Fundamental methodologies and tools for the employment of webometric analyses

Thesis by

Liv Danman Fugl

The Royal School of Library & Information Science, Denmark - May, 2001

Fundamental methodologies and tools for the employment of webometric analyses

- a discussion and proposal for improving the foundation of webometrics

Thesis by

Liv Danman Fugl

OBU, 4th semester, class of 1999

The Royal School of Library and Information Science, Denmark

May - 2001

Abstract

The paper *Fundamental methodologies and tools for the employment of webometric analyses* defines the most important rules to keep in mind before performing webometric analyses. The paper deals with the two basic elements, that constitutes the foundation for webometric analyses: the documents being analysed, and the tools that are applied for the data collection. The concepts of a citation theory and a link theory are discussed through a study of the current litterature. Different methodologies for uncovering motivations for making references in scientific articles are reviewed and discussed. A methodology for uncovering motivations for making links on webpages is proposed and applied on six researchers' websites at the Royal School of Library and Information Science in Denmark, and on all the institutes at the same institution and at selected institutes at The Technical University of Denmark. The paper further contains a review on the linktopology of the Internet and the current status for the tools available for data collection. Finally, alternative possible tools for applying webometric analyses are proposed. The alternative tools are the Researchindex invented by Lawrence and Giles (Lawrence, Bollacker & Giles, 1999b; Giles, Bollacker & Lawrence, 1998), Kleinberg's HITS algorithm employed in the Clever search engine (The Clever Project, n.d.; Kleinberg, 1998), Proposals for possible extensions to the HTTP protocol to facilitate the collection and navigation of backlink information in the world wide web made by Chakrabarti, Gibson and McCurley (Chakrabarti, Gibson & McCurley, 1999c) and finally Link Agent, a program we have developed for this paper. The program makes it possible to uncover the reciprocal linking webpages, that exist in relation to the outgoing links from a chosen webpage.

Keywords: Informetrics, Webometrics, Citation theory, Link theory, Motivations for links, Motivations for references, Search engines, Webometric tools

Table of contents

Chapter 1

1.1 Introduction

In 1964, when the Science Citation Index (SCI) was launched by Eugene Garfield and the Institute of Scientific Information (ISI), it was done primarily to improve the possibilities for information retrieval (Garfield, E., 1998a, p.70). The idea was to make it possible to seek for further litterature about a certain subject, both backwards and forwards in time, starting with one relevant article using the lists of indexed litterature in SCI. One major advantage was the new way to search by using the references in the articles, instead of having to search on words from titles, keywords or subject headings.

Since that time, the SCI has been improved and expanded to include the well known citation databases SciSearch, Social SciSearch and the Arts & Humanities, and the use of the citation databases has also evolved in many new directions. Within the domain of informetrics many different quantitative methodologies have been established and explored. Methodologies that today have reached widely acceptance and recognition especially within research evaluation. Countries, research institutions and researchers are being compared according to their publicationrates, and how many citations they receive, and these are just a few examples of the results of existing informetric analyses.
The citation databases have made it technically possible for the informetric area to take major steps into new research and become a more established research domain.

Some of the more wellknown methods that have evolved from the citation databases are the Cocitation analyses, The Bibliographic coupling analyses and the Journal Impact Factor (JIF) calculations. In short: cocited documents are two or more documents who appear together in the same referencelist. Bibliographic coupled documents are documents who share one or more of the same references in their referencelists. JIF is a method invented and used by ISI to measure a journals impact based on the number of citations the articles published in a given year receive after a timespan of two years.

The JIF calculation and its later varities have been examined and explained in a very complete manner by Ingwersen and Hjortgaard Christensen in 1997 (Hjortgaard Christensen & Ingwersen, 1997).

The Cocitation analysis was invented independently by two different researchers, Small and Marshakova, back in 1973 (Marshakova, 1973; Small, 1973). Marshakova's idea was to invent a system, that was able to periodically adjust a classification system, when taking in consideration that clusters within a domain will change, new ones will appear, and old ones disappear. The clusters were to be uncovered by the cocitation analysis (Marshakova, 1973, p. 53-56). Small's original idea was based on the hypothesis, that if highly cited documents are representing the central ideas and methodologies within a domain, then cocitations can be used for a detailed mapping of the positions between the central subjects (Small, 1973, p. 265-266).

The cocitation methodology has been applied in many later projects. The Institute of Scientific Information has used the method for identifying new research fronts in the ISI-databases, and the method has further been refined to not only cover highly cited documents, but also applied to highly cited authors. The latter was done by White and McCain who mapped an extensive analysis of the domain of Information Science using multidimensional scaling (White & McCain, 1998).

Very close to the cocitation analysis is the method of Bibliographic coupling, which was also invented independently by two different researchers (Fano, 1956; Kessler, 1963). Fano was thinking of a library as a three-dimensional room with dots, where each dot resembles a document in the collection. All the dots are related to each other in different ways, which concludes in a creation of many overlapping clusters. The dots that belongs to a certain cluster are the documents within a specific subject. Fano was sure, that what connected the documents within one cluster of a subject was not common words from the abstract, title or the body of the text. What could instead unambiguously identify their similarity was the common reference in their referencelists (Fano, 1956).

Kesslers idea was somewhat very close to Fano's. In 1958 he had a vague hypothesis stating that the referencelist of a document contained some properties that could characterise the content of the document. In 1963 he had a more precise hypothesis: "A number of scientific papers bear a meaningful relation to each other (they are coupled) when they have one or more references in common" (Kessler, 1963, p. 49).

The advantages for both the cocitation method and the bibliographic coupling method are very clear. Suddenly a new set of tools were available for maintaining classification systems and keeping up to date with new research fronts, and the tools could even be applied without having the documents physically in the hand.

Even though the two methodologies seem very much alike, it is important to note, that some major differences do exist. The mapping of a scientific domain using the cocitation analysis will show the domain as it is interpreted by the researchers citing the highly cited documents or authors, and the clustering of subdomains will be based on how allready known knowledge is used in new dimensions. The mapping of a scientific domain using the bibliographic coupling analysis will show the domain as it is interpreted by the researchers writing the new knowledge, and it is here their own interpretation of their position in the scientific domain that will be shown on the map (Balslev & Fugl, 1999).

Since the start of the worldwide access to the Internet around 1994/1995, millions of websites, all of many different types (e.g. private sites, cooperate sites, institutional sites), have been published, and numerous search engines have tried to solve the problem of indexing and structuring the websites. Only a few of these search engines offer the more advanced possibility of using Boolean searching, and even fewer offer specific search

operators as e.g. link:, domain: and host: etc. Operators that are very similar to the CA:, CW:, CY: etc. in the ISI citation databases.

This new opportunity, for new ways of making informetric analyses and develop new laws and methods that could apply to the documents on the Internet, has of course led many researchers within the informetrics to make varied forms of analyses and uncovering networks between countries, between scientists etc. Even a new journal on the subject appeared online, only to exist on the Internet, Cybermetrics. The first article was treating the problems of Lotkas distribution, sitations[1] and self-sitations - concepts that all previously have been treated in the classical citationdatabases (Rousseau, 1997).

One of the first persons to practice the informetric methods on the Internet was Larson, who created a cocitation map of the Earth Sciences using multidimensional scaling, based on results of searches performed in the AltaVista's advanced search, using the query "link:page A AND link:page B" (Larson, 1996).

Another important pioneer within the application of informetric methods to the Internet was Crone Almind, who back in 1997 in his master thesis tried to set up some proximity measures to be applied on researchers websites. The proximity measures were based on a combination of the frequency for bibliographic coupling, cocitations and reciprocal linking (Almind, 1997). Later that same year Crone Almind and Ingwersen published an article arguing for the informetric methods to be applied on the Internet – a concept they proposed to be named Webometrics (Almind & Ingwersen, 1997).

In 1998 Ingwersen proposed a calculation of the Web Impact Factors (Web-IF) a measure that quantitatively can give an indication of the relative attractiveness of countries or research sites on the WWW at a given point in time. The Web-IF calculation is based on the same concepts as the allready wellknown JIF-calculation (Ingwersen, 1998).

When performing informetric and webometric analyses, it is of high importance to be aware of the lacks and limitations in the databases that are being used (e.g. the ISI databases or the selected Websearch engines). It is a well known fact that the ISI databases have an anglo-american bias, and that any referenced author or document may appear in many different forms and therefore does not have one unambiguous form in the citationindex, which to some degree complicates the citation analyses.

An important question to put forward in this discussion is also the question of the reasons and motivations for making citations when writing a scientific article. Is it at all possible and correct to make citation analyses based on quantified amounts of references, when we do not know the exact motivations behind the references for the articles? This is a fundamental discussion of the existence of a normative theory of citing. The major

[1] Sitation is an expression used by McKiernan (1996) and later Rousseau (1997). The concept sitation is equal to citations in scientific documents, internal and external webpages that link to a specific webpage.

concerns and problems when doing citation analyses have been outlined in two articles by MacRoberts and MacRoberts (1989; 1996).

When applying the informetric methodologies to the Internet, it is of high importance, keeping the critics of the citation analyses in mind, to be aware of similar and new problems. We find that the trustworthiness to the web search engines is weak, and very much affected by commercial interests, who only seem to concentrate on one thing; searching for the perfect algorithm, matching the perfect user's perfectly well defined need of information.

At the current moment we cannot trust the results from the search engines, because we have only little or no insight in the rules of their indexing and the performance of their web spiders[2], resulting in a biased searchindex favoring some types of sites over others. As one person has stated: "At the current time, the quality and the reliability of most of the available search tools are not satisfactory, thus informetric analyses of the Web mainly serve as demonstrations of the applicablity of informetric methods to this medium, and not as a means for obtaining definite conclusions" (Bar-Ilan, 2001).

This paper takes its starting point within the informetric domain focusing especially on the methodologies and data collection tools for applying webometric studies.

Informetrics is one of the core subdomains of Information Science (Ingwersen, 1995, p. 147) and webometrics are informetric studies applied on the Internet, but as have been noted by Björneborn and Ingwersen: "there is a strong element of re-engineering and clean-up in webometric analyses" (Björneborn & Ingwersen, 2001, p. 66), and that is were this paper takes it point of departure.

1.2 Main research questions for the paper

When making informetric and webometric analyses we find it to be of fundamental importance that some specific aspects are kept in mind in order to state clearly, what has been examined, and what has not? Where are the limitations and boundaries for these types of analyses? and can we at all make analyses on quantified citations and links?

In this paper we will try to define the most important rules to keep in mind before performing webometric analyses. The following questions will be the main subjects for investigation in this paper:

[2] A web spider is a program that crawls on web sites extracting information for search engine databases.

- Is it possible to speak of a citation theory and a link theory?

- What is the difference between making a reference in a scientific article and making a link on a website?
  Which methods are the most reliable to discover the motivations behind making a reference in a scientific article and making a hyperlink on a website?
  What information can we extract about a website by looking at the links, that are on it?
  What are the causes that currently exist and block the way for links to be a more common, established and formalized way of expressing the networks, that surrounds the website?

- How do the current data collection tools for making webometric studies perform?
  What other tools could be useful for these kind of analyses?

1.3 The structure of the paper

The first chapter gives an introduction to the paper and outlines the main research questions and the methodology. Next, there will be a study in chapter two of the existing literature covering the discussion about the concepts of a citation theory and a link theory.

Chapter three holds an analysis on the motivations for making citations and links, their dissimiliarities and similarities, and gives a discussion on the methodologically best way to investigate these motivations.

The fourth chapter gives a state of the art on current data collection tools for making webometric analyses, and it further contains proposals for alternative tools.

Chapter five rounds off the paper with a conclusion and some main statements and proposals for future research, in order to improve the performance and reliability of webometric analyses.

1.4 Some important definitions

This section will give a definition of the most important expressions and concepts that are being used in the paper.

Citation / Reference

It is important to make a distinction between the terms reference and citation. They are quite often being used as equal, even though they denote two different concepts. Figure 1 reflects the difference between the two concepts:

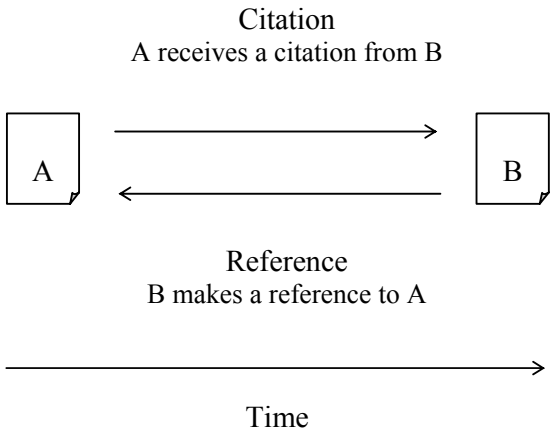

Figure 1: Definition of reference and citation

It is shown that the concept reference is used as a retrospective term, while the concept citation is a forwardlooking term (Egghe & Rousseau, 1990, p. 204). In other words, a reference is the recognition you give another document, while a citation is a recognition you receive from another document. The referencelist at the end of a document contains references to other documents who receive a citation.

Similar variations of the concepts are cited and citing. The figure shows that document A is a cited document, while document B is a citing document.

Incoming link
Incoming links are also known as backward links or ingoing links. The concept covers all links from other servers on other websites pointing to a certain webpage or website. An incoming link is similar to receiving a citation in a document.

Informetrics
"Informetrics is the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists" (Tague-Sutcliffe, 1992, p. 1).

Internal link
Internal links are links that are all pointing to other webpages or nodes within the same website. E.g. a link between the two webpages ix.db.dk and www.db.dk would be considered as an internal link even though they don't belong to the same server, but they do belong to the same institution (db). Internal links are mostly for navigational use within the website.

Link
A relationship between two webpages or nodes. The concept link does not distinguish between giving and receiving a link in the same way as with the concepts citation and

reference. Instead we use the concepts outgoing link which is similar to giving a reference, while an incoming link is similar to receiving a citation.

Node
A unit of information. Node is a concept frequently used in the hypertext world, but has the same meaning as the word 'document' or 'webpage'. In this paper only the words document and webpage are being used.

Outgoing link
Outgoing links are links that are all pointing out to other servers on the Internet, servers that are not physically located on the same website. Outgoing links are also known as external links. An outgoing link is similar to giving a reference in a document.

Reciprocal link
If two webpages or two websites both have a link pointing to each other, we define the link as a reciprocal link. The concept will be further enlarged in chapter three.

Webometrics
Webometrics is the study of quantitative aspects of webpages or nodes.

Webpage
A unit of information, demarcated by one unambiguous URL (webaddress). A webpage is also known as a node.

Website
A collection of one or more webpages or nodes affiliated to the same institution, organisation or the like.

1.5 Methodology and delimitations for the paper

The purpose of this paper, is to provide an indepth analysis of the basic elements, that constitutes the informetric and webometric analyses, expose and discuss their weaknesses, and to bring light on important principles and possibilities, that are essential for the domain to start working on. These principles and suggestions for further steps are important, in order to strengthen the validity and reliability in the analyses, that are being produced for the surrounding world to see and make decisions on.

The basic elements for informetric and webometric analyses fall into two parts. The documents and their references or links respectively and the databases for making the analyses.

A third important element in these types of analyses are the methodologies that are being developed and used on the documents and databases. E.g. the cocitation and bibliographic coupling methods or the Web-IF calculation. This element will not be made an object for

discussion in this paper, since it is not considered to be an element in its original form, but rather an element that is applied. It is however important to be critical and sceptical about any kind of these applied methodologies, but that is not the purpose for this paper.

The first two chapters concern the element that covers the documents and citations that have been used for the analyses, while chapter four concentrates on the element that concerns the databases e.g. search engines and other possible webometric tools.

1.5.1 The question of a citation theory and a link theory

In chapter two, there is a study of the existing literature that focus on the discussion whether it is at all possible to speak of a citation theory. The major arguments in this discussion have been outlined, in order to show the pros and cons that have previously been said on the subject. Further, there is a discussion on the possible existence of a link theory.

1.5.2 Motivations for making references and links

The third chapter gives an outline of the motivations for making citations in scientific articles. The outline is based on various studies previously made. These studies all have different methodological ways for uncovering the motivations for citations, and the chapter therefore also includes a discussion on the pros and cons for the different methodologies.

Further the chapter includes a similar discussion on the problems for uncovering motivations for making hyperlinks on websites, and an outline of a methodology for solving the problem will be proposed.

It is emphasized that the chapter will only have its focus on those citations and links that are not either self-citations or self-links. That is, citations going to other authors or documents than the author or the document itself, and links, going to other websites than the website itself.

1.5.2.1 An exploratory study

An exploratory study has been designed and implemented as a way for testing the idea of a methodology for uncovering the motivations for making hyperlinks.

The exploratory study is chosen because the uncovering of motivations for making hyperlinks is still an unexplored area of research. It is therefore essential, first of all to identify possible problems and hypotheses, that can be made an object for investigation in future studies (Andersen, 1998, p. 25).

It is important to keep in mind, that when conducting an exploratory study, it is based on only small samples of the population combined with litterature on the topic (Hellevik,

1997, p. 77). The small samples make it impossible to perform valid analyses that could be applied for the whole population of websites. The results can only be concluded for the sample websites, that have been examined.

Another argument for using the exploratory study, is based on the nature of the Internet. Websites are of so many different natures and types, so taking a sample that could be applicable for all websites, would be almost impossible.

We are of the opinion, that motivations for making hyperlinks varies greatly depending on the type of website being inspected e.g. personal website, business website or institutional website. It was therefore natural, to select a few, but weldefined types of websites in this study.

We have made a selection of two specific types of websites: websites of selected institutes at respectively The Royal School of Library and Information Science in Denmark and The Technical University of Denmark, and personal websites of researchers at The Royal School of Library and Information Science in Denmark. The personal websites were all professional websites located at the webservers of the institutions. A few of the researchers also had a more personal website. These would usually go back to a time before the appearance of a demand from the institution to have a personal website, and usually, these websites had not been updated for a longer period of time. The websites and their links were though included in the discussion during the interview. The selected websites that are included in the study, serve as examples for the application of the proposed methodology.

1.5.2.2 The purpose of the study

The use of this specific study had various objectives. Besides from demonstrating the proposed methodology, the aim was also to harvest some of the very first results and indications for further studies about motivations for making hyperlinks, based on the outcome of examining the chosen types of websites.

The demonstration of the proposed methodology was an example of which questions, we thought were of importance to investigate in advance, if we had decided to conduct a webometric study, that was aimed to uncover the scientific network existing for the chosen researchers, and to uncover the scientific network existing for the chosen institutes on the scientific institutions.

1.5.2.3 An outline and description of the study

Since questions concerning motivations for making hyperlinks is a matter of going back in time and recalling senses and thoughts from the time of the production of the website, this study has to be a retrospective study. If we should capture the thoughts when they were still fresh in mind, we would have to make a study asking owners of websites that were only a few days old. Since this would have demanded quite an investigation and time

demanding work, this has not been possible. Instead we put a limit for the time period of the last update of the websites to be no later back than August 1st, 2000.

All questions that were asked in the study were primarily open-ended (it was emphasized that each answer was enlarged and reasoned), which was chosen because of the exploratory nature of the study. It was important to collect as many different types of data as possible, giving a possibility for the respondents to bring up new important aspects and dimensions letting their thoughts drift away, instead of limiting them to only a few predefined set of answers, and thereby miss valuable data that could later turn out to be important research issues. This method is also supported by Frankfort-Nachmias and Nachmias who states: "The virtue of the open-ended question is that it does not force the respondent to adapt to preconceived answers" (Frankfort-Nachmias & Nachmias, 1996, p. 254).

Two different variants of the proposed methodology have been used. One was the use of personal interviews with the researchers about their personal websites, while the other was the use of questionnaires sent to the selected institutes at the two institutions. The choice of using two different ways of collecting data (personal interview and questionnaire), was based on two reasons. One was to test if the quality of the answers would vary a lot, since all the participants were presented to mainly the same questions. The other reason was mainly a lack of time, forcing the study to conduct only a very limited amount of interviews before this paper was due to be turned in.

All questions were personalized by taking their starting point in the links on the researcher's or the institute's website, but as a principal rule, all the questions in the interviews and in the questionnaires were based on the same templet[3], and looked very much alike.

The interviews all took place at The Royal School of Library and Information Science, and for all cases except one, in the researcher's own office. The interviews lasted between 30-60 minutes. The respondents were presented to a copy of the questions as they had been aimed towards their own website. The data was collected by notetaking on the questionnaire by the interviewer during the interview. The data was later typewritten, but due to a lack of time, the respondents were not asked to give a feedback on the results. The individual data from the interviews can be acquired upon request to the author.

The investigation of the different websites and the collection of the empirical data was performed during a timeperiod of four weeks in the spring 2001. Within this timeperiod none of the websites changed their apperance or content.

[3] The templet for the questionnaires is available in appendix A. An elaboration and justification of the questions has been done in chapter 3.

1.5.2.4 Criteria for selected websites

The interviewed researchers at The Royal School of Library and Information Science were selected on the following criteria:

-They should have at least one outgoing link on their website.
-They should be the main responsible person for maintaining the content on the website.

Six researchers were asked to participate, and they all gave their consent.

The institutes that received the questionnaires by e-mail were selected on different criteria. The reason for choosing institutes at two different institutions was in order to see, if there would be some kind of overlap or dissimilarities in their answers.

For all the institutes, the questionnaire was sent to the person in charge of the website, that would either be the webmaster[4] as stated on their website or the head of the institute.

The questionnaire was sent to all three institutes that exist at The Royal School of Library and Information Science without no further criteria, while the selection of the four Institutes at The Technical University of Denmark were based on the following criteria:

-Their last update should be no later back than August 1st , 2000.
-They should have at least one outgoing link on their website.

A lot of the websites at the institutes at The Technical University had not been updated since 1999, and others only concentrated on showing aspects of the institute itself, and did not have any outgoing links.

Four institute websites at The Technical University met the criteria outlined above, and two of them responded on the questionnaire. The first institute had a webpage with a list of very broad types of links. These links were located within the mainpages of the website. Unfortunately the more scientific links to e.g. other scientific institutions or partners of cooperation were located within websites for the different projects they were involved with, or the more specialized subsites of the institutes. These webpages were not discovered during the examination - especially due to the difference in the address of the url. The collected data for this particular institute therefore gives a bias for making any valid conclusions.

Only one of the institutes at the Royal School of Library and Information Science responded on the questionnaire. A second institute did however give a reply and a valid excuse for not responding.

[4] The webmaster did in all cases also serve as a researcher at the particular institute.

1.5.2.5 Analysis and interpretation of the empirical data

The empirical data that has been collected was analyzed and interpreted. We were primarily looking for possible tendencies that could give an indication on the motivations for making hyperlinks, and reasons for not making them.

The results of the motivations for making hyperlinks were compared to some of the results for motivations for making references as they have been examined by different researchers. The results could perhaps also give us ideas on where to continue further research.

Further an interpretation of which sociological aspects we can conclude on in webometric studies for the selected websites is discussed.

The study is also evaluated in regards to the proposed methodology. Focus is set on the aspects of the study that worked as intended, and which ones that didn't.

The main results from the study are described in chapter three.

1.5.3 The search engines and other webometric tools

The fourth chapter starts with an outline of the current status on the knowledge of the size and topology (based on hyperlinks) of the Internet.

Further, the chapter contains a critical outline of the demands we need to state for the search engines we use for the webometric analyses, in order to perform valid and reliable results. The need for these demands will be demonstrated by a review on the current knowledge about using the major search engines for webometric analyses.

The chapter is rounded off with a review on other types of data collection tools. Tools that could make up for good alternatives for making webometric analyses besides the ordinary search engines. The alternative tools are the Researchindex invented by Lawrence and Giles (Lawrence, Bollacker & Giles, 1999b; Giles, Bollacker & Lawrence, 1998), The Clever search engine made by members of the Clever Project (The Clever Project, n.d.; Kleinberg, 1998), Proposals for possible extensions to the HTTP protocol to facilitate the collection and navigation of backlink information in the world wide web made by Chakrabarti, Gibson and McCurley (Chakrabarti, Gibson & McCurley, 1999c) and finally Link Agent, a program we have developed for this paper. The latter program makes it possible to uncover the reciprocal linking webpages that exist in relation to the outgoing links from a chosen webpage.

Chapter 2 - A citation theory and a link theory ?

This chapter focus on the discussion, whether it is at all possible to speak of a citation theory and a link theory. The major arguments in this discussion are outlined based on the current literature, in order to show the pros and cons that have previously been said on the subject.

This discussion is important, because informetrics and webometrics do highly rely on a foundation, that is based on the legality of quantifying citations and links. Quantifications that are used for making analyses and conclusions about the scientific research production, performance measures, visibility and sociological networks among researchers etc. These conclusions can very well be called into question, if we fail to proof this legality.

Small has indicated that two theories of citations exists side by side. The normative theory of citations and the social constructivist theory of citations. The normative theory meaning scientists cite to give credit where credit is due, and to cite the best sources for their purposes, while the social construction of citations indicates that scientists cite to gain political advantage, advance their interests, defend their claims against attack, and convince others (Small, 1998, p. 143).

As we see it, this discussion is all about whether researchers make citations from a neutral perspective or whether their personal interest constitutes the greatest portions of their citations.

2.1 Discussion of a citation theory

The normative theory of citations dates back to Kaplan, who stated that scientists are constrained by norms to give credit where credit is due (Kaplan, 1965). Another eager advocate for the normative theory was Merton, in spite of his background as a sociologist. In a foreword to a book written by Garfield, Merton claims that citations are a recognition of intellectual debts and original research findings. Scientists can only lay claim to their works by publishing it to the public domain of science, and let peers make references to the works and thereby recognize and give credit to the author. Merton describes it as 'the reward system of science', where citations and references operate within a jointly cognitive and moral framework: "In their cognitive aspect, they are designed to provide the historical lineage of knowledge and to guide readers of new work to sources they may want to check or draw upon for themselves. In their moral aspect, they are designed to repay intellectual debts in the only form in which this can be done: through open acknowledgment of them" (Merton, 1979, p. viii).

Even though Merton clearly speaks in favour of the normative theory of citations, he is also known as the father of the expression Obliteration by incorporation (OBI). "'Obliteration by incorporation': the obliteration of the source of ideas, methods, or findings by their incorporation in currently accepted knowledge. In the course of this

hypothesized process, the number of explicit references to the original work declines in the papers and books making use of it. Users and consequently trasmitters of that knowledge are so thoroughly familiar with its origins that they assume this to be true of their readers as well. Preferring not to insult their readers' knowledgeability, they no longer refer to the original source. And since many of us tend to attribute a significant idea or formulation to the author who introduced us to it, the altogether innocent transmitter sometimes becomes identified as the originator…..to the extent that such obliteration does occur - itself an empirical question that is only beginning to be examined - explicit citations may not adequately reflect the lineage of scientific work. As intellectual influence becomes deeper, it becomes less readily visible." (Merton, 1979, p. ix).

Three things are talking against the concept 'Obliteration by Incorporation'. First, it is not a 'rule' that happens to be applied for all highly cited documents, and second, Merton states himself that OBI has not yet been proved, and third, the obliteration will only take place after substantial visibility through citations has occured (Merton, 1979, p. ix-x).

Latour made a clear break with the Mertonian tradition, when he elaborated the formulation on the rhetorical function of citations (Luukkonen, 1997, p. 28). According to Latour, references in articles are some of the means that the author's can use in their effort at trying to "make their point firm" and to support their own knowledge claims (Latour, 1987, p. 36, 38). He states, that references have a major function in scientific texts: that of mobilising allies in the defence of knowledge claims.

Cozzens multidimensional approach to the problem of the citation theory, is a combination of the Mertonian approach (citations are given as a reward) and the approach of Latour (citations as persuasive for the stated knowledge claims). Cozzens argues that we should think of citations first as rhetoric (citations as persuasive of own statements) and second as reward (give credit for achievements). The two concepts having each a purpose. Citations made within the rethorical system are first and foremost a portion of a power-seeking text, while citations made as rewards are just a result of a citation etiquette. Cozzens argues further: "If the rhetorical standards are violated, the paper may either not review well, or be ignored; if citation etiquette is violated, objections may be raised" (Cozzens, 1989).

Cozzens viewpoint indicates, that researchers citation behavior is primarily being ruled by personal motivations for reaching power and acknowledgement within the domain. An argument that indicates Cozzens primarily agrees with the Social constructivist theory of citations, while the normative theory of citations is being ranked second.

A small study performed by Zuckerman, shows the direct opposite facts. Zuckerman argues, that if persuasion by authority really was the major motivation to cite, then a large share of all citations should go to such authoritative papers. Based on a table showing the numbers of citations to articles cited between 1975 and 1979 in the Cumulated Science Citation Index, she refutes this statement. If setting a minimum level for an authoritative paper to have received a minimum of 10 citations within the five year period, the table

shows that only 6% of the citations went to such authoritative papers, while about 64% of alle the papers were only cited once (Zuckerman, 1987, p. 333-334).

MacRoberts and MacRoberts have also been eagerly arguing against the quantification of citations, using the results as performance measures. They claim their studies have shown that not all influences by other works are present in the reference lists, e.g. this is especially evident when authors do not make references to common knowledge that has become tacit within the researchers knowledge (OBI), or when researchers are making references within the body of the text and not in the bibliography, or making references to secondary sources like review articles etc. (MacRoberts & MacRoberts, 1986, 1989, 1996). Van Raan is of the direct opposite opinion, stating that the validity of citation analysis is not affected, even though the researchers never cite all the work they used for their research: "Scientists have, like everybody in this world, to make choices, also in their reference lists. They will immediately admit that they have been influenced and stimulated by their parents, teachers, by ideas of scholars in quite other fields of science, by the 'climate' in a research institute, and even by many publications in their own field, and yet not referring to these influences. They'll focus their citations mainly to 'research front work' which is nicely and statistically sufficiently demonstrated by advanced co-citation analysis" (Van Raan, 1998, p. 135).

Van Raan gives critism to the sociological theorising on citations, stating that they focus too much on the role of the citing author and his/her references, instead of focusing on the cited author and the citations that he/she receives. "In a reference analysis (the 'citing side') we have one citer and different cited papers 'per unit of analysis'. However, in citation analysis as used for reseearch performance analysis, there are many citers and just one cited paper" (Van Raan, 1998, p. 136).

To some extent we do agree with Van Raan. It is true, that we should have more focus and examinations on why papers are being cited, but even if this would be the sole problem of investigation, we cannot evade ourself from arguing for the existence of a normative theory of citations. If we do not know the arguments for the existence of a normative theory of citations, and thereby cannot prove its justification, we cannot talk about patterns of motivations for making citations, which leads us to the fact, that we cannot conclude that certain patterns for the cited papers exists. If we were examining a highly cited paper, and asking the citing researchers about the motivations for citing this one paper, without the knowledge of the pattern created by the motivations for citations, that are the basis for the normative theory of citations, we could not use the results for anything, since we cannot be sure, if the high citations for an article is always equal to a well written article, or if perhaps some highly cited articles have high citations, due to criticism of their works. We need to know much more about the individual motivations for making references, and how great a prevalence they each have, before we can start analysing the cited papers, instead of the citing papers.

Blaise Cronin has performed a study, making different experts go through a number of articles from their own domain, where all references had been removed, and indicate with a note each time they thought a reference would be appropriate and to whom. His conclusion showed: "a broad agreement among various groups, though, as one would expect, wide variation at the individual level…My general feeling was that experts in a given field have a tacit understanding as to what constitutes acceptable/required citation behaviour in that field…[The results] suggest that there may be a "norm" of citation behavior" (extract from letter to Garfield, (Garfield, 1989, p. 125)).

Bibliographic coupling is another example brought forth by Garfield as to be proof of an existing norm of citation behavior. He states: "that two papers on the "same" topic rarely cite the identical list of articles…Perhaps one author cites 5 or 10 papers that another does not. Each, however, may cite about the same 50 percent of the references. More than likely an even higher percentage of the core papers or books in the field will be co-cited" (Garfield, 1989, p. 126).

Searching related articles using bibliographic coupling is only possible, if the normative citation behavior exists.

Another argument that entitles the existence of a normative theory of citations, is due to the fact, that scientific documents is a homogeneous type of document, that has existed and evolved through many decades to its existing form. During the years, they have evolved into a firm and structurized shape, that has its own rules for a certain structure, techniques for arguing and norms of making citations. It is evident though, that differences on these rules and norms exists within different scientific domains, especially between the natural sciences and the humanistic sciences. Researchers within the latter domain are primarily communicating to peers through monographs, while the natural sciences are communicating through journal articles. Even within subdomains (e.g. physics and biology), different cultures and norms for writing (and hereby also citing) exists, differences that make it extreemly important that we do not perform informetric analyses comparing different domains etc. This has also been expressed as a serious concern by the founder of the citation indexes Eugene Garfield, in his arguing for the existence of a normative citation theory: "A theory of citation might include a set of commandments of citation analysis. Another commandment that pertains to the evaluation of people, journals, and institutions - always compare or judge equivalent or truly comparable cohorts. Naive administrators, uninformed in citation analysis, will make the mistake of using citation data without regard to the discipline or invisible college involved. Cross-disciplinary comparisons are usually inappropriate. Even in large disciplines, it can be difficult to establish perfect cohort groups of authors or journals" (Garfield, 1998b, p. 73).

This statement very clearly disproves the assumptions made by MacRoberts and MacRoberts in 1996, who said: "Today, in spite of an overwhelming body of evidence to the contrary, citation analysts continue to accept the traditional view of science as a

privileged enterprise free of cultural bias and self-interest and accordingly continue to treat citations as if they were cultre free measures" (MacRoberts & MacRoberts, 1996, p. 442).

Citation analysts are on the contrary extremely aware of the data they are dealing with, and the possible conclusions that they can and cannot infer.

The peer review process that is applied to all scientific publications before they are being published, is the major reason, why we can talk of an existing culture and a normative theory of citations. It is the peer reviewing that makes us able to rely on the content of the document. We can trust that the results are based on a methodologically correct procedure, but we can also be sure, that the norms of citations have been observed. If they had not been observed by the authors, it would be the job of the referees to draw the attention on the problem. This has also been stated by Garfield: "Over 30 years ago, I pointed out that it was the job of patent examiners to refresh the memories of inventors. I can't recall how often I've said the referee's job is similar - to remind authors when they overlook or perhaps deliberately omit relevant references" (Garfield, 1989, p. 123).

It is the presence of a normative theory of citations and the peer review process, that makes it possible for the informetric domain to perform quantitative citation analyses that are based on a reliable foundation.

2.2 Discussion of a link theory

As it is important not to make any comparisons of quantitative measures performed within different scientific domains, the same rule should be applied when it comes to websites on the Internet. The comprehensive number of types of websites that exist, very clearly indicates that we should be careful about comparing any of these websites. At the current moment, websites have not yet developed a common culture and norms for linking to other websites, and furthermore we don't have a peer review process, to assure that the norms are being observed. Some of the people who publish websites, have not been 'raised' within a scientific domain and learned the rules for how to write and publish, in the same way as researchers have been 'raised' to know how to communicate and substantiate new knowledge and ideas within their research domain.

At the moment, motivations for linking are being influenced by many different cultures. It may even be possible, that differences exists within different countries, or within different types of websites e.g. business websites, personal websites or institutional websites. It is still too early to say, due to the young age of the Internet.

2.3 Discussion and main conclusions on citation theory and link theory

We are of the opinion, that to a certain degree, motivations for making citations entitles both theories, the normative and the social construction of citations, to exist side by side. The question of importance for the informetric research is to what degree the normative

theory of making citations exists. Which of the two theories that have the greatest influence on the scientist when he is writing articles, and to what degree the referees will accept a lack of the normative way of making references. Since we have no explicit stated rules on when to cite, it is not possible to state an exact number or measure for fulfilling the normative rules of making references. Whether the 'rules' have been fulfilled is solely based on a personal judgement, but a judgement that is independently made by at least two persons (one author and one referee) working within the domain. As long as we can trust the persons who perform the peer review process, we can trust, that a normative theory of citation does exist.

Imagine a map of a domain, perhaps a multidimensional scaled map based on cocitations or bibliographic couplings. It is the obligation of the referees to assure, that the researcher who is writing an article on a certain subject, is also making references to colleagues within the same area on the map.

The normative theory of citation has evolved within the scientific domains and across borders of institutions and countries. This goes especially for the natural sciences and less for the humanistic sciences. We would say, the more a domain or subdomain is globally homogeneous, the more the domain shares a common normative theory of science not limited by borders of local institutions or countries.

This fact of a normative theory, for those domains being globally homogeneous, is what makes it possible to perform informetric analyses on a macro-level, e.g. comparing citation rates across borders on authors, institutions and countries.

Similar, the norms for making references may differ from the micro level to the macro level. This has also been noted by Leydesdorff (1998, p. 16-17) and Brooks (1985, p. 227).

When comparing references and links, we do not know if we can speak of a normative theory for using the latter, but we can be certain that there is a lack of peers to control the observance of the norms - if they do exist. As it is important to discuss the theory of citations, it is even more urgent to discuss the possibility of the existence of a link theory.

Due to the variety of types of websites, and due to the still very young age of the Internet, we believe, that a normative theory of linking has not yet had the possibility to evolve and grow to a steady level. Whether it will be mostly influenced by cultural factors dominant within the country the website is being produced in, or whether we will see different norms occur across borders within different types of websites e.g. business websites, personal websites, organizational websites etc., the current lack of a normative theory of linking erodes the foundation for webometric analyses on a macro-level.

What could instead be possible for webometric analyses at this time are analyses performed on a micro-level. Analyses that are based on similar webpages within a few selected websites, in order to keep the variables as small as possible. The micro-level

analyses would always have to include qualitative examinations of the selected websites. Examinations that would give us a possibility to investigate and clarify which types of motivations for linking that are present on the website, and thereby indicate the possible types of conclusions available. An example could be, that one wishes to discover networks between researchers on two selected institutions. A lack of links between them doesn't necessarily mean that they are not networking, The answer may be just as simple, as a lack of motivation for showing the network on their websites.

In order to perform webometric analyses on a macro-level, two things would be required:
1. A normative theory of linking that is globally homogeneous. E.g. if it turns out that we can verify that a normative theory does exist within types of websites, as it is similar with scientific domains.
2. A control process similar to the peerreview process.

Since the last point seems rather utopian, other ideas need to be suggested. One way possible is, that the webometric analysis would need to include a qualitative examination of the type of websites that are included in the analysis. An examination that could assure, that the norms for linking are being observed - the examination could perhaps be based on a test sample.

This chapter has clarified, that making informetric analyses on a macro-level can be done due to the existence of a globally normative theory of citations within especially the scientific domains, and a peerreview process to control that the norms are being observed. A possible similar normative theory of links has still not grown to a steady level, which makes it impossible to perform webometric analyses on a macrolevel. Instead, it is suggested, that we keep the webometric analyses on a micro-level and that they should always be followed by a qualitative analysis, indicating the motivations for linking on the selected websites.

It is still too early to say, if it will be possible to perform webometric analyses on a macro-level, this would depend on a globally applicable normative theory to evolve, and a control process similar to the peer review process.

Chapter 3 - Motivations for citations and links

The chapter gives an outline of the motivations for making citations in scientific articles. The outline is based on various studies previously made. These studies all have different methodological ways for uncovering the motivations for citations, and the chapter therefore also includes a discussion on the pros and cons for the different methodologies.

Further the chapter includes a similar discussion on the problem for uncovering motivations for making hyperlinks on websites, and an outline of a methodology for solving the problem will be proposed.

Whether citations are given as a credit where credit is due (and this goes for both positive and negative references) or to persuade for ones knowledge claims, the important thing that matters, is that we can be sure that to some extent, the unwritten norms for citing are being followed, creating a certain pattern behind the motivations for making references. A pattern which gives the informetric domain the possibility to make analyses and draw conclusions upon the performances within the scientific domains.

Most researchers have been 'raised' in the specific pattern of how to communicate research results in journal articles, monographies etc. within their domain. They instinctively know how to construct the right layout of their documents, and they know by 'instinct' when to make a reference when they are writing. Therefore, it is a natural reflex for a scientist to react when reading articles, that are not following this pattern. This happened to Kidd, who has written an article on references, an article that was originally initiated, because a thematic set of review articles in a major journal had an unusual pattern of referencing (Kidd, 1990).

The first part of this chapter will have its focus on discovering parts of this pattern of citations, and how we methodologically best can examine the problem.

3.1 Why do we make citations?

3.1.1 Methodological approaches for uncovering motivations for references

When trying to uncover motivations for making hyperlinks, it is important that we do it the correct way and avoid certain biases, that can influence the results. The following reviews the most important pitfalls, we should be aware of.

3.1.1.1 The angle of analysis

As it was briefly discussed in chapter two, it is important that the test sample, that is being used for analysis of the motivations for making references, covers all possible types of motivations. This is important in order to obtain a realistic pattern of the normative theory to be uncovered, and to achieve a realistic sense of the size of statistical distribution on

each type of motivation. This is not possible if we focus our analyses on the motivations from the perspective of the cited documents. We have drawn two figures to indicate the meaning.

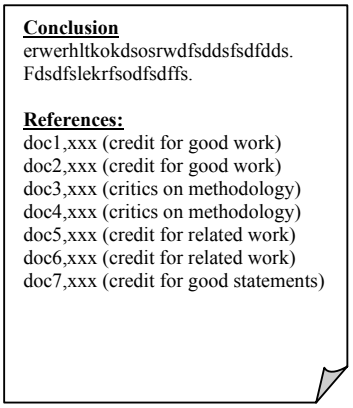

Figure 2: Example of motives for making references

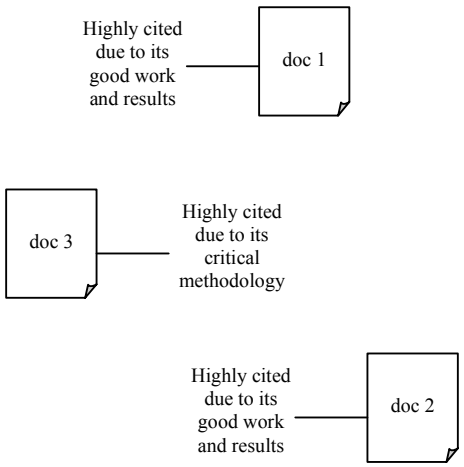

Figure 3: Example of motives for being cited

Informetric analyses are mostly based on quantified analyses of e.g. highly cited documents or authors. When uncovering the motivations and norms for making references, it is of importance to base the test sample on all possible references (motivations) as shown in figure 2, and not only on citations to selected documents, as shown in figure 3, as it can give a serious bias in the end results, giving an overweigth of some types of motivations or missing other ones.

3.1.1.2 The reliability and dispersion of the test sample population

When examining the motivations for making references, one has to be extremely careful about, what is actually being examined, and to what extent the results can be useful outside the testpopulation. When selecting a test sample one should be very careful to make sure that the units of analysis do not represent different populations. This could e.g. happen, if we chose articles coming from two or more scientific domains that do not share the same norms for making references, or if we chose documents from one domain, but distributed on two or more countries, when it is obvious that the selected domain is very likely not to have norms for making references that are globally homogeneous.

3.1.1.3 Open-ended or closed questions

The next thing to keep in mind, is the way the authors of the citing documents are being asked about their motivations. Open-ended questions give the authors an opportunity to naturally put their own words that come to mind on the motivations they had, when making

each reference. By using closed questions, we run the risk of putting a major constraint on the authors, forcing them to make their answers fit into pre-made 'tick off' boxes, and at the same time we run the risk of missing the discovery of new and alternative motivations, because we did not think of these ourselves, when preparing the questionnaire. Further we may risk, that the authors misinterpret the predefined categories, giving us results that suddenly can be analysed and interpreted in more than one way. Making predefined categories for motivations for references that are clear to understand, is an extremely difficult thing to do. We would recommend to allways use open-ended questions, and afterwards it is our own problem to categorize the many statements.

3.1.1.4 The collection of data

When investigating motivations for references, the authors of the documents containing the references have to be asked themselves. There is no way that we as researchers can examine the documents ourself, and classify the references in different categories by ourself. It is only the authors that can give an indication of what exactly their thoughts and feelings were, when they made each reference.

3.1.1.5 The aspect of time

Even though the analysis of motivations for making references almost inevitable will be a retrospective analysis, it is important to keep in mind, that the more time that has passed since an article was produced, the harder it is to make an analysis that can reveal the true motivations, as they were at the time of the production of the article. In order to make this retrospective analysis as correct as possible, it is therefore essential to use articles that recently have been published. We would recommend that this type of analysis should not cover articles more than three years old.

To sum up, use only data that is based on the citing articles, and carefully choose the testsample assuring it is representative of the population. Use open ended questionnaires, and ask the authors themselves about their motivations for the references and only ask them about references in articles less than three years old.

3.1.2 Review on previous studies

Moravcsik and Murugesan were one of the first researchers to start investigating and classify different types of references according to the context they appeared in. They proposed a classification scheme with eight categories, arranged in four dichotomous groups. Each category contains polar opposites, but a reference may belong to more than one group, just not both categories withing a single group (Moravcsik & Murugesan, 1975, p. 88).

1. Is the reference conceptual or operational?
2. Is the reference organic or perfunctory?
3. Is the reference evolutionary or juxtapositional?
4. Is the reference confirmative or negational?

Table 1: Categories for classifying contexts of references (Moravcsik & Murugesan, 1975, p. 88).

Moravcsik and Murugesan's work was one of the first to enter the debate on reasons for citations, and it was initiated as a reaction on the growing critism that had appeared against citationanalyses, for not making indepth analyses of the contexts of the references, and not trying to give an "explicit demonstration or quantitative estimate of the extent of the ambiguities and inconsistencies presumably encountered in the use of citations as measure of science" (Moravcsik & Murugesan, 1975, p. 86-87). Their analysis was based on 30 articles within the area of physics. Their results indicated that 41% of their citations fell into the 'perfunctory' category (not a truly needed reference, but mainly an acknowledgement to other works within the same general area), and 14% fell into the 'negational category' (disputing correctness of others works) (Moravcsik & Murugesan, 1975, p. 90).

Moravcsik and Murugesan did make one mistake though, a mistake that can be assigned to section 3.1.1.4. They classified the references themselves according to the four categories, and did not ask the authors who had initially made the references. They did though make this classification independently, and made an intercomparing afterwards, finding a high rate of overlap in the results (Moravcsik & Murugesan, 1975, p. 89). Using predefined categories and categorising the references themselves certainly can make the results biased from the real population.

In 1989 Cano adopted the model of Moravcsik and Murugesan to examine 42 papers, but this time asking the authors to judge their references themselves and categorize them within the eight categories of the model. Twenty-one scientists within the area of structural reliability were asked to classify the references they had made in two of their recent papers (Cano, 1989, p. 284). Among other results, 26% of the references were selected as to be perfunctory, 14% to be evolutionary (built on the foundations provided by the reference), and 2% to be negational (Cano, 1989, p. 285).

Even though Cano's study was a step in the right direction, it is still critical that the authors are being forced to categorize their references into categories that may not be applicable for their motivations, and other motivations than the proposed categories may very likely

exist. By using predefined categories Cano also runs the risk of the authors misinterpreting the definitions of them.

By investigating motivations for references made in recent papers, Cano does actually get the opportunity to uncover the true motivations, as they were at the production of the papers.

Brooks made an analysis of different models proposed for citer's motivations, and from these he identified seven citer motivations forming a basis for an interview instrument. He presented the motivations to 26 authors at the University of Iowa, whom had been selected on the criteria for recently having published an academic article. The authors came from various university departments representing both the humanitarian and scientific domains (Brooks, 1985, p. 225). The authors also had the possibility to disregard the proposed motivations and nominate their own, and references could be categorized within more than a single motivation. The results were interpreted in two ways. First as a full data set, and second when the dataset was divided into a science subset and a humanities subset. The results from the full data set showed that persuasiveness was the major motivation for citations, while referring to a 'social consensus' (to demonstrate ones knowledge of the area), and making references for 'negative credit' were the lowest ranked motivations for references. When looking at the subsets the results are not as homogeneous. The scientific subdomain having 'currency' (citations made to indicate that the author is up-to-date) and 'reader alert' (background reading, alerting to new work, providing leads, and identify original publication) at the top of the rank, while 'negative credit' is still ranked at the low end. The humanities subset had 'persuasiveness' ranked on top, while 'social consensus' for this subset ranked lowest (Brooks, 1985, p. 226-227).

Brooks noted that despite some similarity it seemed that motivations were varied, and differed among disciplines. As a critique to Brooks we must note, that his selection of the testsample included too many different subdomains, making it extremely difficult to uncover the major norms for referencing within each subdomain, which resulted in only general motivations with no really high indicators of any type.

Brooks proved one year later, that references are often made on the basis of more than a single motivation. He showed that 70.7% of the references were attributed to more than 1 motive. His study further indicated that the citer motives showed 3 general groupings: (1) persuasiveness, positive credit, currency, and social consensus (2) negative credit, and (3) reader alert and operational information (Brooks, 1986).

Vinkler made an extensive study on references in chemistry articles written by 20 selected authors from the Central Research Institute for Chemistry of the Hungarian Academy of Sciences (CRIC). Vinkler tried to uncover the motivations for citing and to find out why some papers were cited while others were neglected (Vinkler, 1987, p. 50). The selection of the authors and the respective paper to participate in the study, were based on several criteria. The criteria for the authors was e.g. to represent the research field of CRIC,

represent the average of the qualified scientists of CRIC, have a total of at least 10 publications and publish at least one paper annually. The criteria for the selection of the individual papers was e.g. to be published in a scientific periodical and may not be a review, research note or other special publication (should be an average publication of the field) and should not be older than 2-3 years (Vinkler, 1987, p. 50-52).

Vinkler categorized the possible motivations into two major groups: professional motivations and connectional (non-professional) motivations. Professional motivations covers the theoretical and practical aspects of the authors research, while the connectional motivations are related to the authors personal, social or external factors. The authors were presented to three questionnaires, each containing possible motivations for references. Even though the questionnaires contained predefined categories, all three also contained the possibility to state motivations other than the suggested ones. The three questionnaires were divided into categories covering 'motivations for citing professional motivations', 'motivations for citing connectional motivations' and 'motivations for neglecting references' (Vinkler, 1987, p. 53-59).

The study of Vinkler resulted in 81% of the references were exlusively made for professional reasons, 17% were made due to a combination of both professional and connectional reasons, while only 2% were made exlusively to connectional reasons (Vinkel, 1987, p. 53-54). Within the professional motivations, the documentary reason is the most frequent (the cited article is a part of a review in the citing article), followed by the applicational reason (the paper is based entirely or in part on the cited document) and the confirmative motivations (documents are cited to confirm the authors own results). Within the connectional reasons 40% specified an existing relationship or a possible future relationship as the motivation for citing the documents. As Vinkler indicates: "Personal relations play, obviously, important part in citing, since the papers of authors with whom some relationship exists are better known. Such connections may be formed during study trips, visits, conferences or other events. Presumably, one pays more "attention" to papers written by known persons" (Vinkler, 1987, p. 64).

The primary reason for neglecting possible references was 'professionally not relevant enough' (Vinkler, 1987, p. 65). Other reasons were based on authors taking over commonly known or incorporated pieces of information (OBI), authors making references to reviews instead of the original papers, primarily because of the convenience (Vinkler, 1987, p. 66-67). Vinkler further concludes that the use of perfunctory citations hardly played any role in the references examined (Vinkler, 1987, p. 68).

The selected methodology of Vinkler does to a great extent avoid the pitfalls indicated in the first part of this chapter, by selecting a single subdomain for analysis and using articles less than three years old. The possibility of indicating other motivations than the ones proposed by Vinkler, can to some extent be a useful combined solution instead of the timeconsuming process of using only open-ended questions.

The last study on citer's motivations to be examined is a study of White and Wang from 1997 (White & Wang, 1997). They conducted a long-term qualitative study of document usage among 12 agricultural economists faculty and graduate students, and they focused their study to discover both behaviors for citing and for not citing. Each participant was contacted and interviewed in 1992 and later in 1995 using open-ended questions. The data was this way obtained using the participants own language without predefined categories. A total of 314 documents were examined and divided into three categories: those who remained uncited (the documents were read, but not cited), cited documents (documents were read and cited), and new cited documents (documents that appeared after the initial interview in 1992, and then were cited). They identified 28 different factors for making references, who were further divided into three types of categories: internal, self-related, and external. The 'internal criteria' covers specific elements of the document such as the author, recency and reputation. The 'self-related criteria' is applied for the readers intellectual or physical capabilities, while the 'external criteria' relates to how the participant's own paper will be received by journals, referees and other persons in formal roles as judges and peers (White & Wang, 1997, p. 133). White and Wang draw a comparison of their categories to Vinklers categories, suggesting the 'internal criteria' to be similar to Vinkler's professional motivations, while the 'external criteria' is similar to Vinkler's connectional motivations (White & Wang, 1997, p. 134). The results indicate that motivations belonging to the internal criteria were the most noticeable reasons. This result seems also very similar to Vinkler's result that indicated 81% of the motivations was exclusively for the professional motivations. The major reasons for not making references were statements as 'too old' or 'too specific' (White & Wang, 1997, p. 138).

As the different surveys has shown above, the motives for making references differ quite a bit individually. This is due to both various domains being examined, but is also due to different methodologies applied. It is uncertain to directly derive a set of globally norms for making references, but being aware and obeying the pitfalls described above, as it has been very fine demonstrated by White and Wang in their study, one can make individual demarcated studies layered on a methodological stable ground.

We need to see more of White and Wangs type of analysis. A qualitative study, that clearly tells us the primary motives for making citations within a selected scientific domain. Based on these results, we can safely go ahead and make quantified citation analyses, because we now posses the knowledge of the nature of the quantified citations within that specific domain, and with this knowledge, we can draw conclusions on the data, and be ensured that they are based on a methodological correct foundation.

3.2 Why do we make links?

The nature of references in scientific documents and hyperlinks on the Internet seem to a certain degree very much alike. They are both indicators of an existing relationship between two documents or nodes. Even though it may be tempting to put a sign of equation between the two concepts, it is strongly adviced not to do so.

A certain combination of the two concepts has started to appear in electronic articles using hyperlinks. Only a single study by Kim in 2000 has been applied to discover the motivations for this specific type of a hyperlink (Kim, 2000). He made qualitative interviews with fifteen Indiana University faculty and graduate students, who had published at least one electronic article, containing at least one external hyperlink. Kim discovered 19 different hyperlinking motivations, which he grouped into three motivational groups (Scholarly, social, and technological). The technological became the group to contain the motivations being the most different from motivations for making references in printed articles. Kim concluded that scholars use hyperlinks for a variety of purposes, and that hyperlinking behavior frequently results from a complex interplay of motivations (Kim, 2000). The latter being similar to Brooks results of the complex citer motivations (Brooks, 1986).

Kim's study indicated, that only a few new types of motivations (belonging to the technological group) for making hyperlinks in electronic articles had occured in addition to the motivations for making references in printed articles, and this was probably due to the scholars hyperlinking behavior being influenced by their conventional practices for making references (Kim, 2000, p. 897).

When shifting the focus to websites not having a scholarly purpose of publishing new scientific results, the motivations for making hyperlinks will presumably differ even more. It is a research area that we still do not have much knowledge about, but it is essential for us to start investigations, and try to avoid a déjà vu on the same criticism as have been applied to the citation analyses.

The remains of the chapter will present a suggested methodology for uncovering motivations for hyperlinks, located on various websites existing on the Internet.

3.2.1 Uncovering motivations for making hyperlinks on websites

Websites are of so many different natures and types, so taking a sample to collect data that would apply for all websites would be an impossible task. We do not know yet how the possible norms of making references will evolve. Whether they will be dominated by cultural factors due to the country the website is produced in, or whether norms for making links will perhaps evolve within different types of websites, e.g. Private websites, Company websites, Public institution's websites (e.g. Scientific research institution's websites) and Webportals.

Due to the possible variations of linking between websites, we cannot perform webometric analyses that could e.g. compare all the universities in Denmark based on the number of outgoing and incoming links on their websites. Within the different universities different attitudes towards linking exist, depending on whatever guidelines the faculty has stated, and the extent of confidence with the Internet and its use, the faculty and each employee

may have. Some universities choose to have an official website, but forget to exploit the hyperlinking nature of the media, and only have internal links making their website a dead end road, while other universities actively take part in exploiting the hyperlinking nature and become a mark on the map of interwoven websites.

While the motivations and norms for linking is still an unexplored research area, some researchers, especially within the more technical domains e.g. computer science, have tried to categorize and automate some of the various link types that exist (e.g. summary and expansion links, equivalence links, comparison and contrast links, tangent and aggregate links (Allan, 1996, p. 44)), especially in order to invent systems that automatically can gather documents for a hypertext, and automatically can produce links and give them annotations within the hypertext (Allan, 1996, p. 42).

Agosti & Melucci have proposed a similar approach, but state that we should start by distinguish between two types of objects involved in IR on the web: Web pages and web auxiliary data (e.g. directory entries, keywords or metadata). The distinguish is to be used to automatically generate links of different types. The link type is in contrast to Allans proposal to be defined by the type of the node (the node is typologized to be either an ordinary webpage or a webpage containing auxiliary data) (Agosti & Melucci, 2000).

Other researchers have examined linking styles within and between websites, trying to establish metrics for determining good and bad linking practice (Carr, Hall & Miles-Board, 2000).

Haas and Grams made an analysis of 75 webpages, and discovered major categories of link types to be: Navigation, Expansion and Ressource. They further used the information of the anchor (the clickable part of the link) and its context, to determine what reasons the author offered the reader to follow a certain link (Hass & Grams, 1998, p. 485). As for the latter, the authors of the examined websites were not asked themselves, but the results were based on the labels surrounding the anchors, and categorized within the three mentioned categories, placing the Navigational reason in top followed by Expansion and Ressource respectively (Haas & Grams, 1998, p. 492).

Another way of defining the linktaxonomy could be on a purely meta level, by defining linktypes using the addresses of the websites from where the link starts and ends. This is very similar to the existing definitions of citations, which we usually categorize in two categories: Self-citations and citations to documents by other authors. Likewise, hyperlinks can be categorized in the categories Internal links on a website, and Outgoing links. Additionally for the webmedia, an extra category can be defined, the Reciprocal links.

Reciprocal links can exist in many natures, and often appear within the same website. What is more interesting and to be discussed a bit later in chapter four, are the reciprocal links going between websites. In short, reciprocal links between two websites can appear in three different forms: Between two webpages linking exactly to each other, which we

define as a true reciprocal linking (illustrated on figure 4), and between three or four webpages located within exactly two websites (illustrated on figure 5 and 6).

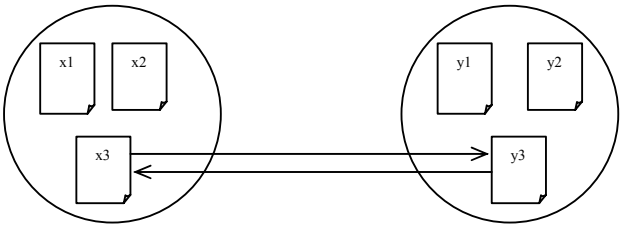

Figure 4: True reciprocal linking between two websites and two webpages

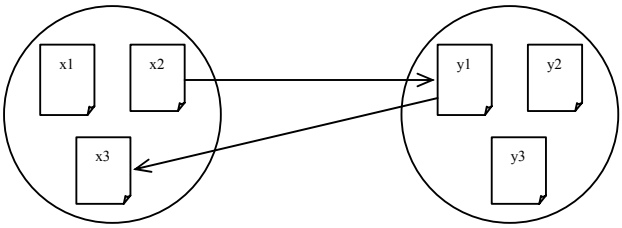

Figure 5: Reciprocal linking between two websites and three webpages

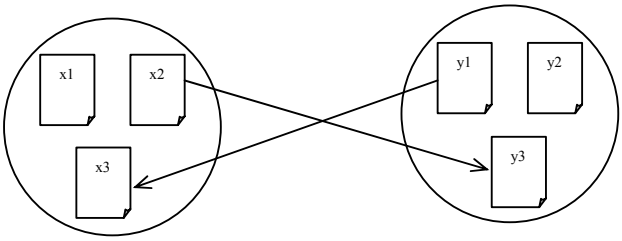

Figure 6: Reciprocal linking between two websites and four webpages

3.2.2 Proposed methodology for uncovering motivations for links

For the situation of using webometric studies, it is essential to make qualitatitive analyses, to explore the motivations for making outgoing links on websites. Whether we allready know a bit about the linktaxonomy, we have no right to say we possess the knowledge of the authors personal motivations for producing the links. The knowledge of the linktaxonomy can only be of value in the process of defining the borders of the webometric analyses e.g. have the focus on the outgoing links on websites, or focus on reciprocal links.

3.2.2.1 The purpose of the proposed methodology

It is our intention, that this methodology should be applied in connection with any kind of webometric study, that is based on the quantification of hyperlinks. Since our knowledge of the motivations for making hyperlinks is still scarce, it is important to make these qualitative analyses in connection with the quantitative ones. The qualitative analyses are the foundation, and they make it possible for us to make the right conclusions on the data from the quantitative analyses.

The purpose is to make a design for a methodology that can help reveal motivations for making hyperlinks on any type of website. Further the methodology will focus on discovering the unknown barriers for not making hyperlinks. The latter is important in order to avoid making incorrect conclusions, that would solely be based on the lack of certain datatypes within the collection of data for the webometric study (e.g. to make a conclusion that two companies have no cooperation, just because they do not have links to each others websites).

3.2.2.2 Methodology for qualitative studies uncovering motivations for linking

The methodology for making qualitative studies of motivations for websites contains the following points, that should all be carefully considered:

- Use outgoing links and not incoming links as the angle of analysis
  This is the same problem as the angle of analysis for uncovering the motivations for references explained in section 3.1.1.1 - we refer to that section for further elaboration.

- Define the population to be examined, and make sure the test sample is representative
  Due to the possible variations in the motives for making hyperlinks, variations that can depend on the culture in different countries or types of websites, it is important to keep the analysis at a micro-level. Choose a welldefined population that contains as small variations as possible.

- Use open-ended questions
  By using open-ended questions we give the respondents a possibility to express their thoughts about the motivations in their own words. If we chose to use closed questions we would run the risk of forcing the respondents into making answers that are incorrect. Either because they misinterpret the meaning of the question, or because they feel they have to state an answer, even if none of the boxes to tick do apply.

- Use only the personal statements from the authors/webmasters of the websites
  It should be obvious to say, but we cannot as researchers answer on behalf of the respondents own motivations. Only they can truly put words on their thoughts and motivations for making each link.

- Use only websites younger than 6-12 months
  Since the study is a retrospective study, it is necessary to use fairly new websitescome, in order to get responses that are as close to real life as possible. As time goes by, the testperson will tend to forget their original thoughts and motivations for choosing each link.

- Ask questions that take their departure in the purpose of the selected webometric study
  If you're planning on calculating the Web-IF or calculate the degree of reciprocal linking between websites to uncover scientific networks between organisations, institutes or researchers, the questions to ask in the qualitative study must vary and focus on the motivations for the essential links being used. E.g. focus on the motivations for the outgoing links or the motivations for the reciprocal links.

- Ask questions that focus on the lack of certain types of links
  This point is important to uncover what possible reasons that may exist for not producing the types of link that is perhaps needed for performing the webometric study.

- Be aware of answers that could indicate possible new areas for webometric studies
  Since we know so little yet about the motivations for hyperlinking, we should be aware, that new research areas for the webometric studies could occur when we listen to the different respondents and their statements.

In the following section, an application of the proposed methodology is demonstrated in relation to the examination of motivations for linking at the selected websites for the study.

Since we were in doubt whether this type of methodology would best be applied using personal interviews or sent as questionnaires, both methods have been used. The researchers were personally interviewed about their websites, while the webmasters of the institutes received the questionnaire by mail.

3.2.3 Application of the proposed methodology

The examination was an example of which questions we thought would be of importance to investigate in advance, if we had decided to conduct a webometric study, that was aimed to uncover the scientific network existing for the selected researchers and to uncover the scientific network existing for the selected institutes on scientific institutions.

3.2.3.1 Questions from the interviews and the questionnaires, and their justification

All the questions that were asked (this goes for both the interview and the questionnaire), were based on the same templet, since these questions all fullfilled the requirements as stated in the proposed methodology.

The purpose of the exemplified webometric study was to uncover the scientific network existing for the selected researchers and to uncover the scientific network existing for the selected institutes on scientific institutions. The questions and their variations (shown below) as they were asked to the researchers, had the following justifications[5]:

1, What was the primary purpose and target group for the website?
This question served to uncover whether the scientific network of the researcher was also a part of the target group for the website.

2, You have the following outgoing links on your website. Please indicate the motivations you had for choosing each of them to be a hyperlink on your site:
outgoing link 1, outgoing link 2, outgoing link 3 etc.
The question served to collect empirical evidence on motivations for making hyperlinks. It was our hope, that the motivations could also give us an indication of which type of data and conclusions, we could be eligible to use in webometric studies for the specific types of websites.

3, Do you correspond with researchers at other institutions? (e.g. using phone, e-mail, ordinary mail, personal contact etc.) (Please enlarge your answer)
The purpose of this question was to ensure, that the lack of links to personal websites was not due to a lack of a personal network within the scientific domain.

4, When examining the outgoing links on your website (as outlined above), it appears that none of them / only a few of them are going directly to personal websites of researchers at other institutions.
Have you considered to make links directly to researchers personal websites within your own professional network? (Please enlarge your answer)
This question was asked in order to see if the researchers perhaps had particular reasons for making or not making personal outgoing links.

5, Do you think it could be relevant and useful for yourself, if researchers within your own network had more links to each others websites? (Please state the reasons for your answer)
This was an indirect question to uncover whether the researchers actually use the websites of their fellow colleagues to seek information.

6, Do you think it could be relevant and useful for yourself, if researchers within another scientific network had links to each others websites? (Please state the reasons for your answer)
This is very similar to the question above, but served also indirectly to explore to what extent the researchers were using the web and personal websites of other researchers, when seeking other types of scientific information.

[5] All questions were asked in Danish, and have only been translated, in order to appear in this paper.

By making a search in the search engines AltaVista and Google, I have found that the following webpages contains a link to your website:
webpage 1, webpage 2, webpage 3 etc.
7, Do you know of any additional webpages that have a link to your website? (If yes, please indicate their address)
Since search engines are not exhaustive, it was important for us to both state clearly to the researchers, that we knew that the webpages above were not exclusively all webpages linking to their site. At the same time, we gave the researchers an opportunity to add webpages that they perhaps knew of.
The question was also merely a test of the search engines. To uncover whether there would perhaps be a lack of specific types of webpages indexed.

8, Do you immediately recognize the webpages that contain a link to your website?
This was an indirect test to see, whether the network to the website (and thereby the researcher or the institute) as it was interpreted by the search engines, was in an immediate congruence with the researcher's or the webmaster of the institute's own conviction.

9, Are you surprised by the number of webpages containing a link to your website? (Please state the reasons for your answer)
This was to see if there was an immediate congruence between the researcher's or the webmaster of the institute's own conviction of the websites visibility and the search engines interpretation of this concept.

10, Can you define the major categories of types of webpages, that link to your site? (e.g. conferences, institutions, private webpages etc.) (Please enlarge your answer)
The question mainly served to give a quick overview of which types of webpages linking to the researcher's and the institute's website, that was indexed in the search engines. Since the search engines are not exhaustive, and we do not know about the criteria of selection for their spiders, we cannot exclude that some types of webpages do not link to the chosen website. They may just as well not have been indexed.

11, Are you surprised, that none/only a few of the webpages, that link to your website are personal webpages? (Please state the reasons for your answer)
This was to uncover how the researcher's attitude was towards receiving personal links.

The interchange of links between researchers (or institutes) within your own research domain can be of valuable use for e.g. researchers from a different domain or students that are not yet fully confidential with the domain, and who may be visiting your website in order to achieve a more indeepth knowledge of your scientific network with other researchers (or institutes), and by that way, acquire ideas for supplementary litterature.
12, Have you considered to make an interchange of links with researchers within your own research domain? (Please state the reasons for your reasons)
This question was based in a line of thoughts concerning the problem of why links are still as unstructured as we see them today. The question tried to uncover some of the causes

that currently exist and block the way for links to be a more common, established and formalized way of expressing the networks that surrounds the researcher or institute.

13, Are there any websites that you have deliberately not made a link to? (Please enlarge your answer)
This question was asked to see if there could be an indication stating that critical links are rather omitted than mentioned.

14, Do you consider a link to be something positive to give and receive? (Please enlarge your answer)
We wanted to collect data in this exploratory study, to see whether there could be a support for a hypothesis on the motivations for hyperlinks to be considerably positive.

The questions for the institutes sent as questionnaires were mainly the same as for the researchers, but a few additional questions were added or substituted for some of the above.
Since outgoing links were very limited at the institutes at the Royal School of Library and Information science, question two was substituted for this question:

2, There are no outgoing links on the institutes website. Is this a deliberate choise? (Please enlarge your answer)
The question was asked in order to uncover some of the reasons for not making hyperlinks at the institutional level.

In connection to question four, an additional question was asked:

4a, When examining the outgoing links on your website (as outlined above), it appears that none of them / only a few of them are going directly to websites of institutes at other institutions.
Have you considered to make links directly to websites of other institutes within the professional network? (Please enlarge your answer)
This question was asked in order to see, if the webmasters perhaps had particular reasons for making or not making links directly to institutes at other institutions.

3.2.3.2 Analysis and results

The purpose of all the websites (both the researchers and the institutes) was to give a presentation of the researchers or the institutes. The target groups for all the websites were colleagues and students in and outside the institution, organisations and other contacts cooperating with the institution. Some of the websites appeared in both an English version and a Danish version, while others only appeared in one of the languages. The definitions of the target groups of the websites are very broad, and the personal scientific network is only a part of the target groups.

Motivations for linking on the researchers personal websites
The researchers at the Royal School of Library and Information Science all had webpages that looked quite alike. This is due to a templet they had all been asked to use from the institution when preparing the website. The templet was mainly focused on showing the CV-information of the employees (e.g. education, boardmembers, publications etc).

The total number of outgoing links examined were 60, with an average of 10 outgoing links on each, and a standard variation of 8.9 outgoing links.

The templet limits the types of websites being linked to quite a bit. The primary motivations for linking were the (Some links were based on more than one motive):

Major motivations for making hyperlinks on researchers personal websites at The Royal School of Library and Information Science in Denmark.	# of times stated	Percentage of amount of links
Linking to various projects they've been involved with	17	28.3 %
Linking to associations and journals where they hold board member positions	9	15.0 %
Linking to other institutions because they've been cooperating with them	8	13.3 %
Linking to websites of journals because they've published articles within them	8	13.3 %
Linking to show websites of prior schools of education	7	11.6 %
Linking to various associations with whom they cooperate	4	6.6 %
Linking to conferences because they have been involved in the preparation	3	5.0 %
Linking to online documents they've published	3	5.0 %
Linking to show websites of previous employments	3	5.0 %

Table 2: Distribution of main statements for motivations for making hyperlinks on personal websites at The Royal School of Library and Information Science in Denmark.

Even though the data are made on a rather small sample of only six researchers, we can see a clear tendency of linking to projects they've been involved with is at the top position, followed by associations and journals where they hold board member positions on a 2nd place, while 3rd place is shared by links to institutions they've cooperated with and journals they've published within. The reason why linking to online documents have a rather low rank, may be due to the copyright restrictions that most authors experience to be constrained by. As one of the respondents noted: "One can make many philosophical thoughts about, for what reasons we actually need the commercial publishing firms: They take our copyrights, make other researchers do the peer-reviewing for free and finally they sell both the reprints and the journals back to ourselves!".

Motivations for linking on different institutes
The institutes at the Royal School of Library and Information Science did not have any outgoing links on their websites, and only one of the institutes responded on the questionnaire. They were planning on a redesign of the website, and will afterwards be linking to cooperating partners (researchers and institutes) from the different projects they are involved with.

The type of outgoing links on the institutes at The Technical University of Denmark differed quite a bit between the two websites. The first institute had a webpage with a list of very broad types of links (e.g. academic institutions in Denmark, telephone directories and the journey planner). These links were located within the mainpages of the website. Unfortunately the more scientific links to e.g. other scientific institutions or partners of cooperation were located within specific websites for the different projects they were involved with, or the more specialized subsites of the institutes. These webpages were not discovered during the examination, and the collected data for this study therefore gives a bias for making any valid conclusions.

The other institute had a very long webpage with 257 outgoing links, and the motivations and subjects for the links varied quite a bit. The webmaster indicated, that the links were not selected on a basis of a thorough judgement, but had been added as he had come to know of them, or were told of their existence by other people, or when the cooperating companies finally got a website.

The dispersion of the percentage on the different motivations has not been calculated, because they would be based on only a single website, and therefore not be indicative for the population at all.

Due to the large amount of links, the webmaster only stated the motivations for the major categories as they were allready prepared on the website:

Links made for showing the partners of cooperation e.g.:
Companies like www.bang-olufsen.com, www.brunata.dk and www.danfoss.com
Links made for internal use that could resemble a bookmarklist for the employees e.g.:
Different universities in the scandinavian countries and around the world Sharewaresites (e.g. www.microsoft.com, www.tucows.com, www.shareware.com) Transportation sites (e.g. www.krak.dk and www.dsb.dk) Phone- and addressdirectories (e.g. oplysning.cybercity.dk) Search-engines (e.g. www.altavista.com, www.lycos.com, www.yahoo.com) Student magazine from the university The local pizzadelivery
Links made for the usefulness for both internal and external users e.g.:
Different electricity companies in Denmark and around the world Different mediasites and public websites (e.g. www.folketinget.dk, www.berlingske.dk) Websites within the same researchdomain (e.g. www.risoe.dk)
Links made for students at the faculty and for future exchange students e.g.:
Studentorganizations and websites about studying abroad
Links made for masterstudents who are soon to be jobseeking e.g.:
Various jobindexes

Table 3: Motivations for making hyperlinks at one institute at The Technical University of Denmark

Reasons and barriers for not making links
Question no.4 was focusing on the barriers for not making outgoing links to personal websites or to other institutes (asked to researchers and institutes respectively).

Lack of time seemed to be the major reason for not spending more time on making personal links on the researchers websites. It would require quite a lot of maintenance every time the peers change jobs or the address of their webpages change. Another reason is the question of the criteria to choose from. Who should the researchers choose to make a link to, and who should they not? and how will the peers who have been omitted react? it is quite a sensitive area to be working with. As one of the respondents indicated "it would be like maintaining a book collection - using criterias for selection, and on what conditions should these criterias be founded?". The researchers seem to prefer linking to official institutions, journals and conferences - all types of websites that have more stable addresses. These types of websites are also the ones who are primarily linking to the examined websites.

Another reason why the researchers did not link to personal websites was due to the templet they had all been asked to use from the institution when preparing the website. It was mainly focused on showing the CV-information of the employees (e.g. education, publications etc).

Question no.12 tried to uncover some of the causes that currently exist and block the way for links to be a more common, established and formalized way of expressing the network that surrounds the researcher or institute.

Most of the researchers had not considered to exchange links with peers or institutions in order to establish reciprocal links. Some had never thought of doing it, while others stated, that they would perhaps employ it in a future update of their website.

When trying to uncover the networks between the researchers, it is important to remember, that not all researchers have a website yet. That is, one cannot receive an ingoing link if one has not published a website, just as one cannot receive a citation if one has not published an article.

The institutes at the Royal Danish School of Library and Information Science only had very few outgoing links, which is why they had a seperate question (no.2) focusing on these reasons. Since only one institute answered the questionnaire, the result is not of much use. But the institute did mention, that the outgoing links would primarily exist on the individual researchers websites, and the institutes website was still under construction due to a reconstruction of the organizational structure at the school, but links to cooperating partners (institutions and researchers) would appear in the future.

The institutes at The Technical University of Denmark stated that their links to other institutes or cooperating partners were mostly located within websites for the different projects, they were involved with, or the more specialized subsites of the institutes.

The results for question no.3, concerning the researchers correspondance with peers at other institutions, clearly indicates that a lack of personal outgoing links to peers is not due to a lack of a scientific network. All the researchers have a scientific network, and it's quite often also international and not only national. The researchers don't have an obvious personal need to display their personal network, since they allready communicate with their peers via e-mail, fax etc.

The answers to question no.5 and no.6 concerning the usefulness of peers showing their networks on their websites, was actually 50% to both yes and no. If the answers had primarily been in favour of a no, it could have given an indication, that the lack of personal outgoing links had to do with the respondents own nature of surfing and retrieving information about peers by using the peers webpages and their networks on the Internet.

What conclusions can we make on the revealed motivations for linking?
Since the personal websites primarily link to projects they've been involved with, associations and journals, where they hold board member positions, and cooperating institutions they are involved with, and since we know that the researchers do have a scientific network, often on a international level, we cannot conclude in the webometric analyses, that researchers do not cooperate personally with colleagues within the domain,

based on the lack of personal outgoing links on the examined websites. Instead, a webometric study based on the researchers websites could primarily give indications of the type of networks that applies from a CV perspective.

A webometric study based on the institutional level would for the institutes at the Royal School of Library and Information Science not be useful for any indications, since they hardly show any kinds of network on their websites. For the webometric study for the institutes at the Technical University of Denmark, it is difficult to make any significant conclusions. The webmaster of the first institute did indicate, that they had links to cooperate partners, but they were only located on the subservers and therefore not found. The second of the institutes who possesed a long list of links, could perhaps be useful for a webometric analysis of focusing on cooperate partners, but also on networks displaying websites with similar interests.

The difference between motivations for linking on private and institutional websites
Due to the few responses from the institutes, we cannot say, if there are any differences between the two types of institutes.

Apparently all the websites had the same purpose, and primarily the same target groups. The institute websites usually had broader subjectgroups of links as a service aimed for their targetgroups, than the personal websites, that primarily stayed within a CV-type of links. Both types of websites had primarily links going to an institutional level, and they both hardly had any personal outgoing links.

The personal websites had more links to selected conferences and journals, while the website of the second institute at the Technical university of Denmark would be more exhaustive in their linking to all types of institutions within a certain subject category or all types of universities around the world.

The nature of the motivations for choosing the links indicates, that the links on the personal websites were primarily to give inspiration for further information, while the links on the institutional websites were more of a navigational nature, indexes to be used on a more structured basis for seeking further information.

3.2.4 Identified possible problems and hypotheses for future studies

3.2.4.1 Evaluation of the methodology

The questionnaire had the advantage, that the interviewer could not influence on the answers from the respondents, but the method requires high standards to simplicity in the formulation of the questions, in order not to be misinterpreted by the respondents. The respondents usually gave less varied answers, answers that could have been enlarged if an interviewer had been present.

The interviews gave more varied and detailed answers, but the method has the disadvantage, that the interviewer could perhaps influence on the respondents answers. This has also been noted by Frankfort-Nachmias and Nachmias: "The lack of standardization in the data collection process also makes interviewing highly vulnerable to interviewer bias. Although interviewers are instructed to remain objective and to avoid communicating personal views, they nevertheless often give cues that may influence respondents' answers" (Frankfort-Nachmias & Nachmias, 1996, p. 238).

Using the interview as a datacollection method therefore requires a strict discipline from the interviewer to be aware of this possible bias.

Based on our experience, we would recommend to stick with the personal interviews in the future when using the proposed methodology to examine the motivations for making hyperlinks.

3.2.4.2 Evaluation of the analysis

Analysing three different types of websites is a more complex and time-consuming task than at first expected. As a consequence, we recommend to only analyse one type of website at a time, e.g. the personal websites, or at least, make sure to have plenty of time available for the project.

Seen in the backmirror, the questionnaire had a bit too many questions to be answered. Instead we should have concentrated more on the motives for making the hyperlinks. E.g. by repeating the questions about the motivations during the interview, to see if they could think of other motivations for each link than the allready stated. This could especially be valuable in order to uncover the degree of the existence of a complex linking motivation, similar to Brook's complex citer motivations (Brooks, 1986).

The possibility of misunderstanding the questions did to some extent occur with question no.2. When the respondents were presented to the list of the outgoing links on their websites, they would rather categorize them into different types of websites, than actually state the motivations that they had for each of them. The question of whether we can speak of a difference between the type of a link and the motivation of a link can be discussed, at times they will be equal, which can be the reason for the misunderstanding. The misunderstanding of this question mainly occured in the questionnaires that were sent by mail.

The questions no.8, 9, 10 and 11 should have been omitted from the study, because the search results, which the respondents were asked to state their judgements on, were all based on data that originated from search engines from the Internet. Search engines of which we have only a vague knowledge of their indexing policies, their coverage of the Internet etc. Therefore, it was like asking the respondents to bring judgements on insufficient data material. This conclusion is also supported by the answers to question

no.7, if the respondents knew of any additional webpages linking to their websites. The majority of the respondents did actually know of additional webpages, that had not been found by the two search engines used.

We could have made the introduction to the purpose of the study more clear to the respondents, in order to focus the answers better. Since working so much with this subject, one tend to forget, that the respondents do not have the same knowledge of all the different aspects in the study, and they may not possess the same interest in it either.

3.2.4.3 Prospects for future studies

The exploratory study has shown an indication of a hypothesis, that motivations for linking are made from a less personal interest (e.g. demonstrate ones knowledge of certain subject), and instead appear for a more practical navigational reason and for using the technology available.

Another hypothesis that could be interesting to investigate, is the use of links made from a positive, negative or neutral point. The study indicates that the respondents mainly link for a neutral or positive reason, while the opportunity to link for a negative/critical reason is rather being ignored than used.

A field that could be of interest to examine is the mere sociological background that is hidden behind the reciprocal links. The study has showed us, that the selected websites do not use reciprocal links as a way of exposing their networks. We do not know if this conclusion is applicable for all types of websites, and if they do exist, we first need to evolve a useful tool to uncover these specific types of links.

To investigate further what this might reveal, we have developed the program Link Agent. The program is able to show the user surfing on a specific webpage, the information about which outgoing links that exists on the webpage, and which of the webpages that are linked to, that have a link back. The program will be further described in chapter four.

3.3 Discussion of differences between citations and links

When comparing the uncovered motivations for linking with some of the major types of motivations for making references, it is clear, that motivations for linking are made on a much more varied foundation than for making references. References are made within a scientific context and restricted by norms for making references, while links are made within a much broader context, and usually to ease users navigation. In short, when writing a scientific paper, it is mandatory to make references to other documents, while producing a website gives you the opportunity to make links.

When making a reference in a scientific article, the author has every right to make references to articles he finds to be of relevance for writing his article, no matter how

obscure the subject of the cited article seems, and seen from the contrary side, an author of an article has no right to claim whether or not his article should be cited in the future, that is a decision to be made solely by his peers.

This aspect has changed slightly when discussing hyperlinks. The dynamic and widespread nature of the media has evolved the possibility and an awareness of which websites one is receiving links from, and thereby the discussion of the rights and expectations of a website that does or does not receive one or more links. To what extent can the authors of a website claim to have a right to receive a link?, or claim to have a right not to be linked to?. A few cases about the latter have allready had their way to the media. E.g. when a Danish real estate website named 'Home' was complaining about two major webportals linking to their website with misleading information (Skovmark, 2001).

The possibility for similar discussions in the future are not only fiction. Imagine, perhaps a researcher on a scientific institution starts linking to Nazi propaganda websites because he in his personal life has an interest in Nazi material, or the opposite, that Nazi propaganda websites start to link to one or more researchers websites on a scientific institution. This link, that to some extent indicates a connection between the two websites, may not be appreciated by the faculty of the scientific institution. Further, some search engines even give lower ranks to websites being linked to by e.g. Nazi webpages, which to some degree would lower the respect for the institution, and make it harder to find the institutions website using the web search engines.

This could call for the urgence and importance of making linkpolicies for websites. What types of webpages should it be allowed or tolerated to be linking to from our website? and who do we tolerate to make links to our website? the latter is very hard to control, since everyone can still make links to whoever they like, but it is important to start the discussion of how the authors of a website wish their website should appear on the map of intervowen links, and what the consequences of a wrong appearance could end up with.

An example of trying to make restrictions on who should be allowed to make links to a website has been seen at the NEC Research Institute in connection to their Researchindex database (NEC Research Institute, n.d.).

3.4 Main conclusion on reasons behind citing and linking

It is recommended to use the proposed methodology for qualitative analyses, when one wants to perform quantitative webometric analyses.

It has been shown, that the type of network that can be concluded on, varies quite a bit, depending on whether the analysis is based on using citations or links. Even within each type differences do occur, and it is therefore important, to continue making qualitative analyses. This way we can uncover the major motivations for making the references or links, and we can safely go ahead and make quantified analyses, draw conclusions on the

data, and be assured that they are based on a methodological correct foundation, because we now possess the knowledge of the nature of the quantified citations and links within that specific area.

Chapter 4 - The right tools for making the right research

When conducting webometric analyses, one part is to be aware of the specific nature of motivations for making hyperlinks within different types of websites, and the other part is to be aware of the quality of the search engines and other data collection tools that are being used. These two basic elements constitutes the foundation for performing webometric analyses. Working within the informetric domain, conducting webometric analyses, we have an obligation to investigate the nature of the motivations for making hyperlinks, and assure the quality of the data collection tools applied in order to perform valid and realiable results. The first element has been treated in chapter 3 and the second element will be covered in this chapter.

The chapter starts with an outline of the current status on the knowledge of the topology and size of the Internet based on hyperlinks. Further, the chapter contains a critical outline of the demands we need to state for the search engines, we use for the webometric analyses, in order to perform valid and reliable results. The need for these demands will be demonstrated by a review on the current knowledge about using the major search engines for webometric analyses. The chapter is rounded off with an examination of other types of possible tools available for performing webometric analyses.

4.1 Link topology and size of the Internet

4.1.1 Link topology

Kleinberg was one of the first persons to investigate the properties and structure of the Internet. He developed the small world phenomenon covering the concepts of hubs and authorities, that can be identified by applying his special search technique HITS (Hyperlink-Induced Topic Search) (Kleinberg, 1998).

Authorities are webpages, that have a high rate of ingoing links, while hubs are pages, that have many outgoing links (See figure 7). Hubs and authorities represent a mutually reinforcing relationship (Kleinberg, 1998). Kleinberg assigns the hubs and authorities to be the most relevant pages, when conducting a query on a subject (Kleinberg, 1998, p. 670).

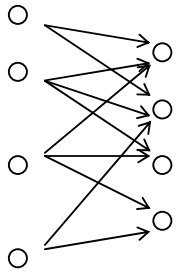

Figure 7: Hubs (left) and Authorities (right) (Kleinberg, 1998, p. 670).

The output of the HITS search technique is a list of pages with the largest hub weights and a list of pages with the largest authority weigths within a certain subject. The webpages are found using the following technique: 200 webpages on the same keywords are found using one of the available search engines on the Internet. These webpages are called 'the root set'. Then the set is expanded by finding any page on the Internet, that has a link to/from one of the pages in the set. The new set is called the 'base set'. Each page in the base set is assigned an authority weight and a hub weight after an iterative process has been carried out by going through all pages. (Kleinberg, 1998, p. 699-670, 674). The highest weight results within the authorities and hubs is defined to be 'the community'. These results are to be presented to the user as a response to his query on the subject (Kleinberg, 1998, p. 671).

Several problems can be assigned to the HITS technique. First of all, the root set is based on results from existing search engines about which we do not know their rules for indexing, their limitations etc. Secondly, the use of keywords for finding the root set can be very sensitive, and result in a root set not exactly covering the same subject. This would easily occur when searching on subjects, that are also homographs.

Gibson, Kleinberg and Raghavan have further studied the properties of the different communities that can be revealed by the HITS technique (Gibson, Kleinberg & Raghavan, 1998). They report that the HITS algorithm is rather robust, thus the communities uncovered on different root sets, that were based on only small samples of relevant pages were mainly the same. They also concluded, that topics that are covered by both commercial and individual involvement will be dominated by the commercialized pages in the uncovered authorities covering the topic. Further, some pages are so highly interlinked with other pages (e.g. AltaVista and Yahoo), that they often appear on the list of authorities on different topics (Gibson, Kleinberg & Raghavan, 1998, p. 229-232).

In 1999 Albert, Jeong and Barabási also conducted a study on the topological structure of the Internet made on the basis of data from a complete map of the nd.edu domain. Their data consisted of 325,729 documents and 1,469,680 links. Their study indicated that "the web is a highly connected graph with an average diameter of only 19 links" (Albert, Jeong, Barabási, 1999, p. 130), meaning that two randomly chosen pages will be connected within 19 clicks on links.

In contrast, Broder et al. made an extensive analysis of the topology on the Internet based on link structures. Their study covered 200 million webpages and 1,5 billion hyperlinks (Broder et al., 2000). Their results showed a significant different pattern of the link connectivity than the results of Albert, Jeong and Barabási.

Broder et al. revealed a new structure of the web, indicating that the web can mainly be divided into four parts forming the shape of a bow tie (see figure 8). The first part is the center of the web. All pages within the central part can reach one another using directed links. This part is also named the strongly connected component (SCC) and consists of about 56 million webpages (27.7%) of the testsample (Broder et al., 2000, p. 10-11).

Within the 'SCC', navigation between websites is fairly easy, and the path between two webpages consists of only a few links.

Another large part of the sample is the one labelled 'IN', which contains webpages that link to the center of the bow tie, but cannot be reached from it, because no links are pointing back out. The webpages within the 'IN' group are usually new webpages that have not yet been established within a network on the Internet. The 'IN' group constitutes for about 21.3% of the sample webpages (Broder et al., 2000, p. 11).

The 'OUT' group consists of about 43 million webpages (21.2%) of the sample. This part consists of webpages, that are linked to from the center of the web, but do not have any links pointing back. Webpages within this group are usually corporate websites containing only internal links (Broder et al., 2000, p. 11).

Two other groups of webpages fall outside the bow tie shape. The 'Tendrils' and the 'Tubes'. The 'Tendrils' are webpages that are only connected by either receiving links from the 'IN' group or by pointing to a website within the 'OUT' group. If the 'Tendril' is both receiving a link from the 'IN' group and pointing to a webpage within the OUT group and is not a part of the strongly connected websites within the center of the bow tie (SCC), the 'Tendril' forms a 'Tube'. The 'Tendrils' constitutes for about 21.5% of the sample, while the rest of the sample, the 'Disconnected' (8.2%), are websites that are not connected in any way to the rest of the webpages (Broder et al., 2000, p. 11).

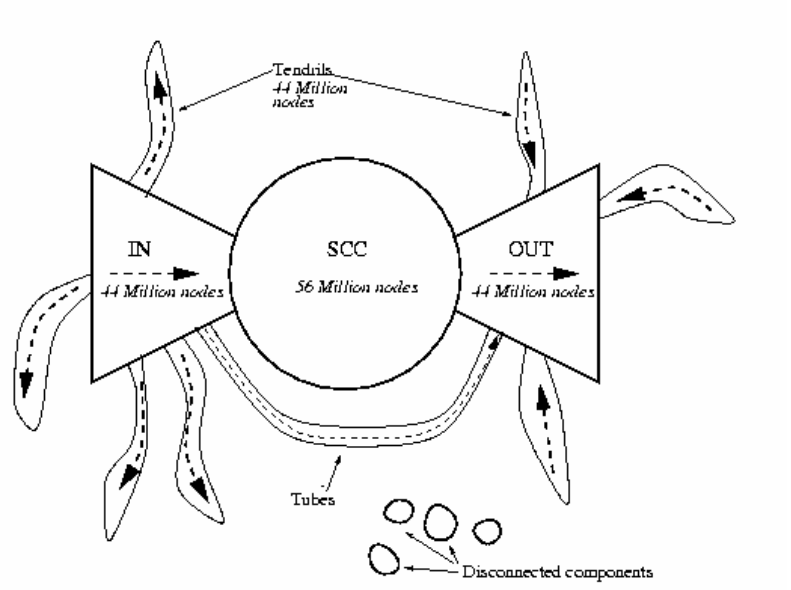

Figure 8: The Bow tie shape of the Internet, based on linkstructures between nodes. (Broder et al., 2000, p. 10)

When comparing Kleinbergs small world communities algorithm to the map (figure 8) of Broder et al.(2000) it appears, that the authorities would be located within the 'SCC' or the 'OUT' group, while the hubs would be located within the 'IN' or the 'SCC' group.

If we compare the map to the flow of how surfing is supported by links, then the flow would definitely go from the left towards the right side. One could imagine, that you would read about a new and interesting webpage in a newsletter or newspaper. You would check out the webpage (located in the 'IN' group) and follow the links to other webpages (located within the 'SCC' group), and your surfing would be stopped, when you ended up on a webpage with no outgoing links (located in the 'OUT' group).

Broder et al. further found the following results (Broder et al., 2000):
- The probability, that a directed path[6] exists from a random source document to a random destination document is only 24%.
- If a directed path does exist, its average length will be about 16 links.
- If an undirected path exists, it's average length will be about 6-7 links.
- Over 90% of the webpages in the sample are reachable from a random webpage, by following an undirected path (either forward or backward links).

4.1.2 The size of the Internet

Estimations of the total number of webpages on the Internet are quite difficult to make and only few have made studies based on a high level of testsamples. Further, the number of webpages continues to grow every day making the estimations incomparable.

In 1998 Lawrence and Giles estimated the size of the 'indexable web' to be 320 million pages, based on data retrieved in December 1997. The 'indexable web' is defined as the types of webpages being indexed by the ordinary search engines, and it does therefore not include e.g. webpages hidden behind search form. They based their estimation on an analysis of the overlap among different pairs of search engines (Lawrence & Giles, 1998).

A year later Lawrence and Giles published a new estimation of the 'indexable web' based on a different and rather more reliable methodology. They first tested 3.6 million IP addresses, and found a web server for one in every 269 requests, which gave them an estimate of 16 million web servers in total. Next, they crawled all the pages on the first 2,500 random web servers. The average of webpages per server was 289 pages, which gave an estimate of the publicly 'indexable web' to be around 800 million webpages. (Lawrence & Giles, 1999a, p. 107).

In January 2000 the Inktomi Corperation and NEC Research Institute (represented by Lawrence) announced they had estimated and verified (by indexing) the size of the indexable web to be 1 billion unique documents (Inktomi; NEC, 2000). Inktomi is one of the leading databases on the web, powering many wellknown search engines as HotBot, MSN Web Search and GoTo.

[6] Directed path refers to a path using only forward pointing links, as a contrast to an undirected path which can follow links going in both backward or forward direction.

The BrightPlanet Internet Content Company published a rather remarkable article on 'the deep web' authored by Bergman. They claimed, that ordinary search engines only cover the surface of the web, while their newly developed search engine 'Lexibot' was able to search the content of the deep web. That content that resides in seperate searchable databases on the net, and can usually only be retrieved by a direct query (e.g. library catalogues). Bergmans definition of the surface of the web is very equal to Lawrence and Giles definition of the 'indexable web', that did not include webpages hidden behind search forms. Bergman describes the Lexibot as a directed query engine, an engine with an "ability to query multiple search sites directly and simultaneously that allows deep Web content to be retrieved" (Bergman, 2000, p. 4).

Bergman claims, that the size of the deep web is estimated to be 500 times larger than the surface web (Bergman, 2000, p. 13, 18). Their estimation is based on a profound inspection of the largest known 60 websites containing information characterized within the category of the deep web (e.g. content within individual search engines). Bergman further estimated, that the total number of deep web sites at the time of inspection was roughly 100,000 (Bergman, 2000, p. 17).

As the various studies have shown, the estimation of the size of the web is hard to give an exact number about, but the above estimates are based on well founded methodologies. Their estimates differ primarily due to the difference in the time of the investigation, and due to their definition of what exactly is a webpage. We think, that Bergmans definition of a surface web and a deep web is a good terminology in the distinction of different types of webpages.

It is important to stay up to date with new results on the mapping of the structure of the Internet and its size, due to two reasons. We need to have a solid knowledge and overview of the structure and size of the Internet, in order to be able to evaluate the search engines, that are indexing the webpages. Only this way can we be aware of the possible problems and opportunities for the search engines. The next thing we can use the knowledge of the structure and size for, is when we develop our own tools. We have an obligation to make sure, that the tools are cabable of taking full advantage of all the webpages on the Internet.

4.2 Outline of demands for search engines

When using search engines to conduct webometric analyses, the results can only be useful, if they are based on tools that are valid and reliable. We have experienced too many times, that different search engines give different results for the same query, and even within the same search engine can the number of results indicated vary quite a bit. Based on previous experiences, we outline the following demands to be stated for the search engine to be used next time one wants to perform a webometric study:

- Must have a high coverage of the websites to be analysed.
- Search results must be 100% reliable. A repetition of a query should give the same result, if performed within a short distance of time.
- Must give free access to knowledge about the methods of indexing, current coverage of the Internet and whether some types of webpages are not being indexed and why.
- Must give access to perform boolean search queries and to use advanced field searching, enabling us to find e.g. reciprocal linking webpages, webpages that are colinked or bibliographic coupled webpages.

Working within the informetric domain, producing and analysing quantified search results, we must not forget, that it is also our own obligation to make sure the analyses are founded on valid data. If we use search engines of which we have no or only very little control and knowledge about how work, we cannot guarantee that our results are valid results. Results, that can clearly be said to be based on a representative sample of the whole population that is under examination. A demand, that is a fundamental methodological demand! (Frankfort-Nachmias & Nachmias, 1996, p. 181).

As a matter of fact, it has been proved, that webpages with many ingoing links have a higher chance of being indexed in the search engines (due to the spiders who usually finds new webpages by following links between them) than webpages with a low number of ingoing links (Lawrence & Giles, 1999a, p. 109). This clearly points to a possible bias for webpages within the 'SCC' part of Broder's bow tie (2000) to have a greater share of the search engines, than other lower connected webpages.

4.3 Review on previous critics of search engines

The need for stating the outlined demands above, will be demonstrated by a review on the current knowledge about various search engines, that are used for webometric analyses.

4.3.1 Coverage and overlap of search engines

The amount of webpages on the Internet, that is indexed in the search engines is a well used parameter for competion between search engines. The question remains though, if we can trust the amount they state. Instead, researchers have developed different methods to calculate the search engines coverage of the Internet.

In 1998 Bharat and Broder proposed an advanced but standardized and statistical way to measure search engine coverage and overlap through random queries. Their technique was tested twice on more than 10,000 queries for each test in four search engines (Bharat & Broder, 1998, p. 380). The queries were using a lexicon of words drawn from a wide range

of topics on the Web. The queries made up for a sample of random webpages from each of the selected search engines. Each page retrieved was checked for availability in the other three search engines. Based on this result of overlap between each pair of search engines, they calculated each search engines coverage of the Web (Bharat & Broder, 1998). Their results showed, that at the time, AltaVista ranked first with a coverage of 62% of the estimated total 200 million webpages existing at the web. The overlap between the search engines was rather low, only 1.4% of the webpages had been indexed in all four search engines (Bharat & Broder, 1998, p. 380).

Lawrence and Giles analysed in 1998 the response of 11 search engines to 1,050 queries based on a very similar technique. They retrieved the entire list of documents for each query in each engine and analysed every individual of them. To estimate each engines coverage of the Internet, they used an absolute value for the number of pages, that had been indexed in one of the search engines (Lawrence & Giles, 1999a, p. 108). The result showed, that no engine covered more than 16% of the entire web. Further, they calculated the overlap between the engines and showed, that a combination of all the selected search engines could reach up to 42% of the estimated number of the total size of the web (Lawrence & Giles, 1999a, p. 108).

Notess has invented an 'effective size' estimate of the major search engines (Google, Fast, AltaVista, MSN search, Northern Light, iWon). The estimate is calculated due to various reasons. Some of the computers holding the index for the search engines may be down for backup or other maintenance, and it is also well known, that AltaVista on some searches make use of a time out, and therefore only deliver partial results. Notess' estimate is an attemp to show the true size of the search engines as they are at the time of the search being run (Notess, 2001a). The estimate of the 'effective size' is calculated on the basis of another study by Notess (Notess, 2001b), that covers the 'relative size showdown' between the same major search engines. The percentage of each search engine's total hits from the 'relative size showdown' is multiplied with the exact counts obtained from Fast and Northern light. The final estimate is an average of those two numbers (Notess, 2001a). The results of the estimate as calculated by Notess, shows that even though Google ranks first in both the estimated (625 million webpages) and the selfreported (700 million webpages indexed) amount of webpages, the estimated number is a good size below the selfreported number. The 2nd ranked search engine was Fast[7], that reported 607 million webpages had been indexed, while the estimated amount was 539 million webpages. AltaVista ranked 4th with 500 million webpages claimed to be indexed, while the estimated amount was 423 million webpages (Notess, 2001a).

In February 2000 Notess also conducted a study on the overlap between fourteen different search engines. He compared the results of five small queries, that resulted in 795 hits of which 298 were unique webpages. 110 webpages (36.9%) of the 298 pages were found by only one of the fourteen search engines while another 79 webpages were found by only

[7] Fast is an index powering search engines like www.alltheweb.com and www.lycos.com. All the web was the one applied in the study.

two search engines. Over 76% were found by no more than three search engines (Notess, 2000b). These numbers clearly indicates, that the overlap between search engines is not very high.

All the above studies, even when looking apart from the year they've been conducted, show that none of the search engines even come close to index all webpages on the Internet, and it is further remarkable, that the low estimation of the overlap among search engines is a rather stable result through all the studies.

4.3.2 Freshness / Recency of the indexes

As it is important to perform webometric analyses based on data retrieved within a short period of time, due to the dynamic media, changing the content of webpages often and not based on regular intervals, it is evident, that we need to state the same demand to data in the search engines, to have been retrieved and checked very recently. If a search engine contains many dead links or webpages with content, that is no longer current, our analyses would not be reliable, even though we did consider to perform the test within a short period of time. In an analysis by Lawrence and Giles, the percentage of invalid links were on average 5.3 % varying between 14.0% at the highest and 2.2% at the lowest (Lawrence & Giles, 1999a, p. 108).

4.3.3 Variation in number of returned hits

The number of hits returned on a query and the accuracy has been examined by Notess (Notess, 2000a). His results show (based on the search engines HotBot and Altavista), that the number is often either missing or inaccurate. Further, there is a lack of a definition on what the counter actually does count. Whether it is the total number of websites found or the total number of webpages, and whether the number of possible results under the 'more pages from this site' are included in the calculation (Notess, 2000a).

Bar-Ilan found a similar inconsistency, when using AltaVista limiting the search to only the English language. The number of returned pages was more than twice as big as for the same query, that was not limited by language (Bar-Ilan, 2001, p. 16). The same happened, when she used a limitation on the calendar year. The result of the sum of each year would add up to a much higher number than the result of a single query without the limitation on the year (Bar-Ilan, 2001, p. 16).

4.3.4 Reliability over time

During a five month period of using the query "informetrics OR informetric" within six major search engines in 1998, Bair-Ilan observed quite a variation within each search engine, especially for Exite and AltaVista. URLs, that had been retrieved by them at one time, did not occur on a succeeding search, even though they were still available on the Web. Further, the URLs sometimes re-occured on the list in even later search results (Bar-

Ilan, 1999, p. 12). Supposedly, AltaVista has now implemented a new technology to assure more stable results.

4.3.5 Boolean operators and field operators

Most of the major search engines provide the possibility to apply Boolean operators within their advanced search (e.g. AltaVista, Google, All the web and Northern Light), but at the same time, some of them are quite restricted, preventing more complex and combined search strings to be fully applied (Google and All the web).

Again, the possibility of using advanced field operators, does not guarantee a valid product. Thomas and Willet conducted a study to investigate for correlation between sitation[8] data (for departments of librarianship and information science) and the peer evaluations of research excellence embodied in the RAE rankings. Their results could not support this correlation, but it was mainly due to incorrect sitation counts from the search engine (AltaVista), that had been used. The counts were easily disproved by manual inspection of the selected websites (Thomas & Willett, 2000, p. 424).

Snyder and Rosenbaum (1999) also draws the attention to the obscureness of using Boolean operators and field searching. To demonstrate, they did a search in AltaVista on 'host:osu.edu' which returned 1,408 pages at the host. Then they checked the first twenty pages to assure that they each had one or more links to an 'edu' site. The search was then narrowed to 'host:osu.edu AND link:edu'. Only four pages were returned, and none of them included the twenty pages, they had first retrieved (Snyder & Rosenbaum, 1999, p. 380). This result clearly indicates, that something very fundamental is wrong with the application of the Boolean operator or the field operator.

Since the study was performed almost three years ago, and since AltaVista supposedly should have become more stable, we re-conducted the search. This time the result was 7,048 webpages from the 'host:osu.edu'. We checked and confirmed manually that at least 20 pages within the first 100 results did contain a link to an 'edu' site. After narrowing the search string to 'host:osu.edu AND link:edu', 15 webpages were returned, so evidently, the problem is still the same. One might just ask, what is the use of Boolean operators and field operators, if we cannot trust the results?

4.4 Prospects for the future of webometric tools

The above review on the various flaws and shortcomings of the current search engines, clearly indicates, that they are not of any valuable use to perform valid and reliable webometric analyses. Instead, the review has shown, that search engines are not made to provide exhaustive and reliable results of all webpages, but were only built and maintained to provide average results for ordinary people searching for various subjects.

[8] Sitation is an expression used by McKiernan (1996) and later Rousseau (1997). The concept sitation is equal to citations in scientific documents, internal and external webpages that link to a specific webpage.

In order to perform valid webometric analyses, it is important, that we start to participate more in the development of the data collection tools we want to use, and detach ourselves from the current dependence on the commercial search engines. We need to look further, and be seriously aware of our responsibility to produce reliable analyses and results, and not just make half-hearted and unreliable analyses, that cannot be used for anything, but confuse the outside world with misleading research results. It is about time, that we take the necessary steps to build our own tools, either by ourselves, or by cooperating with people, who know how to, and who will listen to our needs and demands. This has also been proposed by Snyder and Rosenbaum (1999, p. 382), Lawrence and Giles (1998, p. 100) and Bar-Ilan (2001, p. 22).

The following sections cover a portion of those types of tools. Tools, that are all potential to be used for webometric analyses.

4.4.1 The Clever Project

Kleinberg's HITS algorithm (Kleinberg, 1998) to uncover hubs and authorities, as was discussed in the beginning of this chapter, has later been improved and implemented in Clever - a search engine that is a part of The Clever project at the IBM Almaden Research Center (The Clever project, n.d.). By implementing the HITS algorithm to the Clever search engine, they could prove, that the advantage of retrieving webpages using link structures for searching, could easily compete with the ordinary search engines, that are based on heuristics (e.g. frequency of specific words) to determine the rank of the documents. Their search results had a much narrower focus, and could even discover highly relevant webpages, that did not contain the search word (e.g. AltaVista or Google don't contain the expression 'search engine' on their webpage) (Chakrabarti et al., 1999a, p. 3).

Their further results have shown that their algorithms can discover Web communities, with highly specific interests, that even a human indexed search engine like Yahoo may not be able to find (Chakrabarti et al., 1999b, p. 62).

From a webometric viewpoint, these very specific communities will be of high interest, since their network through links can reveal certain properties of ways of linking within different types of communities.

So far, the only disadvantage is the foundation for the Clever search engine. It still finds the 'root set' of webpages to start from, by using a standard text index such as AltaVista (Chakrabarti et al., 1999a, p. 4). By doing this, they run the risc of restricting the results to certain areas of the Internet, due to the indexing methods, that are applied by AltaVista (which are unknown). If they instead would base the Clever search engine on an index made by themselves, or at least, use an index with an open access to the policy of indexing, they could apply webometric analyses and produce results of high value, because results would be based on well defined parts of the Internet.

4.4.2 Researchindex.com / (CiteSeer)

Lawrence, Bollacker and Giles at the NEC Research Institute have developed the Researchindex (previously known as CiteSeer) - an index, that collects academic research publications available in electronic format on the Internet (Giles, Bollacker & Lawrence, 1998, p. 89). Their idea arose due to a need for fast and easy access to scientific litterature that is up to date - something the Internet would be evident to provide.

The project is an index based on automatic indexing of academic literature existing on the Internet and found through various sources (e.g. search results from queries sent to multiple search engines, monitoring maillists and indexing posted documents, crawling webpages and selfreporting directly from authors) (Lawrence, Bollacker & Giles, 1999b, p. 140).

When the documents have been retrieved, they are parsed into different segments to be saved in an SQL database. The segments are mainly the document text and the section of references (Bollacker, Lawrence & Giles, 1998, p. 118). The idea is to make the documents available for various ways of searching e.g. keywords, cited documents, cocitations and documents originating from the same website etc.

The index is in many ways similar to ISI's Science Citation Index, but they distinct themselves from the ISI index by being more comprehensive on the number of journals and more up to date on the current literature (e.g. by retrieving articles from proceedings available online before they are published in print) (Lawrence, Bollacker & Giles, 1999b, p. 141).

When an item has been retrieved and is shown on the screen, information about how it is connected to other documents (e.g. cited documents, cocited documents, citing documents etc.) in the database is shown at the same time, providing an easy access for further browsing among related documents (Lawrence, Bollacker & Giles, 1999b, p. 143).

The potentials for using the Researchindex for webometric studies are quite obvious, but the types of analyses would be very similar to the ones we allready know from using informetrics within the ISI databases, since the documenttypes are very similar in nature. One major advantage seen in contrast to the ISI databases, is the broad coverage of various journals, and not to be limited to perform analyses that have a bias towards anglo-american journals as it is known to be in the ISI databases.

Further, The Researchindex has the advantage, that it is based on an open source code, and the software is available at no cost for non-commercial users. This way, their program could be a valuable base to build new implementations on with the focus for a use within the webometric studies.

4.4.3 Proposal for employing backlinkdata into servers of webpages

Chakrabarti, Gibson and McCurley had the idea, that if one could get access to the backlinks (ingoing links) to a webpage, it would add significant value to the information discovery for users on the webpage. Further, it would give a possibility to discover less easy accesible webpages e.g. new webpages, that mainly have outgoing links, but no ingoing links (Chakrabarti, Gibson & McCurley, 1999c, p. 1). If compared to Broder's bow tie (2000) (see figure 8), then the program is able to reveal webpages within the 'IN' group and the 'SCC' group. The user is no longer restricted to be surfing from the left towards the right side of the model, but is also able to go from the right to the left.

They invented an applet displaying two windows to be running along a webbrowser. The upper window is displaying the users current location (the URL) and stores the links to webpages that have been visited, while the lower window displays the ingoing links to the page. To begin with, the program was produced to retrieve search results from HotBot, but was planned later to be based on results retrieved directly from the server of the webpage. The server would contain a record of the backlink information from the Referer field[9] in the header of the HTTP protocol, based on previous visits on the page (Chakrabarti, Gibson & McCurley, 1999c, p. 2).

The backdraw of their idea is, that it requires all servers to collect and redistribute information derived from the Referer header of the HTTP protocol, something that may not be approved and implemented by all website owners (e.g. if a webpage has many ingoing links that are critical, the website owner may not want this to be displayed) (Chakrabarti, Gibson & McCurley, 1999c, p. 3-4).

Their own user tests showed for some topics, that the use of backlink navigation produced measurable improvement in the quality of the information, that was discovered. (Chakrabarti, Gibson & McCurley, 1999c, p. 12).

Seen from a webometric viewpoint, the program will have potentials on the micro level of webometric studies. It could be a valuable tool for investigating the link topology for selected websites and maybe reveal certain patterns in types of structures - structures that could further be investigated through qualitative studies to uncover motivations for hyperlinking in this particular way.

4.4.4 Link-agent, the program to reveal reciprocal links

A final tool, that could be used for webometric studies is a program we've invented for this project. We needed a tool, that was independent of the existing search engines, and that could reveal some of the more particular networks that exist around a webpage. We have

[9] "The Referer field allows the client to tell the server the URL of the document that triggered this request, permitting savvy servers to trace clients through a series of requests" (Baccala, 1997).

therefore started to develop the program Link Agent, a program that can reveal which of the links on a given webpage that are also reciprocal links.

The program has been developed in the programming language Visual Basic and has been produced in collaboration with programmer Lars Kamp Mortensen, an employee at the DTV (Technical knowledge Center of Denmark).

The program can reveal the types of reciprocal links that are defined in figure 4 and 5 in section 3.2.1 (page 32). That is, either true reciprocal links going exactly between two webpages, or reciprocal links that covers three webpages. The last type that includes four different webpages, as illustrated on figure 6 in section 3.2.1 is not possible to discover using the Link Agent.

The latest beta-version of the program is enclosed in the back of this paper. Due to a lack of time for developing the program, it is though a very simple version, and errors may occur during use.

The program has two main functions:

- Finding reciprocal links between two selected webpages.
- A generation of a report that displays a total of which links from the selected webpage that are also reciprocal links.

The program consists of two browserwindows (see figure 9). The first window is the one that displays the selected webpage. E.g. when typing in the URL: www.dsr.dk and pressing the 'Analyse this!' button using the mouse. The first browserwindow then displays the selected webpage, and the window to the right displays a list of the external links (outgoing links) from the webpage. In this case, the programs definition of an outgoing link, is based on a comparison between the domainname of the selected webpage (displayed right below the first browserwindow) and the start of the URL on each link on the webpage. This will in some cases result in links being displayed in the 'External links' window, that are not actually outgoing links. E.g. a link from www.dtv.dk to lrc.dtv.dk will be interpreted as an external link, despite its location is actually within the same organisation.

Below the first browser is additionally two small windows showing the total number of links on the webpage ('URL's'), and the total number of outgoing links ('URL' ext.').

The second browser window displays the webpage that has been selected by marking one of the links in the 'External links' window from browser one. To the right of the second browser a window displays all the links that have been found on the webpage.

Below the second browser three small windows are displayed. One is counting the total number of links on the webpage ('Total links'), and the other two are showing the total number of links that point back to the domain ('Links back to domain') or webpage ('Links back to page') for browser one. If the page is only pointing back to the domain of the

webpage in browser one, but not directly to the webpage, it is not a true reciprocal link, but belongs to the definition as illustrated in figure 5 in section 3.2.1 (page 32).

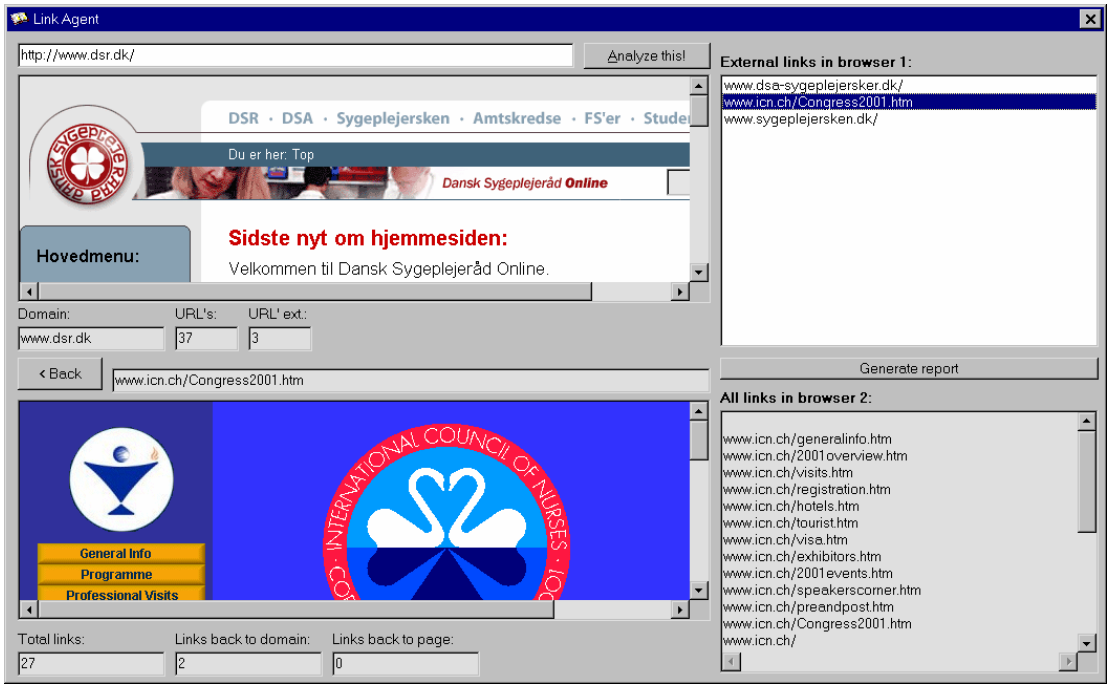

Figure 9: Illustration of the program Linkagent. The first browser displays the selected webpage (www.dsr.dk), and the second browser displays the webpage (www.icn.ch/Congress2001.htm) that has been marked in the 'External links' window for browser one.

The second function for the program is a generation of a report (see figure 10). When clicking the button 'Generate report' a new window pops up, and a routine going through each of the external links begins. The routine is similar to the one that happens in browser number. For each external link in browser one, a report is written for the number of links the webpage has back to the domain or webpage in browser one. At the end of the report, a sum of the total number of webpages, that have a link back to either the domain or the webpage, is displayed.

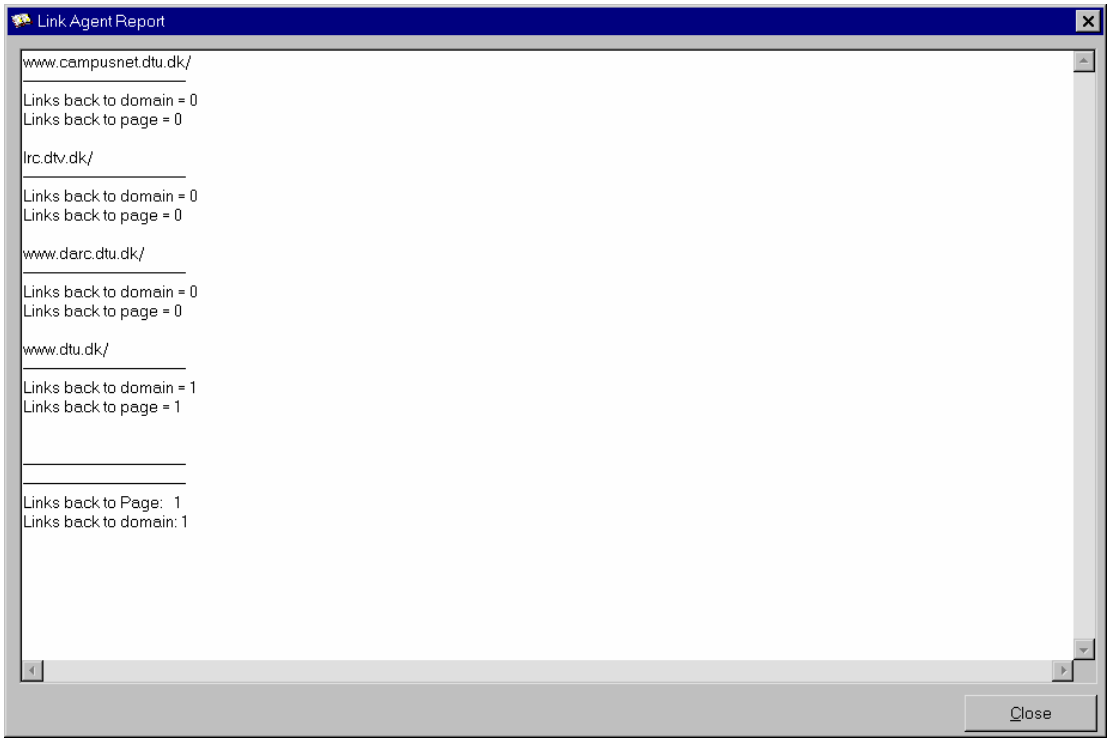

Figure 10: A generation of a report from Linkagent, based on the webpage www.dtv.dk. The window shows the relation to each of the webpages it has an outgoing link to. At the bottom is a sum of the total number of the webpages that have a link back to either the domain or the webpage (in this case, the domain name is equal to the URL of the webpage).

The Link Agent has not yet been used for any empirical studies, since it is not yet developed to function fully. The program still has a lot of bugs, that need to be worked on, in order to obtain a proper function. Currently, the program can only function on very simple made webpages. Webpages that are made with frames or other kinds of more advanced technology (flash etc.) prevents the program from extracting links on the webpage. The same problem currently exists, when a generated report is made. The program will stop in its generation of the report, when it finds a dead link or if the link is pointing to a webpage with frames, flash etc.

It is a major advantage, that the program can work independently from the commercial search engines that currently exist on the Internet. The program operates directly on the webpages as they appear at the time of the examination, and we can truly say that all the webpages within the network are accessible, and not limited by the more or less random and opaque indexing methods applied by the search engines.

If the program is developed to function on all types of webpages, it will have a good potential to be applied as a webometric tool. A tool that can assist in the discovery of reciprocal links. Finding out where they exist, and whether they occur more often on some types of webpages than others (e.g. more on private sites than organizational sites). We could make qualitative analyses based on the foundings, and examine if the reciprocal links exist by coincidence or, if they are created due to a mutual agreement between the owners

of the webpages. We could make graphs of the different types of networks and present it to members of the network, and see if they could recognize their place in the network. This would give us indications on, to what extent the owners of the webpages are aware of the network that they are a part of. Further, by interviewing the owners of the webpages, we will learn more about the motives behind the reciprocal links. Learn what kinds of information we can conclude about the network, when we find that two webpages have reciprocal links to each other.

Another area to explore, is whether the program can be a useful tool for searching webpages on a certain subject. Webpages being members of webrings on different subjects allready use this phenomenon for searching related sites on different subjects.

At the moment, the use of reciprocal links is not a widespread use. That was also shown in the results of the analysis in this paper, but the future may bring a difference for this area. Different articles have allready discussed the valuable possibilities for improving traffic to websites by exchanging links between companies and organisations (Banks, 2000; Ellegaard, 2000). This could in the future lead to even more new areas for an application of the full developed Link Agent.

4.5 Main conclusions on search engines and tools for webometric studies

Seen in the light of the knowledge about the linktopology of the web and the knowledge about the deficiencies of the current search engines, we can only draw the conclusion, that in order to make valid and reliable webometric analyses, we need to work on inventing our own data collection tools. Tools, that we have a knowledge about how function, and have a knowledge about for what types of analyses they can be applied and for which they can't. That is the next step we need to take within the webometric domain.

Chapter 5 - Conclusions

The focus for this paper has been to outline the most important rules to keep in mind before performing webometric analyses.

Chapter two contained a study of the current literature about whether or not we can speak of a citation theory and a link theory. It was shown, that making informetric analyses on a macro-level can be done due to the existence of a globally normative theory of citations within especially the scientific domains, and due to a peerreview process that controls, that the norms are being observed. A possible similar normative theory of links has still not grown to a steady level, which makes it impossible to perform webometric analyses on a macrolevel. Instead, we have suggested to keep the webometric analyses on a micro-level, and that they should always be followed by a qualitative analysis, indicating the motivations for linking on the selected websites.

In chapter three, we proposed a methodology to uncover motivations for linking, based on previous studies that have been made to uncover motivations for making references. The methodology was put into practice by an examination, to uncover the motivations for making hyperlinks on institute's websites and researcher's websites at The Royal School of Library and Information Science in Denmark and at selected institute's websites at The Technical University of Denmark.

The first empirical findings by applying the methodology showed, that the researchers primarily make links that are related to their CV's (projects the've been involved with, board member positions, and institutes they've cooperated with), while linking to personal websites to researchers within their scientific network did hardly exist. This was primarily due to a lack of time, and due to the need for frequent updates, when peers change to new jobs, or their websites move to different addresses.

Due to the few responses from the institutes at both institutions, we could not say, if there were any major norms for making hyperlinks on each of the two types of institutes.

The examination also showed, that it is adviced to use personal interviews in preference of mailed questionnaires. Especially to uncover possible complex citer motivations, as have been proved by Brooks for references in scientific articles (Brooks, 1986).

We recommended to use the proposed methodology for qualitative analyses in advance, when one wants to perform quantitative webometric analyses.

It was shown, that the type of network that can be concluded on, varies quite a bit, depending on whether the analysis is based on discovering motivations for references or links. Even within each one of them, differences do occur, and it is therefore important, to continue making qualitative analyses. This way we can uncover the major motivations for making the references or links, and we can safely go ahead and make quantified analyses,

draw conclusions on the data, and be assured, that they are based on a methodological correct foundation, because we posses the knowledge of the nature of the quantified citations and links within the specific area.

When comparing the uncovered motivations for linking with some of the major types of motivations for making references, it was clear, that motivations for linking are made on a much more varied foundation than for making references. References are made within a scientific context and restricted by norms for making references, while links are made within a much broader context, and usually to ease users navigation. In short, when writing a scientific paper, it is mandatory to make references to other documents, while producing a website gives you the opportunity to make links.

We further recommended, that due to the possible effects of linking, webmasters should take precautions and start making linkpolicies, and give some serious thoughts to the location of where they would like to be within the network of links.

The last chapter discussed the various tools available for implementing webometric studies. It was shown, that the current commercial search engines are not valid for a foundation for webometric analyses. They do not cover all webpages on the Internet, their indexing rules and frequencies are opaque, and their offers to use Boolean operators and field operators are useless.

Seen in the light of the knowledge about the linktopology of the web and the knowledge about the deficiencies of the current search engines, we can only draw the conclusion, that in order to make valid and reliable webometric analyses, we need to work on inventing our own data collection tools. Tools, that we have a knowledge about how function, and have a knowledge about for what types of analyses they can be applied and for which they can't.

The chapter was rounded off by a review of other types of tools that could make good alternatives to the current search engines. We see great possibilities in the future developments of Kleinbergs HITS algorithm (1998), Chakrabarti, Gibson and McCurley's proposal for the implementation of backlinkdata into servers of webpages (Chakrabarti, Gibson & McCurley, 1999c) and Lawrence, Giles and Bollacker's Researchindex (Lawrence, Bollacker & Giles, 1999b; Giles, Bollacker & Lawrence, 1998). Developments that they either could propose themselves, or peers within the domain could take an active part in future proposals for development.

Finally we proposed the program Link Agent, that we have developed for this paper. A tool to discover reciprocal links between webpages. We see many future potential applications for the idea in this program, such as to discover whether reciprocal links occur more often on some types of webpages than others, make qualitative analyses based on the foundings, and examine if the reciprocal links exist by coincidence or, if they are created due to a mutual agreement between the owners of the webpages. We could make graphs of the different types of networks and present it to members of the network, and see if they

could recognize their place in the network. This would give us indications on, to what extent the owners of the webpages are aware of the network that they are a part of. Further, by interviewing the owners of the webpages, we will learn more about the motives behind the reciprocal links. Learn what kinds of information we can conclude about the network, when we find, that two webpages have reciprocal links to each other.

When conducting webometric analyses, one part is to be aware of the specific nature of motivations for making hyperlinks within different types of websites, and the other part is to be aware of the quality of the search engines and other data collection tools that are being used. These two basic elements constitutes the foundation for performing valid and reliable webometric analyses, and working within the informetric domain, conducting webometric analyses, we have an obligation to investigate the nature of the motivations for making hyperlinks and assure the quality of the data collection tools that are applied.

5.1 Main statements to keep in mind when doing webometric research

Our research has shown, that our statements on doing webometric research is of high importance. The statements focus on two aspects for these kind of research:

1. Make sure to collect a good knowledge of the websites that are to be studied. The websites and their links are the basic elements of the webometric research. This knowledge could be collected through qualitative investigations using the proposed methodology, uncovering what types of links they have chosen and the primary motives behind these choices.

2. Allways make sure to know the tool completely before using it. Seek information about its ways of indexing, its limitations, its coverage of the Internet etc.

5.2 Proposal for future work to be done

Two main proposals should be outlined for future work to be done within the domain:

Start making more qualitative research about the motivations for hyperlinking. The Internet is too enormeous for us to be able to make a single or a few unified statements about why different websites choose to make hyperlinks. We need to uncover the hidden motivations and main differences for making hyperlinks, that exist among these many different types of websites.

Develop better tools that are produced to be used within the informetric domain, so we can be much more sure about the accurate boundaries and limitations for our research, when performing webometric studies.

References

Agosti, M; Melucci, M. (2000). Information retrieval on the web. In: ESSIR 2000. European Summer School in Information Retrieval, September 11-15, 2000 - Villa Monastero, Varenna, Italy.

Albert, R.; Jeong, H.; Barabási, A-L. (1999). Diameter of the World-Wide Web. In: Nature. Vol. 401, p. 130-131. http://www.nd.edu/~networks/Papers/401130A0.pdf (04/03/01)

Allan, J. (1996). Automatic hypertext link typing. In: Proceedings for the Hypertext '96 conference. p. 42-52, March. Washington, D.C.: ACM.

Almind, T.C. (1997). Lænker på World Wide Web : - lænker set som citationer på WWW. Copenhagen: Royal School of Library and Information Science. (Master thesis).

Almind, T.C.; Ingwersen, P. (1997). Informetric analyses on the world wide web : methodological approaches to 'Webometrics'. In: Journal of Documentation. Vol. 53, no. 4, p. 404-426.

Andersen, I. (1998). Den skinbarlige virkelighed : -om valg af samfundsvidenskabelige metoder. Frederiksberg C: Samfundslitteratur.

Baccala, B. (1997). Connected: An Internet Encyclopedia. 3rd ed. http://www.freesoft.org/CIE/index.htm and http://www.freesoft.org/CIE/Topics/102.htm (04/05/01)

Balslev, A; Fugl, L.D. (1999). Cocitationer & Bibliografisk Kobling : en sammenlignende analyse af metodernes anvendelighed til afdækning af fagområder, belyst med et eksempel i faget Information Science & Library Science. Copenhagen: Royal School of Library and Information Science. (Hovedopgave).

Banks, P. (2000). Give and take : our E-commerce marketing expert shows you how to increase traffic with reciprocal links. In: Entrepreneur.com. May 5th. http://www.entrepreneur.com/Your_Business/YB_PrintArticle/0,2361,274248-----,00.html (03/19/01)

Bar-Ilan, J. (1999). Search Engine Results over Time : a case study on search engine stability. In: Cybermetrics. Vol. 2/3, no. 1, paper 1. http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html (01/20/01)

Bar-Ilaan, J. (2001). Data collection methods on the web for informetric purposes - A review and analysis. In: Scientometrics. Vol. 50, no. 1, p. 7-32.

Bergman, M. (2000). The Deep Web : Surfacing Hidden Value. USA: BrightPlanet, the Internet content company. (White Paper).
http://128.121.227.57/download/deepwebwhitepaper.pdf (09/28/00)

Bharat, K.; Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. In: Computer Networks and ISDN Systems. Vol. 30, p. 379-388.

Björneborn, L.; Ingwersen, P. (2001). Perspectives of Webometrics. In: Scientometrics. Vol. 50, no. 1, p. 65-82.

Bollacker, K.D.; Lawrence, S.; Giles, S.L. (1998). CiteSeer : an autonomous web agent for automatic retrieval and identification of interesting publications. In: K.P. Sycara, M. Woolridge (eds). Proceedings of the 2nd international Conference on Autonomous Agents. p. 116-123. New York: ACM Press.
http://www.neci.nj.nec.com/~lawrence/papers/cs-aa98/cs-aa98.pdf (10/20/00)

Broder, A. et al. (2000). Graph structure in the web. In: Proceedings of the 9th International World Wide Web Conference. Amsterdam, Netherlands. http://www9.org or
http://www.almaden.ibm.com/cs/k53/www9.final/ (03/14/01)

Brooks, T.A. (1985). Private acts and public objects : an investigation of citer motivations. In: Journal of the American Society for Information Science. Vol. 36, no. 4, p. 223-229.

Brooks, T.A. (1986). Evidence of complex citer motivations. In: Journal of the American Society for Information Science. Vol. 37, no. 1, p. 34-36.

Cano, V. (1989). Citation behavior : classification, utility, and location. In: Journal of the American Society for Information Science. Vol. 40, no. 4, p. 284-290.

Carr, L.; Hall, W.; Miles-Board, T. (2000). Writing and reading hypermedia on the web. United Kingdom: Univ. of South Hampton. (Technical Report No. ECSTR-IAM00-1).
http://www.bib.ecs.soton.ac.uk/data/3368/html/WRWH.html (04/04/01)

Chakrabarti, S. et al. (1999a). Hypersearching the Web. In: Scientific American. no. 6, June. http://www.sciam.com/1999/0699issue/0699raghavan.html (10/20/00)

Chakrabarti, S. et al. (1999b). Mining the Web's Link Structure. In: Computer. Vol. 32, no. 8, p. 60-67.

Chakrabarti, S.; Gibson, D.A.; McCurley, K.S. (1999c). Surfing the Web Backwards. In: Proceedings of The Eigth International World Wide Web Conference. Toronto, Canada.
http://www8.org/w8-papers/5b-hypertext-media/surfing/surfing.html (01/30/01)

Clever Project, The; Project overview. [not dated].
http://www.almaden.ibm.com/cs/k53/clever.html (04/04/01)

Cozzens, S.E. (1989). What do citations count? the rethorical-first model. In: Scientometrics. Vol. 15, no. 5-6, p. 437-447.

Cybermetrics. available: http://www.cindoc.csic.es/cybermetrics (04/04/01)

Egghe, L.; Rousseau, R. (1990). Introduction to informetrics : quantitative methods in library, documentation and information science. Amsterdam: Elsevier Science Publishers.

Ellegaard, M. (2000). Klumme : strategisk linksamarbejde. In: SOL ComON. May 14th. http://www.comon.dk/20/printview.asp?ID=5765 (12/28/00)

Fano, R.M. (1956). Information theory and the retrieval of recorded information. In: Documentation in action. New York: Reinhold Publ. Corp., p. 238-244.

Frankfort-Nachmias, C.; Nachmias, D. (1996). Research Methods in the Social Sciences. 5th ed. London: Arnold.

Garfield, E. (1989). Citation behaviour : -an aid or a hindrance to information retrieval?. In: Essays of an information scientist: creativity, delayed recognition, and other essays. Vol.12, p. 123-128. (Current contents. Vol. 18, p. 3-8)

Garfield, E. (1998a). From citation indexes to informetrics : is the tail now wagging the dog? In: Libri. Vol. 48, p. 67-80.

Garfield, E. (1998b). Random thoughts on citationology : its theory and practice. In: Scientometrics. Vol. 43, no. 1, p. 69-76.

Gibson, D.; Kleinberg, J.; Raghavan, P. (1998). Inferring Web Communities from Link Topology. In: Proceedings 9th ACM Conference on Hypertext and Hypermedia. http://www.cs.cornell.edu/home/kleinber/ht98.pdf (02/03/01)

Giles, C.L.; Bollacker, K.D.; Lawrence, S. (1998). CiteSeer : An automatic citation indexing system. In: I. Witten, R. Akscyn and F. Shipmann III (eds.). Digital Libraries 98 : Third ACM Conference on Digital Libraries. p. 89-98.
http://www.it-uni.sdu.dk/mmp/Library/BollackerEtAlCiteSeer99.pdf (20/10/00)

Haas, S.W.; Grams, E.S. (1998). A link taxonomy for web pages. In: Proceedings of the 61st ASIS annual meeting. Vol. 35, p. 485-495. Medford, NJ: Info. Today.

Hellevik, O. (1997). Forskningsmetode i sosiologi og statsvitenskap. 5th ed. Oslo: Universitetsforlaget.

Hjortgaard Christensen, F.; Ingwersen, P. (1997). Online determination of the journal impact factor and its international properties. In: Scientometrics. Vol. 40, no. 3, p. 529-540.

Ingwersen, P. (1995). Information and Information Science. In: Encyclopedia of Library and Information Science. Vol. 56, suppl. 19, p. 136-174. Allen Kent & Carolyn M. Hall (editors). New York: Marcel Dekker, Inc.

Ingwersen, P. (1998). The calculation of web impact factors. In: Journal of documentation. Vol. 54, no. 2, p. 236-243.

Inktomi Corporation, NEC Research Institute (2000). Web Surpasses One Billion Documents. Press release issued January 18th. http://www.inktomi.com/new/press/2000/billion.html and http://www.inktomi.com/webmap/ (01/20/01)

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. In: American Documentation. Vol. 16, no. 3, p. 179-184.

Kessler, M.M. (1963). An experimental study of bibliographic coupling between technical papers. In: IEEE transactions on information theory. PTGIT IT-9, p. 49-51.

Kim, H.J. (2000). Motivations for hyperlinking in scholarly electronic articles : a qualitative study. In: Journal of the American Society for Information Science. Vol. 51, no. 10, p. 887-899.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In: 9th proceedings of the annual ACM-SIAM symposium on discrete algorithms. p. 668-677. Full version available at: http://www.cs.cornell.edu/home/kleinber/ (02/03/01)

Larson, R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. ASIS96. Available: http://sherlock.berkeley.edu/asis96/asis96.html (04/04/01)

Latour, B. (1987). Science in action : how to follow scientists and engineers through society. Open University Press, Milton Keynes.

Lawrence, S.; Giles, C.L. (1998). Searching the World Wide Web. In: Science. Vol. 280, p. 98-100.

Lawrence, S.; Giles, C.L. (1999a). Accessibility of information on the web. In: Nature. Vol. 400, p. 107-109.

Lawrence, S.; Bollacker, K.; Giles, C.L. (1999b). Indexing and Retrieval of Scientific Literature. In: Eighth International Conference on Information and Knowledge Management, CIKM 99. Kansas City, Missouri. p. 139-146. http://www.neci.nec.com/~lawrence/papers/cs-cikm99/cs-cikm99.pdf (10/20/00)

Leydesdorff, L. (1998). Theories of citation?. In: Scientometrics. Vol. 43, no. 1, p. 5-25.

Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? : discussion of sociological interpretations of citation analysis. In: Scientometrics. Vol. 38, no. 1, p. 27-37.

MacRoberts, M.H.; MacRoberts, B.R. (1986). Quantitative measures of communication in science : a study of the formal level. In: Social Studies of Science. Vol. 16, p. 151-172.

MacRoberts, M.H.; MacRoberts, B.R. (1989). Problems of citation analysis : a critical review. In: Journal of the American Society for Information Science. Vol. 40, no. 5, p. 342-349.

MacRoberts, M.H.; MacRoberts, B.R. (1996). Problems of citation analysis. In: Scientometrics. Vol. 36, no. 3, p. 435-444.

McKiernan, G. (1996). CitedSites(sm): Citation Indexing of Web Resources. http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm (03/21/01)

Marshakova, I.V. (1973). A system of document links constructed on the basis of citations (according to the "Science Citation Index"). In: Nauchno-Tekhnicheskaya Informatisiya: Scientific and Technical Information Processing. Series 2, no. 6, p. 49-57 (p. 3-8 in the Russian edition.)

Merton, R.K. (1979). Foreword, p. vii-xi. In: E. Garfield. Citation indexing : -Its theory and application in science, technology, and humanities. New York: John Wiley & Sons. (Information Sciences Series).

Moravcsik, M.J.; Murugesan, P. (1975). Some results on the function and quality of citations. In: Social studies of Science. Vol. 5, p. 86-92.

NEC Research Institute. (n.d.). Terms of Service. http://citeseer.nj.nec.com/terms.html (05/03/01)

Notess, G. (2000a). Search Engine Inconsistencies. In: Online. Vol. 24, no. 2. http://www.onlineinc.com/onlinemag/OL2000/net3.html (06/13/00)

Notess, G. (2000b). Search Engine Statistics : Database overlap. In: Search Engine Showdown : The user's guide to web searching. February 21st. http://www.searchengineshowdown.com/stats/overlap.shtml (02/02/01)

Notess, G. (2001a). Search Engine Statistics : Database total size estimates. In: Search Engine Showdown : The user's guide to web searching. April 7th. http://www.searchengineshowdown.com/stats/sizeest.shtml (02/02/01)

Notess, G. (2001b). Search Engine Statistics : Relative size showdown. In: Search Engine Showdown : The user's guide to web searching. April 7th. http://www.searchengineshowdown.com/stats/size.shtml (02/02/01)

Rousseau, R. (1997). Sitations : an exploratory study. In: Cybermetrics. Vol. 1, no. 1, paper 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html (04/04/01)

Skovmark, H. (2001). Cyberkrig på boligmarkedet. In: Bitconomy, April 2nd. http://www.bitconomy.dk/nyheder.asp?id=1952 (04/04/01)

Small, H. (1973). Co-citation in the scientific literature : a new measure of the relationship between two documents. In: Journal of the American Society for Information Science. Vol. 24, no. 4, p. 265-269.

Small, H. (1998). Letter to the editor : Citations and consilience in science. In: Scientometrics. Vol. 43, no. 1, p. 143-148.

Snyder, H.; Rosenbaum, H. (1999). Can search engines be used as tools for web-link analysis? : a critical review. In: Journal of Documentation. Vol. 55, no. 4, p. 375-384.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. In: Information Processing & Management. Vol. 28, no. 1, p. 1-3.

Thomas, O.; Willet, P. (2000). Webometric analysis of departments of librarianship and information science. In: Journal of Information Science. Vol. 26, no. 6, p. 421-428.

Van Raan, A.F.J. (1998). In matters of quantitative studies of science : The fault of theorists is offering too little and asking too much. In: Scientometrics. Vol. 43, no. 1, p. 129-139.

Vinkler, P. (1987). A quasi-quantitative citation model. In: Scientometrics. Vol. 12, no. 1-2, p. 47-72.

White, H.D.; McCain, K.W. (1998). Visualizing a discipline : an author co-citation analysis of information science, 1972-1995. In: Journal of the American Society for Information Science. Vol. 49, no. 4, p. 327-355.

White, M.D.; Wang, P. (1997). A qualitative study of citing behavior : contributions, criteria, and metalevel documentation concerns. In: Library Quarterly. Vol. 67, no. 2, p. 122-154.

Zuckerman, H. (1987). Citation analysis and the complex problem of intellectual influence. In: Scientometrics. Vol. 12, no. 5-6, p. 329-338.

APPENDIX A

Templet for questions that were asked to selected researchers with a personal website at The Royal School of Library and Information Science in Denmark, and to the webmasters of the selected institute websites at The Royal School of Library and Information Science in Denmark and Technical University of Denmark. The questions were originally asked in Danish, but have been translated for the purpose of this paper.

1, What was the primary purpose and target group for the website?

2, You have the following outgoing links on your website. Please indicate the motivations you had for choosing each of them to be a hyperlink on your site:
outgoing link 1, outgoing link 2, outgoing link 3 etc.

Since outgoing links were very limited at the institutes at the Royal School of Library and Information science, question two was substituted for this question:
2, There are no outgoing links on the institutes website. Is this a deliberate choise? (Please enlarge your answer)

3, Do you correspond with researchers at other institutions? (e.g. using phone, e-mail, ordinary mail, personal contact etc.) (Please enlarge your answer)

4, When examining the outgoing links on your website (as outlined above), it appears that none of them / only a few of them are going directly to personal websites of researchers at other institutions.
Have you considered to make links directly to researchers personal websites within your own professional network? (Please enlarge your answer)

In connection to question four, an additional question was asked to the webmasters of the institutes:
4a, When examining the outgoing links on your website (as outlined above), it appears that none of them / only a few of them are going directly to websites of institutes at other institutions.
Have you considered to make links directly to websites of other institutes within the professional network? (Please enlarge your answer)

5, Do you think it could be relevant and useful for yourself, if researchers within your own network had more links to each others websites? (Please state the reasons for your answer)

6, Do you think it could be relevant and useful for yourself, if researchers within another scientific network had links to each others websites? (Please state the reasons for your answer)

By making a search in the search engines AltaVista and Google, I have found that the following webpages contains a link to your website:
webpage 1, webpage 2, webpage 3 etc.
7, Do you know of any additional webpages that have a link to your website? (If yes, please indicate their address)

8, Do you immediately recognize the webpages that contain a link to your website?

9, Are you surprised by the number of webpages containing a link to your website? (Please state the reasons for your answer)

10, Can you define the major categories of types of webpages, that link to your site? (e.g. conferences, institutions, private webpages etc.) (Please enlarge your answer)

11, Are you surprised, that none/only a few of the webpages, that link to your website are personal webpages? (Please state the reasons for your answer)

The interchange of links between researchers (or institutes) within your own research domain can be of valuable use for e.g. researchers from a different domain or students that are not yet fully confidential with the domain, and who may be visiting your website in order to achieve a more indeepth knowledge of your scientific network with other researchers (or institutes), and by that way, acquire ideas for supplementary litterature.
12, Have you considered to make an interchange of links with researchers within your own research domain? (Please state the reasons for your reasons)

13, Are there any websites that you have deliberately not made a link to? (Please enlarge your answer)

14, Do you consider a link to be something positive to give and receive? (Please enlarge your answer)