Web Indicators for Scientific, Technological and Innovation Research

Report

"Survey of Practice"
(previous title: Best practice manual for Web data use in indicator research)

Isidro Aguilo, Natalia Arroyo (CINDOC, Madrid)
Viv Cothey, Mike Thelwall, David Stuart (Editor) (University of Wolverhampton, UK)
Sylvan Katz (SPRU, UK)
Hildrun Kretschmer (NIWI-KNAW, Amsterdam)

Deliverable: DL 23-1/1 **December 2005**

Survey of Practice

Isidro Aguilo
Natalia Arroyo
Viv Cothey
Sylvan Katz
Hildrun Kretschmer
Mike Thelwall

Contents

List of Tables and Figures

1

Introduction to the Survey of Practice

1.1 Web indicators

The Internet and the World Wide Web play an increasingly important role in today's society, and even more so for the research and innovation process. They have effected many of the day to day activities of researchers e.g., the searching and finding of information, the way they communicate with colleagues, and the whole scholarly publishing process. Some of these changes pose problems for the measurement of some traditional science and technology indicators, e.g., changes in the publication process may make it more difficult to assess the number and quality of academic articles; but these changes also offer the opportunity for indicators that were not possible previously.

One of the traditional indicators of research productivity, the number of published journal articles, has become much more complicated due to recent changes in scholarly publishing. Journal articles are now published in a variety of new ways, e.g., subscription electronic journals, open access electronic journals, personal self publishing, institutional self publishing; these articles will be of varying quality, with some being peer reviewed, and some not. Although there has always been the need for value judgements when quantifying a researcher's outputs, for example, only counting those journals which are ISI indexed, the quantity of people involved in the publishing process makes it much more difficult to assess. The situation is likely to become more complicated in coming years; organisations that once implemented the publishing of journal preprints on a individual basis have started implementing organisational wide institutional repositories, and certain funding is conditional on the results being made available as free preprints. Establishing web indicators will not negate quality issues, but will allow outputs that would be missed in traditional indicators to be taken into account.

Web indicators also have the potential of extending the ways the impact of a researcher's work is assessed. Traditional citation studies only take into account the number of times an article is cited in journals that have been indexed by the citation database, on the web many different types of citation could be taken into account, e.g. mentions on web pages and on reading lists. Not all scientific disciplines use scholarly papers as their main scholarly communication channel, and one of the criticisms that have been levelled at citation databases in the past is the little attention that is paid to non-journal works, if the planned mass digitization projects of the major search engines are successful it would be possible to incorporate them in any assessment of a researcher's work. Webometrics also provides the possibility of indicators for subjects that are not directly linked with any publication effort.

As well as having the possibility of extending the traditional indicators, there are potentially new indicators that have no traditional equivalent. There are many ways that a researcher may try to find the information they require: they may directly approach the web site of a journal, or collection of journals, that are relevant to their area of interest;

they may utilise an abstracting and indexing service; or use one of the many search engines that may cater for the general user (e.g. Google, Yahoo, or MSN) or for the more academic searcher (e.g. Google Scholar or Scirus). Much of the data created by academics searching for, and looking at, certain information is not available in the public domain. The information stored on these web servers has the potential to provide indicators that were not available for the traditional media: emerging areas of interest before any academic papers are written; the number of times a paper has been looked at, and under certain circumstances how much time the user looking at a page.

The web also has the potential to provide information on the structure of communication within the sciences. Not only the formal collaborations, that may be visible through traditional bibliometrics, but the more subtle informal collaborations.

The web is massive, and huge amounts of data can be produced about it relatively easily. Whilst there is a lot of potential for some of this data, it is not necessarily the case that this data will lead to meaningful indicators. Unlike scientific papers the web pages have little or no quality control. Information scientists who have an intellectual grounding in information issues, and citation analysis in particular are the best equipped to cut through the excess noise.

1.2 The contents of this manual

This is a companion volume to *Link Analysis: An Information Science Approach,* which presents a best practice for establishing and validating a type of web indicator. Link analysis is only one type of web indicator, and some feel that it does not fully address all issues of reliability, reproducibility, and validity. This volume is a *Survey of Practice,* designed to compliment the best practice volume; with a survey of the tools that can currently be used to measure the structure and nature of the web, the aspects that need to be taken into consideration when establishing web indicators, and some empirical examples.

This manual contains:

Methodological considerations
Empirical examples
A glossary of terms
An extensive bibliography of current research in the field.

There are three broad stages in quantitative analysis of the web: determining the units of analysis; collecting the data, and analysing the data. Chapter 2 focuses on determining the unit of analysis; this is a central feature of the research question and design; especially with regards to the web because of the general lack of formality with which most people engage with it. Without clarity of definition the data cannot be collected reliably. Chapter 3 contains a survey of many of the tools available for the collection of data about the web: custom crawlers; commercial crawlers; and search engines. These tools are constantly changing and the features are often commercial secrets; the survey of these tools demonstrates the range of features available, and their capabilities. The differences between the different tools may have an important effect on results, so it is necessary to understand them fully to ensure reliability. Chapter 4 covers the main aspects that need to be taken into consideration when analysing the data; most

importantly emphasising that the three sections can not be dealt with separately, rather that the three stages are interdependent aspects of the overall research design.

As the web is used for a variety of different reasons, intuitions regarding it are often wrong; therefore empirical examples are included to emphasise the importance of the methodological considerations. They highlight some of the problems that emerge with three approaches to establishing web indicators:

Link indicators (chapter 5: link analysis)
Co-link indicators (chapter 6: co-link indicators)
Web visibility of collaboration (chapter 7: small scale, in-depth case studies).

Problems include: the need for the reassessment of methodologies as trends in web publishing change; and the limitations of current tools.

The last chapter of this volume provides information to facilitate future research. Terminology plays an important role in the establishment of web indicators. Without a clear understanding of what a person means by terms such as 'web site' and 'web page' how are we meant to use these as units of analysis? Chapter 8 provides a definition of some of the important terms, to aid in the establishment of web indicators and progression of ideas within the field. Also included at the end of this work is a comprehensive bibliography of the current research in the field.

1.3 The future of web indicators

With the current sets of tools available link analysis provides a popular practice for establishing web indicators. However, the range of tools available for the gathering of information from the web is changing all the time, and as they do, the web indicators available will also increase. For example, recent addition of search engine APIs allows the automatic sending of queries to the search engine databases, databases that are larger than it would be possible for individual researcher or research group to create. It should also be remembered that as well as the tools available for measuring the web changing, so does the web itself; it is not just the information that is on it that is changing, so are the ways the information is presented e.g. dynamic pages. Every new format that appears requires new methodologies and approaches, and as such research into web indicators is an ongoing task.

Any current practices in establishing web indicators, including link analysis, will need to be reassessed in light of future changes on the web. The broad framework of this survey of practice provides the methodological considerations that need to be taken into account when establishing future web indicators.

2

Units of analysis

by Sylvan Katz and Viv Cothey

2.1 Introduction

How ‘big’ is the World Wide Web?

There are two prerequisites for providing a satisfactory answer to this and similar such reasonable questions. First we must understand the context of the question and what the questioner means. Is this big in an economic or business sense: how much money is involved? Is this big in an audience sense: how many people does the World Wide Web reach? Is this big in some computer storage sense: how much warehouse space might be taken up by all the information that is available on the World Wide Web? Or is this big in one of potentially many other senses?

Secondly, given that we have settled on a particular meaning for the question, say the size of the audience, then we need investigative techniques, that are both valid and reliable, that allow us to say how big is the World Wide Web's audience. It is clear in this instance that our investigation should make use of units of analysis such as households or individual users: counting the number of personal computers that can access the World Wide Web is not likely to be regarded as valid. Even when a valid unit of analysis is selected it may not be reliable. For example, as here, counting households reliably demands very careful definition and specification.

The unit of analysis is thus a central feature of the research question and the research design. The form of analysis in units of analysis does not have to be quantitative. The research design may be qualitative or a hybrid of both qualitative and quantitative.

The explicit need for valid and reliable units of analysis is especially true for research involving the World Wide Web (or just 'web') because of the general lack of formality (or precision) with which most of us engage/discuss the web. One measure of the web's success is the extent to which it has so quickly become embedded in popular culture. Most of us are neither interested nor care very much about the technical intricacies that underpin it. And yet few of us are inhibited from using a ‘technospeak’ that refers to the web: “visit our web site at www.somewhere.com", “logon to www.somewhere-else.biz" etcetera.

In this chapter we survey how units of analysis for investigating research questions about the web are designed and constructed. In particular we consider units of analyses that may be relevant in the context of policy.

2.2 Models of the web

The web is an evolving complex network of distributed hypertext.

That the web comprises hypertext gives it its most important characteristic. In contrast to the linear form of text the reader is not obliged to traverse the text sequentially as determined by the author. The hypertext form contains textual links that allow the reader to traverse the text in any sequence. Earlier instances of hypertexts shared the computer that supported the user's particular traversal of the text. Thus the hypertext was closed in that it was intrinsically bounded. The web is a distributed hypertext because the textual links can link to another computer. This property also makes the web an open hypertext.

The web is a network. That is, it comprises an ensemble of connected entities which can be realised as an instance of distributed hypertext. This network is complex because its statistical properties cannot be explained by classically random processes. In consequence the web possesses emergent or self-organising properties. Lastly, the web is an evolving network in contrast to being static since the number of connected entities is increasing. The evolutionary property of the web invites enquiry of the generative processes involved. This evolving complex network of distributed hypertext can be modelled or thought about in a range of ways from the extremely abstract to the entirely concrete. The uniform resource locator (URL) is central to any discussion or description of the web. For example the textual links mentioned above make use of a URL. The URL features in both representations of the web discussed here.

2.3 Abstract representations of the web

In abstract terms the web is most conveniently represented as a mathematical graph, or more precisely a directed graph or 'digraph' of nodes and arcs. Each node and arc in this mathematical model represents a web page and hyperlink respectively. Hence operationalising the model requires only for us to operationalise what we mean by web pages and hyperlinks. As we will see this is a relatively straightforward process that is assisted by one of the web's characteristic components, the URL.

The web digraph model supports a web evolution theory that is synthetic. That is, roughly speaking, as each new node is created it has an arc that links to an existing node with a certain probability. Such evolutionary models originate with Simon (1955) and Price (1976). Barabasi and Albert (1999) used the term ‘preferential attachment’. An analytic theory of Information Process Production (IPP) is also applicable. This also gives rise to an evolutionary theory for the web.

Neither theoretical approach explains individual behaviour. The goal here is to explain (and predict) the effect on the web of the collective behaviour of many individuals. In particular, theory attempts to explain the existence of characteristic distributions found in the web and the value of their parameters.

2.4 Concrete representations of the web

In concrete terms the web is a collection of human, computer hardware, software and telecommunications components and infrastructure that participate in web transactions.

These transactions are constructed and regulated by the hypertext transfer protocol (HTTP). A minimal concrete model is illustrated in Figure 2.1.

Operationalising the web and web transactions in respect of any concrete model is much affected by mediation by the technology involved. In order to construct a consistent and reliable representation a particular paradigm, network, server or user must be chosen. These are fundamental differences that arise and which depend upon the observational perspective adopted.

In principle a user requests a copy of a web resource by commanding the web client software to fetch it. The client software makes use of the resource's URL and HTTP to send an instruction addressed to the appropriate web server as specified by the URL as being where the resource is located. The network delivers the instruction which is processed by the web server software. The web server responds by providing a copy of the requested resource addressed to the client. The network delivers the resource so that the client software can process it and render it for consumption by the user. Hence there is a steady flow of (copies of) web resources from servers to clients.

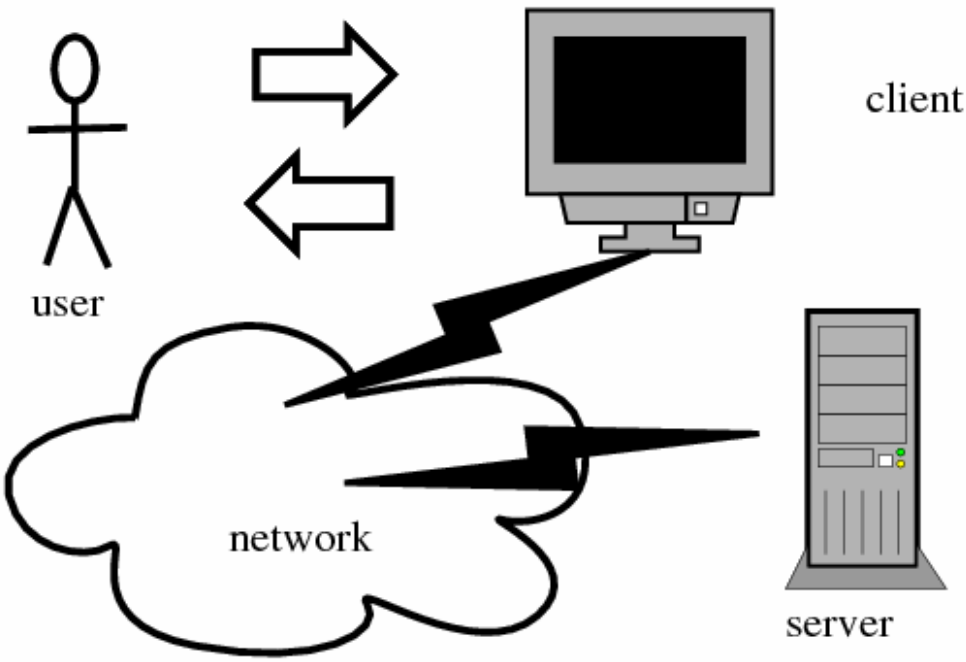

Figure 2.1. Concrete web - user, network and server.

2.5 The uniform resource locator

The URL is centrally important to any investigation of the web. Each URL is an address that points to where a particular web resource can be found. It is thus not the resource per se, but can be thought of as a label for the resource. Canonically it consists of several components but two of these are key. These are the so-called host (and port) part and the path (and file name) part as shown in Figure 2.2.

The host part is used by the Domain Name Servers (DNS), which forms part of the network in Figure 2.1 to describe in a human readable form the electronic Internet address of the web server. It is thus more accurately the domain name of the server. The server's address is concluded by the port number. For HTTP the default port number is 80. Because this is implied it is usually not stated (except of course when it differs from 80). The host and port part must be terminated by a '/'. Software will usually insert this if it is missing.

The DNS is not case sensitive. Hence <sussex.ac.uk> and <Sussex.ac.uk> are seen as identical by the DNS which regards all domain names as being in lower case when resolving domain names to their corresponding electronic Internet address. The DNS can be set so as to resolve different domain names to the same Internet address, which is then shared. In this way another domain name such as <www.sussex.ac.uk> can share the same Internet address as <sussex.ac.uk>.

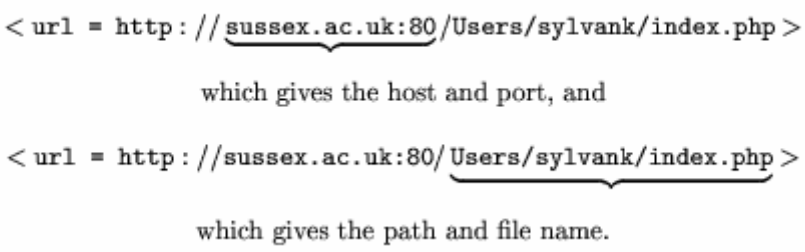

Figure 2.2. Uniform resource locator components.

It may be that many domains and their associated (default) servers all share a single Internet address. In addition several servers each having their own port may share a single domain name and therefore share a single Internet address even though each server will be serving clients from its own particular collection of web resources.

The host and port component of the URL thus defines the web server where the resource is located. The path and file component of the URL defines for the web server which particular resource is being referred to. The path tells the web server where to look for file. The path and file component of the URL is intended to be case sensitive. For example the path <Users/sylvank> is not equivalent to the path <users/sylvank>. (Not all web servers honour case).

When regarded as a character string a URL for a particular resource can have multiple different representations. It is therefore necessary, for example when counting different URLs, to standardise how a URL is represented as a character string. This is the canonical form of the URL.

Although a URL locates or labels a web resource by fully specifying the file name of the principal resource so that a web server can locate it, the web client can also make use of the content of the file provided to obtain copies of the online or subsidiary resources such as image files that are also specified. These may be located at other web servers. The notion of a web page can correspond to this collection of resources or files delivered to the web client and which the client renders and presents to the user for consumption.

Alternatively the term web page can mean just the principal resource labelled by the URL. This counter intuitive usage facilitates operationalising the abstract web page digraph. Each node in the digraph is the principal resource described here. Since this is

a single file (although possibly created dynamically) then the node entails attributes such as size, age and author. The URLs in the web page (principal resource) that point to non-online resources provide the hyperlink structure that is represented by the digraph's arcs.

2.6 Network, server and user paradigms

The terminology used to discuss the web is generally vague and imprecise. The elite engineers who do have a precise understanding do not engage in general discussion. In consequence misconceptions are rife and historically these have led to fallacious conclusions. For example hits as measured by the web server and recorded in the server log were used to indicate popularity or worse still the volume of users.

We show here some examples of how the concrete web mediates units of analyses and affects their validity. In consequence, particularly when aggregating or comparing units of analyses it is desirable that there is a consistent use of the paradigms. Additionally the reliability of units of analyses may be affected when using multiple paradigms. We focus on two mediating processes, the DNS and caching.

2.6.1 The network paradigm

Here the web is constructed from the point of view of the network. That is network traffic from sources to destinations. This may be either a physical network or a logical network. Users do not appear in this paradigm however entities such as web servers and client Internet addresses do. Malformed URLs including any URLs without DNS resolvable domain names are also excluded. Hence the collection of sources and destinations and their associated web transactions define both structural and usage characteristics for the web.

Sources and destinations are defined in terms of their Internet addresses as given by the DNS. In the Internet the DNS does not distinguish between web transactions and other types of Internet transaction. Hence individual web servers are not distinguished.

Caching is the temporary storing of web resources (files). Both the client and components within the network exploit caching in order to improve the overall efficiency of the Internet. In particular caching reduces network delays, traffic, and also server delays. Because of client caching some user requests can be satisfied from the cache and thus not every user request entails an HTTP. In consequence when client caching is effective, user activity is underrepresented by the network activity. Network caching is also used in order to avoid repetitive requests of the same server from different clients for the same resource. This means for example that the number of requests logged by a server under represents the number of such HTTP actually made by clients.

2.6.2 The server paradigm

The web constructed by reference to the server paradigm is based on the traffic received from clients. For example server logs may be used as a data source. Necessarily the only external entities known to the server are the clients possibly augmented by cookies. It is not possible to know remotely whose hands are on the keyboard.

The effect of caching is that servers are not aware of all the occasions on which a URL request from that server has been satisfied.

Web servers are configured to serve default web pages in response to requests for partial URLs. In addition two or more URLs at the same server may locate physically the same resource (file).

This together with the mediating effect of the DNS means that, from the servers' point of view many typographically different URLs can actually locate the identical resource.

2.6.3 The user paradigm

A user's perception of the web is different to that of either the network or the server. Users will not be explicitly aware of caching or the DNS. However user's can distinguish typographic differences in URLs that are not apparent to either the network or to servers. In addition users are generally unaware of identical resources being available from completely different URLs.

The notion of a 'web page' is user centric. That is, the web page is the particular ensemble of resources that come together to be rendered as a page by a web browser in response to a request by the user for a particular URL. Such a web page may comprise several files that are associated by virtue of the HTML of the requested resource. For example the page may contain an embedded image. It is not necessary for the image file to be at the same server as the requested URL.

This challenges any presumption that a 'web page' has particular singular attributes such as, a server, an age, a size etcetera. In consequence the term 'web page' is more usually used to mean just the single file or resource located by the URL in question.

2.7 Units of analysis

These are many and varied depending on the goals of a particular research project. Here we concentrate on units of analysis where the object of study is the web itself. We thus do not consider for example issues such as the demographics of web usage other than to suggest that even in such research there needs to be reliable operationalisation of the concept 'web usage'. (Does this include other Internet services such as email, IRC, etcetera?)

2.7.1 Macro analyses

Macro analyses of the web are analyses where the whole of the web is studied as a single object. In practice we only sample part of the web but in principle the form of analyses is the same.

Many units of macro analysis relate to characteristics or features of the web digraph. For example, the diameter of the graph. Mathematical graph theory provides us with a growing number of other characteristics. However analysing large (multi-million node) graphs becomes computationally infeasible. Computational science nevertheless continues to make inroads into the problems.

Web page characteristics such as the distribution of indegree provide more tractable units of analysis. Although still computationally and theoretically challenging several of these are suggestive of further study within a policy context.

Characteristics or units of analysis that can be considered here include:

Age
Size
Outdegree
File type

Comparative macro analyses is possible when the macro analytic results of different samples are compared. For example one might compare the diameter of the web with respect to one innovation system to that of another.

2.7.2 Micro analyses

Micro analyses of the web ignore the whole but focus instead on particular components within the web. For example interest is now on the characteristics and features of particular domains or web servers.

Given that the fundamental unit of the web is the web page (meaning just the file located by a URL) then compound units of analysis such as pages per server or pages per domain can be constructed. Clearly such units can be constructed almost endlessly employing a variety of qualifications; hence '.pdf' files per server etcetera. Compound units such as domains per GDP or servers per capita appear to be obvious candidates for policy relevance.

2.8 Conclusion

It is now a cliché to say that the web is important in today's society and even more so for research and innovation practice. However establishing either a theoretical or empirical grasp of the web remains a challenge.

In order to make progress good empirical data and good theory must be complemented by soundly based analysis. This will be founded upon a sufficient technical understanding of both empirical data collection and the theoretical/analytical issues arising.

In this chapter we have demonstrated some of these issues. We cannot exhaustively examine every possible unit of analysis. These depend upon the particulars of each research project. However we do identify a fundamental unit, that of each web page (i.e. located file) and its role as a node within the web page digraph.

We conclude by stressing the importance of clarity of definition of whatever unit of analysis is chosen and of ensuring the reliability of how the definition is operationalised.

References

Barbasi, A. L., & Albert, R. (1999). Emergence of scaling and random networks. *Science*, 286, 509-512.

Price, D.J. de S. (1976). A general theory of bibliometric and other cumulative advantage process. *Journal of the American Society Information Science*, 27, 292-306.

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.

3

Data collection

3.1 Introduction

Information can be gathered about the content and structure of the web in a number of ways, each of which has advantages and disadvantages depending on aspects such as the sample size and the unit of analysis. This chapter provides a survey, from a webometric perspective, of some of the tools currently available for data collection:

Custom crawlers
Commercial and academic crawlers
Search engines

It also discusses the advantages and disadvantages of manual analysis for the validation of information gathered by automatic tools.

3.2 Custom crawlers

by Viv Cothey

Web crawlers (sometimes called web spiders) provide an automated means to carry out data collection from the web. The nature of the data that is collected varies. Web crawlers have become strongly associated with search engines and web information retrieval. This is because search engines populate their information databases from data collected by crawling. Burke (2002) provides a good technical but accessible explanation of how a web crawler is constructed. Patterson (2004) emphasises that the simple principles belie the practical realities.

Cothey (2004) discusses web crawling and shows diagrammatically the architecture of a web crawler. The three principal functional components that are of most concern here are the:

Fetcher
Link extractor
Controller

The operation of the crawler and the nature of the data collected depend on the interaction of these functional components. For example, content crawling can be differentiated from link crawling (Cothey, 2004). The terminology used to describe the crawler functions differs although the underlying architectural model is equivalent; see for example “An introduction to Heritex” (Mohr, Kimpton, Stack, & Rantovic, 2004).

Web crawlers are more or less configurable. That is the administrator of the crawler (the administrator) can define some overall parameters for the crawl within the operational constraints of the crawler prior to commencing the web crawl. The GNU wget manual provides excellent documentation in respect of how one might configure a crawler. However a custom crawler allows whatever level of access is desired to modify any of the key functional components. This is generally not the case with crawlers that are just configurable.

The obvious prerequisite for any modification is that the administrator is also informed as to the specification of the functional components. Unfortunately specification documentation is usually absent (for any crawler) and in any event is not to be trusted in the absence of thorough testing. In consequence it is more frequently the case that the specification is the code.

This distinction implies that custom crawlers must be open source. Also, unfortunately, crawler administrators need to have some familiarity with the code in order fully to exploit the potential for customisation. However even if there are no modifications made, there is the advantage of being informed as to the detail of the data collection procedure.

Open source crawlers allow the administrator to inspect the source code of the crawler, and modify (customize) the code to meet the exact needs of the data collection exercise.

Lists of web user agents or robots associated with the fetcher component can be obtained from the web. There is no shortage. However there are relatively few open source crawlers available.

Open source crawlers can be classified according to the programming language employed and the design intent. Each crawler is assembled from standard or quasi-standard smaller software components.

Three programming languages dominate: C/C++, java and Perl. For each of these languages there are libraries of the smaller software components needed to construct a crawler. For example, the W3C libwww. These softwares have been written by experts in the field and have been well tested. In some cases the software provides the reference case to define how the protocols that support the web should work.

Each designer of a crawler will have particular design intent. Seven illustrative examples are given below:

heritix. The heritex crawler (java) is designed to support the endeavours of the Internet Archive. It is therefore intended to offer ‘archival quality’ web crawling by remedying the deficiencies in other crawlers that fail to perfectly mirror collections of web pages.

nutch. The nutch (java) is an open source crawler/search engine. It is designed to be both a public search engine where the search and retrieval algorithms used are published and to offer individual users a customisable crawler service.

ht://Dig. This is an older enterprise crawler/search engine (C++).

Harvest/Harvest NG. Harvest (C) is a federated crawler/search engine making use of information brokers. Harvest NG is a Perl implementation of a crawler/search engine that is compatible with the original brokered information retrieval architecture. Neither project is current although the code is still available.

blinker. blinker (web link crawler) uniquely here is not associated with information retrieval and it makes no attempt to index the content of web pages. Instead it is designed to generate web-page crawl-graphs. Like heritix, the design/implementation details attempt to remedy the shortcomings found in other crawlers in particular content crawlers. blinker is written in Perl and is structured for ease of comprehension of functionality and modification (such as including content summarisation) rather than for the high performance (equals a large web-page download rate) of most crawlers. blinker uses LWP the Perl implementation of libwww.

larbin. This is a high performance web crawler written in C that is capable of being extended to support a search engine.

GNU wget. Also written in C, wget is a standard utility distributed by GNU that provides a non-interactive download of files from the web. Functionally it is therefore similar to larbin. wget is highly configurable and open source. It is unlikely that anyone would modify the code in practice.

All crawlers should allow the administrator to configure the fetcher component, which makes the requests for web pages from the web servers that are being crawled. These parameters will include restrictions such as the maximum server hit rate and the maximum web pages from any one server. There is a consensus on the etiquette for how the fetcher should behave with respect to the network and servers. This includes 'politeness' and conformance to the informal robots exclusion protocol. It should not be possible for the administrator to configure the fetcher to misbehave.

The link extractor component is currently the most dynamic aspect of crawler development. This component is responsible for identifying within the content of web pages that have been fetched the URLs of the web resources to which the web page links either explicitly or implicitly. Clearly the more extensive the capability of the extractor, the more URLs are going to be identified. As an aside, it is equally clear that data collected using materially different link extractors will not be compatible.

The controller component controls which URLs the fetcher is required to download. There are several levels of control. Given that the administrator defines the crawl space within which the crawler is permitted to operate then the controller implements this constraint. In addition the controller will also filter URLs from being fetched. The bases for doing this are many and varied. Clearly once again unless these are known to the administrator then the data collected may not be what is believed or was intended.

In addition to the opportunity that custom crawlers provide the administrator to take charge of the data collection procedure, custom crawlers also permit the administrator to control the output from the crawler.

The primary output from 'search engine' crawlers is the search/retrieval database that is the product from the data collection. Constructing this database may have involved considerable processing of the web page data collected by the crawler. In addition any research on the data will require a data extraction step to obtain the data from the database. Hence access to the data collected is mediated by the database management system. This may further exacerbate the corrupt nature of the data collected compared to what the data collection procedure was thought to achieve.

Custom crawlers enable the administrator to control and modify the form of output generated from the data collection. This facilitates further analysis of the data. For example the output from blinker described above is not mediated by a database

management system. It is a web-page crawl-graph in valid XML. This can be processed by standard XML tools.

3.3 Evaluation of commercial and academic software for webometric purposes

by Natalia Arroyo

3.3.1 Introduction

Traditionally, the most frequently used tools in webometrics for extracting quantitative web data are search engines (Bar-Ilan, 2001), but others, such as commercial and academic software, have also been used. Some commercial tools (Aguillo, 1998b; Arroyo, Pareja & Aguillo, 2003) have proved to be a useful method for collecting web data since they have been employed in several research projects, —the last of them being EICSTES (European Indicators, Cyberspace and the Science-Technology-Economy System). Some academic crawlers have been designed for the same purpose as well (Thelwall, 2001; Adams & Gilbert, 2003).

The scope and objectives of these tools are different. Search engines have been designed to search in the web, and the unit of analysis they employ are web pages. Commercial crawlers have been created by companies for webmasters to manage their web sites, and they employ web sites or institutional web sites as a unit of analysis, as well as the academic crawlers analysed in this study, that have been created for academic purposes in order to investigate the web. The use of commercial and academic crawlers is much more suitable to work at a microscopic level, because it entails a great amount of work, while search engines are much more suitable for macroscopic information. The depth of the crawling of search engines is not precise enough to know the entire web, but only the part that is visible to them. Commercial and academic crawlers surf the web until the level indicated by users (that sometimes can be restricted), so they are much more precise in this sense (Arroyo, 2004).

This section aims to deepen the understanding of the use of commercial and academic crawlers, evaluating a selection of them to study the possibility of them being employed to extract web data for webometric purposes in further investigations. It is important to clarify that the intention is not to compare the software, but to investigate how they work, their capabilities, and how to interpret the results obtained from them.

3.3.2 The experiment

To achieve this, an experiment has been conducted. Eleven programs have been initially selected according to their availability. Six of them are web analysis tools developed for webmasters to control their web sites: Astra SiteManager 2.0 (trial version); COAST webMaster 6.0 (trial version); Funnel web Profiler 2.0; Site Analyst 2.0; Content Analyzer 3.0; and webKing 2.0. Two are crawlers developed for academic purposes: SocSciBot 1.8.88; and webcount. One is a link checker: Xenu's Link Sleuth 1.2e. One a log analysis tool: Web Trends 7.0c. The last one a file retrieving tool: Teleport Pro 1.29.1981, that can be used to make copies of web sites to enable offline browsing.

In order to test the software, twenty web sites were extracted from the European academic web space, trying to represent different types of institution, all EU countries,

different research areas, and different programming languages employed to build web sites in order to observe the behaviour of each crawler.

Each program was run against every web site on two dates, firstly on October 22nd, 2003, and secondly seven days later. Just before running the software, on the same day in order to avoid changes in the web, all web sites were copied to a hard disk by using Teleport Pro. The aim of this last procedure was to enable running the software online as well as offline in order to study its behaviour in a known environment. Therefore, the final outputs are four samples: two of them online and two offline.

Although the options available are very different from one crawler to another and they have been assumed not to be comparable, attempts have been made to select the same options wherever possible for the results to be as homogeneous as achievable. Program options are very important because some of them affect the results obtained, while many others only refer to secondary aspects of the running of the software or other utilities.

Supporting this experiment is the concept of the institutional web site, introduced by Aguillo (1998a), and can be defined as a "web page or a set of them, hierarchically linked to a home page, identifiable by an URL, and considered as a documentary unit that might be recognized by its subject, authorship or institutional representativeness" (Arroyo & Pareja, 2003).

3.3.3 The results

For analysing the results of this experiment the number of internal pages of the web site will be used as an indicator of the coverage of the results of each software, since other resources embedded on them can only be retrieved through links in site pages. The coverage can be defined in this paper as the capability to crawl more or less resources on the web site analysed

Having a quick look at the data obtained (annex 3.1, at the end of the chapter), great differences in the same web sites and date from one crawler to another can be observed. There are mainly two reasons to explain them: program options and hidden features. Program options are not the same for all the software, as explained before, so although the most similar options were chosen it is almost impossible to get the same parameters. The rest of the features of each crawler cannot be controlled by the user and they are protected by commercial secrets, so it is very difficult to find them out.

The only agreements between crawlers happened when JavaScript or Flash files were found: none of these programs can access the hyperlinks in them (see the results for the web sites denoted as 2 and 11 in annex 3.1), so the data are identical, except the ones obtained from Astra, which counts the sum of resources, not only pages, and webcount, whose results are always different from the others.

The most similar data was achieved in web sites mainly structured in HTML (non-dynamic) files and the reason is that crawlers are developed for counting HTML pages, so there is no problem with them (see web sites 1, 5, 15 and 16).

If results over time are compared, there are not great variations, only those produced by the instability of the web, so only big and dynamically-built web sites look unstable.

The greatest differences between crawlers mainly occur when dynamic web pages are analysed, as can be deduced from the great differences in the results of this kind of web sites. This is due both to hidden features of the software and the structures of the web sites analysed. This last one means a part of what has been called the

invisible Internet, that was defined from a crawler perspective in a previous work as "the part of the web that commercial and academic crawlers can not see" (Arroyo, 2004).

In order to study how the software works when it is analyzing dynamic web pages some cases have been examined in detail. The cases selected are web sites including ASP, PHP and CGI files. For this purpose the hypertextual structure of each web site, that is drawn in most of the reports excepting those generated by webcount and COAST webMaster, has been checked.

An important point to take into account is the section where every resource is included, that is to say how it is classified by the software to be shown in the reports. Some types of web objects are not classified correctly because the software cannot recognize its extension, so it is added to other section. This is just what happened when Site Analyst and Content Analyzer have included PHP pages in the section 'Other resources' because it did not recognised them as if they were web pages. Sometimes this can be solved changing the option 'definition of extensions of web pages', but when the reason is on hidden features of the software the only solution is to sum the number of those files to the right section.

The problem called by Cothey (2003) user interaction absent, that means that "it is not usually possible for a web-crawler to be able to progress beyond the initial web object which invites a user response since a null response usually generates some form of default or error reaction from the server". It has been also identified in web sites 6 and 12, both of them including CGI files with forms. When the form is not completed an error page is obtained. This is clearly shown in COAST webMaster reports. Quantitative results are not very different from one crawler to another in the case of web site number 12, excepting the ones provided by SocSciBot, which adds web pages in other sub-domains (like www.cst.dk) due to the truncation in the left of the URL performed by this program.

Hyperlinks to web pages dynamically generated are found on web site number 6, so crawlers can access them through the links. Funnelweb, SocSciBot, webTrends and Xenu (with higher results than the others) can crawl this kind of CGI files.

Web sites including ASP or PHP files can be mapped depending on the capabilities of the software to recognize them, but only if they are linked by any web page and no interactivity is needed. Therefore, examining carefully the reports generated it could be observed that the software evaluated can be classified in 3 groups depending on the coverage of dynamic web sites:

> Wide coverage, if dynamic objects can be scanned. *Funnelweb*, *SocSciBot* and *Xenu* are on this group.
>
> Basic coverage, if only the first or a few pages of dynamic sites can be fetched.
>
> No coverage, if there are no dynamic web pages analysed by them, like Teleport Pro in some cases.

Webcount has not been included in any of those groups because it is not transparent enough to be analysed.

The software performs differently depending on the kind of dynamic web pages it has to analyse. Only Funnelweb, SocSciBot and Xenu have been able to scan every type, so they can be considered very appropriate for scanning dynamic web pages. Other crawlers can only reach medium coverage, just like Microsoft Site Analyst and

Content Analyzer. And finally, there is a third group of crawlers with high coverage for some kind of dynamic web pages but medium in other cases, like Astra SiteManager, COAST webMaster, webKing, webTrends and Teleport Pro.

3.3.4 Software evaluation

Finally, the software will be evaluated taking into account the following criteria. They are also described in annex 3.2.

> Additional utilities like link checking, log files analysis or words analysis.
> Coverage of the results.
> Difficulties found in running, like crashing computers or the impossibility to crawl offline.
> Graphical environment of the tools, like maps, *cyberbolic* views or graphics.
> Limitations like the number of resources that can be retrieved, the impossibility of measuring some resources, and so on.
> Outputs obtained (statistics, reports, maps, graphics, etc).
> Interest of the program options offered.
> Resources of the computer consumed, like RAM memory, etc.
> Transparency, understood as the facility to decipher what are the resources they crawl.

Both Blueprint and COAST are interesting tools from this point of view and can be very useful tools for webometric purposes if only a few brief statistics are needed. Quite the opposite of this is the main goal of webKing, its extensive options make it a very complete tool.

SocSciBot obtains very good results even in dynamic web sites, through its simplicity, in a brief space of time.

Xenu is suitable for analysing small or medium-sized web sites while big ones cause computer crashes. Web Trends often generates program errors, regardless of web sites' sizes, so it should be avoided.

Otherwise, if the point of the coverage is not especially important or has been assumed both Site Analyst and Content Analyzer are very useful crawlers that gather complete statistics.

3.3.5 Conclusions

Due to the great differences between crawlers, outputs obtained from them are not comparable because of hidden features protected by commercial secrets, program options and their own characteristics. Researchers must therefore know, as far as possible, what they measure and assume that each one offers different options and can be used for different purposes. These great differences may represent an important problem for webometrics whenever they are not controlled by the user, so it is really necessary to face up to this point before using this kind of tools.

A great challenge for commercial and academic software is the measurement of dynamic web sites. This represents one of the main reasons for differences in the

results because each crawler treats dynamic web pages in different ways; therefore pages that may be considered part of the invisible web for one crawler are found by another. Further investigations should be carried out into this area for a better understanding of the way these tools work and a better control of their outputs. This is becoming a more important question since the number of dynamic web pages is increasing day after day.

The lack of control over the results is perhaps one of the greatest holes in webometric research today. This must be solved by investigating the way the tools employed by both search engines and commercial and academic crawlers work to guarantee the reliability of the data obtained.

3.4 Search engines: Uses and abuses in webometric analysis

by Isidro Aguillo

There are five different phases in the building of a modern commercial search engines. Each one is different and needs some kind of customisation. This can explain the relevant characteristics of the engine but also their main shortcomings.

As many of the features of the engines are commercial secrets, the following comments derive from empirical tests and some fragmentary information. Moreover, they are constantly changing so their validity is more or less restricted to the final quarter of 2005, when this analysis took place.

This section looks at largest commercial search engines with independent databases currently available. The three largest and more powerful are Google, Yahoo Search and MSN Search, while other less popular contenders with smaller databases like Teoma/Ask Jeeves, Gigablast and Exalead were also examined.

3.4.1 Crawling

Robots (or web crawlers) attempt to crawl the entire web according to certain criteria, some depending of the internal program, some influenced by external factors. As a general rule, all the robots intend to index all the pages they can collect following a hypertext tree, the links in a starting page. If a page is linked to by another one there is a higher probability that it will be indexed by the robot. Unfortunately there are a lot of permanent or temporarily 'orphaned' web pages that are not indexed. A very striking situation occurs in Google, which adds to its database all the linked web pages even when these are not reachable or not longer available.

When the geographical coverage of the engines is observed, several striking patterns appear. Some regions appear under-represented in some search engines, whereas North America and Western Europe looks well covered. The problem is especially relevant regarding the Far East and Southeast Asia suggesting a topology problem of the Internet infrastructure there. Language could be also involved in these observed biases.

The robots only index the sites that are operative when visited. If the machine is down, has any problem or simply is very busy, the contents are not indexed. Besides, most of the engines have several net etiquette rules designed so they do not overload the server. Commercial engines state that they respect the robots.txt exclusion system, although many webmasters indicate the contrary. Other self-regulatory systems include a time limit and a depth of indexing limit. If a server gets busy, only a fraction of their

contents will be indexed. The depth level a robot can crawl is unknown, but sometimes there are pages in levels lower than ten (subdirectories), so we can assume a fraction of pages are never reached.

Among the most relevant problems robots face are the design mistakes in the target pages. Providing a flash movie as the only start point in a home page can be a serious barrier for a robot. A navigation system with a complex JavaScript, map, or any other system is also a matter of concern in compiling web pages. Robots are unable to follow links from these pages so many complete sites are not indexed. This is not applicable to all the situations as a page could be friendly for one robot and not for another.

It has been described that there is also a size limit regarding to the volume of information that is collected from each single page. Yahoo covers about 500 KB for each page and the Google limit seems to be lower. It is difficult to say, but there are lots of single pages with text content exceeding those limits.

Deep or invisible web is the term coined to refer to information that can only be accessed through a gateway. As the contents are not accessible by robots because they cannot automatically extract the records, such information is said to be invisible to the search engines. Library catalogues, bibliographic and alphanumeric databases and document repositories represents a huge number of resources that are not usually indexed. During recent months some of the engines, especially Google, have been able to recover and index these records. Preliminary analysis suggests an irregular pattern; not all the databases are scanned and those included are only partially indexed. From a quantitative point of view each record in a database is counted as a single page, with relatively low content. This situation can artificially inflate the size of a domain.

The time between consecutive visits is unknown although the faster engines are probably situated between 15 and 30 days. Newspapers, news agencies and similar frequently updated web sites have a different collecting scheme, being visited at least once each day. The number, composition and distribution of these specially controlled web sites are unknown.

As a result of the limitations described, the size of the databases of the major search engines is only a fraction of the total web space. Moreover, overlap between engines is fairly limited, so it could be desirable to combine these databases. It is difficult to say, but probably overlap is higher nowadays than previously.

Image databases should be linked to the web crawling processes but it appears they are updated less frequently than their web counterpart. Probably, image data is collected only a few times per year given the huge bandwidth resources required.

3.4.2 Indexing

The main database of search engines is textual based, so the indexing of web pages requires they are available in an HTML or derived format or in other format that can be easily translated into text format. The number of formats an engine can index has increased during recent years, although Google clearly covers more types than their competitors. Tests show that engines identify rich files even without the correct suffixes, e.g. Adobe Acrobat files not ending in .pdf.

The indexing process also consists of two additional assignments, language and country of origin. Given the large size of the databases both tasks are done automatically that means several errors and biases are produced. As high as 20% of the language tags are probably incorrect, including multilingual pages and pages where the

language is mistakenly assigned. The national assignment is probably done by recovering data about the owner of the register of the IP of the server. Obviously there are a huge number of hosted pages in hired foreign servers worldwide and sometimes the information provided by the register is not correct.

Most of the search engines have a family filter, an option that can be enabled for avoiding recovery of information tagged as inadequate. Criteria used for this filtering are not public and the automatic system that performs the assignation is probably as faulty as in similar situations. For webometric purposes it is recommended to switch off these filters.

It is suspected that additional information is collected by search engines that are not usually available to the users e.g., IP address, file size, server date. In some cases application programming interfaces (APIs) provided by some search engines are offering a back door access to this kind of data. Unfortunately webometric data obtained from these APIs can be different, fewer records are returned than those obtained from the commercial search interface.

3.4.3 Recovering

The range of search options available in the main search engines is not very large, with similar approaches and even syntaxes common to all of them (see annex 3.3, at the end of the chapter). However, the way in which these operators really work is not transparent for the user and for similar requests the results obtained from several search engines are strikingly different.

Comparing with standard database management systems the Boolean capabilities of the search engines are limited. The Boolean AND is implicit when two or more terms are considered. When an institution has several domains, it can be practical to use the operator OR. In Google this operator does not work adequately, especially when three or more are used in the same strategy. Other relevant operators are those that filter records according to certain criteria. They are called delimiters and several authors have described inaccuracies when using them. Nevertheless they are extremely useful for webometric purposes, especially when relative comparisons are intended.

It is relevant to point out that images databases can be also filtered according to URL-related criteria so it is possible to extract quantitative information in these databases too.

There is a limit to the length of the request but usually this is high enough to not represent a problem in webometric analysis. Until recently Google only supported ten terms per search, but now the limit has climbed to 32 words.

A special option only available in Google is the number range. Separating two natural integers with two consecutive points the system will search all the numbers between both limits. Currently it does not working with numbers larger than 10^7.

Taking into account the way the date stamp is collected from web pages, none of the date related filters can be considered reliable and, at the moment, this kind of analysis should be discarded for webometric analysis.

A point to be taken into account for the future is the disparity of the units counted, especially when we are considering the disparate sizes among pages, even when only text is considered. In the same way all the links are put together even though we know there is a great diversity of motivations for pointing to another page.

3.4.4 Ranking

The rank in which the results are presented is not usually interesting from a webometric point of view as there are a great number of criteria involved and usually their combination is a commercial secret. The public formula for Google's PageRank is no longer applied and the values its toolbar provides (natural numbers from 0 to 10 in a logarithmic scale) are not very useful for fine analysis.

At least six big factors are considered for sorting results when a search is done: the number of times the search term appears; the position in which the term appears along the page; the linking structure (specially the external inlinks) of the page; the local aspects involved (country and language); the freshness (update frequency) of the page; and the number of visits to that page intercepted by the engine in a certain period. In Google the order is different when the language option is chosen, favouring pages written in that language, and also the results change.

For webometric purposes, the link related algorithm of Google, PageRank, can be used as a measurement of visibility as it has been discovered that using neutral terms it is possible to discard extra factors in the ranking. The results obtained are sorted by their PageRank.

Popularity, the number of visits, can be used indirectly by search engines as the number of times a certain page has been visited from the results of a search engine request. This information is clearly collected by most of the engines, including data about the origin of the request. Unfortunately if this information is really used and to what degree this is taken into consideration for the ranking of results is mostly unknown. The only engine that provides absolute and relative values about popularity is Alexa. Unfortunately this information is provided only for institutional domain level and for the rest of purposes they use the Google database. As Alexa and A9 are both properties of Amazon we can expect a new contender in the search engines arena in the next few years.

3.4.5 Presentation

None of the engines provide a complete list of the results. The limit is usually to list only the first 1000 pages, but frequently even this figure is not reached.

The low quality of the number the search engines provide about the results of a strategy is very surprising. When a bibliographic database is queried the result is an exact stable number that represents the true value requested. However, most of the search engines round the results, so it is not strange to obtain values ending in '00 or '000. And when low results strategies are analysed, sometimes the number of pages is far below than the number of results.

The numbers can change even after two consecutive requests, so the same strategy performed several times during the same day can provide far different numbers. The unreliability of the results may be related to the retrieval of the results from different databases. Yahoo provides a number in its first page that is usually slightly higher than the ones provided in further pages, as if it is narrowing the results to a limit that represent the exact value. In a few cases numbers are provided for strategies not correctly built or malfunctioning.From the results pages it is difficult to know if the grouped results, usually by domain, are counted separately.

3.5 Conclusion

All methods of data collection have problems with transparency. In the case of commercial crawlers and search engines this is due to the commercial secrets; whilst with the custom crawlers it is due to the lack of documentation meaning their true workings are really only intelligible to the advanced programmer. This means that any of the tools that are used need to be rigorously tested, as do the results. Only then a researcher be sure that the data collected, is what they state has been collected, and that it proves what they state it proves.

References

Adams, K., & Gilbert, N. (2003). Indicators of intermediaries' role and development. Deliverable 6.2. EICSTES Project.

Aguillo, I. (1998a). Hacia un concepto documental de sede web. El Profesional de la Información, 7(1-2), 45-46.

Aguillo, I. (1998b). Herramientas de segunda generación. Anuario SOCADI, 85-112.

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

Arroyo, N. (2004). What is the Invisible Web? A Crawler Perspective. Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods, Brighton, 19-22 september. Retrieved September 1, 2005, from http://cybermetrics.wlv.ac.uk/AoIRASIST/arroyo.html.

Arroyo, N., & Pareja, V. (2003). Metodología para la obtención de datos con fines cibermétricos. III Taller de Indicadores Bibliométricos; Madrid.

Arroyo, N., Pareja, V., & Aguillo, I. (2003). Description of Web data in D3.1. Deliverable 3.2. EICSTES Project.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes: a review and analyses. Scientometrics, 50(1):7-32. Björneborn, L. & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics,* 50(1), 65-82.

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

Burke, S.M. (2002). *Perl & LWP*. Sebastopol, CA: O'Reilly.

Cothey, V. (2003). Web-crawling reliability. 9th ISSI International Conference on Scientometrics and Informetrics; Beijing.

Cothey, V. (2004). Web crawling reliability. *Journal of the American Society for Information Science and Technology*, 55(14), 1228-1238.

Cothey, V. & Kretschmer, H. (2004). Does the link structure of the web provide evidence of a collaborative hypertext? *Journal of Information Management and Scientometrics*. 1(2), 9-12.

Garfield, E. (1964). Science Citation Index: A new dimension in indexing. *Science*, 144, 649-654.

Internet Glossary. Retrieved September 1, 2005, from http://www.svn.net/helpdesk/glossary.html.

Kostoff, R. N. (2002). Citation analysis of research performer quality. *Scientometrics,* 53(1), 49-71.

Kretschmer, H. & Aguillo, I.F. (2004). Visibility of collaboration on the Web. *Scientometrics*, 61(3), 405-426.

Kretschmer, H., Kretschmer,, U., & Kretschmer, T. (in press). Reflection of Co-authorship Networks in the Web: Web Hyperlinks versus Web Visibility Rates. *Scientometics.*

Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I. (2004). Introduction to Heritrix, an archival quality web crawler. 4th International Web Archiving Workshop (IWAW04), September 16 2004, Bath, UK, http://www.iwaw.net/04/.

Parasoft. Retrieved September 1, 2005, from http://www.parasoft.com/.

Anna Patterson (2004). Why writing your own search engine is hard. *ACM Queue,* 2(2).

Quest software. Retrieved September 1, 2005, from http://www.quest.com/

Rousseau, R. (1997). Sitations: An exploratory study. *Cybermetrics*, 1(1), Retrieved July 30, 2002, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

SocSciBot3. Retrieved September 1, 2005, from http://cybermetrics.wlv.ac.uk/socscibot/.

Teleport Pro. Retrieved September 1, 2005, from http://www.tenmax.com/teleport/pro/

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

Thelwall, M. (2002/3). A Free Database of University Web Links: Data Collection Issues. Cybermetrics, 6/7(1). Retrieved September 1, 2005, from http://cybermetrics.cindoc.csic.es/cybermetrics/articles/v6i1p2.html.

Thelwall, M., F. Barjak, & Kretschmer, H. (in press). Web links and gender in science: An exploratory analysis. *Scientometrics.*

Wilkinson, D., Harries, G., Thelwall, M. & Price, E. (2003). 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), 59-66.

Xenu. Retrieved September 1, 2005, from http://home.snafu.de/tilman/xenulink.html.

Annex 3.1. Results of crawling with different commercial and academic software.

		Teleport Pro		Astra Site Manager				Blueprint				COAST WebMaster				FunnelWeb		Microsoft Site Analyst			
		031022	031029	Online		Offline		Online		Offline		Online		Offline		Online		Online		Offline	
				031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029
1	www.cindoc.csic.es/cybermetrics	82	82	201	201	200	200	82	82	81	81	82	82	82	82	76	79	82	82	81	81
2	www.ite.upv.es	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	www.upct.es/~de	510	510	212	212	344	344	2243	2243	108	108	70	70	510	510	312	321	—	—	109	109
4	www.ceu.es	8965	8959	4367	4366	11836	11821	8833	8630	—	—	9037	9035	8966	8960	390	392	309	309	204	204
5	www.shef.ac.uk/dentalschool	58	58	143	143	131	131	56	56	56	56	59	59	58	58	58	58	71	71	57	57
6	www-cryst.bioc.cam.ac.uk	884	880	—	—	3104	3098	5042	5038	633	633	629	631	884	880	119	118	1019	1002	613	613
7	www.pem.cam.ac.uk	971	968	1885	1885	1940	1940	905	905	970	965	946	946	971	968	12	12	980	980	970	965
8	www.teipat.gr	345	649	776	776	681	683	309	309	307	309	617	618	643	647	—	314	321	321	313	315
9	www.dsg.unito.it	543	537	982	981	1143	1134	532	529	532	526	493	493	543	537	102	103	541	540	543	537
10	www.kun.nl/phil	511	511	783	783	767	765	349	349	510	510	511	511	511	511	351	352	528	526	511	511
11	www.mat.chalmers.se	3	3	7	7	7	7	3	3	3	3	3	3	3	3	3	3	3	3	3	3
12	cst.dk	461	460	888	888	859	859	376	378	460	459	411	411	461	460	—	—	—	431	425	425
13	www.montefiore.ulg.ac.be	1612	1704	5306	5347	5122	5626	14153	14164	1470	1534	1399	1401	1611	1703	380	380	1601	1592	1608	1671
14	wwwai.wu-wien.ac.at	19258	19284	50073	50790	53212	—	8391	8707	—	—	14353	14667	19257	19283	501	504	15713	16208	15787	16037
15	www.uni-saarland.de/fak8/iaua	30	30	121	121	135	135	26	29	30	30	30	30	30	30	29	30	30	30	30	30
16	www.medizin.uni-greifswald.de/humangen	14	14	27	27	28	28	13	13	14	14	14	14	14	14	14	14	13	13	14	14
17	www.math.jyu.fi	3267	3715	10360	—	9168	12504	—	—	—	—	2633	2767	3266	3714	310	313	2469	2518	2551	2828
18	www.alltheweb.com	1250	1155	225	—	1069	1075	—	—	967	963	49	49	1250	1155	150	151	7336	7328	51	51
19	www.etsia.upv.es	181	28470	815	815	753	—	171	171	177	73	180	180	181	28470	161	160	259	259	178	134
20	www.usal.es	553	104	240	241	822	226	84	84	518	83	86	86	553	104	83	83	89	89	96	83

		Microsoft Content Analyzer				SocSciBot				WebCount				WebKing		WebTrends				Xenu			
		Online		Offline		Online		Offline		Online		Offline		Online		Online		Offline		Online		Offline	
		031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029
1	www.cindoc.csic.es/cybermetrics	82	82	81	81	80	80	81	81	232	232	37	37	82	82	—	—	2491	2491	82	82	81	81
2	www.ite.upv.es	2	2	2	2	2	2	2	2	1	1	1	1	2	2	2	2	2	2	2	2	2	2
3	www.upct.es/~de	69	69	109	109	68	68	107	107	—	23	4	4	67	69	70	0	118	118	70	68	109	109
4	www.ceu.es	293	293	204	204	187	187	5093	3460	—	2182	5	5	253	253	516	—	—	—	8833	8844	—	—
5	www.shef.ac.uk/dentalschool	71	71	57	57	56	56	57	57	—	—	14	14	57	57	70	—	157	157	58	58	57	57
6	www-cryst.bioc.cam.ac.uk	1049	1049	613	613	538	469	—	—	25	25	5	5	—	601	—	—	—	—	6783	6435	639	639
7	www.pem.cam.ac.uk	978	974	970	965	1017	1018	905	900	19	19	12	12	904	900	—	—	1333	1328	950	950	970	965
8	www.teipat.gr	321	321	313	315	330	326	376	311	6	6	6	6	314	314	5	5	14	14	314	314	312	314
9	www.dsg.unito.it	544	544	543	537	452	453	415	413	38	38	25	25	496	496	187	—	729	718	537	537	—	536
10	www.kun.nl/phil	528	527	511	511	495	495	486	485	3	3	3	3	507	507	539	551	1385	1388	515	515	511	511
11	www.mat.chalmers.se	3	3	3	3	3	3	3	3	2	2	2	2	3	3	3	3	3	3	3	3	3	3
12	cst.dk	428	428	426	425	714	1017	419	417	—	557	15	15	124	6	621	—	—	0	512	512	—	457
13	www.montefiore.ulg.ac.be	779	1612	1601	1664	1680	3371	4063	—	32	32	27	27	1583	1579	—	—	—	3935	—	—	1582	1647
14	wwwai.wu-wien.ac.at	15792	16224	15786	16036	5106	5051	3668	3673	89	89	12	12	15103	13612	—	—	—	—	10534	—	—	—
15	www.uni-saarland.de/fak8/iaua	30	30	30	30	28	28	29	29	379	380	9	9	30	30	32	—	110	110	31	31	30	30
16	www.medizin.uni-greifswald.de/humangen	13	13	14	14	12	12	13	13	8	8	8	8	13	13	14	—	19	19	14	14	14	14
17	www.math.jyu.fi	2466	2520	2551	2828	1437	1462	4946	—	2739	2879	13	13	1826	1853	—	—	5397	5901	2858	614	—	—
18	www.alltheweb.com	95	95	51	51	57	57	977	—	14142	14142	13	13	49	49	360	341	725	1	—	—	—	—
19	www.etsia.upv.es	248	248	178	130	164	164	135	—	11	11	11	11	177	601	330	327	709	141	181	181	181	78
20	www.usal.es	89	89	96	83	83	3514	478	1	—	78	32	30	83	83	12	—	34	62	155	130	516	83

Annex 3.2. Evaluation of the commercial and academic software.

	Astra	**Blueprint**	**Microsoft**	**COAST**	**FunnelWeb**
Additional utilities	Scheduler	Scheduler Link checker	Site copy Link checker	Scheduler	Traffic statistics Word analysis Evaluation of quality
Coverage	Only first pages of dynamic programming	Dynamic sites	Only some dynamic objects		Some dynamic objects
Difficulties	Program errors in conflictive sites		Program errors in conflictive sites		
Graphics	Cyberbolic view	No graphics	Cyberbolic view	No graphics	Very good graphics
Limitations	Unclear data	Only one window per computer Expires in 15 days	Great web sites make it collapses	Expires in 15 days	Only 1000 items per site Can't save reports Only one window per computer
Output	Reports including some statistics (local or external pages, resources & dynamic) and link structure	Reports including statistics (total number of links, HTML files and images) and link structure	Reports including complete statistics (pages, links, objects and kind of objects) and link structure	Statistics (Dynamic HTML, GIF & JPG images, Flash, internal & external files…)	Graph view, word analysis, link structure, a few statistics: items, size, links (total, broken and external) & words
Options	Similar to other web analysers	Similar to other web analysers Includes files extensions definition	Similar to other web analysers	A big range of options Files extensions definition	Similar to other web analysers
Resources consumption	Great		Great (specially RAM)	Yes	Yes
Transparency	No	Yes	Yes	No	Yes

	SocSciBot	WebKing	WebTrends	WebCount	Xenu	Teleport
		Link checker	Log analysis Web site improvement suggestions	WayBack Machine data extractor	Link checker	Search of keywords
Coverage	Dynamic web sites	Dynamic web sites	Dynamic web sites	Unclear	Dynamic web sites	Only HTML
Difficulties	Specify URL containing Statistics extraction	**Additional utilities**	A lot of program errors		It collapses generating big reports	
Graphics	No	No	No	No	No	No
Limitations	Only until 10000 links Sub-domains are included as internal resources	LITE mode restrictions				Only 50000 files
Output	TXT reports including site structure, content and statistics	Complete statistics by file extension	HTML reports including a few statistics (HTML files, links, kind of files, size & errors)	TXT reports including some statistics (internal & external pages & links; words; images; CGI…) and errors found	HTML reports including some statistics (internal URLs by MIME type) and site structure	Copy of files
Options	Simple options	Similar to other web analysers	Similar to other web analysers Includes files extensions definition	Simple options	Simple options	
Resources consumption	No	Yes	Yes	No	Only big web sites	Specially disk space
Transparency	Yes	Yes	Unclear	No	Yes	Yes

Annex 3.3. Search engine delimiters.

DELIMITERS	WEBOMETRICS OPTIONS (AUGUST 2005)					
	GOOGLE	MSN SEARCH	YAHOO SEARCH	TEOMA	GIGABLAST	EXALEAD
National Domain	site:xx	site:xx	site:xx	site:xx inurl:xx	NO	site:xx
Institutional Domain Subdomain/Site	site:aaa.xx	site:aaa.xx	site:aaa.xx	site:aaa.xx inurl:aaa.xx	site:aaa.xx	site:aaa.xx
Subsite/ Directory	site:aaa.xx/bb	site:aaa.xx/bb	url:http://aaa.xx/bb	inurl:aaa.xx/ bb	url:aaa.xx/bb	site:aaa.xx/bb
Term in URL	inurl:bb	NO	inurl:bb	inurl:bb	suburl:bb	NO
Links to Domain/Site	NO	linkdomain:aaa.xx	linkdomain:aaa.xx	NO	link:aaa.xx	link:aaa.xx
Links to Subsite/ Directory	link:aaa.xx/bb	link:aaa.xx/bb	link:http://aaa.xx/bb	NO	link:aaa.xx/bb	link:aaa.xx/bb
Countries	form (adv) (200)	loc:XX	form (adv) (24)	NO	NO	NO
Languages	form (adv) (35)	language:xx	form (adv) (32)	lang:XX	NO	language:xx
File type	filetype:yyy	filetype:yyy	originurlextension:yyy	NO ?	type:yyy (filetype:yyy)	filetype:yyy

4

Data analysis

by Viv Cothey

4.1 Introduction

The intended unit of analysis will have been determined as described in chapter 2. This will have informed the web data collection (see chapter 3). In this chapter we now consider the practice of web data analysis. That is how the detail within the data analysis phase is operationalised. Web data presents some challenges in this respect that may not be encountered in other contexts. The interested reader should investigate also the topic of web data mining.

The particular areas of interdependent practice that are considered here relate to:

Validity: does the data truly represent what is claimed of it?

Scaling: how does practice accommodate the scale-free effects of the power law relationships that much web data appears to exhibit?

Sampling: all data collection from the web is necessarily a sample. How is this recognised in practice? How are issues of reliability addressed?

Analytical paradigm: the data analysis will be predicated upon some mental model of the web and its analytical structures. What impact do these have upon practice?

4.2 Defining/clarifying the unit of analysis

In the chapter units of analysis we emphasised the need for reliability of operationalisation and the need for a sufficient technical understanding of the data collection/analysis issues arising. Consider for example a macro analysis of degree (inlink and outlink). Then if the link extractor extracts from .pdf and .doc application files as well as from html text files, should these application files be included in the analysis? Are multiple arcs going to be simplified to a single arc? Are loops that are self-links going to be included? Are dead links going to be included? And so on.

Such questions together with their answers serve to define and clarify the unit of analysis being used. As well as improving reliability this enhances validity by focussing interest on the precise relationship between the empirical data and the data analysis. Another example could relate to server response time by determining the difference in time between the URL request and the server response. This would be fraught with problems. Not only do the client and server clocks need to be

synchronised but one would have to consider the effect of proxy servers. It is doubtful whether a valid study into server response could be undertaken.

The question of data validity is especially relevant in the context of testing theory. Clearly the data analysis should validly correspond to the conditions anticipated by the theory. For example if one wishes to test a particular web evolution theory that presumes a simplified web digraph, that is no loops and no multiple arcs, then the empirical data should be in this form also. On the other hand if the theory presumes no simplification then the empirical data should likewise not be simplified.

As commented previously the list of examples here could be endless. In each case the interdependence between how the data collection is operationalised, how the unit of analysis is refined and clarified and how the data analysis is operationalised should be noted. Continual vigilance is needed to ensure validity.

4.3 Scaling (non-linearity)

More and more phenomena or systems are being discovered that exhibit scale-free properties or that have emergent properties. Scale-free means that no matter how much we magnify the object, its features remain similar. An emergent property (of a system) is a stable characteristic that describes the whole system even though the components comprising the system are all independent.

A characteristic of these phenomena is that they exhibit counter intuitive properties such as the small world effect. It is also found that the distribution of parameters from the system follow power laws and are heavy tailed. In consequence the occurrence of unlikely or rare events is much greater than would be implied were the distributions to be classical. An important technical point that flows from this is that the underlying conditions for the application of much statistical analysis and hypothesis testing in particular are breached.

Several aspects of the web have been discovered to be scale-free, although positive identification is problematic. One identification criterion is a power law distribution. That is, if a parameter is distributed as a power law then the phenomenon is scale-free. The usual test for a power law is to plot the logarithms of the distribution and then determine the corresponding straight-line function. The distribution is a power law if the straight line is an appropriate approximation.

Determining the corresponding straight-line function in respect of power laws is itself problematic. The usual technique (least squares) is known to be error prone. An alternative technique exists (maximum likelihood estimation) but is little used. There are several variants in the detail of how least squares techniques are used. These include Barabasi's log-binning (1999), Crovella's complementary distribution (Crovella, Taqqu, & Bestavros, 1998), gap removal (Levene, Fenner, Loizou, & Wheeldon, 2002) and Katz's histogram methods (Katz & Katz, 2005).

Not only does scaling, that is power law relationships, need to be taken into account when analysing data collected from the web, scaling may also need to be considered when interpreting conclusions. This may be particularly relevant in the context of policy relevant indicators. Typically such indicators attempt to illustrate and illuminate disparate performance between, for example, innovation systems. These indicators themselves may be subject to scaling. A scale-free indicator is a way of accounting for either the cumulative advantage (or success breeds success) or cumulative disadvantage that may be present.

4.4 Sampling

It is inevitable that the data collection procedure will represent a sample. Therefore one needs to question what effect the sampling procedure has on the data collection and on the data analysis.

The concept of a sampling frame is well known. For web crawling the equivalent concept is the crawl space. Other data collection techniques should have their equivalents although in some cases such as commercial search engines little or no information is available.

There have been attempts to devise random, that is unbiased, sampling procedures. One such attempt was based on sampling the numerical IP address space within, say the crawl space. However technical advances in the way that IP addresses are exploited have vitiated this method.

A popular technique for operationalising sample spaces is by 'domain', for example by country-code top-level domains (ccTLD). File extension and MIME type are also used although both reliability and validity are in doubt. This is because one of the guiding engineering principles of the web is for both client and server software to be conservative in what they provide but liberal in what they accept. In consequence only the loosest of controls exist; files/resources and even entire servers can masquerade as whatever they like.

In some cases there may be a robust regime of entitlement to a domain name for example the .uk academic domain '.ac.uk'. However even here the criteria for inclusion is broad and notable institutions such as the British Library are not included. Many educational institutions also make use of servers within the '.net' or '.org' domains.

Identifying all the domain names within a domain is no longer possible and in any event a server may not have a DNS entry and therefore may not be eligible for inclusion within a sample space.

Hence even operationalising the sampling frame involves bias. Whilst this may not be serious, it should be recognised and the analyst should be mindful of any adverse consequences.

The data collection by search engines is biased. Details are not revealed but in general for example web pages from larger servers are under represented. The bias in TLD and language is not known but is changing as the search engines' competition in global reach intensifies.

In the event that a crawler is used then the design of the sample data collection from within the crawl space needs to take account of the intended analysis. Polar opposite procedures are 'broad' crawling and 'deep' crawling. Since crawling is expensive in both time and in other ways then any sample has to be a compromise between extensive sampling from a few servers to less extensive or light sampling from each server but including many servers. Clearly this prompts the question of how particular servers should be chosen or how light sampling might be organised.

One cannot and should not be prescriptive. Suffice to say that there should be an appropriate rationale for the sampling decisions that are made. Reliability in sampling rests on making these decisions explicit.

4.5 Analytical paradigm

Given appropriate data collection and sampling procedures in support of relevant units of analyses then the final step discussed here is the analytical paradigm. Possibly the fundamental paradigm or mental model of the web is the complex network of hypertext as represented by the web page digraph. This is only tractable in the case of small crawl spaces. The notion of clustering, node condensation and simplification can all be brought to bear in order to transform the digraph to more manageable proportions. Clearly the validity and utility of such transformations will depend upon the particular circumstances of the analysis.

A common paradigm is that of the 'web site'. It is unclear whether this should refer to a single (user-centric) web page, a collection of web pages within a server or an entire server. Popular usage seems to favour the collection of web pages within a server however it is unlikely that this usage can be operationalised. Research into identifying 'web sites' continues.

Transforming the web page digraph to a server graph has been proposed (Bharats, Chang, Henzinger, & Ruhl, 2001) together with varying amounts of simplification (Thelwall, 2004). Note however that this may well exclude resources that have dotted decimal IP addresses in place of domain names. A similar issue arises when a top-level domain analysis is carried out, that is not all nodes in the digraph will have a DNS entry.

The ccTLD can be used as a proxy for political country in the case of a by-country analysis. In addition to the DNS issues just noted, the ccTLD is not a valid indicator of the geographic locality of a server. Contra-wise a country's web presence is not validly represented by the collection of servers having a particular domain name. In particular larger companies may well use the .com domain.

A more reliable indicator of a server's national affiliation may be obtained from the geographical IP address assignment. However as an example of the complexities that arise consider the case of the server <www.heaven.li> that is registered in Switzerland on behalf of the Liechtenstein country domain administration. The machine hosting the server is physically sited in California while the administration of the server is carried out from London. This is not an exceptional instance. Administrations such as Niue (.nu) and Tuvalu (.tv) have turned the opportunities to their commercial advantage.

Analyses can be driven by content or by link structure as well as by the URL. Examples of a content paradigm include mining web data for communities or topics, that is collections of web pages that have semantic similarities. However as noted earlier these analyses are computationally intensive and thus can only be applied to small samples. It should be clear that in order to support such analyses extra care may needed as regards the data collection and sampling techniques used. For example, using topological clustering such as identifying graphical cliques requires that all arcs that is links and nodes have been collected; note the distinction between link-crawling and content-crawling.

4.6 Conclusion

We have shown that the data analysis phase cannot be considered separately from the data collection phase. They are interdependent aspects of the overall research design.

In addition we have shown how a reliable and valid research design necessitates an appropriate understanding of the web and a thoughtful consideration of the units of analysis. The abuse of 'hit rates' in order to generate the marketing hype that fuelled the dot-com boom should be a salutary example of how web data can be misused and misinterpreted.

References

Barbasi, A.L., & Albert, R. (1999). Emergence of scaling and random networks. *Science*, 286, 509-512.

Bharat, K., Chang, B., Henzinger, M., & Ruhl, M. (2001). Who links to whom: Mining linkage between web sites. In IEEE International Conference on Data Mining (ICDM '01), San Jose, California, November 2001. pg 51-58.

Crovella, M., Taqqu, M., & Bestavros, A. (1998). *Heavy-tailed probability distributions in the world wide web*. In R. E. Feldman R. J. Adler and M. S. Taqqu (Eds.), *A Practical Guide to Heavy Tails*. (pp. 3-26). Boston: Birkhauser.

Levene, M., Fenner, T., Loizou, G., & Wheeldon, R. (2002). A stochastic model for the evolution of the web. *Computer Networks*, 39,277--287.

Katz, L., & Katz, J.S. (2005). Personal communication

Thelwall, M. (2004). *Link Analysis: An information science approach.* San Diego: Elsevier.

5

Empirical Examples: Link analysis

by Mike Thelwall

5.1 Historical background

In the early years of the web, several information scientists recognised the structural similarity between hyperlinks and citations, noticing that both are inter-document connections and pointers (Larson, 1996; Rodríguez i Gairín, 1997; Rousseau, 1997). This underpinned the creation of a new field, webometrics (Almind & Ingwersen, 1997), defined to be the application of quantitative techniques to the web, using methods drawn from informetrics (Björneborn & Ingwersen, 2004).

The power of the web could first be easily tapped for link analysis when commercial search engines released interfaces that allowed link searches (Ingwersen, 1998; Rodríguez i Gairín, 1997). For example, from 1997 it was possible with AltaVista to submit extremely powerful queries, such as for the number of pages in the world that linked to Swedish pages (Ingwersen, 1998). This meant that with a few hours work submitting search engine queries, the 'impact' of sets of web sites could be compared, assuming links, like citations, to measure the impact of published information. In citation analysis, researchers typically need to pay for access to the citation database of the Institute for Scientific Information, but for link analysis the web 'database' is free, potentially giving it a wider set of users. Using commercial search engines the impact of many entities were compared, including journals, countries, universities or departments within a country and library web sites (An & Qiu, 2004; Harter & Ford, 2000; Ingwersen, 1998; Smith, 1999; Tang & Thelwall, 2005, to appear; Thomas & Willet, 2000). The early studies showed that care was needed to conduct appropriate link analyses because of many complicating factors such as duplicate web pages and sites, errors in search engine reporting, incomplete search engine coverage of the web, link replication within a site, and spurious or trivial reasons for link creation (Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; Egghe, 2000; Harter & Ford, 2000; Smith, 1999; Snyder & Rosenbaum, 1999; van Raan, 2001). Nevertheless, link analysis has produced interesting and useful results and has been adopted by several non-information science fields, as shown below.

A review of some current types of link analysis is given below, preceded by a brief methodological discussion and speculation about the range of types of information that this new informetric technique may be employed to help measure.

5.2 Data sources

Link data can be obtained from commercial search engines, borrowed from web link databases or obtained directly with a link crawler.

At the time of writing, Google's link command could be used to count the number of pages in Google's database that contain a hypertext link to any given page. For example, link:www.wlv.ac.uk/disclaimer.htm reports the number of pages linking to this URL. AltaVista's linkdomain: command, in contrast, counts the number of pages known by AltaVista to link to any page in the specified domain, a more general search. For example, the results of linkdomain:www.wlv.ac.uk would include all pages linking to any page in the www.wlv.ac.uk domain, not just to the home page. A researcher can conduct a relational analysis of the links between a set of web pages (Google) or sites (AltaVista) by obtaining link counts for all pairs individually from the search engine.

Commercial search engines have problems of coverage (i.e. not crawling some sites and crawling others incompletely), and so are not optimal for link analysis, although their use is often unavoidable (Thelwall, 2004). There is a collection of web link databases online at http://cybermetrics.wlv.ac.uk/database that includes the university web site link structures of five countries and is free for use by other researchers, including tools to analyse the results in various ways. A free web crawler is available at http://linkanalysis.wlv.ac.uk for those who need to gather their own data. This can crawl sites of up to 5,000 pages, but is not suitable for very large sites.

5.3 Review

5.3.1 Interdepartmental link analysis

Most departmental link analyses have aimed to validate link counts as a measure of research impact. A common hypothesis is that the number of links to a department correlates with an established research measure, such as citation counts. Link counts are often normalised by dividing the number of links to a department by the number of pages or researchers in the target department, versions of Ingwersen's (1998) Web Impact Factor. Typically also, links within a departmental site are excluded, assumed to be for internal navigational purposes.

Although early results were discouraging (Thomas & Willet, 2000) subsequent studies have demonstrated a correlation between research measures and link counts, supporting the use of links to track research (Li, Thelwall, Musgrove, & Wilkinson, 2003; Li, Thelwall, Wilkinson, & Musgrove, 2004a, 2004b). Links should not be used as a significant part of research assessment in the way that citations sometimes are, however, because only a small percentage of links reflect research achievements directly (e.g., links to online articles acting like online citations). Most are more indirect, such as teaching related or reflecting membership of a shared organisation or research group (Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004; Wilkinson, Harries, Thelwall, & Price, 2003).

Departmental link analyses have demonstrated enormous differences in web publishing. Even aside from natural web users like computer scientists, one study showed physicists to publish and interlink at least a thousand times more than historians in the U.S. (Tang & Thelwall, 2003). Li's (Li et al., 2004a, 2004b) studies of

disciplinary differences are the most detailed yet, covering similar hard science disciplines in similar countries (physics, chemistry and biology in Australia, Canada and the UK). She found that even similar disciplines used the web in very different ways, such as the extent to which they interlink. One logical explanation is differing national portfolios of sub-specialisms within each subject: for example the balance of organic and inorganic chemistry. This finding is also consistent with some theories from social informatics (Kling & McKim, 2000) which claim that the use of communication technology does not lead to easy universal patterns: small-scale practical needs can determine the way in which technologies are adopted and adapted. Perhaps more surprising are Li's international differences within the same subject, for instance Biology links in Australia were significantly less international (60%) than those of the UK (74%) and Canada (80%). From a functionalist perspective, and given the international nature of science, broad similarities in web use might be expected. Nevertheless the differences support organisational sociologies of science that emphasise the importance of multiple social factors on the practice of science (Fuchs, 1992). From a practical, informetric perspective the lesson is that link counts are perhaps most valuable for identifying unexpected differences and since link pages can be traced, identifying their cause. Thus, link analysis seems a natural partner to sociologies of science, as is citation analysis (e.g. with Merton).

A good example of an interesting application of link analysis is its use in conjunction with other sources of information about connections between researchers, such as European collaborative project membership, to investigate patterns of collaboration in Europe for a specific field (Heimeriks, Hörlesberger, & van den Besselaar, 2003). This shows that in scientometrics, link analysis can be used in conjunction with other techniques as part of a multiple method approach.

5.3.2 Social network analysis

A second set of researchers that have identified links as interesting because they are similar to their normal object of study are the social network analysts (Wasserman & Faust, 1994). These study groups of individuals or organisations, focussing on the connections between them. For example, one famous study analysed information flows in a village by word of mouth, using networks of social acquaintances to explain the communication patterns found (Rogers & Kincaid, 1981).

The phrase hyperlink network analysis has been coined for the use of social network analysis (SNA) techniques for networks formed by the links between web pages (Park, 2003). Research questions from social network analysis tend to focus on the network properties of sets of web sites, using social network analysis measures. These measures assess properties such as the centrality of individual nodes (i.e. web pages or sites) within a network, using metrics such as inlink counts, outlink counts and the frequency with which a node appears on the shortest chain of links between pairs of nodes, 'betweenness centrality'. An illustrative instance of this type of research is Garrido and Halavais's (2003) investigation of web sites supporting the Zapatista Mexican peasant revolutionary movement which found gender politics to play a surprisingly central role in the web network, second only to official Zapatista information sites. A criticism of SNA style hyperlink analysis is a tendency in some research to assume that hyperlinks are always communication devices, whereas in reality they play a variety of roles (Park & Thelwall, 2003).

In information science the potential for SNA techniques to assess information networks has now been recognised (Otte & Rousseau 2002). Björneborn (2004) has applied SNA metrics to UK academic web sites at the domain level, finding interesting patterns of cross-disciplinary connections. This is a new area of research that may yield new insights into information structures and academic communication patterns, perhaps inspired by sociological and mathematical theories of networks (Granovetter, 1973; Watts & Strogatz, 1998).

It is interesting that some computer scientists have produced analysis of social networks using web links (Adamic & Adar, 2003). Their curiosity about social groups may be related to the computer science need to develop software applications to support group activities such as online collaboration, and the potential to exploit group knowledge to improve computer systems (e.g., 'collaborative filtering'). Social analyses of the web seem to be a promising research direction for many different scientific fields.

5.3.3 Other social link analysis

Links have been exploited in social science research in a non-informetric context, with some promising results. The Internet and Elections Project (politicalweb.info) is an international collaborative attempt to compare web use in political elections across the world. Hyperlinks are a key part of the analysis. They are used to help identify relevant sites and the practice of linking is also an object of interest in its own right. A preliminary study of a US election concluded that link creation was not yet recognised sufficiently to be put on a formal footing in the sense of developing accepted codes of practice. Nevertheless, there was some convergence in their use, particularly in terms of the tendency to link to like-minded sites rather than differing opinions on a topic of interest (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003).

Link creation has also been of interest in qualitative social science research, indicating its importance as phenomenon in its own right. Examples include Hine's (2000) discussion of the linking practices of web authors contributing to a public debate and Beaulieu's (2005, in press) investigation into links targeted at the PubMed Central medical resources web site. Beauleau found that link creation was surprisingly rare, even for scientists that had their research papers in the PubMed central site and so could logically use a link to it in order to make their research more accessible.

5.3.4 The social sciences link analysis methodology

The following has been proposed as a generic framework for link analysis in social sciences research (Thelwall, 2004, p.3)

> Formulate an appropriate research question, taking into account existing knowledge of web structure.
> Conduct a pilot study.
> Identify web pages or sites that are appropriate to address a research question.
> Collect link data from a commercial search engine or a personal crawler, taking appropriate safeguards to ensure that the results obtained are accurate.

Apply data cleansing techniques to the links, if possible, and select an appropriate counting method.
Partially validate the link count results through correlation tests.
Partially validate the interpretation of the results through a link classification exercise.
Report results with an interpretation consistent with link classification exercise, including either a detailed description of the classification or exemplars to illustrate the categories.
Report the limitations of the study and parameters used in data collection and processing (stages 3 to 5).

This framework echoes many of the points made above, but notice that there is a preliminary stage with a pilot study. This is important because the variety of uses made of the web (Burnett & Marshall, 2002) means that our intuitions about what web links ought to be used for in a particular context can be wrong. The pilot study allows a research problem that will not yield an informative link analysis to be aborted before too much effort has been given to it. Perhaps the most important message of the framework, however, is the centrality of link type classification studies for results interpretation. If we have no idea why links are created then we can only make the most abstract inferences from link counts.

5.4 Future directions for link analysis

Within information science there are several promising future research directions at the moment. The first three are discussed briefly, and the fourth, blog link analysis is described in more detail.

Investigations into why links are created, particularly in academic contexts. Although there have been a few such studies already (Bar-Ilan, 2004b, 2004c; Harries et al., 2004; Wilkinson et al., 2003), the findings have emphasized the variety of link creation motivations used. This variety makes link classification studies difficult, but it would be interesting to know more about differences in link creation motivations.

Time series analyses. One problem endemic to web link analyses is that the web is continuously evolving and hence any web study may be out of date by the time it is published in the academic literature. Hence it is very important to know how all types of web link analysis results vary over time. A low rate of variation would lengthen the 'shelf-life' of webometric results.

Applying social network analysis measures to information collections. Following Björneborn (2004), this is a type of research that needs to be applied to web information in order to fully assess its value and give new insights into the structure of information and online groupings such as invisible colleges of academics (Caldas, 2003). One issue that will need to be resolved – perhaps differently in every study – is the fact that link creation is not endemic: the lack of a link between two web sites or pages does not imply that they are unrelated.

Supporting wider social sciences research. Since the web is not exclusively an academic space, it can be used in wider social science research both as an object in its own right (e.g. to study online communities) and as an easily accessible source of information about offline phenomena that happen to be reflected in the web.

5.5 Blog link analyses

Web logs (blogs) are online diaries maintained by millions of web users (BBC, 2005; Nardi, Schiano, Gumbrecht, & Swartz, 2004). Their main attraction is their ease of use. An inexperienced web user can create and maintain an attractive blog without the need to know any of the normal technical details of web publishing. Blogs are enormous repository of information of varying quality. There are many information-centred blogs, created by people wishing to provide frequently updated expert information on a given topic. Such blogs play the role of a specialist newsletter (Bar-Ilan, 2004a). An example is *The Shifted Librarian* (www.theshiftedlibrarian.com), which is full of facts related to libraries. There is extensive linking within and between blogs (Kumar, Novak, Raghavan, & Tomkins, 2004; Marlow, 2004). Many blogs allow visitors to post comments on a blog entry, and it is also easy for one blogger to post a follow-up comment on their own blog, linking to the original via its 'permanent URL'. Another link feature is the 'blogroll' list of links to other similar or recommended blogs. Since link creation is so easy and natural within blogspace, it seems a particularly promising medium for link analysis. In fact link counts are already used to compile a daily list of the top 100 'most popular' blogs (www.blogstreet.com/top100.html). It seems likely that link counts could also be used in ways analogous to citation analysis because blog links can connect individual posts 'documents' and bloggers ('authors') create many posts. There are also differences: the lack of quality control over blog posts; the fact that, unlike scientific publications, they are probably rarely central to their author's job; and the lack of the natural topic organisation that journals give articles. Nevertheless, the following seem to be likely applications of blog link counts.

Lists of the most popular blogs (by analogy with most highly cited authors).
Lists of the most popular individual blog posts (most highly cited articles).
Relational analysis/network diagrams of the links between blogs (author cocitation/citation graphs).
Relational analysis/network diagrams of the links between individual blog posts (article citation diagrams).

Apart from deciding what is possible for future blog link analyses it is important to discuss what is likely to be useful and how link analysis can best be exploited. Clearly there is no pressing need for evaluational blog link analysis with the same importance as the need to use citations to evaluate scientists' productivity: it might be useful but will not significantly help to direct billions of euros of government research funding. It seems that the key findings will be most useful for social sciences research by providing information about the phenomenon of blogging, and by providing data about the spread of individual topics (e.g. presidential debate topics could be of interest to political scientists) or, more generally, to analyse information diffusion in blogspace by finding spreading patterns that are common across many topics (Gill, 2004; Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). Topic-centred blog link analysis will probably need to employ some kind of text analysis to identify topic-relevant blogs or blog posts, and will probably need to be semi-automated with a program used to gather and filter blog data.

5.6 Case study: The Shifted Librarian

The Shifted Librarian Blog is heavily linked to: according to Google, on 21 January 2005, there were at least 148,000 web pages linking to its home page (via the query link:www.theshiftedlibrarian.com). This includes internal links (i.e. from pages within the same blog site), but browsing Google's results pages did not reveal too many of these, so it is clearly a highly linked-to site. This was crosschecked with AltaVista, which allows internal links to be suppressed in its results. AltaVista reported 167,000 site inlinks (linkdomain:www.theshiftedlibrarian.com NOT domain:theshiftedlibrarian.com) to any page on the site. This is an enormous number of links for a single librarian. Cambridge University's main web site, in contrast, could only manage 98,000 in AltaVista (linkdomain:www.cam.ac.uk NOT domain:cam.ac.uk). The huge number of links for this librarian is due to the structure of blogs and the fact that they are database driven, a kind of very light content management system. This means that links can easily be replicated throughout all pages on a blob. If the shifted librarian becomes a favourite for a prolific blogger, she may add a link on her main blogroll, which is automatically replicated throughout the site, perhaps creating tens of thousands of links with a single keystroke. Similar phenomena have been observed away from blogs, such as the automatic replication of links in link bars that are replicated to all pages of a web site (Thelwall, 2002), but this seems to be far more common in blogspace. This is a problem for link analysis because link counts will tend to reflect replication patterns in the source blog rather than the value of the blogs linked to. The solution may well be the same for an effective link analysis: to crawl the sites using a special purpose crawler and aggregate links into sites so that if there is more than one link between a pair of blogs, all links after the first are ignored.

5.7 Conclusion

There is still plenty of scope for more methodological research to explore how best to use links and the contexts in which they are most useful. This kind of research should be carried out by information scientists, who have an intellectual grounding in information issues, and citation analysis in particular.

The main strength of link analysis is that it can be applied in a wide variety of contexts, not just in its intellectual home of information science, but also more widely in the social sciences to help address problems that relate to the web or have a reflection in the web. Blog link analysis seems particularly promising in this regard because of the likely wide social background of bloggers. A second strength of link analysis is that both the data and tools to gather it are free and the tools (either search engine searches or the link crawler) are easy to learn. These two strengths make link analysis a practical new research tool. It would be logical for information scientists to use link analysis as part of collaborative research with other social scientists: this is another information-centred role that we can play (Beaulieu, 2004; Wouters, 2000). The principal weakness of link analysis is that link creation is an unsystematic phenomenon and partly dependent upon factors that are not of interest in most research, such as the design choices of individual web authors. For large-scale research, such decisions tend to even out, but this does not necessarily happen on a smaller scale. Hence, unless links themselves are the object of study, they will often need to be used in conjunction with other data sources (e.g. citations) for triangulation in studies with small web sites.

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*(3), 211-230.

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

An, L., & Qiu, J. P. (2004). Research on the relationships between Chinese journal impact factors and external web link counts and web impact factors. *Journal of Academic Librarianship, 30*(3), 199-204.

Bar-Ilan, J. (2001). Data collection methods on the web for informetric purposes - A review and analysis. *Scientometrics, 50*(1), 7-32.

Bar-Ilan, J. (2004a). Blogarians – A new breed of librarians. Proceedings of the American Society for Information Science & Technology.

Bar-Ilan, J. (2004b). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59*(3), 391-403.

Bar-Ilan, J. (2004c). Self-linking and self-linkded rates of academic institutions on the web. *Scientometrics, 59*(1), 29-41.

BBC. (2005). Blog reading explodes in America. Retrieved December 1, 2005, from http://news.bbc.co.uk/1/hi/technology/4145191.stm.

Beaulieu, A. (2004). From brainbank to database: the informational turn in the study of the brain. *Studies in History and Philosophy of Biological and Biomedical Sciences.*

Beaulieu, A. (2005, in press). Sociable hyperlinks: An ethnographic approach to connectivity. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet*. London: Berg.

Björneborn, L. (2004). Small-world link structures across an academic web space - a library and information science approach. Royal School of Library and Information Science, Copenhagen, Denmark.

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

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

Burnett, R., & Marshall, P. (2002). *Web theory: An introduction*. London: Routledge.

Caldas, A. (2003). Are newsgroups extending 'invisible colleges' into the digital infrastructure of science? *Economics of Innovation and New Technology, 12*(1), 43-60.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. *Journal of Information Science, 26*(5), 329-335.

Foot, K., Schneider, S., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere. *Journal of Computer Mediated Communication, 8*(4). Retrieved December 1, 2005, from http://www.ascusc.org/jcmc/vol8/issue4/foot.html

Fuchs, S. (1992). *The professional quest for truth: A social theory of science and knowledge*. Albany, NY: SUNY Press.

Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying Social Network Analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165-184). London: Routledge.

Gill, K. E. (2004). *How can we measure the influence of the blogosphere?* Paper presented at the WWW 2004 Workshop on the weblogging Ecosystem: Aggregation, Analysis and Dynamics.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*, 1360-1380.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through Blogspace.* Paper presented at the WWW2004, New York. Retrieved December 1, 2005, from http://www.www2004.org/proceedings/docs/1p491.pdf

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5).

Harter, S., & Ford, C. (2000). Web-based analysis of e-journal Impact: Approaches, problems, and issues. *Journal of American Society for Information Science, 51*(13), 1159-1176.

Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics, 58*(2), 391-413.

Hine, C. (2000). *Virtual Ethnography.* London: Sage.

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

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science, 51*(14), 1306-1320.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35-39.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. Paper presented at the AISS 59th annual meeting.

Li, X., Thelwall, M., Musgrove, P. B., & Wilkinson, D. (2003). The relationship between the WIFs or inlinks of computer science departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics, 57*(2), 239-255.

Li, X., Thelwall, M., Wilkinson, D., & Musgrove, P. B. (2004a). National and international university departmental web site interlinking, part 1: Validation of departmental link analysis. *Submitted*.

Li, X., Thelwall, M., Wilkinson, D., & Musgrove, P. B. (2004b). National and international university departmental web site interlinking, part 2:Link patterns. *Submitted*.

Marlow, C. (2004). Audience, structure and authority in the weblog community. *International Communication Association Conference*. Retrieved December 1, 2005, from http://web.media.mit.edu/~cameron/cv/pubs/04-01.pdf.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41-46.

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

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections, 25*(1), 49-61.

Park, H. W., & Thelwall, M. (2003). Hyperlink analyses of the world wide web: A review. *Journal of Computer-Mediated Communication, 8*(4). Retrieved December 1, 2005, from http://www.ascusc.org/jcmc/vol8/issue4/park.html.

Rodríguez i Gairín, J. M. (1997). Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red. *Revista Española de Documentación Científica, 20*(2), 175-181.

Rogers, E. M., & Kincaid, D. L. (1981). *Communication Networks: Toward a New Paradigm for Research*. New York: Free Press.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics, 1*(1), http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html.

Smith, A. G. (1999). A tale of two web spaces; comparing sites using web Impact Factors. *Journal of Documentation, 55*(5), 577-592.

Snyder, H. W., & Rosenbaum, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation, 55*(4), 375-384.

Tang, R., & Thelwall, M. (2003). US academic departmental web-site interlinking in the United States disciplinary differences. *Library and Information Science Research, 25*(4), 437-458.

Tang, R., & Thelwall, M. (2005, to appear). A hyperlink analysis of US public and academic libraries' web sites. *Library Quarterly*.

Thelwall, M. (2002). Conceptualizing documentation on the web: an evaluation of different heuristic-based models for counting links between university web sites. *Journal of American Society for Information Science and Technology, 53*(12), 995-1005.

Thelwall, M. (2004). *Link Analysis: An Information Science Approach*. San Diego: Academic Press.

Thomas, O., & Willet, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science, 26*(6), 421-428.

van Raan, A. F. J. (2001). Bibliometrics and Internet: Some observations and expectations. *Scientometrics, 50*(1), 59-63.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, NY: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). 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), 49-56.

Wouters, P. (2000). *Cyberscience: The informational turn in science.* Paper presented at the Lecture at the Free University, Amsterdam.

6

Empirical Examples: Co-link indicators

by Sylvan Katz

6.1 Introduction

This chapter describes the methods that were used to gather co-link data and to prepare co-link indicators of 862 natural science and engineering group web sites at 50 major European universities.

Two documents are considered to be co-linked if the links to them occur in the same web document. Citation theory says that if a pair of documents is frequently co-cited then the contents of the documents are related in some manner. This property was investigated within the framework of co-linked web documents and then it was used to build example co-link web indicators. The examples illustrate how co-link indicators can used to investigate questions such as 'How frequently are ERA university group web documents co-linked with web documents on servers in other ERA countries or on US servers?' and 'Are linguistic and cultural biases present in web indicators data like they are in bibliometric data?' Google was used to identify source documents that contained links to university group web sites. A custom web crawler was used to collect links from source documents and the links from these documents were used to build co-link indicators.

6.2 Findings

Indicators were constructed using geographical information about the location of hosts of the source documents and co-linked documents. These indicators showed that ERA group web sites were most often linked to by source documents that reside on European (48%) and US (39%) hosts. They showed that documents in ERA group web sites tended to be co-linked most frequently with documents residing on European hosts. ERA university group web sites tended to be most often co-linked with documents on servers in their own country and the second most frequent co-link are with documents on servers in the UK or Germany. These findings have the potential of being indicators of European cohesion and language and cultural biases on the Internet and/or Google.

Interpreting co-link indicators and web indicators can be problematic. Unlike bibliometric indicators such things as basic units for counting and the limitations of these units are not well defined for web indicators. Also, unlike scientific publications there is little, if any, quality control on the content of web pages and the content can change with time. We are just beginning to understand the usefulness and difficulties of

creating and interpreting web indicators for innovation systems. The information gained from this project will help clarify the problems associated with producing and using co-link web indicators of science, engineering and innovation research and provide a focus for future research.

Indicators were used to measure characteristics about the geographical distribution of hosts on which the source and co-linked documents resided. An examination of source documents showed that the authors of these documents tended to link more frequently to documents located on servers in the same country as the source document server. Documents from European university group web sites tended to be co-linked with documents residing on European servers. Depending on the discipline between 69% and 80% of the co-linked host domains were located on European servers and between 12% and 21% were located on US servers. An in-depth examination of European co-linked host patterns showed that a group's documents tended to be co-linked more frequently with documents on servers in their home country. The location of the second highest co-linked documents was on a UK or Germany server. These data suggest such things as social networks, culture and language may be influencing the linking and co-linking patterns. Perhaps this type of web indicator can used to shed more light on the question of the cohesiveness of the ERA.

6.3 Conclusions

The web is huge. It is growing at an astounding rate. It is also very noisy. Unlike scientific papers, web pages have little or no quality control. This fact alone makes it difficult or impossible to construct web indicators that are directly comparable to bibliometric indicators. Perhaps the best we can hope for is to look for correlations between web and bibliometric indicators. However, few people would dispute the fact that university web sites provide a valuable and significant way to disseminate research information and attract attention to a research group's activities. From this perspective it makes sense to explore how to collect data to build web indicators of this segment of an innovation system. Interpreting indicators built from such a noisy environment as the web is problematic and it will take a lot of research to understand what these web indicators are telling us about the innovation community.

In summary, the field of web indicators research is a very young. It will take a lot of time and effort to design and test these indicators before we really begin to understand what can be measured and how to interpret what we are measuring. Many fundamental issues concerning the construction and interpretation of web indicators have yet to be resolved. Best practice data collection and analytical techniques are still evolving.

7

Empirical Examples: Small scale, in-depth case studies

by Hildrun Kretschmer

7.1 Introduction

A search for a new method to study the extent to which collaboration structures visible on the web follow similar rules to collaboration networks measured by bibliometric data was proposed in Kretschmer and Aguillo (2004). Although at present the interest in web hyperlink research is increasing (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2001; Kretschmer & Thelwall, 2004) there also needs to be indicators independent of hyperlinks: hyperlink structures do not provide more evidence of collaboration than is shown by traditional bibliometric analyses (Kretschmer & Aguillo, 2004; Kretschmer, 2004; Kretschmer, Kretschmer, & Kretschmer, 2005a); and it is argued that hyperlinks are not a promising source of quantitative information about gender differences in communication strategies or online visibility, at least for senior researchers or research groups (Thelwall, M., F. Barjak & H. Kretschmer, submitted).

In this way a new indicator, Web Visibility of Collaboration, was developed and tested in pilot studies, i.e., in small samples (Kretschmer & Aguillo, 2004; Kretschmer & Aguillo, 2005; Kretschmer, Kretschmer, & Kretschmer, 2005a, 2005b). Results of the pilot studies are shown here in part. However the sample sizes of the pilot studies are not sufficient enough for the main purpose; although trends can already be seen, discovery of laws (rules) requires larger samples and a variety of different scientific disciplines. There are only a few general rules in bibliographic co-authorship networks available from the literature. Thus, general rules in bibliographic collaboration structures that are suitable for testing in web networks had to be collected. Along these lines two approaches are proposed in two other papers (Kretschmer, 2004; Kretschmer & Kretschmer, 2004) for future testing in the bibliographic and web networks.

7.2 Web visibility indicators of collaboration

According to Vaughan and Shaw (2003) 'web citations' refer to web text citations or mentions of published papers on the web. These authors searched for citations to articles on the web using the Google search engine. The search strategy was to enter the article's title in quotation marks (i.e., phrase search in Google). There are different

categories of citing items, for example the citation of a publication in the on-line version of an article or lists of bibliographies for the students or publication lists in own home page, etc.

Web citations are different from web hyperlinks. The former refers to web text citations or mentions of published papers on the web, while the latter refers to hypertext links seen on web pages. There are many studies on web hyperlinks, but very few studies have examined web citations (Vaughan & Shaw, 2003).

Vaughan and Shaw's (2003) method of searching for web citations was used successfully, albeit in a slightly modified form, to measure the visibility of the collaboration on the web with the following definitions of new web visibility indicators of collaboration (already presented in Kretschmer& Aguillo, 2004):

> The *web visibility rate of a multi-authored publication from bibliographic data (WVP)* is measured as a frequency of the different web sites on which the bibliographic publication is mentioned after entering the full title of a co-authored publication into Google or into another search engine.
> A multi-authored publication obtained by bibliographic data is visible on the web if the following is valid: WVP>0.
> The *web visibility rate of a pair of collaborators (WVC)* is equal to the sum of web visibility rates (ΣWVP), of all of their co-authored publications.
> A pair of collaborators is visible on the web if the following is valid: WVC>0.
> This data was the basis for the social network analysis (SNA) of the co-authorship network from webometric data. The two collaborators are nodes in the network and there is an edge between them if WVC>0.

Example with a total of 2 publications:

Title of the first publication: Cytometric Analyses in Patients with Systemic Autoimmune-Diseases.

Co-authors: Dorner T, Odendahl M, Radbruch A

Institutions: - Institut für Medizinische Immunologie, Charité, Berlin, Germany
-Deutsches Rheuma-Forschungszentrum Berlin, Germany

Entering into Google: "Cytometric Analyses in Patients with Systemic Autoimmune-Diseases."

Google Hits (on September 15, 2004) - 1 Website:

[PDF] rheu433 389..395

PDF/Adobe Acrobat

www.springerlink.com/index/82M1LMR0W0F28KKD.pdf

Note: Both the number of the Google hits and the corresponding mentioned websites can change over time. Thus, the date of the search has to be marked.

Results: WVP=1
WVC (Dorner T / Odendahl M)=1
WVC (Dorner T / Radbruch A)=1
WVC (Odendahl M / Radbruch A)=1

WVC (Institut für Medizinische Immunologie, Charité, Berlin, Germany/Deutsches Rheuma-Forschungszentrum Berlin, Germany)=1

Let us assume there is a second publication with another title:

Co-authors: Dorner T, Odendahl M
Institutions: Institut für Medizinische Immunologie, Charité, Berlin,Germany
Results: Google Hits (on September 15, 2004) - 3 websites
WVP=3
WVC (Dorner T, Odendahl M)=3

Results for the two publications in total:

ΣWVP=4
WVC (Dorner T / Odendahl M)=4
WVC (Dorner T / Radbruch A)=1
WVC (Odendahl M / Radbruch A)=1
WVC (Institut für Medizinische Immunologie, Charité, Berlin, Germany/Deutsches Rheuma-Forschungszentrum Berlin, Germany)=1

Hyperlinks and web visibility rates are not only different from the point of view of the content, as mentioned above, but additionally hyperlink structures may be different from the co-authorship structures based on web visibility rates.

7.3 Possible differences between hyperlink structures and co-authorship structures visible on the web

Using hyperlinks, the comparison of collaboration patterns online and offline needs an offline starting point in the form of lists of institution or lists of individual scientists. The institutions or individual scientists are the nodes (or vertices) of the collaboration networks offline in the form of co-authorship networks. However, using the new web visibility indicators, this comparison of collaboration patterns needs an additional offline starting point in the form of a list of the titles of the co-authored papers. If one starts with this list of co-authored publications or bibliographies, the web visibility rate gives an indication for the visibility of this collaboration online. As a consequence there are some differences between the possible links between the nodes online after using hyperlinks or web visibility rates.

Although in both methods, using hyperlinks or using web visibility rates, the institutions or individual scientists (or their web sites respectively) are the nodes of the collaboration networks on-line, there are some constraints on the existence of edges online when using web visibility indicators (Example in Figure 7.1):

Hyperlinks (in the form of arcs or edges) can occur between any two web sites independently from co-authorship relations offline.

However, using the Web Visibility Indicators, only those co-authorship relations (edges) can become visible on the web which exists in the co-authorship network offline already. Thus, the edges visible on the web are a subset of the set of edges offline. The maximum number of edges online is equal to the number of edges offline.

> When using web visibility indicators, the similarity between the collaboration networks offline and online increases with the number of visible edges in the web network. This not necessarily the case using the hyperlinks.

The bibliographic collaboration network of six nodes is presented in the upper graph, right side, of Figure 7.1. The second and the third graphs below, right side, belong to different web networks after using the web visibility indicators.

An edge exists between two nodes if the related institutions or individual scientists have published at least one publication in co-authorship. Using the web visibility indicators of collaboration either this edge is visible in the web network or it has disappeared. A higher percentage of the bibliographic edges are visible in the second graph than in the third. Thus, the similarity between the second graph and the bibliometric graph is higher than between the third and the bibliometric graphs.

A hyperlink network is presented on the left side of Figure 7.1. Although the number of edges is equal to the number of edges the second graph on the right side, the similarity to the bibliographic graph is less. However it is possible for the graph of a hyperlink network to be similar or equal to a bibliographic network.

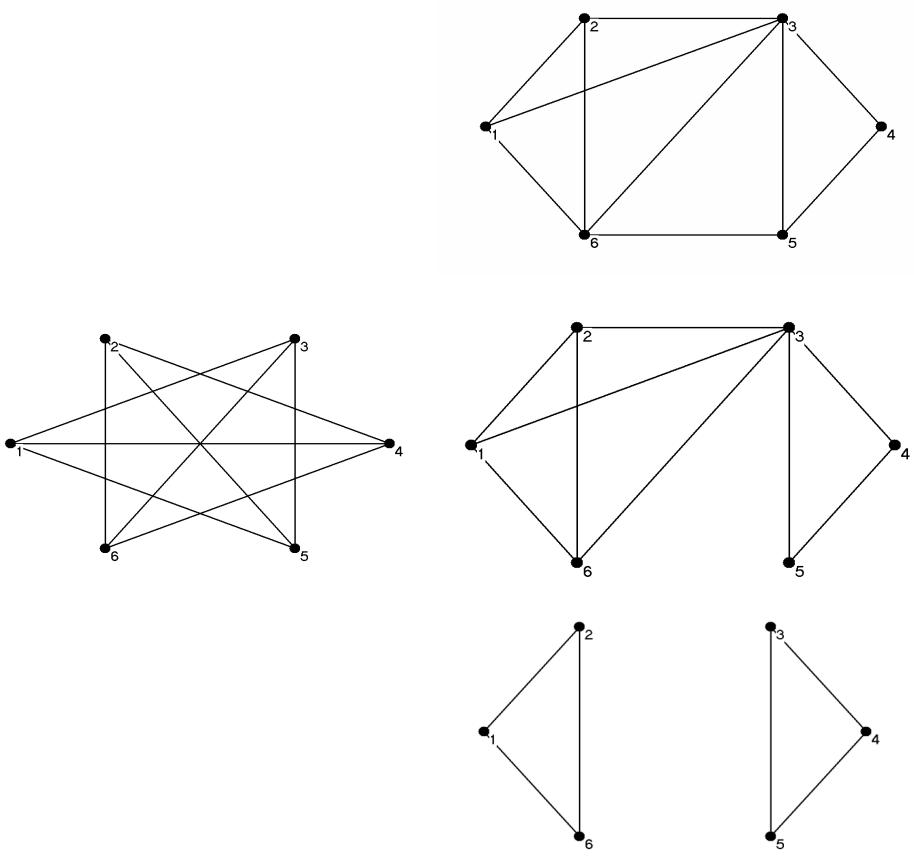

Figure 7.1. Possible differences between hyperlink structures and co-authorship structures visible on the web.

Note: The bibliographic collaboration network of six nodes is presented in the upper graph, right side. The second and the third graphs below, right side, belong to

different web networks after using the web visibility indicators. A hyperlink network is presented on the left side beside the second graph using the web visibility indicators.

Similarities between hyperlink structures and bibliographic co-authorship structures can be based on the same background: collaboration. However similar structures can also be caused by different reasons. Thus, further background investigations have to be done for clarification.

Differences between hyperlink structures and bibliographic co-authorship structures can occur on the basis of the verified hypothesis by Cothey (2004), mentioned above: that the link structure of the web should show more evidence of 'collaboration' than is seen in traditional bibliometric analyses and this is because the web: can provide evidence of 'work in progress'; and the normative constraints of peer review publishing do not apply. However, additional studies of suitable motivation for the creation of reciprocal hyperlinks have to be done for this explanation (Wilkinson, Harries, Thelwall, & Price, 2003). Using the web visibility indicators (WVP>0) increasing similarity between the collaboration networks off-line and on-line means increasing visibility of the bibliographic co-author networks on the web: Additional information can be delivered after checking the amount of the WVP per co-authored paper or the amount of WVC per pair of collaborating institutions, etc.

The disadvantage of web visibility indicators is that the comparison of collaboration patterns needing an additional off-line starting point in the form of a list of the titles of the co-authored papers has to be improved after further research. To this end, the development of a modified method is proposed using the new search engine: http://scholar.google.com.

7.4 Testing web visibility indicators of collaboration in the German Society of Immunology

The extent to which collaboration structures visible on the web follow similar rules to collaboration networks measured by bibliometric data was analysed. Expected results are new indicators for collaboration on the web: strength and structure of collaboration. The two different kinds of web indicators presented above, i.e., web hyperlinks between the web sites of institutions or home pages of individual scientists, and web visibility indicators of collaboration, have to be empirically tested in several studies.

The co-authorship network from bibliometric data (SCI data from 2002) was compared both with the hyperlink structure between 80 research institutions of the German Society of Immunology and the co-authorship network from web visibility rates. Hardly any hyperlink structure could be found between the 80 institutions. However using the web visibility rates the web network visualised by the map drawn with Pajek (Figure 7.3) is almost the same as the bibliometric co-authorship network (Figure 7.2). 50 nodes and 70 edges are presented in the bibliometric graph, 50 nodes and 63 out of these 70 edges are in the web-graph.

From the SCI data (year 2002), all publications were selected that appeared in co-authorship between authors affiliated to at least two institutions of the 80 research institutions of the German Society for Immunology. Thus, it concerns 87 bibliographic co-authored publications.

There are 50 nodes presented in the graph of the bibliographic network (Figure 7.2), i.e. 50 of the 80 institutions (62.5%) are linked with at least one other institution of

immunology by co-authorship between the scientists of these institutions. That means, at least one 'edge' (or line in the graph) is adjacent to each of these 50 nodes. The other 30 institutions are not presented in Figure 7.2 because they are singletons without a link to any of the other institutes in the SCI data, 2002. This situation can change over time.

In general, a pair of institutes can be linked by one or more co-authored papers. In this way the 87 bibliographic co-authored publications are distributed over the pairs of institutes.

There are 70 edges adjacent to the 50 nodes in the bibliographic network.

77 co-authored publications (=88.5% of the 87 bibliographic co-authored publications) became visible at least once in Google. The structure of the network obtained from the web is similar to the structure of the network obtained from the bibliographies. The bibliometric network is slightly reduced on the web. There are only 7 missing edges in the web network. In other words there are 63 edges visible on the web.

Discussion about the central role of special institutions in the collaboration networks offline and online is presented in another paper (Kretschmer et.al. 2005a).

Beyond the network analysis there are also some additional results using the web visibility indicators.

As mentioned above, the web visibility rate of a bibliographic co-authored publication (WVP) is measured by the frequency of the different web sites, on which this publication is mentioned. The curvilinear distribution of the 87 co-authored publications with definite WVP is presented in Figure 7.4. Only 11.5% of the 87 bibliographic papers, i.e. 10 co-authored publications with WVP=0, are not visible on the web. Thus, it can be stated that bibliographic co-authored publications which were investigated in immunology are visible to a high percentage in the web and that it follows, therefore, that collaboration between institutions is well reflected on the web.

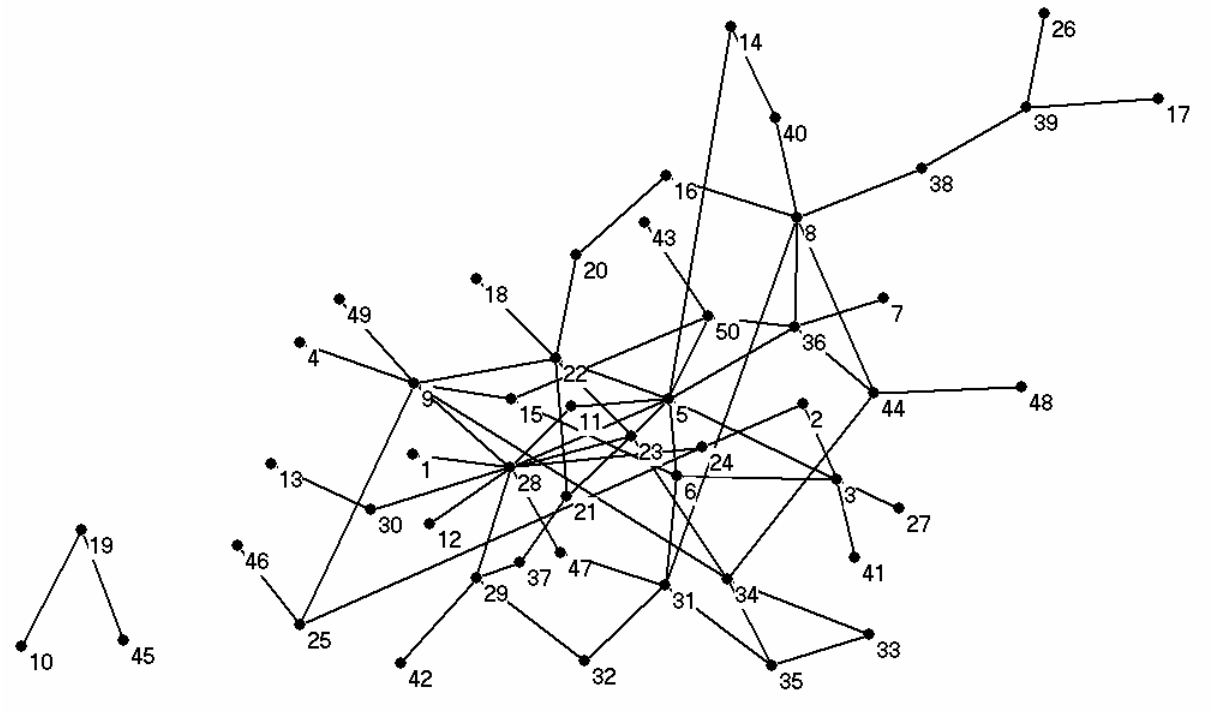

Figure 7.2. Bibliometric co-authorship networks of 50 institutions in the German Society for Immunology.

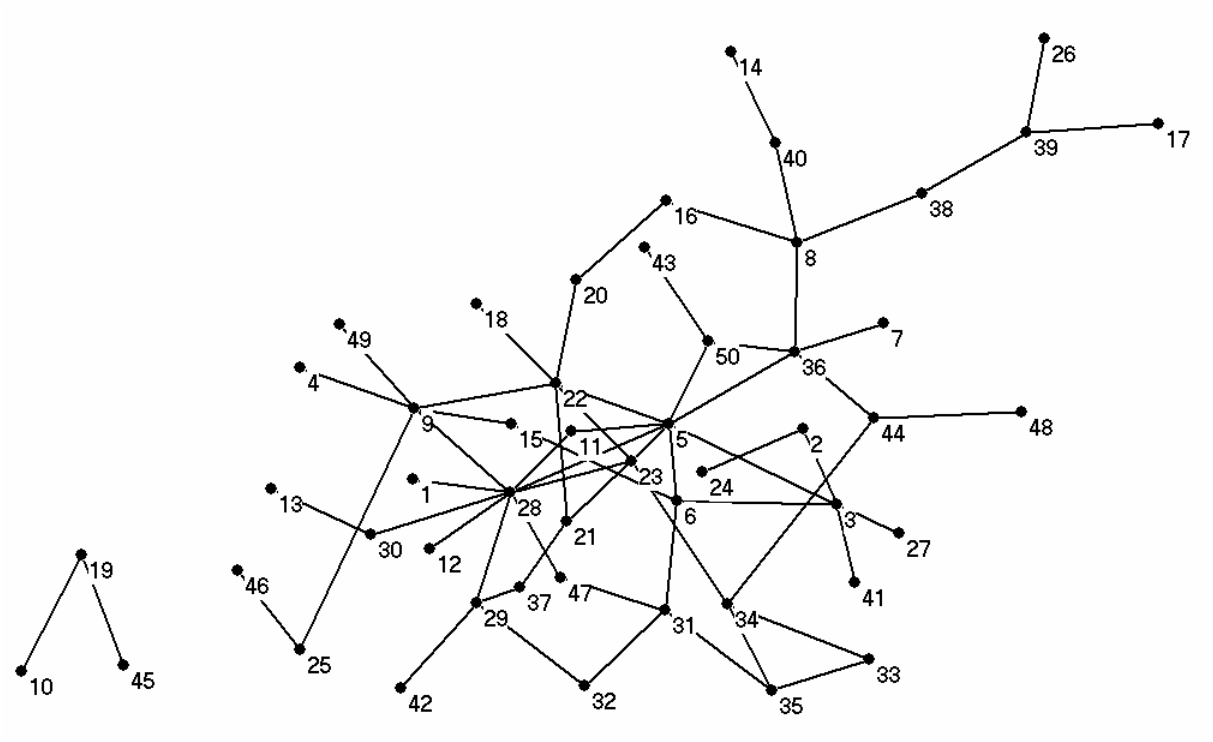

Figure 7.3. Web co-authorship networks of 50 institutions in the German Society for Immunology.

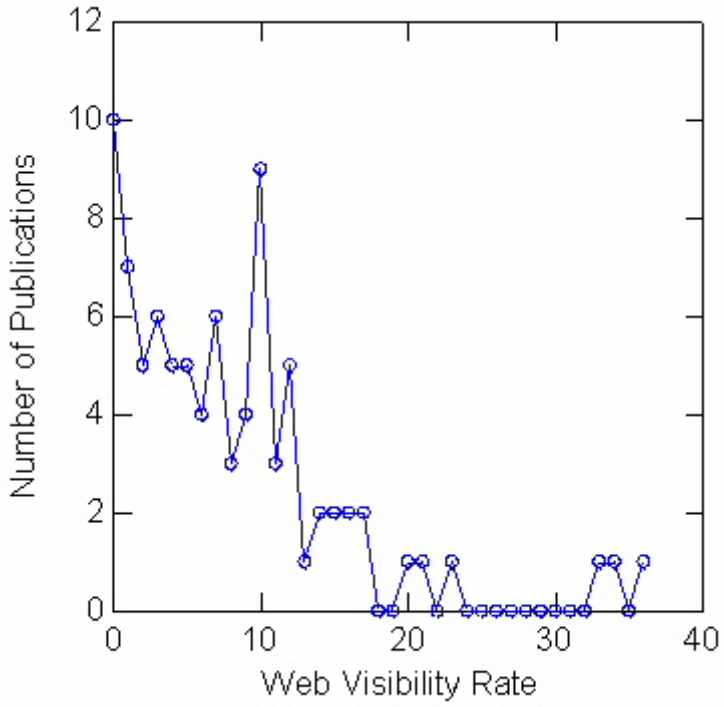

Figure 7.4. Distribution of co-authored publications with definite Web Visibility Rates.

7.5 Conclusion

The results of some empirical pilot studies have suggested that 'web visibility rates of collaboration' can be used as web indicators of collaboration whereas counts of hyperlinks are not useful in reflecting collaboration structures measured by bibliometric data. Although trends could be already shown in the small pilot studies, the discovery of laws (rules) needs studies of larger samples of different scientific disciplines in future.

7.6 References

Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404-426.

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

Cothey, V. & Kretschmer, H. (2004). Does the link structure of the web provide evidence of a collaborative hypertext? *Journal of Information Management and Scientometrics*. 1(2), 9-12.

Kretschmer. H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the web. *Scientometirics*, 60(3), 409-420.

Kretschmer, H. (2004, in Chinese). Visibility of collaboration on the web. In: Liu Zeyuan & Wang Xu-kun (Eds.), Science, Technology, Development. 2003's Research Yearbook on Science Studies and Management of Science and Technology in China. Dalian University of Technology Press: Dalian, 2004, 114-122.

Kretschmer, H., & Aguillo, I. (2004). Visibility of collaboration on the web. *Scientometrics*. 61(3), 405-426.

Kretschmer, H., & Kretschmer, T. (2004). Comparison of rules in bibliographic and in web networks. In: T.A.V. Murthy, S.M. Salgar, G. Makhdumi, P. Pichappan, Y. Patel and J. K. Vijayakumar (Eds.). Proceedings of the Second International Caliber 2004: Road Map to New Generation of Libraries Using Emerging Technologies. February 11-13, 2004, New Delhi, India. INFLIBNET Centre: Ahmedabad, 470-486.

Kretschmer, H., & Thelwall, M. (2004). From librametry to webometrics. *Journal of Information Management and Scientometrics,* 1(1), 1-7.

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

Kretschmer, H., Kretschmer, U., & Kretschmer, T. (2005a). Reflection of Co-authorship Networks in the web: web Hyperlinks versus web Visibility Rates. In: Proceedings of the 5th Triple Helix Conference, May 18-21, 2005, Turin, Italy (CD-ROM).

Kretschmer, H., Kretschmer, U., & Kretschmer, T. (2005b). Visibility of collaboration between immunology institutions on the web: including aspects of gender studies. In: Peter Ingwersen & Birger Larsen (Eds.). Proceedings of the 10th ISSI International Conference on Scientometrics and Informetrics, July 24-28, 2005, Stockholm, Sweden, Volume 2. Published by Karolinska University Press: Stockholm, 750-760.

Thelwall, M., Barjak, F., & Kretschmer, H. (submitted). Web links and gender in science: An exploratory analysis.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology,* 54(14), 1313-1322.

Wilkinson, D., Harries, G., Thelwall, M. & Price, E. (2003). 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), 59-66.

8

Terminology and definitions

by Isidro Aguilo

Basic definitions

During the preparation of web indicators terminology is needed for their definition and description. Some concepts require new coined names and there is a need to review the concepts of some terms. Basically we adopted the terminology proposed by W3C (http://www.w3.org/WCA), but extending and adapting names and definitions for our purposes. The proposed semantics will probably be improved in the future, as a consensus will be reached.

Domain

Definition: The Internet naming scheme consists of a hierarchical sequence of names, from the most general to the most specific (right to left), separated by dots. Technically, each part of a name is a domain, but domain is usually reserved to refer to the highest-level domains.

Example: It is very difficult to establish a classification of domains, but at least three types of domains can be recognised: geographical domains that refers to both countries (fr for France, uk for United Kingdom) or territories of those countries (nc for New Caledonia, fk for Falkland Islands); USA domains, that consists of those exclusively American like us, gov and mil but also edu, that involves a limited number of foreign academic websites); and the general domains like com, net, info or biz, where the US presence is probably over 50% and the more international org and int (with a large contribution of the European Commission sub domain eu.int).

Subdomain

Definition: A second level of naming domains used for identifying group of organisations or activities. Usually they follow the US model, grouping governmental bodies, academic institutions, commercial organisations or military resources, but there is no general agreement on their use or nomenclature.

Examples: Not all the domains have been subdivided into subdomains, e.g. France has not applied this theoretical schema. The situation is futher complicated as many subdomains represents different kinds of institutions in different countries. Academic subdomains are usually ac (like in ac.uk) or edu (like in edu.au), although in

some countries all the education resources are under the same subdomain or both subdomains being used for universities websites.

Institutional domain

Definition: The organisations and some individuals register a domain name for their reserved use, so they can be located on the Internet in a distinctive way. Generally only registered domains are used for counts, although some entities generate virtual (third level) domains, preserving the institutional name. These virtual sites are not considered institutional domains.

Examples: There are two and three unit domains like unam.mx (Universidad Autonoma de Mexico) or cam.ac.uk (University of Cambridge), but virtual domains are ever built on them: dgbiblio.unam.mx and lib.cam.ac.uk are used for their libraries.

Host

Definition: Unfortunately the term host is used in several different contexts. Usually it refers a computer connected to the Internet with its own and distinctive network or IP number. Sometimes it is referred also as a node, as several different computers can have Internet access from this point. Taking into account some services uses dynamic IPs and firewalls and proxies acting as nodes it is not correct consider the number of hosts represents the number of computers connected to the Net, but it is the best estimation and the one used in practice. Host and site are also used for denoting the authority part of a web address, i.e., not including any path or directory. Although W3C uses the term authority, both host and site terms are more commonly applied.

Examples: Your personal computer connected to the Internet with permanent IP number 161.111.200.88 is a host. If this computer is serving web pages using webdata.eicstes.org address, then you can refer this authority part of the URL as the host name.

Resource / web resource

Definition: Data or information that form an identified resource on the web.

Examples: A multimedia file (a movie trailer), a rich file (a pdf document), a page (a home page), a collection of pages (an electronic journal or a personal “page”), one site or a series of sites using the same institutional domain are different kinds of web resources.

Web publisher

Definition: The publisher is the person or organisation with the ultimate responsibility on the authorship of a web resource. It may be that a webmaster is invoked as such; unless it is indicated to the contrary the publisher is the institution.

Web server

Definition: A computer connected to the Internet that provides web pages, using special programs called server software. It is a physical unit that can host several different sites or be part of a network serving a big site.

Example: The *Alltheweb* search engine is only one website, but to maintain the service a lot of different computers are serving web pages under the common name.

Web site

Definition: A location on the World Wide Web made up of web pages containing graphics, text, audio, video and other dynamic and/or static materials. Usually it is hierarchically organised so it can be represented by the address of the home page. A site has its own distinctive host name, although some sites can be reached using alternative or alias names (bioinf.well.ox.ac.uk is the same site as bioinformatics.well.ox.ac.uk).

Institutional web site

Definition: The institutional site can be a web site, if the web publisher has an autonomous presence using a host name or not, if they are using a web collection accessed from a directory or other part of the path in the URL address. An institutional site can be nested or nest another institutional site.

Example: An organisation hosting a group of departments with distinct web presence and an electronic journal has one central institutional site, several departmental ones plus another considering the journal.

Sub-site

Definition: A sub-site is a site hosted by a different organisation without providing distinctive host names. Although some sub-sites could be considered as institutional web sites due to their content, the maintenance of the hosting service name in the address can be misleading. Besides, the volatility of these resources is usually very high, so they are not included in the former category.

Example: All the personal pages in the Geocities server that maintains the hostname geocities.com.

Web page / host page

Definition: A web page is an HTML (or similar) file containing texts and specifications about other files (images, multimedia objects) that can be displayed together as a unit in a browser. Every web page has a unique URL address. Host page is the main page of a web site, usually invoked by the hostname or under the name index.html or similar.

Example: Usually web pages are in HTML format (html or htm), but you can find secure pages (shtml or shtm) or active pages (asp, php).

Orphaned page

Definition: A page not linked from any other web page.

Depth

Definition: Number of levels in a web site, where level 0 or 1 is indicated by the home page and then each nested directory add a new level.

Web collection

Definition: A self-contained resource consisting of pages maintained by the same publisher.

Example: A personal 'page'.

URL address

Definition: A URL 'address' refers to the unique location of either a web site or a specific file (generally web page), indicating the name and location of the data on the involved computer. Usually a URL consists of the protocol (e.g., http://), the host name (e.g., www.ucla.edu), and the path (sometimes invisible or represented only by the slash (/), or a string of directories).

Object

Definition: Each one of the files or elements that can be invoked from a web page including the HTML file itself. For counting objects only different ones are taken into account, so repeated numbers are only used when the different categories are considered separately.

Examples: The HTML file, links appearing on it, images, audio or other multimedia resources invoked, the gateways to databases, dynamic pages or other interactive systems.

Multimedia file

Definition: Although multimedia is usually referred to the integration and display of different types of text, audio or graphic formats, we reserve the concept to the narrow group of video and audio files, including large animated graphics, flash or VRML items.

Example: Video formats include: avi, mpg or mpeg and QuickTime's mov or qt; while sound and music appear as: wav, mov, au, midi or mp3. Flash animations use swf format.

Rich File

Definition: A file in a non-textual format like Adobe Acrobat (pdf), Postscript (ps) or MS Word (doc, rtf), Excel (xls) or Powerpoint (ppt), where the special characteristics of these formats increases its value.

Example: An electronic archive like the Los Alamos' repository of physics papers offer the documents in both pdf and ps format.

Link /bad link /deep link

Definition: A hypertext link is a connection from one word, picture, or information object (including multimedia ones) to another in the same page or to another page in the same or different site. On the web, selection of the link activates the navigation towards the linked object. A bad link appears if the connection cannot be established because the target object is no longer available, because it disappeared, changed the location or there is a problem on the network or in the server. A deep link is a link to a page other than home page.

Outlink (External/Internal)

Definition: All the links that appear in a page or web site. The outlinks are represented by the destination URL of the links. Outlinks can internal (those linking objects in the same site or domain) or external (offlinks).

Example: All the links to the resources in the Yahoo index and those devoted to navigate among the different categories are outlinks.

Inlink

Definition: Other pages that link to a page (backlink) or a site from the rest of the webspace, usually excluding those pages from the same site or institutional domains (external inlinks). Inlinks are represented by the source URL of the links.

Example: Using Altavista link operator you can discover all the pages that have a link to a given site or domain. The same operator in Google only refers to a given page (e.g., home page if a host name is provided).

Anchor

Definition: The highlighted object that executes a link is the anchor.

Example: Semantic value of textual anchors can be very important for data mining, although some anchors are simply “Click here”.

Gateway

Definition: A web gateway is point of entrance to another service that requires a data input by the requester, a search term, password, or choosing an option. This activity generates an answer from the database behind the gateway, providing the corresponding information.

Example: The access to the catalogue in a library, a service called OPAC, is provided through a gateway.

9

Bibliography

Abbott, C. (2001). Some young male website owners: the technological aesthete, the community builder and the professional activist. *Education, Communication and Information, 1*, 197-212.

Adamic, L. A. (1999). The Small World Web. In A.-M. V. E. S. Abiteboul (Ed.), *Research and Advanced Technology for Digital Libraries: Third European Conference* (pp. 443–452). Paris, France.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks, 25*(3), 211-230.

Adamic, L. A., Buyukkokten, O., & Adar, E. (2003). A social network caught in the Web. *firstmonday, 8*(6). Retrieved from http://www.firstmonday.org/issues/issue8_6/adamic/

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the World Wide Web. *Science, 287*(5461), 2115.

Adams, D. (2002). The counting house. *Nature, 415*, 726-729.

Adler, R. J., Feldman, R. E., & Taqqu, M. S. (Eds.). (1998). *A practical guide to heavy tails: statistical techniques and applications*. Berlin: Birkäuser.

Aiello, W., Chung, F., & Lu, L. (2002). A Random Graph Model for Massive Graphs. In J. Abello & P. M. Pardalos & M. G. C. Resende (Eds.), *Handbook on Massive Data Sets*. Norwell, Massachusetts: Kluwer Academic Publishers.

Albert, R., & Barabasi, A.-L. (2000). Topology of evolving networks: local events and universality. *Physical Review Letters, 85*(24), 5234-5237.

Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*.

Albert, R., Jeong, H., & Barabasi, A. L. (1999). Diameter of the world wide web. *Nature, 401*, 130-131.

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

Alvarez-Hamelin, J. I., & Schabanel, N. (2004). An internet graph model based on trade-off optimization. *European Physics Journal B, 38*(2), 231-237.

Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Salinger, M. A., & Stanley, H. E. (1998). Power law scaling for a system of interacting units with complex internal structure. *Physical Review Letters, 80*(7).

Amaral, L. A. N., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2001). A model for the growth dynamics of economic organizations. *Physica A, 299*, 127-136.

Amaral, L. A. N., & Ottino, J. M. (2004). Complex networks: Augmenting the framework for the study of complex systems. *European Physical Journal B, 38*, 147-162.

Amaral, L. A. N., & Ottino, J. M. (2004). Complex systems and networks: challenges and opportunities for chemical and biological engineers. *Chemical Engineering Science, 59*, 1653 – 1666.

Amaral, L. A. N., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *PNAS, 97*, 11149-11152.

An, L., & Qiu, J. P. (2003). Research on the relationships between Chinese journal impact factors and web impact factors and external web link counts. *Journal of the China Society for Scientific and Technical Information, 22*(4), 398-402.

An, L., & Qiu, J. P. (2004). Research on the relationships between Chinese journal impact factors and external web link counts and web impact factors. *Journal of Academic Librarianship, 30*(3), 199-204.

An, Y., Janssen, J., & Milios, E. E. (2002). Characterizing the citation graph as a self-organizing networked information space. In H. Unger & T. Bohme & A. R. Mikler (Eds.), *Lecture notes in computer science* (Vol. 2346, pp. 97-107). London, UK: Springer-Verlag.

Anonymous. (1970). Can Science Afford Scientists. *Nature, 226*, 10.

Antelman, K. (2004). Do Open-Access Articles Have a Greater Research Impact? *College & Research Libraries, 65*(5), 372-382.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2-43.

Ardö, A., & Lundberg, S. (1998). A regional distributed WWW search and indexing service: the DESIRE way. In P. H. Enslow & A. Ellis (Eds.), *Proceedings of the seventh international conference on World Wide Web 7* (Vol. 30, pp. 173-183). Brisbane, Australia: Elsevier Science Publishers.

Bailey, P., Craswell, N., & Hawking, D. (2003). Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management, 39*, 853–871.

Baldi, P., Frasconi, P., & Smyth, P. (2003). *Modelling the Internet and the Web*. Wiley: Chichester, UK.

Barabasi, A. L. (2002). *Linked: The new science of networks*. Cambridge, Massachusetts: Perseus Publishing.

Barabasi, A. L. (2003). Emergence of scaling in complex networks. In S. Bornholdt & H. G. Schuster (Eds.), *Handbook of Graphs and Networks: From the Genome to the Internet*: Wiley.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(October 15), 509-512.

Barabasi, A.-L., Albert, R., Jeong, H., & Bianconi, G. (2000). Power-law distribution of the World Wide Web. *Science, 287*(24 March), 2115.

Barabasi, A.-L., & Bonabeau, A. (2003). Scale Free Networks. *Scientific American, 295*, 50-59.

Barabasi, A.-L., Jeong, H., Ravasz, R., Neda, Z., Vicsek, T., & Schubert, A. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*(3-4), 590-614.

Barabasi, A.-L., Menezes, M. A. d., Balensiefer, S., & Brockman, J. (2004). Hot spots and universality in network dynamics. *European Physics Journal B, 38*(2), 169-175.

Barabasi, A.-L., Ravasza, E., & Vicsek, T. (2001). Deterministic scale-free networks. *Physica A, 299*, 559-564.

Barabasi, A.-L., Reka, A., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A, 272*, 173-187.

Barabasi, A.-L., Reka, A., & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A, 281*, 69-77.

Barabasi, A. U. (2001). The Physics of the Web. *Physics World, 14*(7), 33-38.

Bar-Ilan, J. (1997). The 'mad cow disease', Usenet newsgroups and bibliometric laws. *Scientometrics, 39*(1), 29-55.

Bar-Ilan, J. (1999). Search engine results over time - a case study on search engine stability. *Cybermetrics, 2/3*. Retrieved from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

Bar-Ilan, J. (2000). Evaluating the stability of the search tools Hotbot and Snap: a case study. *Online information review, 24*(6), 439-449.

Bar-Ilan, J. (2000). The Web as an information source on Informetrics? A content analysis. *Journal of American Society for Information Science, 51*(5), 432-443.

Bar-Ilan, J. (2001). Data collection methods on the Web for infometric purposes: A review and analysis. *Scientometrics, 50*(1), 7-32.

Bar-Ilan, J. (2002). How much information do search engines disclose on the links to a web page? A longitudinal case study of the 'cybermetrics' home page. *Journal o f Information Science, 28*(6), 455-466.

Bar-Ilan, J. (2004). Blogarians - A new breed of librarians. *Proceedings of the American Society for Information Science & Technology, 41,* 119-128.

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59*(3), 391-403.

Bar-Ilan, J. (2004). Search engine ability to cope with the changing web. In M. Levene & A. Poulovasilis (Eds.), *Web Dynamics*. Berlin: Springer-Verlag.

Bar-Ilan, J. (2004). Self-linking and self-linked rates of academic institutions on the Web. *Scientometrics, 59*(1), 29-41.

Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology, 38*, 231-288.

Bar-Ilan, J. (2005). Information hub blogs. *Journal of Information Science, 31*(4), 297-307.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management, 41*(3), 973-986.

Bar-Ilan, J., & Peritz, B. C. (1999). The life span of a specific topic on the Web. The case of "informetrics": A quantitative analysis. *Scientometrics, 46*(3), 371-382.

Bar-Ilan, J., & Peritz, B. C. (2002). Informetric theories and methods for exploring the Internet: an analytical survey of recent research literature. *Library trends, 50*(3), 371-392.

Bar-Ilan, J., & Peritz, B. C. (2004). 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), 980 - 990.

Barjak, F. (2004). From the "analogue divide" to the "hybrid divide": no equalisation of information access in science through the Internet. *Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods*. Retrieved from http://cybermetrics.wlv.ac.uk/AoIRASIST/Barjak.html

Barjak, F. (2004). The integration of the Internet into informal communication in science. *University of Applied Sciences Solothurn Discussion Paper No. 2004-W02*,

Retrieved May 23, 2005, from: http://www.fhso.ch/pdf/publikationen/dp2004-2002.pdf.

Barjak, F. (2006, in press). From the "analogue divide" to the "hybrid divide": the internet does not ensure equality of access to information in science. In C. Hine (Ed.), *New infrastructures for knowledge production: Understanding e-science*. Hershey, PA: Idea Group.

Barjak, F. (2006, to appear). The role of the internet in informal scholarly communication. *Journal of the American Society for Information Science and Technology*.

Barjak, F., Li, X., & Thelwall, M. (2005). Which factors explain the web impact of scientists' personal home pages? *Presentation at Internet Research 6.0: Internet Generations*, http://conferences.aoir.org/viewabstract.php?id=117&cf=113.

Barnsley, M. (1988). *Fractals Everywhere*. San Diego: Academic Press Inc.

Barthelemy, M. (2004). Betweenness centrality in large complex networks. *The European Physical Journal B, Vol. 38*(2), 163-168.

Barthélemy, M., Gondran, B., & Guichard, E. (2003). Spatial structure of the internet traffic. *Physica A, 319*(2), 633-642.

Beaulieu, A. (2004). From brainbank to database: The informational turn in the study of the brain. *Studies in History and Philosophy of Biological and Biomedical Sciences., 35*, 367-390.

Beaulieu, A. (2005). Sociable hyperlinks: An ethnographic approach to connectivity. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet* (pp. 182-192). London: Berg.

Bergman, M. K. (2000). *White Paper- The Deep Web: Surfacing Hidden Value*. Retrieved 28 Feb. 2003, from:

Bergmark, D., & Lagoze, C. (2001). *Reference linking the Web's scholarly papers*: Computer Science Department, Cornell University.

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. In M. Agosti & C. Thanos (Eds.), *Proceedings of the 6th European Conference on Research and Advanced Technologies for Digital Libraries* (pp. 91-106). Berlin: Springer.

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, F. H., & Secret, A. (1994). The World-Wide Web. *Communications of the ACM, 37*(8), 76,~78-82.

Berners-Lee, T., & Hendler, J. (2001). Scientific publishing on the 'semantic web'. *Nature, 410*, 1023-1024.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems, 30*(1–7), 379-388.

Bharat, K., Broder, A., Dean, J., & Henzinger, M. R. (2000). A comparison of techniques to find mirrored hosts on the WWW. *Journal of the American Society for Information Science and Technology, 51*(12), 1114-1122.

Bharat, K., Broder, A., Henzinger, M., Kumar, P., & Suresh Venkatasubramanian, a. (1998). The connectivity server: fast access to linkage information on the Web. *Proceedings of the 7th International World Wide Web Conference (WWW-7)* (Vol. 30, pp. 469-477). Brisbane, Australia.

Bharat, K., Chang, B.-W., Henzinger, M. R., & Ruhl, M. (2001). Who links to whom: mining linkage between Web sites. In N. Cercone & T. Y. Lin & X. Wu (Eds.), *Proceedings of the 2001 IEEE International Conference on Data Mining* (pp. 51-58). Washington: IEEE Computer Society.

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillation in a hypertext environment. *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 104-111). New York: USA.

Bianconi, G. (2004). Number of cycles in off-equilibrium scale-free networks and in the Internet at the Autonomous System Level. *European Physics Journal B, 38*(2), 223-230.

Björneborn, L. (2001). *Necessary data filtering and editing in webometric link structure analysis*: Royal School of Library and Information Science.

Björneborn, L. (2001). *Shared outlinks in small-world co-linkage analysis: a Webometric pilot study of bibliographic couplings on researchers' bookmark lists on the Web*: Royal School of Library and Information Science, Denmark.

Björneborn, L. (2001). *Small-world linkage and co-linkage.* Paper presented at the 12th ACM conference on hypertext and hypermedia, Aarhus, Denmark.

Björneborn, L. (2004). *Small-world link structures across an academic web space - a library and information science approach.* Royal School of Library and Information Science, Copenhagen, Denmark.

Björneborn, L., & Ingwersen, P. (2001). Perspective of webometrics. *Scientometrics, 50*(1), 65-82.

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

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers, 34*(2), 286-290.

Blood, R. (2004). How blogging software reshapes the online community. *Communications of the ACM, 47*(12), 53-55.

Boguñá, M., Pastor-Satorras, R., & Vespignani, A. (2003). Cut-offs and finite size effects in scale-free networks. *European Physics Journal B, 38*(2), 205–209.

Bollobás, B., & Riordan, O. M. (2003). Handbook of Graphs and Networks: From the genome to the Internet. In S. Bornholdt & H. G. Schuster (Eds.), *Mathematical results on scale-free random graphs*. Berlin: Wiley-VCH GmbH & Co.

Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks, 19*(3), 243-269.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 36). Medford, New York: Information Today Inc.

Borner, K., Chen, C. M., & Boyack, K. W. (2003). Visualizing knowledge domains. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 179-255). Medford, New York: Information Today Inc.

Börner, K., Maru, J. T., & Gladstone, R. L. (2004). The simultaneous evolution of author and paper networks. *PNAS, 101 (supplement 1)*, 5266-5273.

Bornholdt, S., & Ebel, H. (2001). World Wide Web scaling exponent from Simon's 1955 model. *Physical Review E, 64*(3), article no. 035104 Part035102.

Boudourides, M. A., & Antypas, G. (2002). *A simulation of the structure of the world wide web*. Retrieved, from: http://www.socresonline.org.uk/7/1/boudourides.html

Boudourides, M. A., Sigrist, B., & Alevizos, P. D. (1999). *Webometrics and the self-organization of the European Information Society*. Retrieved 29 May 2002, from:

Bowman, C. M., Danzig, P. B., Hardy, D. R., Manber, U., & Schwatz, M. F. (1995). The Harvest information discovery and access system. *Computer Networks and ISDN Systems, 28*(1–2), 119-125.

Braunmuller, K. (2002). Semicommunication and accommodation: Observations from the linguistic situation in Scandinavia. *International Journal Of Applied Linguistics, 12*(1), 1-23.

Bray, T. (1996). Measuring the Web. *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems* (Vol. 28, pp. 993-1005). Amsterdam, Netherlands: Elsevier Science Publishers.

Brewington, B. E., & Cybenko, G. (2000). How dynamic is the Web?, *the 9th International World Wide Web Conference* (pp. 257-276). Amsterdam: Computer Networks and ISDN Systems.

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

Britt, P. J. (2005). RSS investors - There's gold in supporting them thar standards *EContent, 28*(9), 8.

Brock, W. A. (1999). Scaling in Economies: a Reader's Guide. *Industrial and Corporate Change, 8*(1), 409-446.

Broder, A., & GET THIS. (2004). The many wonders of the web graph. In A. Alejandro López-Ortiz & A. A. l. Hamel (Eds.), *First Workshop on Combinatorial and Algorithmic Aspects of Networking* (Vol. 3405, pp. 154-155). Banff, Alberta, Canada: Springer.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the web. *Journal of Computer Networks, 33*(1-6), 309-320.

Bruun, S. D., & Dodge, M. (2001). Mapping the "worlds" of the world wide web: (Re)Structuring global commerce through hyperlinks. *American Behavioral Scientist, 44*(10), 1717-1739.

Buchanan, M. (2002). The physics of the trading floor. *Nature, 415*(3), 10-12.

Buchanan, M. (2003). *Small world*. London: Phoenix.

Bucher, H. J. (2002). The internet in crisis: Communication in the case of September 11th. *First Monday, 7*(4). Retrieved from http://www.firstmonday.org/issues/issue7_4/bucher/

Burd, A., Chiu, T., & McNaught, C. (2004). Screening Internet websites for educational potential in undergraduate medical education. *Medical Informatics and The Internet in Medicine, 29*(3-4), 185-197.

Burda, Z., Jurkiewicz, J., & Nowak, M. A. (2003). Is Econophysics a Solid Science? *Acta Physica Polonica B, 34*(1), 87.

Burner, M. (1997). Crawling towards eternity: building an archive of the World Wide Web. *New architect: Internet strategies for technology leaders, 2*(5).

Burnett, R., & Marshall, P. (2002). *Web theory: An introduction*. London: Routledge.

Burrell, Q. L. (2004). Fitting Lotka's Law: some cautionary observations on a recent paper by Newby et al. (2003). *Journal of the American Society for Information Science and Technology, 55*(13), 1209-1211.

Butler, L. (2003). Explaining Australia's increased share of ISI publications - the effects of a funding formula based on publication counts. *Research Policy, 32*(1), 143-155.

Caldarelli, G., Capocci, A., De Los Rios, P., & Munoz, M. A. (2002). Scale-Free Networks from VaryingVertex Intrinsic Fitness. *Physical Review Letters, 89*(25), 258702-258701-258704.

Caldarelli, G., Pastor-Satorras, R., & Vespignani, A. (2004). Structure of cycles and local ordering in complex networks. *European Physics Journal B, 38*(2), 183-186.

Caldas, A. (2003). Are newsgroups extending 'invisible colleges' into the digital infrastructure of science? *Economics of Innovation and New Technology, 12*(1), 43-60.

Calder, I. (1999). *The blue revolution: land use & integrated water resources management*. London: Earthscan.

Calishain, T., & Dornfest, R. (2003). *Google Hacks: 100 Industrial-Strength Tips & Tools*. Cambridge: O' Reilly & Associates.

Calluzzo, V. J., & Cante, C. J. (2004). Ethics in information technology and software use. *Journal of Business Ethics, 51*(3), 301-312.

Calvert, P. J. (2001). Scholarly misconduct and misinformation on the World Wide Web. *Electronic Library, 19*(4), 232-240.

Cancho, R. F., & Sole, R. V. (2001). The small world of human language. *Proc. R. Soc. Lond. B, 268*, 2261-2265.

Cancho, R. F., & Sole, R. V. (2002). Zipf's law and random texts. *Advances in Complex Systems, 5*(1), 1-6.

Cancho, R. F., & Sole, R. V. (2003). Least effort and the origins of scaling in human language. *PNAS, 100*(3).

Canning, D., Amaral, L. A. N., Lee, Y., Meyer, M., & Stanley, H. E. (1998). Scaling the volatility of GDP growth rates. *Economics Letters, 60*, 335–341.

Carlson, J. M., & Doyle, J. C. (1999). Highly Optimized Tolerance: A Mechanism for Power Laws in Designed Systems. *Physics Review E, 601*, 1412-1428.

Carlson, J. M., & Doyle, J. C. (2002). Complexity and Robustness. *PNAS, 99*(February 19), 2538-2545.

Case, D. O. (2002). *Looking for information: A survey of research on information seeking, needs, and behavior*. San Diego, CA: Academic Press.

Casserly, M. F., & Bird, J. E. (2003). Web citation availability: Analysis and implications for scholarship. *College and Research Libraries, 64*, 300-317.

Catanzaro, M., Boguñá, M., & Pastor-Satorras, R. (2005). Generation of uncorrelated random scale-free networks. *Physical Review E, 71*, 027103-027104.

Chakrabarti, S. (1999). Recent results in automatic Web resource discovery. *ACM Computing surveys, 31*(4es).

Chakrabarti, S. (2003). *Mining the Web: Analysis of hypertext and semi structured data*. New York: Morgan Kaufmann.

Chakrabarti, S., B., D., Raghavan, P., Rajagonpalan, S., Gibson, D., & Kleinberg, J. M. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems, 30*, 65-74.

Chakrabarti, S., Dom, B., Kumar, R. S., Raghavan, P., Rajagonpalan, S., Tomkins, A., Kleinberg, J. M., & Gibson, D. (1999). Hypersearching the Web. *Scientific American, 280*(6), 54-60.

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., & Kleinberg, J. (1999). Mining the Web's link structure. *Computer, 32*(8), 60-67.

Chakrabarti, S., VanDen Berg, M., & Dom, B. (1999, May 11-14). *Focused crawling: A new approach to topic-specific Web resource discovery.* Paper presented at the the 8th International World Wide Web Conference, Toronto, Canada.

Chau, M., & Chen, H. (2003). Comparison of three vertical search spiders. *IEEE Computer, 36*(5), 56-62.

Chau, M., Chen, H., Qin, J., Zhou, Y., Qin, Y., Sung, W.-K., & McDonald, D. (2002). Comparison of two approaches to building a vertical search tool: a case study in the nanotechnology domain. In W. Hersh & G. Marchionini (Eds.), *JCDL '02* (pp. 135-144). Portland, Oregan: ACM Press.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35*(3), 401-420.

Chen, C. (2004). Searching for intellectual turning points: progressive knowledge domain visualization. *PNAS, 101 (supplement 1)*, 5303-5310.

Chen, C., Newman, J., Newman, R., & Rada, R. (1998). How did university departments interweave the Web: A study of connectivity and underlying factors. *Interacting With Computers, 10*(4), 353-373.

Chen, Z., Tao, L., Wang, J., Wenyin, L., & Ma, W.-Y. (2002). *A Unified Framework for Web Link Analysis.* Paper presented at the The 3rd International Conference on Web Information Systems Engineering, Singapore.

Chien, S., Dwork, C., Kumar, R., Simon, D. R., & Sivakumar, D. (2003). Link evolution: analysis and algorithms. *Internet mathematics, 1*(3), 277-304.

Cho, J., & Garcia-Molina, H. (2000). The evolution of the Web and implications for an incremental crawler. In A. E. Abbadi & M. L. Brodie & S. Chakravarthy & U. Dayal & N. Kamel & G. Schlageter & K.-Y. Whang (Eds.), *Proceedings of 26th international conference on very large data bases, September 10-14, 2000* (pp. 200-209): Morgan Kaufmann.

Cho, J., & Garcia-Molina, H. (2002). *Parallel crawlers.* Paper presented at the WWW2002, Honolulu, Hawaii.

Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. *Computer Networks and ISDN Systems, 30*(1–7), 161-172.

Christensen, K., Danon, L., Scanlon, T., & Bak, P. (2002). Unified scaling law for earthquakes. *PNAS, 99*, 2509-2513.

Christensen, N. H. (2004). *Towards format repositories for web archives.* Paper presented at the IWAW-2004.

Chu, H. (2005). Taxonomy of inlinked web entities: What does it imply for Webometric research? *Library & Information Science Research, 27*(1), 8-27.

Chu, H., He, S., & Thelwall, M. (2002). Library and information science schools in Canada and USA: A Webometric perspective. *Journal of Education for Library and Information Science, 43*(2), 110-125.

Chun, T. Y. (1999). World Wide Web robots: an overview. *Online & CD ROM review, 23*(3), 135-142.

Cilibrasi, R., & Vitany, P. (2004). *Automatic Meaning Discovery Using Google*: University of Amsterdam, National ICT of Australia.

Cimiano, P., & Staab, S. (2004). Learning by Googling. *SIGKDD Explorations, 6*(2), 24-33.

Clausen, L. (2004). *Concerning Etags and datestamps.* Paper presented at the IWAW-2004.

Coch, J., & Masanès, J. (2004). *Language engineering techniques for web archiving.* Paper presented at the IWAW-2004.

Collins, J. J., & Chow, C. C. (1998). It's a small world. *Nature, 393*, 409-410.

Consortium, I. S. (2003). *Internet Domain Survey*. Retrieved 3 Mar. 2003, from: http://www.isc.org/ds/WWW-200301/index.html

Cooke, A. (2001). *A guide to finding quality information on the Internet: Selection and evaluation strategies*. London: Library Association Publishing.

Cooley, R., Srivastava, J., & Mobasher, B. (1997). Web Mining: Information and pattern discovery on the world wide Web. *9th IEEE International Conference on Tools With Artificial Intelligence*, 558-567.

Cooper, C., & Frieze, A. (2001). A general model of undirected web graphs. In F. M. a. d. Heide (Ed.), *Algorithms ESA 2001: 9th Annual European Symposium* (Vol. 2161, pp. 500-511). Aarhus, Denmark: Springer-Verlag.

Cooper, C., & Frieze, A. (2003). Crawling on simple models of Web graphs. *Internet mathematics, 1*(1), 57-90.

Cooper, C., & Frieze, A. (2003). A general model of web graphs. *Random Structures and Algorithms, 22*(3), 311-335.

Cooper, C., Frieze, A., & Vera, J. (2004). Random deletion in a scale free random graph. *Internet mathematics, 1*(4), 4563-4483.

Costa, L. d. F., Rodrigues, F. A., Travieso, G., & Boas, P. R. V. (2005). *Characterization of Complex Networks: A Survey of measurements*. Retrieved, from: http://arxiv.org/PS_cache/cond-mat/pdf/0505/0505185.pdf

Cothey, V. (2004). Web-crawling reliability. *Journal of the American Society for Information Science and Technology, 55*(14), 1228 - 1238.

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information Science, 27*(1), 1-7.

Cronin, B. (2001). Semiotics and evaluative bibliometrics. *Journal of Documentation, 56*(4), 440-453.

Cronin, B. (2005). Vox populi: Civility in the blogosphere. *International Journal of Information Management, 25*(6), 483-586.

Cronin, B., & Shaw, D. (2002). Banking (on) different forms of symbolic capital. *Journal of the American Society for the Information Science, 53*(13), 1267-1270.

Cronin, B., & Shaw, D. (2002). Identity-creators and image makers: Using citation analysis and thick descriptions to put authors in their place. *Scientometrics, 54*, 31-49.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science, 49*(14), 1319-1328.

Crovella, M. E., Taqqu, M. S., & Bestavros, A. (1998). Heavy tailed probability distributions on the world wide web. In R. E. Feldman & R. J. Adler & M. S. Taqqu (Eds.), *A Practical Guide to Heavy Tails.*: Birkhauser.

Crowston, K., Kwasnik, B. H., Nilan, M., & Roussinov, D. (2000, November 13-16). *Identifying document genre to improve Web search effectiveness.* Paper presented at the 63rd Annual Meeting of the American Society for Information Science, Chicago, Illinois.

Csete, M. E., & Doyle, J. C. (2002). Reverse Engineering of Biological Complexity. *Science, 295*(March 1), 1664-1669.

Cui, L. (1999). Rating health web sites using the principles of citation analysis: A bibliometric approach. *Journal of Medical Internet Research, 1*(1).

Cunningham, S. J. (1997). Teaching students to critically evaluate the quality of Internet research resources. *ACM SIGCSE Bulletin, 29*(2), 31-38.

Curro, V., Buonuomo, P. S., Onesimo, R., Rose, P. D., Vituzzi, A., Di Tanna, G. L., & D'Atri, A. (2004). A quality evaluation methodology of health web-pages for non-professionals. *Medical Informatics and The Internet in Medicine, 29*(2), 95-107.

Dahn, M. (2000). Counting angels on a pinhead: Critically interpreting Web size estimates. *Online, 24*(1), 35-40.

Dahn, M. (2000). Spotlight on the invisible Web. *Online, 24*(4), 57-62.

Davenport, E., & Cronin, B. (2000). The citation network as a prototype for representing trust In virtual environments. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield.* (pp. 517-534). Metford, NJ: Information Today Inc. ASIS Monograph Series.

Davenport, E., & Cronin, B. (2001). Who dunnit? Metatags and hyperauthorship. *Journal of American Society for Information Science, 52*(9), 770-773.

Davies, M. (2001). Creating and using multi-million word corpora from web-based newspapers. In R. C. Simpson & J. M. Swales (Eds.), *Corpus Linguistics in North America* (pp. 58-75). Ann Arbor: University of Michigan.

Dean, J., & Henzinger, M. R. (1999). *Finding Related Pages in the World Wide Web.* Paper presented at the WWW8 Conference.

Deo, N., & Gupta, P. (2001). *World Wide Web: A Graph-Theoretic Perspective*. Orlando: University of Central Florida.

Dhyani, D., Ng, W. K., & Bhowmick, S. S. (2002). A survey of web metrics. *ACM Computing surveys, 34*(4), 469-503.

Di Guilmi, C., Gaffeo, E., & Gallegati, M. (2003). Power Law Scaling in the World Income Distribution. *Economics Bulletin, 15*(6), 1-7.

Diligenti, M., Gori, M., & Maggini, M. (2002, May 7-11). *Web page scoring systems for horizontal and vertical search.* Paper presented at the WWW 2002, Honolulu, Hawaii.

Dill, S., Kumar, R., McCurley, K. S., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (2002). Self-Similarity In the Web. *ACM Transactions on Internet Technology, 2*(3).

Ding, Y., Chowdbury, G., & Foo, S. (1999). Mapping the intellectual structure of information retrieval studies: An author co-citaion analysis, 1987-1997. *Journal of Information Science, 25*(1), 67-78.

Doerup, J., Hansen, M. S., Ribe, L. R., & Larsen, K. W. (2002). A comparison of technologies for database-driven websites for medical education. *Medical Informatics and The Internet in Medicine, 27*(4), 281 - 289.

Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2004). Large scale properties of the Webgraph. *Eur. Phys. J. B, 38*(2), 239-243.

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2002). Pseudofractal Scale-free Web. *Physics Review E, 65*(066122).

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2004). Potts model on complex networks. *European Physical Journal B, 38*(2), 177-182.

Dorogovtsev, S. N., & Mendes, J. F. F. (2000). Evolution of reference networks with aging. *Physics Review E, 62*(2), 1842-1845.

Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Language as an Evolving Word Web. *Proceedings: Biological Sciences, 268*(1485), 2603-2606.

Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks. *Advanced Physics, 51*, 1079-1187.

Dorogovtsev, S. N., & Mendes, J. F. F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford: Oxford University Press.

Dorogovtsev, S. N., & Mendes, J. F. F. (2004). The shortest path to complex networks. In N. J. e. a. (eds) (Ed.), *Complex Systems and Interdisciplinary Science* (Vol. 4, pp. 1-25). Singapore: World Scientific.

Dorogovtsev, S. N., Mendes, J. F. F., & Samukhin, A. N. (2000). Structure of growing networks with preferential linking. *Physical Review Letters, 85*(21), 4633-4636.

Dorogovtsev, S. N., Mendes, J. F. F., & Samukhin, A. N. (2000). Structure of Growing Networks: Exact Solution of the Barab´asi–Albert's Model. *Physics Review Letters*(85).

Dorogovtsev, S. N., Mendes, J. F. F., & Samukhin, A. N. (2001). Generic scale of the "scale-free" growing networks. *Physics Review E, 63*, 1-4.

Downey, A. B. (2001). *The structural cause of file size distributions.* Paper presented at the MASCOTS 2001.

Eckmann, J. P., & Moses, E. (2002). Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *PNAS, 99*(9), 5825-5829.

Edwards, J., McCurley, K., & Tomlin, J. (2002). *An adaptive model for optimizing performance of an incremental Web crawler*, New York.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. *Journal of Information Science, 26*(5), 329-335.

Egghe, L. (2001). *Discussions on informetrics of the Internet.* Paper presented at the the 2nd International Symposium on Quantitative Research Evaluation and the 6th National Annual Conference of Scientometrics and Informetrics, Shanghai.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Oxford: Elsevier Academic Press.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: quantitative methods in library documentation an information science*. Oxford: Elsevier.

Egghe, L., & Rousseau, R. (2002). Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scientometrics, 55*(3), 349-361.

Eichmann, D. (1994). *Ethical web agents.* Paper presented at the Second international world wide web conference, Chicago, IL.

Eichmann, D. (1994, May 1994). *The RBSE spider: balancing effective search against Web load.* Paper presented at the The Proceedings of the First International World-Wide Web Conference, Geneva Switzerland.

Einstein, A. (1956). *Investigations on the Theory of the Brownian Motion*. New York: Dover.

Etzkowitz, H., Kemelgor, C., & Uzzi, B. (2000). *Athena unbound: The advancement of women in science and technology*. Cambridge: Cambridge University Press.

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations. *Research Policy, 29*(2), 109-123.

Evans, T. S. (2004). Complex Networks. *Comtemporary Physics, 45*(6), 455-474.

Ewell, R. H. (1955). Role of Research in Economic Growth. *Chemical and Engineering News, 33*(29), 2981.

Faba-Perez, C., Guerrero-Bote, V. P., & De Moya-Anegon, F. (2003). Data mining in a closed Web environment. *Scientometrics, 58*(3), 623-640.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). *On Power Law relationships of the Internet Topology.* Paper presented at the Sigcomm 1999.

Fayed, M., Krapivsky, P., Byers, J. W., Crovella, M., Finkel, D., & Redner, S. (2003). On the emergence of highly variable distributions in the autonomous system topology. *Computer communications review, 33*(2), 41-49.

Feitelson, D. G., & Yovel, U. (2004). Predictive ranking of computer scientists using CiteSeer data. *Journal of Documentation, 60*(1), 44-61.

Fenner, T., Levene, M., & Loizou, G. (2004). A Stochastic Evolutionary Model Exhibiting Power-Law Behaviour with an Exponential Cutoff. *Physica A, 335*, 641-656.

Fenner, T., Levene, M., & Loizou, G. (2005). A Model for Collaboration Networks Giving Rise to a Power Law Distribution with an Exponential Cutoff. *arXiv:physics/0503184*. Retrieved from http://arxiv.org/PS_cache/physics/pdf/0503/0503184.pdf

Fenner, T., Levene, M., & Loizou, G. (2005, to appear). A Stochastic Model for the Evolution of the Web Allowing Link Deletion. *ACM Transactions on Internet Technology, 2*.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003). A large-scale study of the evolution of Web pages. *Proceedings of the 12th International World Wide Web Conference*, http://www2003.org/cdrom/papers/refereed/p2097/P2097%2020sources/p2097-fetterly.html.

Fielding, R. T. (1994, May 25-27). *Maintaining distributed hypertext infostructures: welcome to momspiders web.* Paper presented at the The First International World-Wide Web Conference, Geneva, Switzerland.

Finholt, T. (2002). Collaboratories. *Annual Review of Information Science and Technology, 36*, 73-107.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2000). Efficient identification of Web communities. *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining* (pp. 150-160). New York: ACM Press.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of Web communities. *IEEE Computer, 35*, 66-71.

Fletcher, W. (2002). Making the Web more useful as a source for linguistic corpora. In U. Connor & T. Upton (Eds.), *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*: Amsterdam: Rodopi.

Florescu, D., Levy, A., & Mendelzon, A. (1998). Database techniques for the World Wide Web: a survey. *SIGMOD record, 27*(3), 59-74.

Foot, K. A., & Schneider, S. M. (2002). Online Action in Campaign 2000: An Exploratory Analysis of the U.S. Political Web Sphere. *Journal of Broadcasting and Electronic Media, 46*(2), 222-244.

Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere. *Journal of Computer Mediated Communication, 8*(4). Retrieved from http://www.ascusc.org/jcmc/vol8/issue4/foot.html

Fosmire, M., & Yu, S. (2000). Free scholarly electronic journals: How good are they? *Issues in Science and Technology Librarianship, Summer 2000*.

Fox, E., & Urs, S. (2002). Digital libraries. *Annual Review of Information Science and Technology, 36*, 503-589.

Freeman, C. (1987). *Technology and Economic Performance: Lessons from Japan*. London and New York: Pinter Publishers.

Fry, J. (2004). The cultural shaping of ICTs within acacemic fields: Corpus-based linguistics as a case study. *Literary and Linguistic Computing, 19*(3), 303-319.

Fry, J., & Talja, S. (2004). The cultural shaping of scholarly communication: Explaining e-journal use within and across academic fields., *ASIST 2004: Proceedings of the 67th ASIST Annual Meeting* (pp. 20-30): Medford, NJ.: Information Today.

Fujiki, T., Nanno, T., & Okumura, M. (2005). *Differences between Blogs and Web Diaries, Toshiaki Fujiki, Tokyo Institute of Technology*. Paper presented at the

WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan.

Fukuhara, T. (2005). *Analyzing concerns of people using Weblog articles and real world temporal data, Tomohiro Fukuhara.* Paper presented at the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan.

Furner, J., Ellis, D., & Willett, P. (1999). Inter-linker consistency in the manual construction of hypertext documents. *ACM Computing Surveys (CSUR), 31*(4es).

Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics, 114*(3), 739 - 767.

Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). A theory of power-law distributions in finacial market Fluctuations. *Nature, 423*(May 15), 267-270.

Gao, Y. J., & Vaughan, L. (2005). Web hyperlink profiles of news sites - A comparison of newspapers of USA, Canada, and China. *ASLIB Proceedings, 57*(5), 398-411.

García-Manso, J. M., Martín-González, J. M., Dávila, N., & Arriaza, E. (2005). Middle and Long Distance Athletics Races Viewed from the Perspective of Complexity. *Journal of Theoretical Biology, 233*, 191-198.

Garfield, E. (1998). From citation indexes to Informetrics: Is the tail now wagging the dog? *Libri, 48*(2), 67-80.

Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal, 161*(8), 979-980.

Garg, K. C. (2003). An overview of cross-national, national, and institutional assessment as reflected in the international journal Scientometrics. *Scientometrics, 56*(2), 169-199.

Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying Social Network Analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165-184). London: Routledge.

Ghemawat, P. (2001). Distance still matters - The hard reality of global expansion. *Harvard Business Review, 79*, 137-147.

Ghim, C.-M., Oh, E., Goh, K.-I., Kahng, B., & Kim, D. (2004). Packet transport along the shortest pathways in scale-free networks. *European Physics Journal B, 38*(2), 193-199.

Gibson, D., Kleinberg, J. M., & Raghavan, P. (1998). *Inferring web communities from link topology.* Paper presented at the Hypertext 98: Ninth ACM Conference on Hypertext and Hypermedia, New York, USA.

Gill, K. E. (2004). *How can we measure the influence of the blogosphere?* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.

Gill, K. E. (2005). *Blogging, RSS and the information landscape: A look at online news.* Paper presented at the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan.

Ginsparg, P., Houle, P., Joachims, T., & Sul, J.-H. (2004). Mapping subsets of scholarly information. *PNAS, 101 (supplement 1)*, 5236-5240.

Glanzel, W. (2003). *On some on some principle differences between citations and sitation links. A methodological and mathematical approach*, NIWI, KNAW, Amsterdam, Updated version of a paper presented at the 6th Nordic Workshop on Bibliometrics, Stockholm, October 4-5, 2001.

Gmur, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics, 57*(1), 27-57.

Godin, B. (2005). *Measurement and Statistics on Science and Technology: 1920 to the Present*. London: Routledge.

Goldberger, A. L., Amaral, L. A. N., Hausdorff, J. M., Ch. Ivanov, P., Peng, C. K., & Stanley, H. E. (2002). Fractal dynamics in physiology: Alterations with disease and aging. *PNAS, 99*, 2466-2247.

Goldstein, M. L., Morris, S. A., & Yen, G. G. (2004). Problems with Fitting to the Power-Law Distribution. *The European Physical Journal B - Condensed Matter, 41*(2), 255-258.

Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science literature. *Information Processing and Management, 37*(5), 661-676.

Google. (2005). Google Web APIs (beta). http://www.google.com/apis/.

Gopal, R. D., Sanders, G. L., Bhattacharjee, S., Agrawal, M., & Wagner, S. C. (2004). A behavioral model of digital music piracy. *Journal of Organisational Computing and Electronic Commerce, 14*(2), 89-105.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management, 35*(2), 141-180.

Gorgura, H. (2004). *The war on the terror consensus: Anti-war blogs as an online sphere of dissensus*, Brighton, UK.

Gourley, D., & Totty, B. (2002). *{HTTP}: the definitive guide*. Farnham: O'Reilly.

Gournay, K. (2002). Prescribing: the great debate. *Nursing Standard, 17*(9), 22.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly, 68*(1), 2-31.

Gulli, A., & Signorini, A. (2005, May 10–14). *The Indexable Web is More than 11.5 billion pages.* Paper presented at the WWW 2005, Chiba, Japan.

Haas, S. W., & Grams, E. S. (2000). Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of American Society for Information Science, 51*(2), 181-192.

Hafri, Y., & Djeraba, C. (2004). *Dominos: a new Web crawler's design.* Paper presented at the IWAW-2004.

Hales, A., & Dignam, D. (2002). Nurse prescribing: Lessons from the US. *Nursing New Zealand, 8*(10), 12-15.

Hammersley, B. (2005). *Developing feeds with RSS and Atom*. Sebastopol, CA: O'Reilly.

Hammond, T., Hannay, T., & Lund, B. (2004). The role of RSS in science publishing: Syndication and annotation on the web. *Dlib, 12*. Retrieved from http://www.dlib.org/dlib/december04/hammond/12hammond.html

Harary, F. (1972). *Graph theory* (3rd printing ed.). London: Addison-Wesley.

Harnard, S., & Carr, L. (2000). Integrating, navigating, and analysing open eprint archives through open citation linking (the OpCit project). *Current Science, 79*(5), 629-638.

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5), 436-447.

Harter, S., & Ford, C. (2000). Web-based analysis of e-journal impact: Approaches, problems, and issues. *Journal of American Society for Information Science, 51*(13), 1159-1176.

Harter, S., & Taemin, K. P. (2000). Impact of prior electronic publication on manuscript consideration policies of sholarly journals. *Journal of American Society for Information Science, 51*(10), 940-948.

Haveliwala, T. (1999). *Efficient Computation of PageRank*. Stanford University Technical Report. Retrieved 28 Feb. 2003, from: http://www.cs.cornell.edu/home/kleinber/auth.pdf

Haveliwala, T. (2002). *Topic-Sensitive PageRank.* Paper presented at the the Eleventh International World Wide Web Conference.

Havemann, F., Heinz, M., & Wagner-Döbler, R. (2005). Firm-like Behaviour of Journals? Scaling Properties of Their Output and Impact Growth Dynamics. *Journal of the American Society for Information Science and Technology, 56*(1), 3-12.

Hawking, D., Craswell, N., Thistlewaite, P., & Harman, D. (2001). Results and challenges in Web search evaluation. *Computer Networks, 31*(11-16), 1321-1330.

Hayes, B. (2000). Graph Theory in Practice: Part I. *American Scientist, 88*(1), 9-13.

Hayes, B. (2000). Graph Theory in Practice: Part II. *American Scientist, 88*(1), 9-13.

Heimeriks, G., Hoerlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics, 58*(2), 391-413.

Henzinger, M., Heydon, A., Mitzenmacher, M., & Najork, M. (2000). *On near-uniform URL sampling.* Paper presented at the Proceedings of the 9th international World Wide Web conference on Computer networks, Amsterdam, The Netherlands.

Henzinger, M. R. (2000). Link Analysis in Web Information Retrieval. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 23*(3), 3-8.

Henzinger, M. R. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing, 5*(1), 45-50.

Hernandez, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery, 2*(1), 9-37.

Hernandez-Borges, A., Macias-Cervi, P., Gaspar-Guardado, A., Arcaya, M. L. T.-Ã. l. d., Ruiz-Rabaza, A., & Jimenez-Sosa, A. (2003). User preference as quality markers of paediatric web sites. *Medical Informatics and The Internet in Medicine, 28*(3), 183-194.

Hernandez-Borges, A. A., Macias-Cervi, P., Gaspar-Guardado, M. A., Arcaya, M. L. T.-Ã. l. d., Ruiz-Rabaza, A., & Jimenez-Sosa, A. (1999). Can examination of WWW usage statistics and other indirect quality indicators distinguish the relative quality of medical Web sites? *Journal of Medical Internet Research, 1*(1). Retrieved from http://www.jmir.org/1999/1/e1/index.htm

Herring, H. (2006). *From energy dreams to nuclear nightmares: Lessons from the anti-nuclear power movement in the 1970s*. Charlbury, UK: Jon Carpenter Publishing.

Herring, S. C. (2002). Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology, 36*, 109-168.

Herring, S. C. (2004). Slouching toward the ordinary: Current trends in Computer-Mediated Communication. *New Media & Society, 6*(1), 26-36.

Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. *World Wide Web, 2*, 219-229.

Heylighen, F. (1999). *The Science of Self-Organizination and Adaptivity*. Retrieved, from: http://pespmc1.vub.ac.be/Papers/EOLSS-Self-Organiz.pdf

Hine, C. (2000). *Virtual Ethnography*. London: Sage.
Hine, C. (2001). Web pages, authors and audiences: The meaning of a mouse click. *Information, Communication & Society, 4*(2), 182-198.
Hinman, L. M. (2002). Academic integrity and the World Wide Web. *ACM SIGCAS Computers and Society, 32*(1), 33-42.
Hjerland, B. (2002). Domain analysis in information science. Eleven approaches - traditional as well as innovative. *Journal of Documentation, 58*(4), 422-462.
Holbrook, J. A. D. (1991). The influence of scale effects on international comparisons of R&D expenditures. *Science and Public Policy, 18*(4), 259-262.
Hsu, S. H. (2005). Advocacy coalitions and policy change on nuclear power utilization in Taiwan. *Social Science Journal, 42*(2), 215-229.
Huberman, B. A. (1999). Growth dynamics of the world-wide web. *Nature, 399*.
Huberman, B. A. (2001). *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge, MA: The MIT Press.
Hughes, F., & Lockyer, H. (2004). Evidence and engagement in the introduction of nurse prescribing in New Zealand. *Nurse Prescribing, 2*(3), 131-134.
Hullmann, A. (2002). *Bibliometric and patent indicators by gender: Is it feasible?* Retrieved, from: http://www.cordis.lu/indicators/publications.htm
Hunter, P. (2003, April 21). The Power of Power Laws Are you a new user? *The Scientist, 17*.
Hyland, K. (2000). *Disciplinary discourses: social interactions in academic writing*. Harlow: Longman.
Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of American Society for Information Science, 54*(3), 251-259.
Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation, 54*(2), 236-243.
Introna, L., & Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The Information Society, 16*(3), 1-17.
Irvine, J., & Martin, B. (1983). Assessing Basic Research: The Case of Issac Newton Telescope. *Social Studies of Science, 13*, 49-86.
Irvine, J., & Martin, B. R. (1984). *Foresight in Science: Picking the Winners*. London & Dover: Frances Pinter Pub Ltd.
Jackson, M. H. (1997). Assessing the structure of communication on the world wide web. *Journal of Computer-Mediated Communication, 3*(1). Retrieved from http://www.ascusc.org/jcmc/vol3/issue1/jackson.html
Jaina, S., & Krishna, S. (2002). Graph Theory and the Evolution of Autocatalytic Networks. *arXiv:nlin.AO/0210070, 1*. Retrieved from http://arxiv.org/PS_cache/nlin/pdf/0210/0210070.pdf
Jeong H, Neda Z, & Barabasi, A.-L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters, 61*(4), 567-572.
Jepsen, E. T., Seiden, P., Ingwersen, P., & Bjorneborn, L. (2004). Characteristics of scientific Web publications: Preliminary data gathering and analysis. *Journal of the American Society for Information Science and Technology, 55*(14), 1239-1249.
Jin, B. H., Zhang, J. G., Chen, D. Q., & Zhu, X. Y. (2002). Development of the Chinese Scientometric Indicators (CSI). *Scientometrics, 54*(1), 145-154.
Johnson, B., & Gregersen, B. (1997). *European Integration and National Systems of Innovation*: Department of Business Studies, Aalborg University, Denmark.

Joseph, K. S., & Hoey, J. (1999). CMAJ's impact factor: room for recalculation. *Canadian Medical Association Journal, 161*(8), 977-978.

Jung, S., Kim, S., & Kahng, B. (2002). Geometric fractal growth model for scale-free networks. *Physical Review E, 65*(5), No. 056101.

Kalczynski, P. J., & Chou, A. (2005). Temporal document retrieval model for business news archives. *Information Processing and Management, 4*(3), 635-650.

Kaltenborn, K. F., & Kuhn, K. (2003). The journal impact factor as a parameter for the evaluation of researchers and research. *Medizinische Klinik, 98*(3), 153-169.

Karger, D. R., & Quan, D. (2004). What would it mean to blog on the Semantic Web? *Lecture Notes in Computer Science, 3298*, 214-228.

Katz, J. S. (1999). The Self-Similar Science System. *Research Policy, 28*, 501-517.

Katz, J. S. (2000). Scale-independent indicators and research evaluation. *Science and Public Policy, 27*(1), 23-26.

Katz, J. S. (2005). Indicators for Complex Innovation Systems. *Research Policy (submitted)*.

Katz, J. S. (2005). Scale Independent Bibliometric Indicators. *Measurement: Interdisciplinary Research and Perspectives, 3*(1), 24–28.

Katz, J. S., & Hicks, D. (1997). Desktop Scientometrics. *Scientometrics, 38*(1), 141-153.

Katz, J. S., & Hicks, D. (1998). *Indicators for Systems of Innovation - a bibliometrics-based approach* (Project No. PL951005 under the Targeted Socio-Economic Research Programme).

Katz, J. S., & Katz, L. (1994). Fractal (Power Law) Analysis of Athletic Performace. *Sports Medicine Training and Rehabilitation, 5*, 95-105.

Katz, J. S., & Katz, L. (1999). Power Laws and Athletic Performance. *Journal of Sports Science, 17*, 467-476.

Katz, J. S., & Martin, B. R. (1997). What is Research Collaboration. *Research Policy, 26*, 1-18.

Katz, J. S., & Plevin, J. (1998). Environmental Science in the UK: A Bibliometric Study. *Research Evaluation, 7*(1), 39-52.

Kavcic-Colic, A., & Grobelnik, M. (2004). *Archiving the Slovenian Web: recent experiences.* Paper presented at the IWAW-2004.

Keenan, M. (2003). The shifting sands of foresight. *Innovation Policy Review, 3*(5).

Keller, F., & Lapata, M. (2003). Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics, 29*(3), 459-484.

Kelly, B. (2000). WebWatch: A survey of links to UK university web sites. *Ariadne, 23*. Retrieved from http://www.ariadne.ac.uk/issue23/web-watch

Kelly, T. (2002). Thin-client Web access patterns: measurements from a cache-busting proxy. *Computer communications, 25*(4), 357-366.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of beahioural research* (Fourth ed.). London: Harcourt College Publishers.

Kiernan, V. (2003). Diffusion of news about research. *Science Communication, 25*(1), 3-13.

Kiesling, S. (2003). Prestige, cultural models, norms & gender. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender*. Oxford: Backwell.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics, 29*(3), 333-347.

Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science, 51*(10), 887-899.

Kim, J. H. (2005, October 5-9). *Blog as an oppositional medium? A semantic network analysis on the Iraq war blogs.* Paper presented at the Internet Generations: The 6th International and Interdisciplinary Conference of the Association of Internet Researchers, Chicago, Illinois.

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery, 7*(1), 81-99.

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

Kleinberg, J. M. (2000). *The small-world phenomenon: An algorithmic perspective.* Paper presented at the 32nd ACM Symposium on Theory of Computing.

Kleinberg, J. M. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373-397.

Kleinberg, J. M., Kumar, R., Raghavany, P., Rajagopalan, S., & Tomkins, A. (1999). *The Web as a graph: Measurements, models and methods.* Paper presented at the International Conference on Combinatorics and Computing, 1999.

Kleinberg, J. M., & Lawrence, S. (2001). The Structure of the Web. *Science, November 30*, 1849-1850.

Kling, R., & McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. *Journal of American Society for Information Science, 50*(10), 890-906.

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science, 51*(14), 1306-1320.

Klintman, M. (2002). The genetically modified (GM) food labelling controversy: Ideological and epistemic crossovers. *Social Studies of Science, 32*(1), 71-91.

Kobayashi, M., & Takeda, K. (2000). *Information retrieval on the Web:selected topics*. Tokyo: IBM Research: Tokyo Research Laboratory.

Koehler, W. (1999). Digital libraries and World Wide Web sites and page persistence. *Information Research, 4*(4).

Koehler, W. (2002). Web page change and persistence - A four-year longitudinal study. *Journal of American Society for Information Science, 53*(2), 162-171.

Koehler, W. (2004). A longitudinal study of Web pages continued: a report after six years. *Information Research, 9*(2), 174.

Koku, E., Nazer, N., & Wellman, B. (2001). Netting scholars: Online and offline. *American Behavioral Scientist, 44*(10), 1752-1774.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations, 2*(1), 1-15.

Koster, M. (1993). *Guidelines for robot writers* [[Online]]. Retrieved, from: http://www.robotstxt.org/wc/guidelines.html

Koster, M. (1994). *A standard for robot exclusion*. Retrieved, from: http://www.robotstxt.org/wc/robots.html

Kraak, A. (Ed.). (2000). *Changing modes: A brief overview of the mode 2 debate and its impact on South African policy formulation*. Pretoria: HSRC Publishers.

Krapivsky, P. L., Rodgers, G. J., & Redner, S. (2001). Degree Distributions of Growing Networks. *Physics Review Letters, 86*(23), 5401-5404.

Kretschmer, H. (2003). Author productivity and Erdos distances in co-authorship and in web link networks. In J. Guohua & R. Rousseau & W. Yishan (Eds.), *Proceedings of the 9th International Conference on Scientometrics and Informetrics, ISSI 2003* (pp. 393-400). Beijing: Dalian University of Technology Press.

Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics, 60*(3), 409-420.

Kretschmer, H., & Aguillo, I. F. (2004). Visibility of collaboration on the Web. *Scientometrics, 61*(3), 405-426.

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

Kretschmer, H., Kretschmer, T., & Kretschmer, U. (2005, July 24-28). *Visibility of collaboration between immunology institutions on the web Including aspects of gender studies.* Paper presented at the Proceedings of the 10th ISSI International Conference on Scientometrics and Informetrics, Stockholm, Sweden.

Kretschmer, H., Kretschmer, U., & Kretschmer, T. (2005). Reflection of co-authorship networks in the web: Web hyperlinks versus web visibility rates. *Proceedings of the 5th Triple Helix Conference, Turin, It, May 18-21*, (CD-ROM).

Krogh, C. (1996). The rights of agents. In M. Wooldridge & J. P. Muller & M. Tambe (Eds.), *Intelligent Agents II, Agent Theories, Architectures and Languages* (pp. 1-16): Springer Verlag.

Kuhlthau, C. C. (2004). *Seeking meaning: A process approach to library and information services, 2nd Ed.* Wesport, CT: Libraries Unlimited.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). *On the bursty evolution of blogspace.* Paper presented at the WWW2003, Budapest, Hungary, http://www2003.org/cdrom/papers/refereed/p477/p477-kumar/p477-kumar.htm.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35-39.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). *Trawling the web for emerging cyber-communities*, Toronto, http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html.

Kumar, R., Raghavany, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfalz, E. (2000). Stochastic Models for the Web Graph. *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*: IEEE Computer Society.

Kutz, D., & Herring, S. C. (2005). Micro-longitudinal analysis of Web news updates. *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences (HICSS-38)*, http://ella.slis.indiana.edu/~herring/news.pdf.

Lamirel, J.-C., Al Shehabi, S., Francois, C., & Polanco, X. (2004). Using a compound approach based on elaborated neural network for Webometrics: An example issued from the EICSTES project. *Scientometrics, 61*(3), 427-441.

Lampos, C., Eirinaki, M., Jevtuchova, D., & Vazirgiannis, M. (2004). *Archiving the Greek Web.* Paper presented at the IWAW-2004.

Larkey, L. S., Ogilvie, P. M., Price, A., & Tamilio, B. (2000, June 2-7). *Acrophile: An automated acronym extractor and server.* Paper presented at the The Fifth ACM Conference on Digital Libraries, San Antonio, TX.

Larson, R. R. (1996). *Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace.* Paper presented at the The AISS 59th annual meeting, Baltimore, MD.

Latora, V., & Marchiori, M. (2002, 8-12 Apr 2002). *The Architecture of Complex Systems*. Paper presented at the Interdisciplinary Applications of Ideas from Nonextensive Statistical Mechanics and Thermodynamics, Santa Fe Institute.

Latter, S., & Courtenay, M. (2004). Effectiveness of nurse prescribing: a review of the literature. *Journal of Clinical Nursing, 13*(1), 26-32.

Law, J. (2002). *Aircraft stories: Decentering the object in technoscience*. Durham, North Carolina: Duke University.

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science, 280*, 98-100.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400*, 107-109.

Lawrence, S. L. (2001). Online or invisible? *Nature, 411*, 521.

LeBaron, B. (2001). Stochastic Volatility as a simple generator of financial power laws and long memory. *Quantitative Finance, 1*, 621-631.

Lee, Y., Amaral, L. A. N., Canning, D., Meyer, M., & Stanley, H. E. (1998). Universal features in the growth dynamics of complex organizations. *cond-mat/9804100*. Retrieved from http://arxiv.org/PS_cache/cond-mat/pdf/9804/9804100.pdf

Lehmann, S., Lautrup, B., & Jackson, A. D. (2003). Citation networks in high energy physics. *Physical Review E, 68*.

Leskovec, J., Faloutsos, C., & Kleinberg, J. M. (2005). *Growth Power Law for Time Evolving Networks*. Jozef Stefan Institute, Slovenia.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations.* Paper presented at the KDD'05, August 21–24, 2005,, Chicago, Illinois, USA.

Leta, J., & Lewison, G. (2003). The contribution of women in Brazilian science: A case study in astronomy, immunology and oceanography. *Scientometrics, 57*(3), 339-353.

Levene, M., Fenner, T., Loizou, G., & Wheeldon, R. (2002). A Stochastic Model for the Evolution of the Web. *Computer Networks, 39*(3), 277-287.

Levene, M., & Poulovassilis, A. (Eds.). (2004). *Web Dynamics*. Berlin: Springer.

Levesque, S. (2001). The Yellowstone to Yukon conservation initiative. In J. Blatter & H. M. Ingram (Eds.), *Reflections on water: New approaches to transboundary conflicts and cooperation*. Cambridge, MA: MIT Press.

Levidow, L. (2001). Precautionary uncertainty: Regulating GM crops in Europe. *Social Studies of Science, 31*(6), 842-874.

Levy, M., & Solomon, S. (1996). Power Laws are Logarithmic Boltzmann Laws. *J. Mod. Phys. C*. Retrieved from http://xxx.tau.ac.il/PS_cache/adap-org/pdf/9607/9607001.pdf

Leydesdorff, L. (2004). 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), 991-1001.

Leydesdorff, L., & Curran, M. (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy. *Cybermetrics, 4*.

Leydesdorff, L., & Etzkowitz, H. (2003). Can "The Public" be considered as a fourth helix in University-Industry-Government relations? Report of the fourth triple helix conference. *Science and Public Policy, 30*(1), 55-61.

Leydesdorff, L., & Hellsten, I. (2005). Metaphors and diaphors in science communication: Mapping the case of "stem-cell research". *Science Communication, 27*(1), 64-99.

Leydesdorff, L., & Vaughan, L. (2006, to appear). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science & Technology*. Retrieved from http://www.leydesdorff.net/aca/aca.pdf

Li, X. (2003). A review of the development and application of the Web Impact Factor. *Online Information Review, 27*(6), 407-417.

Li, X., Thelwall, M., Musgrove, P. B., & Wilkinson, D. (2003). *Do Academic Web pages in Western Europe Attract More Links if they are in English?*, Beijing, China.

Li, X., Thelwall, M., Musgrove, P. B., & Wilkinson, D. (2003). The relationship between the WIFs or Inlinks of computer science departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics, 57*(2), 239-255.

Li, X., Thelwall, M., Wilkinson, D., & Musgrove, P. B. (2005, to appear). National and international university departmental web site interlinking, part 1: Validation of departmental link analysis. *Scientometrics*.

Li, X., Thelwall, M., Wilkinson, D., & Musgrove, P. B. (2005, to appear). National and international university departmental web site interlinking, part 2:Link patterns. *Scientometrics*.

Li, Y. F., Zhang, C. Q., & Zhang, S. C. (2003). Cooperative strategy for Web data mining and cleaning. *Applied Artificial Intelligence, 17*(5-6), 443-460.

Lifantsev, M. (2000). Voting model for ranking Web pages. In P. Graham & M. Maheswaran (Eds.), *Proceedings of the International Conference on Internet Computing* (pp. 143-148). Las Vegas: CSREA Press.

Lin, J., & Halavais, A. (2004, May 18th). *Mapping the blogosphere in America.* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, New York.

Lin, X., White, H. D., & Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. *Information Processing and Management, 39*(5), 689-706.

London, A. J. (2005). Undue inducements and reasonable risks: Will the dismal science lead to dismal research ethics? *American Journal Of Bioethics, 55*(5), 29-32.

Lucas, W., & Topi, H. (2004). Training for web search: Will it get you in shape? *Journal of the American Society for Information Science and Technology, 55*(13), 1183-1198.

Lundvall, B. A. (1992). *National Systems of innovation: Towards a theory of innovation and interactive learning*. London: Pinter.

Lyle, J. A. (2004). *Sampling the umich.edu domain.* Paper presented at the IWAW-2004.

Maganga, F., Kiwasila, H., Juma, I., & Butterworth, J. (2004). Implications of customary norms and laws for implementing IWRM: findings from Pangani and Rufiji basins, Tanzania. *Physics and Chemistry of the Earth, 29*(15-18), 1335-1342.

Marlow, C. (2004). Audience, structure and authority in the weblog community. *International Communication Association Conference*.

Martin, B., & Irvine, J. (1983). Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio Astronomy. *Research Policy, 12*, 61-90.

Martin, B. R., & Irvine, J. (1985). Evaluating the Evaluators. *Social Studies of Science, 15*, 558-585.

Masayuki Tatemichi, Tadashi Nakano, Katsutoshi Tanaka, Takeshi Hayashi, Takeshi Nawa, Toshiaki Miyamoto, Hisanori Hiro, & Sugita, M. (2004). Possible association between heavy computer users and glaucomatous visual field abnormalities: a cross sectional study in Japanese workers. *J Epidemiol Community Health, 58*, 1021–1027.

Matheson, D. (2004). Weblogs and the epistemology of the news: Some trends in online journalism. *New Media & Society, 6*(4), 443-468.

Mauboussin, M. J., & Bartholdson, K. (2002). More Power to You. *Credit Suisse First Boston, 1*.

McDonald, S., & Stevenson, R. J. (1998). Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and informatin retrieval in hypertext. *Interating With Computers, 10*(2), 129-142.

McElhinny, B. (2003). Theorizing gender in sociolinguistics and linguistic anthropology. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender* (pp. 21-42). Oxford: Backwell.

McEnery, A. M., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McFedries, P. (2004). The (pre) fix is in. *IEEE Spectrum Online*. Retrieved from http://www.spectrum.ieee.org/WEBONLY/resource/aug04/0804tech.html

McInerney, C., & Bird, N. (2005). Assessing Website quality in context: Retrieving information about genetically modified food on the Web. *Information Research, 10*(2).

McInerney, C., Bird, N., & Nucci, N. (2004). The flow of scientific knowledge from lab to the lay public: The case of genetically modified food. *Science Communication, 26*(1), 44-74.

McMillan, S. (2000). The microscope and the moving target: The challenge of applying content analysis to the world wide web. *Journalism and Mass Communication Quarterly, 77*(1), 80-98.

McNeil, J. (2004). Coding a bridge across the data divide. *The Scientist, 18*(24), 25.

Meghabghab, G. (2001). Google's Web page ranking applied to different topological Web graph structures. *Journal of the American Society for Information Science and Technology, 52*(9), 736-747.

Meghabghab, G. (2002). Discovering authorities and hubs in different topological Web graph structures. *Information Processing and Management, 38*(1), 111-140.

Meho, L. I., & Sonnenwald, D. H. (2000). Citation ranking versus peer evaluation of senior faculty research performance: A case study of Kurdish scholarship. *Journal of American Society for Information Science, 51*(2), 123-138.

Meho, L. I., & Spurgin, K. M. (2005). Ranking the Research Productivity of Library and Information Science Faculty and Schools:An Evaluation of Data Sources and Research Methods. *Journal of the American Society for Information Science and Technology, 45*(12), 1314-1331.

Meibauer, J., Guttropf, A., & Scherer, C. (2004). Dynamic aspects of German -er-nominals: a probe into the interrelation of language change and language acquisition. *Linguistics, 42*(1), 155-193.

Menczer, F. (2004). Correlated topologies in citation networks and the Web. *The European Physical Journal B, 38*(2), 211-221.

Menczer, F. (2004). Evolution of document networks. *PNAS, 101 (supplement 1)*, 5261-5265.

Menczer, F. (2004). Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology, 55*(14), 1261-1269.

Menczer, F. (2005). Mapping the semantics of Web text and links. *Internet Computing, 9*(3), 27-36.

Merton, R. K. (2000). On the Garfield input to the sociology of science: A retrospective collage. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 435-448). Medford, NJ: Information Today, inc. ASIS Monograph Series.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines: fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623-651.

Meyer, C., Grabowski, R., Han, H.-Y., Mantzouranis, K., & Moses, S. (2003). The world wide web as linguistic corpus. *Language and Computers, 46*(1), 241-254.

Meyer, M. (2003). Academic patents as an indicator of useful research? A new approach to measure academic inventiveness. *Research Evaluation, 12*(1), 17-27.

Middleton, I., McConnell, M., & Davidson, G. (1999). Presenting a model for the structure and content of a university World Wide Web site. *Journal of Information Science, 25*(3), 219-227.

Miles-Board, T., Carr, L., & Hall, W. (2002). Looking for linking: associative links on the Web. *Proceedings of ACM Hypertext 2002*, 76-77.

Miller. (1957). Some Effects of Intermittent Silence. *American Journal of Psychology, 70*, 311-314.

Miller, H., & Arnold, J. (2001). Breaking away from grounded identity? Women academics on the Web. *CyberPsychology & Behavior, 4*(1), 95-108.

Miller, R. C., & Bharat, K. (1998). *Sphinx: a framework for creating personal site-specific Web crawlers.* Paper presented at the The seventh conference on World Wide Web, Brisbane, Australia.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science, 298*, 824-827.

Mitkov, R. (2003). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Mitra, A., Hazen, M. D., LaFrance, B., & Rogan, R. G. (1999). Faculty use and non-use of electronic mail: Attitudes, expectations and profiles. *Journal of Computer-Mediated Communication, 4*(3), Retrieved 24 August, 2002, from http://www.ascusc.org/jcmc/vol2004/issue2003/mitra.html.

Mitzenmacher, M. (2003). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics, 1*(2), 226-251.

Moed, H., F. (2005). *Citation analysis in research evaluation*. New York: Springer.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal (p). *Journal of the American Society for Information Science & Technology, 56*(10), 1088-1097.

Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I. (2004). *An introduction to Heritrix: an open source archival quality web crawler.* Paper presented at the IWAW-2004.

Monge, P., & Contractor, N. S. (2000). *Emergence of communication networks*. Thousand Oaks, CA: Sage.

Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *arXiv:cond-mat/0104066, 2*(July 9), 1-6. Retrieved from http://ernie.ecs.soton.ac.uk/opcit/cgi-bin/pdf?id=oai%3AarXiv.org%3Acond-mat%2F0104066

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, co-citation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology, 56*(12), 1250-1273.

Mossa, S., Barthelemy, M., Stanley, H. E., & Amaral, L. A. N. (2002). Truncation of Power Law Behavior in "Scale-Free" Network Models due to Information Filtering. *Physical Review Letters, 88*(13).

Mullay, S., Mason, C., & Frogatt, J. (2003). The progress of nurse prescribing in the United Kingdom. *Nurse Prescribing, 1*(3), 104-105.

Murphy, J., Stramer, K., Clamp, S., Grubb, P., Gosland, J., & Davis, S. (2004). Health informatics education for clinicians and managers "What's holding up progress?" *International Journal of Medical Informatics, 73*(2), 205-213.

Musgrove, P. B., Binns, R., Page-Kennedy, T., & Thelwall, M. (2003). A method for identifying clusters in sets of interlinking Web spaces. *Scientometrics, 58*(3), 657-672.

Najork, M., & Heydon, A. (2001). *High-performance Web crawling*. Palo Alto, CA: Compaq Systems Research Center.

Najork, M., & Wiener, J. L. (2001). *Breadth-First Search Crawling Yields High-Quality Pages.* Paper presented at the WWW10, Singapore.

Naldi, F., Luzi, D., Valente, A., & Parenti, I. V. (2004). Scientific and technological performance by gender. In H. F. Moed (Ed.), *Handbook of Quantitative Science and Technology Research* (pp. 299-314). Dordrecht: Kluwer Academic Publishers.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41-46.

Neary, J. P. (2005). Purchasing Power Parity. In S. E. J.J. McCusker, L.R. Fischer, D.J. Hancock, K.L. Pomeranz (Ed.), *Encyclopedia of World Trade Since 1450*. New York: Macmillan Reference.

Nentwich, M. (2003). *Cyberscience: Research in the age of the Internet*. Vienna: Austrian Academy of Sciences.

NetBig. (2001). *The indicator system and weight assignment after adjustment*. Retrieved 1 Apr. 2003, from: http://rank2001.netbig.com/en/about/03.htm

Newman, M. E. J. (2000). The power of design. *Nature, 405*.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physics Review E, 64*(025102).

Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E, 64*(016131).

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*(016132).

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS, 98*, 404-409.

Newman, M. E. J. (2003). Ego-centered networks and the ripple effect. *Social Networks, 25*, 83-95.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *Society for Industrial and Applied Mathematics, 45*(2), 167–256.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *PNAS, 101*(supplement 1), 5200–5205.

Newman, M. E. J. (2004). Detecting community structure in networks. *Phys. Rev. E, 38*(2), 321-330.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *arXiv:cond-mat, v2*. Retrieved from http://citebase.eprints.org/cgi-bin/fulltext?format=application/pdf&identifier=oai:arXiv.org:cond-mat/0412004

Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *PNAS, 99*, 2566-2572.

Newson, M. (1999). *Land, water and development: Sustainable management of river basin systems*. London: Routledge.

Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001). *Stable algorithms for link analysis.* Paper presented at the SIGIR-2001, New York, NY, USA.

Nicholas, D., Huntington, P., Lievesley, N., & Wasti, A. (2000). Evaluating consumer Website logs: a case study of The Times/The Sunday Times Website. *Journal o f Information Science, 51*(5), 144-154.

Nicholas, D., Huntington, P., Lievesley, N., & Withey, R. (1999). Cracking the code: Web log analysis. *Online & CD-ROM Review, 23*(5), 263-269.

Nicholas, D., Huntington, P., & Williams, P. (2002). Evaluating metrics for comparing the use of web sites: a case study of two consumer health web sites. *Journal of Information Science, 28*(1), 63 - 75.

Nicholls, P. (1988). Erratum: Estimation of Zipf parameters. *Journal of the American Society of Information Science, 39*(4), 287.

Nicholls, P. T. (1986). Empirical validation of Lotka's Law. *Information Processing and Management, 22*(5), 417-419.

Nicholls, P. T. (1987). Estimation of Zipf parameters. *Journal of the American Society of Information Science, 38*(6), 443-445.

Nicholls, P. T. (1999). Bibliometric modeling processes and the empirical validity of {Lotka's Law}. *Journal of the American Society for Information Science and Technology, 40*(6), 379-385.

Nilan, M. S., Pomerantz, J., & Paling, S. (2001). *Genres from the bottom up: What has the Web brought us?* Paper presented at the 64th Annual Meeting of the American Society for Information Science and Technology.

Ntoulas, A., Cho, J., & Olston, C. (2004, May 17–22). *What's New on the Web? The Evolution of the Web from a Search Engine Perspective.* Paper presented at the WWW 2004, New York, USA.

Oldham, G., & Achmad, S. (1999). Gender mainstreaming in science and technology - a global report. *Nature*. Retrieved from http://www.nature.com/nature/debates/women/women_frameset.html

Olsen, S. (2003). Google cache raises copyright concerns. http://news.com.com/2100-1038_2103-1024234.html.

O'Neill, E. T., McClain, P. D., & Lavoie, B. F. (1997). *A Methodology for Sampling the World Wide Web*. Retrieved 1 Apr. 2003, from: http://www.oclc.org/research/publications/arr/1997/oneill/o'neillar980213.htm

Oppenheim, C. (2001). LISLEX: Legal issues of concern to the library and information science sector. *Journal o f Information Science, 27*(4), 277-286.

Ortega Priego, J. L. (2003). A Vector Space Model as a methodological approach to the Triple Helix dimensionality: A comparative study of Biology and Biomedicine centres of two European national research councils from a Webometric view. *Scientometrics, 58*(2), 429-443.

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

Ozmutlu, S., & Cavdur, F. (2005). Neural network applications for automatic new topic identification. *Online Information Review, 29*(1), 34-53.

Page, L. (2001). Method for node ranking in a linked database: United States Patent.

Palmer, J. W., Bailey, J. P., & Faraj, S. (2000). The role of intermediaries in the development of trust on the WWW: The use and prominence of trusted third parties and privacy statements. *Journal of Computer-Mediated Communication, 5*(3). Retrieved from http://www.ascusc.org/jcmc/vol5/issue3/palmer.html

Pandia. (2001). *On the size of the World Wide Web*. Retrieved 22 July, 2002, from: http://www.pandia.com/sw-2001/57-websize.html

Pandia. (2004). *The death of AltaVista and AlltheWeb*. Retrieved 1 April, 2004, from: http://www.pandia.com/sw-2004/08-yahoo.html

Pant, G., & Menczer, F. (2002). Myspiders: evolve your own intelligent Web crawlers. *Autonomous agents and multi-agent systems, 5*, 221-229.

Pao, M. L. (1985). Lotka's Law: a testing procedure. *Information Processing and Management, 21*(4), 305-320.

Paolillo, J. C. (2001). Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics, 5*(2), 180-213.

Pardue, M.-L., Hopkins, N., Potter, M. C., & Ceyer, S. (1999). Moving from discrimination at the Massachusetts Institute of Technology. *Nature*. Retrieved from http://www.nature.com/nature/debates/women/women_frameset.html

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections, 25*(1), 49-61.

Park, H. W., Barnett, G. A., & Kim, C. S. (2001). Internet communication structure in Korean National Assembly: A network analysis. *Korean Journal of Journalism & Communication Studies, Special English Edition*, 185-203.

Park, H. W., Barnett, G. A., & Nam, I. (2002). Hyperlink affiliation network structure of top web sites: Examing affiliates with hyperlink in Korea. *Journal of American Society for Information Science and Technology, 53*(7), 592-601.

Park, H. W., & Thelwall, M. (2003). Hyperlink analyses of the world wide web: A review. *Journal of Computer-Mediated Communication, 8*(4). Retrieved from http://jcmc.indiana.edu/vol8/issue4/park.html

Parrish, J. K., Viscid, S. V., & Grunbaum, D. (2002). Self-Organized Fish Schools: An Examination of Emergent Properties. *Biol. Bull., 202*, 296–305.

Paul, G., Tanizawa, T., Havlin, S., & Stanley, H. E. (2004). Optimization of robustness of complex networks. *European Physics Journal B, 38*(2), 187-191.

Payne, N., & Thelwall, M. (2005). Mathematical models for academic Webs: Linear relationship or non-linear power law? *Information Processing & Management, 41*(6), 1495-1510.

Pennock, D., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences, 99*, 5207-5211.

Petermann, T., & De Los Rios, P. (2004). Exploration of Scale-Free Networks: Do we measure the real exponents? *European Physics Journal B, 38*(2), 201-204.

Peterson, G. D. (2000). Scaling ecological dynamics: self-organization, hierarchical structure and ecological resilience. *Climatic Change, 44*, 3.

Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117-136.

Phillips, B. J. (2005). A comprehensive look at the legislative issues affecting advanced nursing practice. *The Nurse Practitioner, 30*(2), 14-47.

Phillips, B. J., & McQuarrie, E. F. (2003). The development, change, and transformation of rhetorical style in magazine advertisements 1954-1999. *Journal of Advertising, 31*(4), 1-13.

Phillips, B. J., & McQuarrie, E. F. (2004). Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing Theory, 4*(1-2), 113-136.

Pikas, C. K. (2005). Blog searching for competitive intelligence, brand image, and reputation management. *Online, 29*(4), 16-21.

Pini, B., Brown, K., & Previte, J. (2004). Politics and Identity in Cyberspace: A case study of Australian women in agriculture online. *Information, Communication & Society, 7*(2), 167-184.

Pinsky, M. (2003). *Future present: Ethics and/as science fiction*. London: Associated University Presses.

Plerou, V., Amaral, L. A. N., Gopikrishnan, P., Meyer, M., & Stanley, H. E. (1999). Similarities between the growth dynamics of university research and of competitive economic activities. *Nature, 400*, 433–437.

Plerou, V., Gopikrishnan, P., Gabaix, P., & Stanley, H. E. (2004). On the Origin of Power-Law Fluctuations in Stock Prices. *Quantitative Finance 4, 4*.

Ploncynski, D., Oldenburg, N., & Buck, M. (2003). The past, present and future of nurse prescribing in the United States. *Nurse Prescribing, 1*(4), 170-174.

Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences1. *American Journal of Sociology, 110*(4).

Prabowo, R., & Thelwall, M. (2005, submitted). A comparison of feature selection methods for an evolving RSS feed corpus.

Price, D. J. d. S. (1963). *Little Science, Big Science*. New York and London: Columbia University Press.

Price, D. J. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science and Technology, 27*, 292–306.

Price, E., & Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology, 56*(8), 883-888.

Priego, J. L. O. (2003). A vector space model as a methodological approach to the Triple Helix dimensionality: A comparative study of biology and biomedicine centres of two European national research councils from a webometric view. *Scientometrics, 58*(2), 429-444.

Prime, C., Bassecoulard, E., & Zitt, M. (2002). Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics, 54*(1), 291-308.

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA: Morgan Kaufmann.

Qiu, J. P., Chen, J. Q., & Wang, Z. (2004). An analysis of backlink counts and Web Impact Factors for Chinese university websites. *Scientometrics, 60*(3), 463-473.

Rafiei, D., & Mendelzon, A. O. (2000). What this page known for? Computing Web page reputations. *Computer Networks, 33*(1-6), 823-835.

Raghavan, S., & Garcia-Molina, H. (2001). Crawling the hidden Web, *VLDB 01 proceedings of the 27th international conference on very large databases* (pp. 129-138). San Francisco: Morgan Kaufman.

Raghavan, S., & Garcia-Molina, H. (2003). Representing Web graphs. In U. Dayal (Ed.), *Proceedings of the IEEE Intl. Conference on Data Engineering* (pp. 405-416). San Jose: IEEE.

Rall, D. N. (2004). Exploring the breadth of disciplinary backgrounds in internet scholars participating in AoIR meetings, 2000-2003. *Proceedings of AoIR 5.0*. Retrieved from http://gsb.haifa.ac.il/~sheizaf/AOIR5/399.html

Rall, D. N. (2004). Locating internet research methods within five qualitative research traditions. *Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods*, http://cybermetrics.wlv.ac.uk/AoIRASIST/.

Randle, V. (1999). Is the glass ceiling an illusion? *Nature*. Retrieved from http://www.nature.com/nature/debates/women/women_frameset.html

Ravasza, E., & Bararbasi, A. U. (2003). Hierarchical organization in complex networks. *Physical Review E, 67*, 026112-026111-026112-026117.

Raymond, D. R., & Fawcett, H. J. (1999). Playing detective with full text searching software. *ACM SIGDOC Asterisk Journal of Computer Documentation, 14*(4), 157 - 166.

Redmond, M. V. (2000). Cultural distance as a mediating factor between stress and intercultural communication competence. *International Journal of Intercultural Relations, 24*, 151-159.

Reed, W. J., & Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E, 66*, 067103.

Reed, W. J., & Hughes, B. D. (2003). Power-law distributions from exponential processes: an explanation for the occurrence of long-tailed distributions in biology and elsewhere. *Scientiae Mathematicae Japonicae, 58*(2), 473 - 483.

Resnik, P., & Smith, N. (2003). The Web as a parallel corpus. *Computational Linguistics, 29*(3), 349-380.

Rey-Rocha, J., Martín-Sempere, M. J., Martinez-Frías, J., & López-Vera, F. (2001). Some misuses of journal impact factor in research evaluation. *Cortex, 37*(4), 595-597.

Riva, G. (2001). The Mind Over the Web: The Quest for the Definition of a Method for Internet Research. *CyberPsychology & Behavior, 4*(1), 7-16.

Rodgers, G. J., & Darby-Dowman, K. (2001). Properties of a growing random directed network. *Eur. Phys. J. B, 23*, 267-271.

Rodriguez i Gairin, J. M. (1997). Valorando el impacto de la informacion en Internet: AltaVista, el "Citation Index" de la Red. *Revista Espanola de Documentacion Cientifica, 20*(2), 175-181.

Rogers, R. (2002). Operating issue networks on the Web. *Science as Culture, 11*(2), 191-214.

Rogers, R. (2004). *Information Politics on the Web*. Massachusetts: MIT Press.

Rosser, S., & Zieseniss, M. (2000). Career issues and laboratory climates: different challenges and opportunities for women engineers and scientists (survey of fiscal year 1997 POWRE awardees). *Journal of Women and Minorities in Science and Engineering, 6*(2), 95-114.

Rousseau, B., & Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. *Cybermetrics, 4*(1). Retrieved from http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html

Rousseau, R. (1993). A Table for Estimating the Exponent in Lotka's Law. *Journal of Documentation, 49*(4), 409-412.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics, 1*(1). Retrieved from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*. Retrieved from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Ru, Y. B., & Horowitz, E. (2005). Indexing the invisible web: a survey. *Online Information Review, 29*(3), 249-265.

Savaglio, S., & Carbone, V. (2000). Scaling in Athletic World Records. *Nature, 404*, 244.

Schaap, F. (2004). *Multimodal interactions and singular selves: Dutch weblogs and home pages in the context of everyday life.* Paper presented at the AoIR 5.0, Brighton, UK.

Schiano, D. J., Nardi, B. A., Gumbrecht, M., & Swartz, L. (2004). *Blogging by the rest of us.* Paper presented at the Conference on Human Factors and Computing Systems (CHI 2004), Vienna.

Schneier, B. (2004). *Secrets and Lies: Digital Security in a Networked World*. New York: Hungry Minds Inc.

Schroeder, M. (1991). *Fractals, chaos and power laws*. New York: W.H. Freeman and Company.

Schubert, A. (2001). Scientometrics: A citation based bibliography 1997-2000. *Scientometrics, 50*(1), 99-198.

Schubert, A., & Glänzel, W. (1984). A dynamic look at a class of skew distributions: a model with scientometric applications. *Scientometrics, 6*(3), 149-167.

Schuch, K. (1998). *The emergence of the European Innovation System and its impact on the Austrian S&T system.* Paper presented at the Proceedings from the 38th Congress of the European Regional Science Association, Aug - Sept, 28-1, Vienna.

Search Tools Consulting. (2001). *Source code for Web robot spiders* [[Online]]. Retrieved, from: http://www.searchtools.com/robots/robot-code.html

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. *PNAS, 101 (supplement 1)*, 5266-5273.

Shiode, N., & Batty, M. (2000). Power Law Distributions in Real and Virtual Worlds. *Centre for Advanced Spatial Analysis Working Papers*(19).

Shkapenyuk, V., & Suel, T. (2002). *Design and implementation of a high-performance distributed Web crawler.* Paper presented at the ICDE-2002.

SIBIS. (2003). *Internet for research: Topic report no. 2*: University of Applied Sciences.

Silva, A. S. d., Veloso, E. A., Golgher, P. B., Ribeiro-Neto, B., Laender, A. H. F., & Ziviani, N. (1999). Cobweb: a crawler for the Brazilian Web. *Proceedings of the String Processing and Information Retrieval Symposium & International Workshop on Groupware* (pp. 184-191).

Silver, D. (2004). Internet/cyberculture/digital culture/new media/fill-in-the-blank studies. *New Media & Society, 6*(1), 55-64.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6-12.

Small, H. (1999). A passage through science: crossing disciplinary boundaries. *Library Trends, 48*(1), 72-108.

Smith, A. (2004). The use of the internet to support the education of nurse prescribers. *Nurse Prescribing, 2*(3), 127-130.

Smith, A., & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics, 54*(3), 363–380.

Smith, A. G. (1999). *ANZAC Webometrics: exploring Australasian Web structures.* Paper presented at the Information Online and On Disc 99, Sydney, Australia.

Smith, A. G. (1999). *The Impact of Web sites: a comparison between Australasia and Latin America.* Paper presented at the INFO'99, Congreso Internacional de Informacion, Havana.

Smith, A. G. (1999). A tale of two web spaces; comparing sites using Web Impact Factors. *Journal of Documentation, 55*(5), 577-592.

Smith, A. G. (2000). Search features of digital libraries. *Information Research, 5*(3).

Smith, A. G. (2004). Web links as analogues of citations. *Information Research, 9*(4). Retrieved from http://informationr.net/ir/9-4/paper188.html

Smith, A. G., & Thelwall, M. (2001). *Web Impact Factors and University research links.* Paper presented at the the 8th international Conference on Scientometrics & Informetrics 16-20 July 2001, Sydney Australia.

Smith, D. A. (2002). Detecting and browsing events in unstructured text. In K. Järvelin & M. Beaulieu & R. Baeza-Yates & S. H. Myaeng (Eds.), *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 73-80). New York: ACM Press.

Smith, S. (2005). Tapping the feed: In search of an RSS money trail. *Econtent, 28*(3), 30-34.

Smith, Z. (1997). The truth about the Web. *New architect: Internet strategies for technology leaders, 2*(5).

Snyder, H. W., & Rosenbaum, H. (1999). Can search engines be used for Web-link analysis? A critical review. *Journal of Documentation, 55*(4), 375-384.

Solomon, S., & Agay, A. (1997). *Power-Laws and Scaling in the Generalized Lotka-Volterra (GLV) Model.* Paper presented at the 1st Workshop on Econophysics, Budapest.

Sornette, D. (1998). Multiplicative processes and power laws. *Phys. Rev. E., 57*(4), 4811-4813.

Sousa, V. D., Zauszniewski, J. A., & Musil, C. M. (2004). How to determine whether a convenience sample represents the population. *Applied Nursing Research, 2*, 130-133.

Spertus, E. (1997). ParaSite: mining structural information on the Web. *Computer Networks and ISDN Systems, 29*(8-13), 1205-1215.

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the web*. Dordrecht: Kluwer Academic Publishers.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of American Society for Information Science, 53*(2), 226-234.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations, 1*(2), 12-23.

Stafford, T. F., & Stafford, M. R. (2001). Identifying motivations for the use of commercial Web sites. *Information Resources Management Journal, 14*(1), 22-30.

Stanley, H. E., Amaral, L. A. N., Buldyrev, S. V., Gopikrishnan, P., Plerou, V., & Salinger, M. A. (2002). Self-organized complexity in economics and finance. *PNAS, 99*(suppl 1), 2561–2565.

Stanley, H. E., Amaral, L. A. N., Gopikrishnan, P., & Plerou, V. (2002). Scale invariance and universality in economic phenomena. *Journal of Physics: Condensed Matter, 14*(9), 2121–2131.

Stanley, H. E., & Plerou, V. (2001). Scaling and universality in economics: empirical results and theoretical interpretation. *Quantitative Finance, 1*, 563-567.

Stokes, D. E. (1997). *Pascal's quadrant: Basic science and technological innovation*. Washington, D.C.: Brookings Institution.

Strogatz, S. H. (2001). Exploring Complex Networks. *Nature, 410*, 268-276.

Strogatz, S. H. (2005). Romanesque networks. *Nature, 433*(7), 365.

Stuart, D., & Thelwall, M. (2005). What can university-to-government web links tell us about a university's research productivity and the collaborations between universities and government? In P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005* (Vol. 1, pp. 188-192). Stockholm: Karolinska University Press.

Sunstein, C. R. (2004). Democracy and filtering. *Communications of the ACM, 47*(12), 57-59.

Sutton, J. (1997). Gibrat's Legacy. *Journal of Economic Literature, 35*(1), 40-59.

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In E. Yannakoudakis & N. J. Belkin & M.-K. Leong & P. Ingwersen (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 49-56). Athens, Greece.

Tadi, B. (2001). Dynamics of directed graphs: the world-wide web. *Physics A, 293*, 273-284.

Tadi, B. (2002). Growth and structure of the World Wide Web: towards realistic modelling. *Computer Physics Communications, 147*, 586-589.

Tague, J., & Nicholls, P. (1987). The maximal value of a Zipf size variable: sampling properties and relationship to other parameters. *Information Processing and Management, 23*(3), 155-170.

Talja, S., & Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases: a domain analytic study in four scholarly disciplines. *Journal of Documentation, 59*(6), 673-691.

Tang, R., & Thelwall, M. (2002). *Exploring the pattern of links between Chinese university Web sites.* Paper presented at the the 65th Annual Meeting of American Society for Information Science and Technology
2002.

Tang, R., & Thelwall, M. (2003). *Patterns of International and National Web Inlinks to US University Departments: A Webometric Analysis of Disciplinary Specificity*, Beijing, China.

Tang, R., & Thelwall, M. (2003). US academic departmental web-site interlinking in the United States disciplinary differences. *Library and Information Science Research, 25*(4), 437-458.

Tang, R., & Thelwall, M. (2005, to appear). A hyperlink analysis of US public and academic libraries' Web sites. *Library Quarterly*.

Teo, T. S. H. (2001). Demographic and motivation variables associated with Internet usage activities. *Internet Research: Electronic Networking and Applications and Policy, 11*(2), 125-137.

Thelwall, M. (2000). Commercial Web sites: Lost in Cyberspace? *Internet Research, 10*(2), 150-159.

Thelwall, M. (2000). Effective web sites for small to medium sized enterprises. *Journal of Small Business and Enterprise Development, 7*(2), 149-159.

Thelwall, M. (2000). Web Impact Factors and search engine coverage. *Journal of Documentation, 56*(2), 185-189.

Thelwall, M. (2000). Who is using the.co.uk domain? Professional and media adoption of the web. *International of Information Management, 20*(6), 441-453.

Thelwall, M. (2001). Commercial Web Site Links. *Internet Research, 11*(2), 114-124.

Thelwall, M. (2001). Exploring the link structure of the Web with network diagrams. *Journal of Information Science, 27*(6), 393-402.

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

Thelwall, M. (2001). The responsiveness of search engine Indexes. *Cybermetrics, 5*(1). Retrieved from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2001). Results from a Web Impact Factor crawler. *Journal of Documentation, 57*(2), 177-191.

Thelwall, M. (2001). A web crawler design for data mining. *Journal of Information Science, 27*(5), 319-325.

Thelwall, M. (2001). Web log file analysis: Backlinks and queries. *ASLIB Proceedings, 53*(6), 217-223.

Thelwall, M. (2002). A comparison of sources of links for academic Web Impact Factor calculations. *Journal of Documentation, 58*(1), 60-72.

Thelwall, M. (2002). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of American Society for Information Science and Technology, 53*(12), 995-1005.

Thelwall, M. (2002). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation, 58*(5), 563-574.

Thelwall, M. (2002). In praise of Google: Finding law journal Web sites. *Online Information Review, 26*(4), 271-272.

Thelwall, M. (2002). An initial exploration of the link relationship between UK university web sites. *ASLIB Proceedings, 54*(2), 118-126.

Thelwall, M. (2002). Methodologies for crawler based Web surveys. *Internet Research: Electronic Networking Applications and Policy, 12*(2), 124-138.

Thelwall, M. (2002). A research and institutional size based model for national university web site interlinking. *Journal of Documentation, 58*(6), 683-694.

Thelwall, M. (2002). Research dissemination and invocation on the Web. *Online Information Review, 26*(6), 413-420.

Thelwall, M. (2002). Subject gateway sites and search engine ranking. *Online Information Review, 26*(2), 101-107.

Thelwall, M. (2002). The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content. *Journal of Information Science, 28*(6), 485-493.

Thelwall, M. (2003). Can Google's PageRank be used to find the most important academic Web pages? *Journal of Documentation, 58*(6), 205-217.

Thelwall, M. (2003). A free database of university Web links: Data collection issues. *Cybermetrics, 6/7*(1). Retrieved from http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html

Thelwall, M. (2003). A layered approach for investigating the topological structure of communities in the Web. *Journal of Documentation, 59*(4), 410-429.

Thelwall, M. (2003). Web use and peer interconnectivity metrics for academic Web sites. *Journal of Information Science, 29*(1), 11-20.

Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research, 8*(3). Retrieved from http://informationr.net/ir/8-3/paper151.html

Thelwall, M. (2004). Can the web give useful information about commercial uses of scientific research? *Online Information Review, 28*(2), 120-130.

Thelwall, M. (2004). *Link analysis: An information science approach*. San Diego: Academic Press.

Thelwall, M. (2004). Methods for reporting on the targets of links from national systems of university Web sites. *Information Processing and Management, 40*, 125-144.

Thelwall, M. (2004). Vocabulary Spectral Analysis as an exploratory tool for Scientific Web Intelligence. In E. Banissi (Ed.), *Information Visualization (IV04)* (pp. 501-506). Los Alamitos, CA: IEEE.

Thelwall, M. (2004). Weak benchmarking indicators for formative and semi-evaluative assessment of research. *Research Evaluation, 13*(1), 63-68.

Thelwall, M. (2004). Will digital libraries generate a new need for multi-disciplinary research skills? *LIBRES, 14*(2), Retrieved April 18 from: http://libres.curtin.edu.au/libres14n12/index.htm.

Thelwall, M. (2005). Data cleansing and validation for Multiple Site Link Structure Analysis. In A. Scime (Ed.), *Web Mining: Applications and Techniques* (pp. 208-227): Idea Group Inc.

Thelwall, M. (2005). Text characteristics of English language university web sites. *Journal of the American Society for Information Science and Technology, 56*(6), 609-619.

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

Thelwall, M., & Aguillo, I. F. (2003). La salud de las Web universitarias espanolas. *Revista Espanola de Documentacion Cientifica, 26*(3), 291-305.

Thelwall, M., Barlow, A., & Vann, K. (2005). The limits of web-based empowerment: Integrated Water Resource Management case studies. *First Monday, 10*(4), Retrieved April 20, 2005 from: http://www.firstmonday.org/issues/issue2010_2004/thelwall/.

Thelwall, M., Binns, R., Harries, G., Page-Kennedy, T., Price, E., & Wilkinson, D. (2002). European Union associated university websites. *Scientometrics, 53*(1), 95-111.

Thelwall, M., Binns, R., Harries, G., Page-Kennedy, T., Price, L., & Wilkinson, D. (2001). Custom interfaces for advanced queries in search engines. *ASLIB Proceedings, 53*(10), 413-422.

Thelwall, M., & Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and Technology, 54*(7), 594-602.

Thelwall, M., & Harries, G. (2004). Can personal web pages that link to universities yield information about the wider dissemination of research? *Journal of Information Science, 30*(3), 243-256.

Thelwall, M., & Harries, G. (2004). Do better scholars' Web publications have significantly higher online impact? *Journal of American Society for Information Science and Technology, 55*(2), 149-159.

Thelwall, M., & Harries, G. (2004). Do the Web sites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology, 55*(2), 149-159.

Thelwall, M., Harries, G., & Wilkinson, D. (2003). Why do web sites from different academic subjects interlink? *Journal of Information Science, 29*(6), 445-463.

Thelwall, M., & Prabowo, R. (2005, submitted). Identifying and characterising public science-related concerns from RSS feeds.

Thelwall, M., Prabowo, R., & Fairclough, R. (2006, to appear). Are raw RSS feeds suitable for broad issue scanning? A science concern case study. *Journal of the American Society for Information Science and Technology*.

Thelwall, M., & Smith, A. (2002). Interlinking between Asia-Pacific University Web sites. *Scientometrics, 55*(3), 363–376.

Thelwall, M., & Smith, A. G. (2002). A study of interlinking between Asia-Pacific university web sites. *Scientometrics, 55*(3), 335-348.

Thelwall, M., & Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics, 58*(1), 155.181.

Thelwall, M., Tang, R., & Price, E. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics, 56*(3), 417-432.

Thelwall, M., Vann, K., & Fairclough, R. (2006, to appear). Web issue analysis: An Integrated Water Resource Management case study. *Journal of the American Society for Information Science & Technology*.

Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library and Information Science Research, 26*(2), 162-176.

Thelwall, M., & Vaughan, L. (2004). New versions of PageRank employing alternative Web document models. *ASLIB Proceedings, 56*(1), 24-33.

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

Thelwall, M., Vaughan, L., Cothey, V., Li, X., & Smith, A. G. (2003). Which academic subjects have most online impact? A pilot study and a new classification process. *Online Information Review, 27*(5), 333-343.

Thelwall, M., & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies. *Journal of American Society for Information Science and Technology, 54*(8), 706-712.

Thelwall, M., & Wilkinson, D. (2003). Graph Structure in Three National Academic Webs: Power Laws with Anomalies. *Journal of the American Society for Information Science and Technology, 54*(8), 706-712.

Thelwall, M., & Wilkinson, D. (2003). Three target document range metrics for university Web sites. *Journal of American Society for Information Science and Technology, 54*(6), 489-496.

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

Thomas, O., & Willet, P. (2000). Webometric analysis of departments of Librarianship and information science. *Journal of Information Science, 26*(6), 421-428.

Tomlinson, M. (2002). The Academic Robotics Community in the UK: Web based data construction and analysis of a distributed community of practice. *DRUID Working Papers*. Retrieved from http://www.druid.dk/wp/pdf_files/02-07.pdf

Trammell, K. D., & Britton, J. D. (2005). *Gatewatching: The impact of blog content on the mainstream media.* Paper presented at the Internet Research 6.0: Internet Generations, Chicago.

Tsai, D. F. C. (2005). Human embryonic stem cell research debates: a Confucian argument. *Journal of Medical Ethics, 31*(11), 635-640.

Uberti, E. (2004). *Trading flows and internet hyperlinks: A network analysis*.

Uren, V., Shum, S. B., Li, G., Domingue, J., & Motta, E. (2003). *Scholarly publishing and argument in hyperspace*, New York, NY, USA.

Valverde, S., & Solé, R. V. (2004). Internet's critical path horizon. *European Physics Journal B, 38*(2), 245-252.

Van Aelst, P., & Walgrave, S. (2002). New media, new movements? The role of the Internet in shaping the 'Anti-Globalization' movement. *Information, Communication & Society, 54*(4), 465-493.

Van Couvering, E. (2004). *New media? The political economy of Internet search engines.* Paper presented at the Annual Conference of the International Association of Media & Communications Researchers, Porto Alegre, Brazil.

van Leeuwen, T., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Van Raan, A. F. J. (2000). First evidence of serious language-bias in the use of citation analysis for the evaluation of national science systems. *Research Evaluation, 8*(2), 155-156.

van Leeuwen, T. N., Moed, H. F., & Reedijk, J. (1999). Critical comments on Institute for Scientific information impact factors: a sample of inorganic molecular chemistry journals. *Journal o f Information Science, 25*(6), 489-498.

van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Van Raan, A. F. J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics, 51*(1), 335-346.

Van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics, 47*(2), 347-362.

Van Raan, A. F. J. (2001). Bibliometrics and Internet: Some observations and expectations. *Scientometrics, 50*(1), 59-63.

Van Raan, A. F. J. (2005). For Your Citations Only? Hot Topics in Bibliometric Analysis. *Measurement: Interdisciplinary Research and Perspectives, 3*(1), 50-62.

Van Raan, A. F. J. (2005). Measurement of central aspects of scientific research: performance, interdisciplinarity, structure. *Measurement: Interdisciplinary Research and Perspectives, 3*(1), 1-19.

Van Raan, A. F. J. (2005). Statistical Properties of Bibliometric Indicators: Research Group Indicator Distributions and Correlations. *Journal of the American Society for Information Science and Technology*((pre-print)).

Vann, K., & Bowker, G. (2001). Instrumentalizing the truth of practice. *Social Epistemology, 15*(3), 247-262.

Vaughan, L. (2005). Exploring website features for business information. *Scientometrics, 61*(3), 467-477.

Vaughan, L. (to appear). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science & Technology*.

Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and impact factors. *ASLIB Proceedings, 54*(6), 356-361.

Vaughan, L., & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology, 54*(14), 1313-1322.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of American Society for Information Science and Technology, 54*(1), 29-38.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management, 40*(4), 693-707.

Vaughan, L., & Thelwall, M. (2005, to appear). A modeling approach to uncover hyperlink patterns: The case of Candian universities. *Information Processing and Management*.

Vaughan, L., & Wu, G. (2004). Links to commercial websites as a source of business information. *Scientometrics, 60*(3), 487-496.

Vazquez, A. (2001). Statistics of citation networks. *arXiv:cond-mat/0105031, 1*. Retrieved from http://arxiv.org/PS_cache/cond-mat/pdf/0105/0105031.pdf

Viegas, F. B. (2005). Bloggers' expectations of privacy and accountability: An initial survey. *Journal of Computer-Mediated Communication, 10*(3), article 12. Retrieved from http://jcmc.indiana.edu/vol10/issue3/viÃ©gas.html

Voyer, R. (1999). Thirty Years of Canadian Science Policy. *Science and Public Policy, 26*(4), 277-282.

Vreeland, R. C. (2000). Law libraries in hyperspace: A citation analysis of World Wide Web sites. *Law Library Journal, 92*(1), 9-25.

Wagner-Döbler, R. (1997). Self-organization of scientific specialization and diversification: a quantitative case study. *Social Studies of Science, 27*(1), 147-170.

Walker, G., Bloch, S., Hunt, G., & Fisher, K. (2003). Counting on citations: a flawed way to measure quality. *Medical Journal of Australia, 178*(6), 280.

Walsh, J. P., Kucker, S., Maloney, N., & Gabbay, S. (2000). Connecting minds: CMC and scientific work. *Journal of the American Society for Information Science, 51*, 1295-1305.

Wang, Y., & Kitsuregawa, M. (2002). *On combining link and contents information for Web page clustering*, London, UK.

Wang, Y., Wu, J., & Di, Z. (2004). Physics of Econophysics. *arXiv:cond-mat/0401025, 1*. Retrieved from http://arxiv.org/PS_cache/cond-mat/pdf/0401/0401025.pdf

Warhaft , Z. (2002). Turbulence in nature and in the laboratory. *PNAS, 99*, 2481-2486.

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Excercises. *Journal of Information Science, 26*, 453-460.

Warner, J. (2000). Research assessment and citation analysis. *The Scientist, 14*(21), 39.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

Wegrzyn-Wolska, K., & Szczepaniak, P. S. (2005). Classification of RSS-formatted documents using full text similarity measures *Lecture Notes in Computer Science, 3579*, 400-405.

Wei, C. P., & Lee, Y. H. (2004). Event detection from online news documents for supporting environmental scanning. *Decision Support Systems, 36*(4), 385-401.

Weigold, M. F. (2001). Communicating science: A review of the literature. *Science Communication, 23*(2), 164-193.

West, G. B., Woodruff, W. H., & Brown, J. H. (2002). Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *PNAS, 99*, 2473-2478.

Wheeldon, R., & Levene, M. (2003). The best trail algorithm for assisted navigation of Web sites, *1st Latin American Web Congress (LA-WEB 2003)* (pp. 166-179). Sanitago, Chile: IEEE Computer Society.

White, H. D. (2000). Toward ego-centered citation analysis. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 475-496). Medford, NJ: Information Today, Inc. ASIS Monograph Series.

White, H. D. (2001). Authors as citers over time. *Journal of American Society for Information Science, 52*(2), 87-108.

White, H. D. (2003). Author cocitation analysis and Pearson's r. *Journal of the American Society for Information Science, 54*(13), 1250-1259.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford: Oxford University Press.

Wiesenfeld, K. (2001). Scaling laws. *Am. J. Phys., 69*(9), 938-942.

Wikgren, M. (2001). Health discussions on the Internet: A study of knowledge communication through citations. *Library and Information Science Research, 23*(4), 305-318.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). 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), 49-56.

Wilkinson, D., Thelwall, M., & Li, X. (2003). Exploiting hyperlinks to study academic Web use. *Social Science Computer Review, 21*(3), 340.

Willinger, W., Govindan, R., Jamin, S., Paxson, V., & Shenker, S. (2002). Scaling phenomena in the Internet: Critically examining criticality. *PNAS, 99*, 2573-2580.

Wilson, C. S. (1999). Informetrics. *Annual Review of Information Science and Technology, 34*, 107-247.

Wolf, J. L., Squillante, M. S., Yu, P. S., Sethuraman, J., & Ozsen, L. (2002). *Optimal crawling strategies for Web search engines.* Paper presented at the The Eleventh International World Wide Web Conference, New York.

Wolpert, L. (2005). The Medawar Lecture 1998 - Is science dangerous? *Philosophical Transactions of The Royal Society B-Biological Sciences, 360*(1458), 1253-1258.

Wouters, P. (2004). *The Virtual Knowledge Studio for the Humanities and Social Sciences*: The Royal Netherlands Academy of Arts and Sciences.

Wouters, P., Beaulieu, A., Park, H., & Scharnhorst, A. (2002). *Knowledge production in the new digital networks. Research Programme*. Retrieved, from: http://www.niwi.knaw.nl/en/nerdi2/research_programme/new/toonplaatje

Wouters, P., & de Vries, R. (2004). Formally citing the Web. *Journal of the American Society for Information Science, 55*(14), 1250-1260.

Wright, E. (2003). The re-design of an integrated water and pollution management programme using the systems-ware model of the log frame. *Physics and Chemistry of the Earth, 28*(20-27), 973-984.

Xi, W., & Fox, E. A. (2001). *Machine Learning Approaches for Homepage Finding Tasks at TREC-10*. TREC 2001 online proceedings. Retrieved, from: http://trec.nist.gov/pubs/trec10/notebook_papers/VTexplainTREC10.pdf

Xue, G.-R., Zeng, H.-J., Chen, Z., Ma, W.-Y., Zhang, H.-J., & Lu, C.-J. (2003, July 28–August 1, 2003). *Implicit Link Analysis for Small Web Search.* Paper presented at the SIGIR'03, Toronto, Canada.

Yearley, S. (2001). Mapping and interpreting societal responses to genetically modified food and plants. *Social Studies of Science, 31*(1), 151-160.

Yook, S.-H., Jeong, H., & Barabasi, A.-L. (2002). Modeling the internet's large-Scale topology. *99*(21), 13382-13386.

Zach, L. (2005). When is "enough" enough? Modeling the information-seeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology, 56*(1), 23-35.

Zeng, Q. T., Kogan, S., Plovnick, R. M., Crowell, J., Lacroix, E.-M., & Greenes, R. A. (2004). Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics, 73*(1), 45-55.

Zhang, Y. (2001). Scholarly use of internet-based electronic resources. *Journal of the American Society for Information Science and Technology, 52*, 628-654.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2003). Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. *Scientometrics, 56*(2), 259–282.

Zuccala, A. (2006, to appear). Author cocitation analysis is to intellectual structure as web colink analysis is to...? *Journal of the American Society for Information Science & Technology*.

Acknowledgment

This work was supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015) (www.webindicators.org).