Clearly, the visibility on the Web was quite different from that displayed in the citation databases. Norway, for instance, was much more visible on a Web scale than in the printed world

at the time of analysis. A special interest dedicated personal home pages is demonstrated by *Bates* and *Lu* (1997) as well as *Wynn* and *Katz* (1997), mainly concentrating on a US cyberspace platform – later followed up by, for example, *Dillon* and *Gushrowski* (2000).

Obviously, the breakthrough for everybody to express themselves, practically without control from authorities, to become visible world wide, also by linking to what pages one wants to link to, to assume credibility by being "there", and to obtain access to data, information, values and knowledge in many shapes and degrees of truth, has generated a reality of freedom of information – also in regions and countries otherwise poor of infrastructure. The other side of the coin is that the Web increasingly becomes a web of uncertainty to its users; the thin red line between opaqueness, shading truth, misinformation, beliefs, opinions, visions or speculation *and* reliability, validity, quality, relevance or truth becomes increasingly thinner. It becomes a matter of trust. "Web archaeology" will in future go hand in hand with webometric analyses and methods.

Quality watch and assessments are currently in high demand. In particular, the health and medical domains are important areas to investigate for such issues. Recently *Lei Cui* made use of citation analysis methods on the Web to detect the overlapping and high frequency inlinked (cited) sites on medical information (1999) and *Allen* et al. looked into the reliability (and pertinence) of bio-related web pages (1999). In *Cui*'s paper one is referred to several other recent studies of health and art issues treated on the Web for which Web citation analyses have been applied as a rudimentary quality indicator, e.g. *Eysenbach* (1998). The Bradford distribution of the thousands of links pointing out from the 25 top medical US Schools is used as strong ties by *Cui* to determine the central sites concerned with specific health topics. The *Allen* et al. contribution is an expert-assessed survey of the reliability of scientific web sites. As was the case for *Rousseau*'s longitudinal study mentioned above (1999) the survey is based on the retrieval of sites according to three exemplary queries on 1) evolution, 2) genetically modified organism, and 3) endangered species. For each query the first 500 web sites were examined sequentially by two expert referees until each had independently reviewed approximately 60 sites containing information pertinent to the *topic*. This methodology is close to that used in the current worldwide TREC IR evaluation experiments. These 60 topical sites per query were then scored as "inaccurate" if they contained factually incorrect information, "misleading" if they misinterpreted science or blatantly omitted facts supporting an opposing position, and "un-referenced" if they presented information without any peer reviewed references. The latter score is purely objective. The overall agreement values for the referees' scores for the categories of "inaccurate" and "misleading" were 87.8 % for the Evolution sites, 82.8 % for the Genetically modified

organism sites, and 73.6 % for the Endangered species web sites retrieved and assessed. Un-referenced sites accounted for more that 48 % for each query (p. 722). These results give cause for doubt as to the trustworthiness of the information and points to the exploitation of portals that provide outlinks to sites that have been reviewed by scientists for accuracy, relevance and currency. Such portals and indeed entire digital libraries may then – as in the traditional textual scientific databases – offer valid peer reviewed information and act as platforms for webometric analyses.

Web Impact Factor studies

In his classic webometric article on "sitations", that is inlinks, *Rousseau* (1997) analyses the patterns of distribution of web sites and incoming links. Although *Rousseau*, like later *Ingwersen* (1998) made use of the "old" version of AltaVista, his study operated with 343 downloaded sites (named data points) retrieved from a query on informetrics + bibliometrics + scientometrics. The analyses are hence more independent of the Web engine characteristics and more robust. The study shows that the distribution of top level domains for the sites follows the ubiquitous Lotka distribution. Similarly, *Rousseau* demonstrates that the distribution of sitations to those 343 sites also follow a Lotka distribution. The proportion of self-sitations was estimated to 30 %.

The difference between links and link pages is illustrated by *Ingwersen* (1998) in his attempt to calculate the Web Impact Factors (Web-IF) for national domains and individual sites.* The underlying idea was that the Web-IF might inform about the awareness or recognition of national sites (on average) or individual sites. The study found three interesting results. 1) Since the AltaVista search engine cannot count the actual number of inlinks to particular sites, but only the number of pages carrying an inlink (or sitation) at least once, the self-linking will not influence the overall Web-IF. The external inlinking hence becomes the important score to observe. The average domain self-link score in *Ingwersen*'s study was approximately 0.5 with the ".com"-domain at 0.59. On average the external link-page IF was 0.39. 2) The Web-IFs for individual web sites were more unreliable than that of the domains. 3) The variance in the calculations could be applied as a *Web engine evaluation measure*. The latter issue of variance also suggested using quite complex methods for calculating the IF and the introduction of detailed query formulations. As demonstrated above by *Rousseau* (1999) the AltaVista engine at the time of Web-IF analysis was indeed unstable compared to

* Note that prior to *Ingwersen*, *Rodriguez i Gairin* (1997) had introduced the concept of information impact on the Internet in a Spanish documentation journal.

later versions from October 1999. Hence, the Web-IF calculation may function as an indicator of engine performance. The reason for applying the AltaVista engine at all was its coverage and retrieval command abilities to search for domain pages in a controlled manner as well as for link-pages.

In connection to the second result in *Ingwersen*'s study on the high variations of individual site Web-IF *Smith* (1999) as well as *Thelwall* (2000) further investigated this phenomenon, unfortunately still applying the unstable AltaVista version (1999). However, exactly due to the observed variations they both became suspicious as to the coverage and retrieval properties of the engine(s). Had the results been stable etc. one might not necessarily immediately have questioned the methodology. *Smith* (1999) demonstrates some periodic and robust data collection methods and shows how results become distorted due to retrieval of noise pages, e.g., Indonesia (domain code: .id) shows very high Web-IF because of the retrieval of the URL element "id" in many sites other than Indonesian. He also shows that the longer the URL-string searched for, the more reliable the result. The context of the string should assure its uniqueness. However, later unpublished studies of the actual coverage of the engines – including AltaVista – with respect to our own known pages and links *on our local server* (ax.db.dk) demonstrate that they do not penetrate to all pages and links. This negative result is confirmed by *Thelwall* (2000) who applied AltaVista, Hotbot and Infoseek in the analysis. The coverage is not random in such a way that the Web-IF denominator and numerator are influenced in identical ways. In short, at the present state of search engine coverage and retrieval modes, "the exiting concept of Web-IF appears to be a relatively crude instrument in practice" (*Thelwall,* 2000; p. 188). Thus far, the outcome is highly problematic and, as stated both by *Rousseau* (1999), *Smith* (1999) and *Thelwall* (2000), one would have to apply dedicated web robots to download samples for local analyses.

Knowledge discovery and issue tracking on the Web

In the last decade has emerged a multidisciplinary research field labelled Knowledge Discovery in Databases, or KDD. This field is concerned with developing methods to exploit the exponentially growing reservoir of contents registered in databases with business, administrative, scientific and other types of data. *Frawley* et al. (1991) define KDD as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". In order to identify and extract new patterns and relationships, that could yield new knowledge, KDD uses a broad variety of methods. Combining computer power and human expertise, techniques of KDD include, for

instance, information retrieval, statistics, machine learning, pattern recognition, multidimensional scaling and visualisation. The objectives and methodologies of KDD and bibliometrics thus have many points in common. For example, bibliometric clustering techniques could be viewed as an application of KDD in bibliographic and citation databases. There are also approaches in KDD research incorporating bibliometrics, for instance, in the area of textual data mining (*Losiewics* et al., 2000). The concept of "data mining" is used in relation to, sometimes synonymously with, KDD. KDD refers to the overall process of discovering useful knowledge from data, whereas data mining is a particular step in this process focusing on pattern recognition (*Fayyad* et al., 1996). Areas that apply KDD include, e.g., consumer behaviour, stellar surveys, cancer diagnosis, chemical structure identification, population analysis, quality control, and modelling of global climatic change (*Vickery*, 1997).

The Web is an obvious environment for applying knowledge discovery as argued by *Etzioni* (1996). The Web can be conceived as an exponentially growing distributed database, containing now in its indexable part well over one billion web pages (not including database request-generated document formats) and roughly 10 billions links. Besides containing distributed data on millions of servers, the Web has other dimensions that differentiate it from ordinary databases. Most importantly, the Web is multi-agent constructed. Millions of diverse web actors, such as laymen, researchers, institutions, etc., dynamically create, adapt and remove web pages and links. 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". And, as indicated above by *Allen* et al. in their study (1999) the Web information may be unreliable. On the other hand, the heterogeneity of the Web can be a fertile source for making discoveries. As stated by *de Jong* and *Rip* (1997), discoveries often result from "making *unexpected* combinations of heterogeneous resources" (italics in original), implying that it is not possible in advance to tell what resources are required.

There are three main directions to perform knowledge discovery on the Web. They are concerned with exploiting 1) web page contents, 2) link structures, and 3) users' information behaviour (i.e., searching and browsing). The focus in this section is on the exploitation of link structures for knowledge discovery on the Web. This approach has close kinship with bibliometric citation analysis, but not only by means of strong ties.

Links weave web documents together in a complex, structured hypertext corpus. Link structures represent implicit human "annotations" that can be exploited for knowledge discovery, for example, inferring web communities (*Gibson* et al., 1998), identifying authoritative web pages (*Kleinberg*, 1998; *Cui*, 1999), topic distillation (*Bharat* and *Henzinger*, 1998), or improving search engine ranking algorithms (*Brin* and

Page, 1998). The creator of the WWW at CERN in 1989-1990, *Berners-Lee*, envisaged this development. One important incentive for him to develop the Web was the possibility to keep track of "the complex web of relationships between people, programs, machines and ideas" (*Berners-Lee*, 1997).

Graph theoretic methods are excellent tools for investigating link structures on the Web. The topology of link structures affects possibilities for human and digital agents to traverse and explore the Web – and make discoveries.

Graph theory and the Web

In graph theory a graph is a mathematical representation of a network consisting of vertices (or nodes) connected by edges. The nodes can be humans (in social networks), ecological actors (in a food web), Internet servers, documents (in a citation network), concepts (in a thesaurus or semantic network), etc. In a directed graph the edges represent directional relations between the nodes. The Web is an example of a directed graph with web pages corresponding to nodes and hyperlinks to edges.

Graph theoretic methods can be used to analyse structural aspects of the Web. *Broder* et al. (2000) give a "bow tie"-looking model of the graph structure of the Web. Using a web crawl consisting of roughly 200 million web pages and 1.5 billion links, they built a database model of a Web graph. The research team showed that the Web consists of five distinct regions characterised by whether nodes have just outlinks, just inlinks, both link types or no links at all – and whether nodes are connected to the "bow tie knot" (consisting of nodes with both in- and outlinks) or not. The "bow tie"-model gives a valuable understanding of the intricate structure of the Web. This intricacy affects possibilities for knowledge discovery when human or digital agents explore and analyse link structures.

In the graph theoretic framework an intriguing dimension of link structures deals with so-called "small-world" phenomena and "small-world" networks. These phenomena could enhance possibilities for knowledge discovery.

"Small-world" networks

In "small-world" networks nodes are highly clustered as in regular graphs, yet the path lengths between any pairs of nodes are short as in random graphs. In a "small-world" network it is sufficient with a very little percentage of links functioning as "short cuts" connecting "distant" parts of the network. *Watts* and *Strogatz* (1998), elaborated in *Watts* (1999), showed that "small-world" topology in the shape of short path lengths

occur in biological, technological and social networks, for example, the neural network of a nematode worm, the electrical power grid of the western United States, and the collaboration graph of film actors. The "small world" theory stems from work by *Milgram* (1967) and *Kochen* (1989), popularised by the notion of "six degrees of separation" concerning short distances between two arbitrary persons through intermediate chains of acquaintances.

There is still a lack of research in Library and Information Science on "small-world" phenomena concerning short distances and their consequences in informational networks such as the WWW, citation databases, semantic networks, associative thesauri, etc. There can be possible "small-world" phenomena when *nodes* in an informational network are defined as corresponding either to documents, terms, authors, cited authors, journals, scientific domains, institutions, or countries, etc., and distance-reducing, "small-world"-creating *links* correspond to either references, related terms, co-term occurrence, descriptors, co-authors,* co-cited authors or journals, etc.

One "small-world" consequence could be the enhanced possibility of discovering unexpected but useful information on the Web. In a current project one of the authors is developing this approach amalgamating graph theory and bibliometrics. A central concept in this framework is so-called "transversal links" (*Björneborn*, 2000).

Transversal links

Transversal links function as short cuts between heterogeneous web clusters. Web clusters consist of closely interlinked web pages and web sites, reflecting cognate subject domains and interest communities. Web pages within a web cluster can be more or less centrally or peripherally positioned, dependent on the intra-cluster link structure. An example of a web cluster could be researchers making links from their personal home pages to other researchers, institutions, projects and papers within their own scientific discipline.

A human or digital agent exploring the Web by following links from web page to web page has the possibility to move from one web cluster to another "distant" cluster using a single transversal link as a short cut. In the above example, a researcher in

* Cf. the so-called Erdős-numbers. Before his death in 1996 the Hungarian mathematician Paul Erdős wrote almost 1500 papers with 472 different co-authors (*Bar-Ilan*, 1998). Erdős-numbers are measured as the number of intermediate co-authorships between a given scientist and Erdős. For example, a scientist has Erdős-number 2 if he/she has been co-author to someone being co-author with Erdős. (Erdős had Erdős-number 0). Cross-disciplinary co-authorships could create "small-world" phenomena in the shape of short paths between authors (and their scientific disciplines) through steps of intermediate co-authors.

information science could have made a transversal link on a web page with favourite links targeted to another of his interest fields, "creativity stimulation", thus creating a trail between two heterogeneous web clusters. Such divergent trails running transversely to well-travelled paths on the Web can affect possibilities for human serendipity and computer-supported knowledge discovery, where unexpected but potentially useful information is encountered and extracted.

The creative "knowledge discovery" potential of link structures has played an influential role in the development of hypertextual information systems, including the Web. In this historical process an important catalyst has been *Vannevar Bush*'s (1945) seminal vision of Memex, an information system, where disparate text paragraphs (on microfilm) could be connected by so-called "trails" (i.e. links) transversely to classificational hierarchies in order to stimulate innovativeness of researchers. The cross-linking dimension of Memex – and of transversal links – can be seen as analogous to how ideas diffuse between heterogeneous social groups through peripherally, weakly affiliated persons, so-called "weak ties" (*Granovetter*, 1973). Transversal links may also be seen as to how academic authors often cite a few sources outside their own scientific domains, so-called "boundary crossings" (*Klein*, 1996; *Pierce*, 1999). Transversal links function as "weak ties" and "boundary crossings" between heterogeneous web clusters. Of course, many transversal links reflect idiosyncrasy, for example, personal non-scientific hobbies. But then again, other transversal links on scientists' web pages could reflect emerging "research fronts" in scientific domains, or cross-disciplinary "invisible colleges". Revealing such "hidden" connections – by human serendipity or by computer-supported knowledge discovery – could render useful information about new directions in the evolving interconnectedness of science, discovering new relationships and patterns. Transversal links crossing scientific boundaries could provide creative insights, thus giving a new signification to the notion of "the strength of weak ties" from social network analysis (*Granovetter*, 1973).

Methodological considerations

Trying to localise transversal links on the Web is concerned with handling data at the low-frequent end of a probably Bradford-like distribution of target web pages for outlinks made in an interest community or scientific domain. This is a hypothesis yet to be tested, but probably most outlinks are targeted to popular and authoritative web pages within the same subject domain (cf. the findings by *Cui*, 1999, mentioned earlier).

As stated earlier, methodological problems in webometrics are concerned with collecting non-biased data from the Web as a basis for empirical investigations. As shown above, there is a lack of reliability in secondary data collected in the big commercial search engines due to great uncertainty about coverage, update frequency, indexing rules, computing performance, ranking algorithms, etc. The alternative approach is to use primary data downloaded directly from the Web. A method in the latter category is so-called random walks used in graph theory. In order to localise transversal links on the Web graph a large number of lengthy random walks can be used by following links in a randomised way from web page to web page. Conducting path analysis of the lengthy link paths thus created (sometimes ending in "cul-de-sacs") transversal links are identified by using criteria of heterogeneity between subject domains reflected on the web pages. The concept of "heterogeneity" (as well as the related reverse concept of "similarity" in classification theory) is not uncomplicated and criteria for definition may be operationalised using word co-occurrence similarity measures.

Randomising starting points for the random walks is another problem due to the distributed nature of the Web mentioned before. Using so-called IP-numbers* or lists of Domain Name Servers in, for example, the ".edu"-domain are possible methods for non-biased sampling of starting points.

Path analysis and "undiscovered public knowledge"

Using path analysis in citation databases, *Small* (1999) investigates pathways crossing disciplinary boundaries in science and the cross-fertilising creativity that can emerge at such boundary crossings. *Small* is concerned with "strong ties" (rather than with "weak ties") in the shape of strong co-citations for creating indirect multi-step pathways through the scientific literature. According to *Small*, in scientific literature it is possible to travel from any topic or field to another because of the interconnected fabric of scientific disciplines. This is illustrated by using a specific path starting in economics and ending in astrophysics. *Qin* and *Norton* (1999), commenting on *Small*, anticipate that "in future retrieval systems, a user could pick two topics or documents and generate a path of documents or topics that connect them, which could be used for information discovery and hypothesis generation". This approach might also be used to localise transversal links on the Web. In a computer-generated Web graph model, using

* In the Internet Protocol (IP) every web host is assigned a significant IP-number with an alias to the URL (e.g., 130.226.186.6 is the IP-number of the Royal School of Library and Information Science: www.db.dk)

algorithmic methods, the shortest link path could be identified between two selected start and end web pages belonging to heterogeneous scientific domains. Depending on how well-connected the cluster neighbourhoods of the chosen web pages are, this method would fail perhaps in about 75% of the time (*Broder* et al., 2000), according to the aforementioned "bow tie"-model, because of the big regions of web pages with just outlinks, just inlinks, or no links at all. But in some 25% of the cases it would be possible to identify the shortest path between the selected two heterogeneous scientific domains, thus revealing shortcutting transversal links along the path.

Small's (1999) idea of creating multi-step *strong* co-citation paths could also be applied to the Web conducting co-link analysis of outlinks co-occurring, e.g., on scientists' web pages. This approach could be a fruitful direction for achieving knowledge discovery on the Web.

Another knowledge discovery method used in bibliographic databases, with possible applicability to the Web, is *Swanson's* research on "undiscovered public knowledge" (1986), developed over the years (e.g., *Swanson* and *Smalheiser*, 1997, 1999). *Swanson* (1986) states, "Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted". *Swanson* uses a systematic trial-and-error strategy for finding transitive relations between two literatures (*Davies*, 1989). Using *Swanson*'s (1986) own example, if literature A is concerned with fish oil and literature C is about Raynard's disease, then literature B on blood platelets can be the missing, transitive relation. If A → B and B → C, then A → C. This literature-based knowledge discovery method is used to find "interesting but previously unknown implicit information" within the scientific literature (*Swanson* and *Smalheiser*, 1999), revealing connections between ideas or concepts that were not considered before (*Garfield*, 1994). Using this method on the Web, transversal links may give useful hints for finding transitive relations between scientific disciplines.

The concept of "systematic serendipity" is useful in the context of computer-supported knowledge discovery – and scientific discovery. Several times since 1966, *Garfield* has used the concept to describe the organised process of discovering previously unknown scientific relations using citation databases. In fact, "systematic serendipity" is a rather precise description of the human-computer collaboration necessary in conducting knowledge discovery either in traditional databases or on the Web. The importance of human-computer collaboration for knowledge discovery in science is stressed by *Valdés-Perez* (1999).

The last approach in this section concerned with the use of bibliometric methods in knowledge discovery on the Web is inspired by *Lancaster's* (1985) concept of "issues management" and *Wormell's* (2000) "issue tracking".

Issue tracking

In his 1985 case study of how the new emerging issue "acid rain" was developed and disseminated in society, *Lancaster* tracked the issue through several different databases showing how this new research issue moved over time into the applied sciences and later on to mass media and legislation. Building on *Lancaster*, *Wormell* (2000) applied informetric methods to trace the pattern of international debate about the modern Welfare State through various domain databases, thus showing how a concept moves through a path of different publication forms. *Wormell* concludes that data and text mining technologies offer great possibilities for informetric analysis in order to extract previously unknown and potentially useful knowledge from bibliographic data. A variant of issue tracking on the Web was applied by *Bar-Ilan* and *Peritz* (1999), who investigated the chosen topic of "informetrics" for a certain period of time using bibliometric methods to analyse data from six major search engines. The dynamic nature of the Web resulted in web documents on the topic, which disappeared over time, whereas new documents were added and some were changed. Following the same methodology, but applying one brief observation window, *Bar-Ilan* (2000) traced the same topic in detail, comparing the patterns on the Web to those of bibliographic and citation databases.

Conclusions

This article has attempted to point to selected areas of webometric research that demonstrate interesting progress and space for development as well as to some currently less promising areas. As stated above, the feasibility of using bibliometric methods on the Web is highly affected by the distributed, diverse and dynamical nature of the Web – and by the deficiencies of search engines. That is the reason that so far the Web Impact Factor investigations based on secondary data from search engines cannot be carried out. The diversity of people creating web documents and links of course affects the quality and reliability of these web elements. The lack of metadata attached to web documents and links – and the lack of search engines exploiting metadata – affects filtering options, and thus knowledge discovery options, whereas field codes in traditional databases support KDD (Knowledge Discovery in Databases).

However, as suggested above, the diversity of the Web also could enhance the possibilities of knowledge discovery. The "anarchistic", local behaviour of millions of web actors is usually considered having negative consequences on the global performance of the Web as an information system. The abovementioned transversal links could be a hitherto neglected positive impact of this "imperfect" behaviour resulting in short paths on the Web and thus causing better exploratory possibilities for human and digital agents – and thereby causing better possibilities for human serendipity and computer-supported knowledge discovery, i.e. "systematic serendipity". Using this approach for discovering useful material for scientific purposes of course requires handling the problem of reliability on the Web. One way of doing this could be the abovementioned method of selecting quality start and end web pages from two heterogeneous scientific domains, and then identify transversal links on the link path connecting the domains. This method could enhance the probability of encountering quality contents in the intermediate web pages along the link path.

The Web consists of both "convergent" and "divergent" link structures, with web clusters corresponding to the former type and transversal links to the latter. These different link structures can support exploration of the Web conducted both in convergent (i.e., rational, goal-directed) *and* divergent (i.e., creative, intuitive) ways. The use of the terms "convergent" and "divergent" is inspired by *Ford* (1999) and the related work of *Bawden* (1986) on stimulation of creativity in information systems. Understanding the complementarity of convergence and divergence – both in the link structures being explored and in the behaviour of human or digital agents being exploratory – is important in order to develop creative methods of computer-supported knowledge discovery on the Web, as well as in bibliographic, citation and other databases. Such methods also could have implications for improving harvesting programs of web robots, ranking algorithms of search engines and visualisation/navigation features of browsers.

Webometrics is a new research field now passing through a necessary tentative and exploratory phase. The novelty of the field explains the substantial amount of descriptive webometric papers on different dimensions of the Web provided in the last 4-5 years. In the years to come, a challenge for researchers in webometrics will be to analyse and synthesise the findings and to further develop theories and methodologies in order to provide a better understanding of the complex topology, functionalities and potentials of the Web.

References

E. S. ALLEN, J. M. BURKE, M. E. WELCH, L. H. RIESEBERG (1999), How reliable is science information on the Web? *Science*, 402 : 722.

T. ALMIND, P. INGWERSEN (1997), Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics", *Journal of Documentation*, 53 : 404–426.

J. BAR-ILAN (1998), The mathematician, Paul Erdos (1913-1996) in the eyes of the Internet, *Scientometrics*, 43 : 257–267.

J. BAR-ILAN (1999), Search engine results over time: A case study on search engine stability, *Cybermetrics*, 2/3, paper 1. ISSN: 1137-5019 (http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html; visited 08.11.2000).

J. BAR-ILAN (2000), The Web as an information resource on informetrics? A content analysis, *Journal of the American Society for Information Science*, 51 : 432–443.

J. BAR-ILAN, B. C. PERITZ (1999), The life span of a specific topic on the Web. The case of "informetrics": A quantitative analysis, *Scientometrics*, 46 : 371–382.

M. BATES, S. LU (1997), An exploratory profile of personal home pages: Content, design, metaphors, *Online & CDROM Review*, 21 : 331–340.

D. BAWDEN (1986), Information systems and the stimulation of creativity, *Journal of Information Science,* 12 : 203–216.

T. BERNERS-LEE (1997), Realising the full potential of the Web. World Wide Web Consortium, (http://www.w3.org/1998/02/Potential.html; visited 08.11.2000).

K. BHARAT, M. HENZINGER (1998), Improved algorithms for topic distillation in a hyperlinked environment. In: W. B. Croft et al. (Eds.). *Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, pp. 104–111.

L. BJÖRNEBORN (2000), *Verdensvævet som "small-world"-netværk og mulighedsrum : omridset af en forståelsesmodel for transversale links på World Wide Web.* ["Small-World" Web and Possibility Space : outlining a conceptual framework for transversal links on the WWW]. Master's Thesis. Royal School of Library and Information Science, Copenhagen.

S. BRIN, L. PAGE (1998), The anatomy of a large-scale hypertextual Web search engine, WWW7 Conference, (http://www-db.stanford.edu/~backrub/google.html; visited 08.11.2000).

A. BRODER et al. (2000), Graph structure in the Web, WWW9 Conference. (http://www.almaden.ibm.com/cs/k53/www9.final; visited 08.11.2000)

V. BUSH (1945), As we may think, *The Atlantic Monthly*, 176 (July) 641–649.

S. J. CLARKE, P. WILLETT (1997), Estimating the recall performance of Web search engines, *Aslib Proceedings*, 49 : 184–189.

M. P. COURTOIS, M. W. BERRY (1999), Results ranking in Web search engines, *Online*, (May/June) 39-46.

B. CRONIN, G. MCKIM (1996), Science and scholarship on the World Wide Web: A North American perspective, *Journal of Documentation*, 52 : 163–172.

L. CUI (1999), Rating health Web sites using the principles of citation analysis: A bibliometric approach. *Journal of Medical Internet Research*, 1(1) e4 (ISSN: 1438-8871) (http://www.jmir.org/1999/1/e4/index.htm; visited 08.11.2000).

R. DAVIES (1989), The creation of new knowledge by information retrieval and classification, *Journal of Documentation,* 45 : 273–301.

H. DE JONG, A. RIP (1997), The computer revolution in science: steps towards the realization of computer-supported discovery environments, *Artificial Intelligence*, 91 : 225–256.

A. DILLON, B. A. GUSHROWSKI (2000), Genres and the Web: Is the personal home page the first uniquely digital genre? *Journal of the American Society for Information Science*, 51 : 202–205.

O. ETZIONI (1996), The World-Wide Web: quagmire or gold mine?, *Communications of the ACM*, 39 (Nov.) 65–68.

G. EYSENBACH (1998), Towards quality management of medical information on the Internet: Evaluation, labelling, and filtering of information, *British Medical Journal*, 317 : 1496–1502.

U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH (1996), The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39 (Nov.) 27–34.

N. FORD (1999), Information retrieval and creativity : towards support for the original thinker, *Journal of Documentation*, 55 : 528–542.

W. J. FRAWLEY, G. PIATETSKY-SHAPIRO, C. J. MATHEUS, Knowledge discovery in databases: An overview, In: G. PIATETSKY-SHAPIRO,W.J. FRAWLEY (Eds). *Knowledge discovery in databases*. Menlo Park, Cal.: AAAI Press, 1991

E. GARFIELD (1966), The who and why of ISI, *Essays of an Information Scientist*, 1 (1962-73) 33–37. Originally printed in *Karger Gazette*, March 5, 1966.

E. GARFIELD (1994), Linking literatures: An intriguing use of the citation index, *Current Contents,* 21 (May 23) 3–5.

D. GIBSON, J. KLEINBERG, P. RAGHAVAN (1998), Inferring web communities from link topology, *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*. (http://www.cs.cornell.edu/home/kleinber/ht98.pdf; visited 08.11.2000).

M. S. GRANOVETTER (1973), The strength of weak ties, *American Journal of Sociology*, 78 : 1360–1380.

P. INGWERSEN (1998), The calculation of Web Impact Factors, *Journal of Documentation*, 54 : 236–243.

J. T. KLEIN (1996), *Crossing boundaries : knowledge, disciplinarities, and interdisciplinarities*, Charlottesville, Virg.: University Press of Virginia.

J. M. KLEINBERG (1998), Authoritative sources in a hyperlinked environment, *Proceedings of the 9th annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 668–677.

M. KOCHEN (Ed.) (1989), *The small world*. Norwood, N. J.: Ablex Publishing Corporation.

F. W. LANCASTER, J.-L. LEE (1985), Bibliometric techniques applied to issues management: A case study, *Journal of the American Society for Information Science*, 36 : 389–397.

R. LARSON (1996), Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In: S. HARDIN (Ed.) *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, 33 : 71–78.

S. LAWRENCE, C. L. GILES (1998), Searching the World Wide Web. *Science*, 280 : 98–100.

P. LOSIEWICZ, D. W. OARD, R. N. KOSTOFF (2000), Textual data mining to support science and technology management, *Journal of Intelligent Information Systems*, 15 : 99–119.

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

C. OPPENHEIM, A. MORRIS, C. MCKNIGHT (2000), The evaluation of WWW search engines. *Journal of Documentation*, 56 : 190–211.

S. J. PIERCE (1999), Boundary crossing in research literatures as a means of interdisciplinary information transfer, *Journal of the American Society for Information Science,* 50 : 271–279.

J. QIN, M. J. NORTON (1999) (Eds). Introduction (In issue: Knowledge Discovery in Bibliographic Databases). *Library Trends*, 48 (Summer) 1–8.

J. M. RODRIGUEZ I GAIRIN (1997), Volorando el impacto de la informacion en Internet: Altavista, el "Citation Index" de la Red. *Revista Espanola de Documentacion Scientifica,* 20 (2) : 175–181.

R. ROUSSEAU (1997), Sitations: An exploratory study. *Cybermetrics*, 1 paper 1. ISSN: 1137-5019. (http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html; visited 08.11.2000).

R. ROUSSEAU (1999), Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3, paper 2. ISSN: 1137-5019. (http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html; visited 08.11.2000).

H. SMALL (1999), A passage through science: Crossing disciplinary boundaries, *Library Trends*, 48 (Summer) 72-108.
A. G. SMITH (1999), A tale of two web spaces: Comparing sites using web impact factors. *Journal of Documentation*, 55 : 577–592.
H. SNYDER, H. ROSENBAUM (1999), Can search engines be used as tools for web-link analysis? A critical view, *Journal of Documentation*, 55 : 375–384.
D. R. SWANSON (1986) , Undiscovered public knowledge, *Library Quarterly,* 56 : 103–118.
D. R. SWANSON, N. R. SMALHEISER (1997), An interactive system for finding complementary literatures: A stimulus to scientific discovery, *Artificial Intelligence*, 91 : 183–203.
D. R. SWANSON, N. R. SMALHEISER (1999), Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery, *Library Trends*, 48 (Summer) 48–59.
M. THELWALL (2000), Web impact factors and search engine coverage, *Journal of Documentation*, 56 : 185–189.
R. E. VALDÉS-PEREZ (1999), Principles of human-computer collaboration for knowledge discovery in science, *Artificial Intelligence*, 107 : 335–346.
B. VICKERY (1997), Knowledge discovery from databases: an introductory review, *Journal of Documentation*, 53 : 107–122.
D. J. WATTS (1999), *Small worlds : the dynamics of networks between order and randomness*, Princeton University Press, Princeton, N.J.
D. J. WATTS, S. H. STROGATZ (1998), Collective dynamics of "small-world" networks, *Nature*, 393 (June 4) 440–442.
I. WORMELL (2000), Critical aspects of the Danish welfare state – as revealed by issue tracking, *Scientometrics*, 4 : 237–250.
E. WYNN, J. E. KATZ (1997), Hyperbole over cyberspace: Self-presentation and social boundaries in Internet home pages and discourse, *Information Society*, 13 : 297–327.

Received November 11, 2000.

Address for correspondence:
PETER INGWERSEN
Center for Informetric Studies (CIS),
Department of Information Studies,
Royal School of Library and Information Science,
Birketinget 6, DK 2300, Copenhagen S (Denmark)
E-mail:pi@db.dk