Mining Author Cocitation Data with SAS Enterprise Guide

Mining Author Cocitation Data with SAS Enterprise Guide

By

Sean B. Eom

Cambridge
Scholars
Publishing

Mining Author Cocitation Data with SAS Enterprise Guide

By Sean B. Eom

This book first published 2015

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-7422-1
ISBN (13): 978-1-4438-7422-9

TABLE OF CONTENTS

PREFACE

The earlier version of my author cocitation analysis (ACA) book (Eom 2009) is based on SAS version 9.1. SAS continuously upgraded and released new versions. With the increasing use of the SAS Enterprise Guide, I realized the need for a new ACA book using SAS Enterprise Guide. Each computer today is part of global telecommunication network. Prior to the era of client server computing, all computing activities (input, process, and output) were performed by a single computer. The essence of client server computing is the division of labour between two or more networked computers. Input and output activities are handled by the client computers, while processing activities are performed by the server computers.

SAS Enterprise Guide is a Windows client application installed on your personal computer or laptop, which communicates with SAS. SAS software can be installed on the user's local computer or it can be installed on another server computer. SAS Enterprise Guide is a visual and customizable user interface that provides users with many features needed to access the functionality of SAS to access and process data. The term "user interface" in information systems is often used as a subsystem using menus, icon-based graphical user interface (GUI), or other types of input (e.g., command-driven) to allow the user to communicate and interact with other parts of the information system. The user of SAS Enterprise Guide always communicates and interacts with SAS Enterprise Guide, which in turn communicates with SAS software. As you perform specific tasks including statistical analysis (factor analysis, cluster analysis, etc.) with your data set, you need to write SAS code, and SAS Enterprise Guide also generates SAS code. When you run a task, the generated and user-written codes are sent to SAS for processing and the results are returned to SAS Enterprise Guide for you.

The second important aspect of the revision is the new chapter on visualization of ACA analysis with introduces a social network analysis tool, UCINET/NetDraw. Traditionally, ACA includes multidimensional scaling and plotting to visualize the intellectual structure of an academic field. Over the past decade, we have seen an increasing use of new visualization software as part of ACA analysis. As part of scientometrics, which we briefly introduced in this book, a new interdisciplinary field of science mapping has emerged to enrich ACA visualization tools and techniques. Science mapping tools include Microsoft Academic Search, HistCite by Thompson Reuters, Gephi, and many others. Moreover, due to an increasing popularity and use of social networks in our society, the study of social networks and social network analysis has gained popularity. This book introduces a social network analysis tool, UCINET.

ACA is is a subfield of informetrics, which is a broader term referring to the quantitative study of retrieval and processing bibliometric data collected from all types of communication media including journals, books, conference proceedings, and electronic communication media. The development of the Internet has expanded the scope of bibliometrics into the new areas of webometrics, cybermetrics, technometrics, and scientometrics. The terms bibliometrics, librametry scientometrics, and informetrics are frequently used interchangeably. Now, the library and information science area seems to have accepted “informetrics” as the umbrella term enveloping all subfields of study of all the quantitative aspects of various information resources including journals, books, and information resources on the World Wide Web and the Internet.

This book focuses on ACA, which is one of the research methodologies that transcend the individual field of inquiry. Despite its usefulness and capabilities that reveal a larger vista hidden in the bibliographic databases, ACA has not been a popular research tool in some academic disciplines including management information systems areas. The huge body of knowledge that exists today is the result of a cumulative research tradition. It is necessary to identify, examine, and trace the intellectual linkage within a given academic field as a basis of assessing the current state of its field to guide future development. These intellectual linkages can be systematically examined by means of counting and analyzing the various facets of intellectual activity outputs in the form of written communications.

This book covers all essential ACA topics for graduate students and researchers who want to learn the basics of ACA and the research techniques and tools to delineate the intellectual structure of various academic disciplines, compare cumulative research traditions, demonstrate theoretical differences between competing approaches, and to trace a paradigm shift in various academic disciplines over time. The basics of ACA included in the book cover the step-by-step procedures of ACA using the factor, cluster, multi-dimensional scaling procedures, and visualization tools using SAS Enterprise Guide and UCINET/NetDraw.

I would like to thank Southeast Missouri State University for granting me a sabbatical leave during the fall semester of 2014 to write this book. This book is dedicated to the two individuals who have changed the course of my life. I am indebted perpetually to them for their influence on me.

Jong Ye Won, my mother who is in heaven
Dr. Sang M. Lee, my doctoral advisor and mentor

References

Eom, Sean B. 2009. *Author cocitation analysis: quantitative methods for mapping the intellectual structure of an academic discipline*. Hershey, PA: Information Science Reference.

CHAPTER ONE

AN INTRODUCTION TO BIBLIOMETRICS AND INFORMETRICS

Author cocitation analysis (ACA) is a branch of bibliometrics. Bibliometrics/informetrics is one of the older areas of library and information science. The terms bibliometrics, scientometrics, and informetrics are frequently used synonymously. This chapter briefly overviews bibliometrics, including basic concepts, scope, and study areas of bibliometrics. The areas of study cover bibliometric distribution, citation and cocitation analyses, and library use studies. This chapter also discusses assumptions, purposes, benefits, limitations, and criticism of ACA.

Introduction

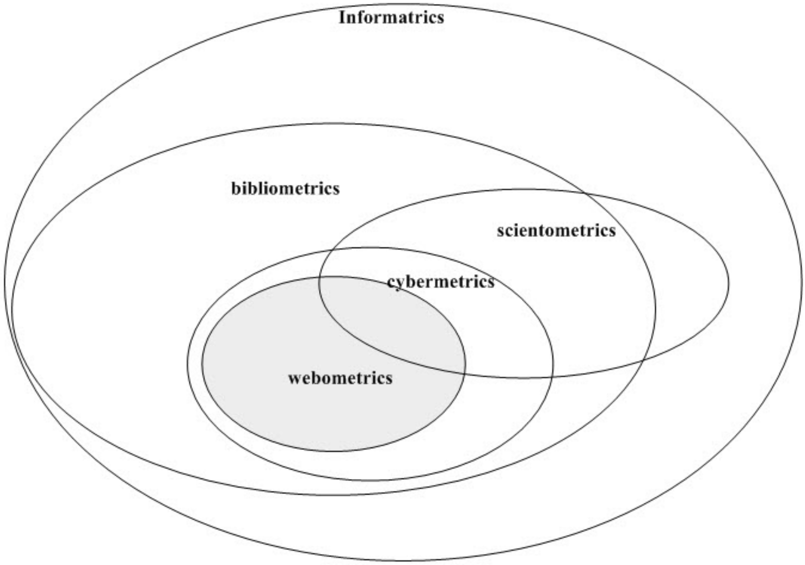
The library and information science (LIS) field consists of informetrics, bibliometrics, scientometrics, cybermetrics, and webometrics. The terms bibliometrics, librametry scientometrics, and informetrics are frequently used interchangeably. Even in the late 1980's, all these terms were not clearly distinguishable from each other. Wormell described the chaotic state of terminologies and the acceptance of the term "informetrics" this way (Wormell 1998, p. 258):

> The individual identities of the subfields "bibliometrics", "informetrics", "scientometrics" and "technometrics" are unfortunately not very clear, and there is chaos in the terminology. At the 1987 international conference some thoughts were given to changing the name of the discipline to "informetrics", and since the late 1980s there has been some support for use of this term. But alongside or parallel with this, both "bibliometrics" and "scientometrics" are frequently used terms. The field is becoming a scientific discipline including all the statistical and mathematical aspects connected with library, documentation and information problems with strong links to the theoretical aspects of information retrieval.

As figure 1.1 shows, nowadays, informetrics is a broader term that encompasses electronic communication of media including the Internet and World Wide Web, books, and journals. Informetrics is defined (Tague-Sutcliffe 1992) as "the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists." The development of the Internet has expanded the scope of bibliometrics into electronic communication media. These new areas are often called webometrics and cybermetrics. Scientometrics is the application of quantitative tools to the study of scientific communications (Leydesdorff 2001).

This chapter aims to provide a bird's-eye view of bibliometrics/ informetrics. This chapter provides basic concepts and the scope of bibliometrical studies. The last section discusses the purposes, assumptions, and limitations of ACA.

Figure 1.1: Relationships among Many Subfields of Library and Information Science Fields



Source: Björneborn & Ingwersen (2004, p. 1217).

What is Bibliometrics (Statistical Bibliography)?

The term statistical bibliography was coined by E. Wyndham Hulme (1923). The purposes of statistical bibliography are:

> 1. to shed light on the processes of written communication and of the nature and course of development of a discipline (in so far as this is displayed through written communication), by means of counting and analyzing the various facets of written communications (Prichard 1969).
>
> 2. the assembling and interpretation of statistics relating to books and periodicals … to demonstrate historical movements, to determine the national or universal research use of books and journals, and to ascertain in many local situations the general use of books and journals (Raisig 1962).

Pritchard (1969) suggested using the term bibliometrics instead of statistical bibliography. He believed that the term statistical bibliography was vague and could be confused with statistics itself or bibliographies on statistics. According to Pritchard, bibliometrics is defined as "the application of mathematics and statistical methods to books and other media of communication."

Scope of Bibliometric Studies

The huge body of knowledge existing today is the result of research publications in the form of journal articles, conference proceedings, books, etc. According to Ravichandra Rao (1983, p.216), bibliometric techniques are extensively used in the identification of trends in subjects such as the identification of core journals and the patterns of library use. They are also used to build models of the study of scientific communication. Most of these models are tested and used primarily at the local level (institutional level) to:

1. Describe scientific productivity;
2. Describe the growth of publications;
3. Identify core journals;
4. Weed out documents;
5. Identify the patterns of library use.

Of the characteristics of documents which have been hypothesized in library use studies, the following are of particular interest:

1. The age of documents—the number of years since they have been published, or the number of years since they have been available for use in a library;
2. The number of citations to documents;
3. Past usage of a given document—the number of times it is circulated or number of times it is used in the library.

The basic units of bibliometric studies are authors and documents (journal articles, conference proceedings and books). The trends and patterns of scientific communications can be detected by analyzing (quantitatively as well as qualitatively) the aggregated periodical data. Ravichandra Rao (1983, p.179) defines bibliometrics and libarametry as an area in which one studies: "Information process and information handling in libraries and information centers by quantitatively analyzing the characteristics and behavior of documents, library staff, and library users." The study areas of bibliometrics and libarametry include bibliometric distribution, citation analysis, library use studies, etc.

Bibliometric Distribution

One of the sub areas in bibliometric research is distribution. The study of bibliometric distribution has led to the following important laws in bibliometrics. They are Lotka's law of scientific productivity, Bradford's law of scatter, and Zipf's law of word occurrence. The term "law" used in bibliometrics is to be interpreted differently from immutable laws found in the physical sciences. According to Wolfram (2003), the term "law" is used by informetricians in its loosest sense to describe a mathematical generalization of an observed regularity in information.

Zipf's Law on Word Frequency

Philologist G. K. Zipf (1935, 1949) found that the relationship between the frequency of a word within a document and its rank can be represented as:

$$r \times f = C$$

where r is the rank of a given word,
f is its frequency, and
C is a constant.

If the words contained within a lengthy document are listed in order of decreasing frequency, the rank of a word on that list multiplied by the frequency in the document equals a constant. See (Wolfram 2003) for detailed descriptions, extensions, and variations of this model. Potter used a simple example of this relationship in Zipf's Law using an analysis of James Joyce's Ulysses by Zipf (Potter 1988).

> He showed that the tenth most frequent word occurred 2,653 times, the hundredth most frequent word occurred 265 times, the two hundredth word occurred 133 times, and so on. Zipf found, then that the rank of the word multiplied by the frequency of the word equals a constant that is approximately 26,500.

Lotka's Law on Productivity of Authors

Alfred Lotka, chemist and mathematician, analyzed the number of publications that appeared in *Chemical Abstracts* during the period of 1907 to 1916. Based on the computation of the theoretical frequencies of publications of authors using the least square method, Lotka suggested the following inverse square law of scientific productivity (Lotka 1926). The following two equations are taken from (Ravichandra Rao 1983).

$$y_x = 6/(\pi^2 x^\alpha)$$
$$x = 1, 2, 3. \ldots. , \alpha > 0$$

where y_x denotes the relative frequency of authors publishing x number of papers. This equation can be rewritten in the following form:

$$y_x = k/ x^\alpha$$
$$x = 1, 2, 3. \ldots. , k = 6/\pi^2 \text{ for } \alpha = 2$$

k and α are constants depending on the specific field; α is approximately 2.

Lotka's law says that the number of authors who wrote n papers in a discipline over time is proportional (equal) to $1/ n^2$.

> The proportion of all contributors that make a single contribution is about 60% of all authors contributied to a field ($60\%/1^2$).

> The proportion of all contributors that make two contributions to a discipline over time is about 15% of all authors contributed to a field ($60\%/2^2$).

The proportion of all contributors that make three contributions is about 6.7% of all authors contributed to a field (60%/3^2).

The proportion of all contributors that make four contributions is about 3.75% of all authors contributed to a field (60%/4^2).

Bradford's Law of Core and Scatter in Journals

Samuel C. Bradford was a mathematician and librarian at the Science Museum in London. He formulated a general relationship between the number of articles published in a given field and the distribution of the journal that published articles in that field (Bradford 1934, 1948). It was Vickery (Vickery 1948) who coined the term, "Bradford's Law of Scattering". Later Garfield also coined *Garfield's Law of Concentration,* which specifically addresses the differences in demand for scientific journals.

In layman's terms, Bradford's law says that in a well-established field of study, a small number of journals publishes a sizeable portion of the total publications in that area and an increasing number of journals publish fewer and fewer articles in the area. Bradford found that based on the study of a bibliography of geophysics, 9 journals contained 429 articles, 59 contained 499 articles, and 258 contained 404 articles.

Bradford classifies all journals in a given field into three categories (Bradford zones). The first (core) zone contains a core of a few journals. The second zone contains more journals than the number of the first zone. The third zone contains the rest of the journals in that field. The core contents of Bradford's law can be stated as follows.

- The ratio among journals in the three zones is found to be about 1: n: n2 , where n is referred to as the Bradford multiplier.
- Each zone is found to publish approximately the same number of articles.

Library Use Studies

Library use studies is one of three main bibliometrics used to measure the adequacy of a library collection. Then future library programs can better serve and satisfy the needs of the users. Further, they aim to formulate mathematical models for patterns of library use in relation to different types of users and documents. These studies include the analysis of circulation statistics, obsolescence study in the use of documents over time, study of the relationships between circulation and acquisition, etc.

Readers are referred to pp. 201-215 of (Ravichandra Rao 1983) for the overview of these research areas.

Researchers build on each others' and their own previous work. Definitions, topics and concepts are shared and interesting lines of inquiry need to be continuously followed up. To facilitate the progress of an academic field, it is important to build such a cumulative research tradition. In this process of knowledge creation, it is necessary to identify, examine, and trace the intellectual linkages to each other in a given academic field as a basis of assessing the current state of its field to guide future development. The intellectual linkages are established through the process of referencing and citation. These intellectual linkages can be systematically examined by means of counting and analyzing the various facets of intellectual activity outputs in the form of written communication.

Citation Analysis

Knowledge creation and dissemination in a discipline are facilitated through the circulation of ideas among "invisible colleges" (Crane 1972). Each individual contributes to the body of knowledge by building on what others have already accomplished. In this process, referencing and citation are important tools to link one''s writing with others. The majority of published research is never cited. Citation researchers are interested in identifying the patterns of how published articles are read and cited over time. According to Derek John de Solla Price (1965, p.511), an information scientist who is credited as the father of scientometrics:

> It seems that, in any given year, about 35 percent of all the existing papers are not cited at all, and another 49 percent are cited only once (n=1). This leaves about 16 percent of the papers to be cited an average of about 3.2 times each. About 9 percent are cited twice; 3 percent, three times; 2 percent, 4 times; 1 percent, five times; and a remaining 1 percent, six times or more.

Citation analysis can be basically classified into two types. The first type is the counting of citations of a document or set of documents authored by an individual without considering intellectual linkages. The second is the co-citation analysis of authors or documents to identify intellectual linkages among authors/publications. For examples of the first type of analysis, see (Eom and Lee 1993, Eom 1994). Such citation analysis is often used to compare the research productivity of an individual faculty member and/or a university's academic program measured by citation counts.

Citation analysis of specific journals created the concept of the impact factor, which is defined as the number of citations to the journal divided by the number of articles published in that journal. To be specific, the impact factor for a journal is calculated based on a three-year period. It is the average number of times published papers are cited up to two years after publication. For example, the impact factor for year 2005 for a journal is computed as A/B:

Where A = the number of times articles published in
2003-4 were cited in indexed journals
during 2005;

B = the number of articles, reviews, or notes
published in 2003-4.

The next type is the cocitation analysis of multiple authors or multiple documents, which was developed under the name of "co-mentions analysis" in 1968 (Rosengren 1968, Rosengren 1990). Systematic analysis of co-citation can be done using many different methods including bibliographic coupling, document co-citation analysis, author co-citation analysis, and co-word analysis. Since the primary focus of the book is author cocitation analysis, we will focus on only one of the many available tools here. For a methodological review of these four different methods, see Baker (1990).

Document Co-citation Analysis

There are two primary types of cocitation analysis to map the intellectual structure of an academic field: document cocitation analysis and author cocitation analysis (ACA). Document cocitation analysis involves the analysis of a set of selected documents (e.g., journal articles, books, proceedings, etc.) in terms of which pairs of documents are cited together. Readers are referred to (Garfield 1979) for a detailed description of the process of document co-citation analysis. Small and his colleagues at the Institute for Scientific Information (ISI) conduct their research on the document co-citation clustering and mapping techniques using ISI citation databases (Small and Griffith 1974, Griffith et al. 1974, Small and Garfield 1985, Small and Sweeney 1985, Small, Sweeney, and Greenlee 1985). ACA and document co-citation analysis share all the technical and methodological procedures. The only difference is the unit of analysis. ACA's unit of analysis is the author, whereas document cocitation analysis uses the document as the unit of analysis.

Small (1973, p.265.) compared the differences between document co-citation and bibliographic coupling in this way.

> To be strongly co-cited, a large number of authors must cite the two earlier works. Therefore, [document] co-citation is a relationship which is established by the citing authors. In measuring [document] co-citation strength, we measure the degree of relationship or association between papers as perceived by the population of citing authors. Furthermore, because of this dependence on citing authors, these patterns can change over time, just as vocabulary co-occurrences can change as subject fields evolve. Bibliographic coupling, on the other hand, is a fixed and permanent relationship because it depends on references contained in the coupled documents. Co-citation patterns change as the interests and intellectual patterns of the field change.

Bibliographic Coupling

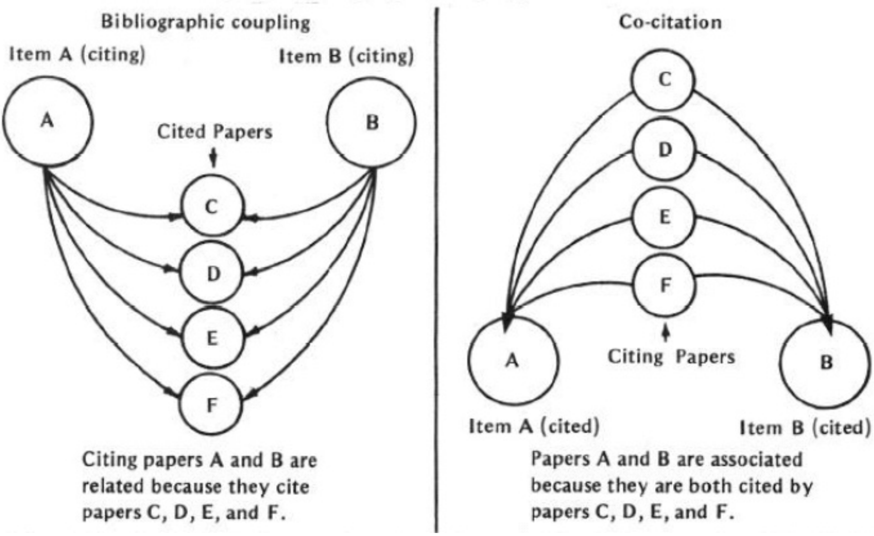
Bibliographic coupling is a technique for measuring the similarity of two *source* documents by counting the number of common bibliographic references (Kessler 1963). If documents share one or more bibliographic references (already published, of course), they are said to be *bibliographically coupled (connected)*. The strength of this connection is measured by the number of shared references. The more shared references they have, the stronger their connection is. On the other hand, document co-citation analysis counts the number of times two documents are to be cited together in *later* publications (see Figure 2). According to Garfield (2001), "Bibliographic coupling is retrospective whereas co-citation is essentially a forward-looking perspective."

Co-word Analysis

This is a technique of analyzing a set of documents to evaluate their strength of linkage by measuring the extent to which they share important key words or terms (Rip and Courtial 1984). The co-word analysis examines co-occurrences of key words and terms extracted from publication titles or their full text. The co-occurrences of key words measure the degree of cognitive linkages among a set of documents. The co-word frequency array (matrix) can be further analyzed via cluster analysis, multi-dimensional scaling, and network analysis to construct a co-word map (Callon et al. 1983).

There are two major differences in ACA, document co-citation, and co-word analysis. First, during a specific research period, co-citation analysis needs citing sources (citing documents, authors in citing documents) and cited references (cited authors, cited documents) of citing documents. However, co-word analysis requires only a set of journal articles in a specific research area such as decision support systems, information retrieval research, etc. Second, the content of the data matrix for each technique is different. The input to cocitation analysis is the author cocitation frequency matrix. On the other hand, document cocitation analysis and co-word analysis process the document co-citation frequency matrix and the co-word frequency matrix, respectively.

Figure 2: Bibliographic coupling vs. co-citation


Bibliographic coupling
Item A (citing)
Item B (citing)
A
B
Cited Papers
C
D
E
F
Citing papers A and B are related because they cite papers C, D, E, and F.
Co-citation
C
D
E
F
A
B
Citing Papers
Item A (cited)
Item B (cited)
Papers A and B are associated because they are both cited by papers C, D, E, and F.

Source: Garfield (2001)

The analysis process and tools are identical. All these techniques process these matrices using hierarchical clustering and multidimensional scaling to produce an empirical map of an academic discipline or a sub-discipline. Co-citation (document and author) analysis results do not provide the details of actual contents of all sub-specialties identified by co-citation analysis. But co-word analysis provides the content of research topics. For a good example of co-word analysis, see (Ding, Chowdhury, and Foo 2001).

Author Cocitation Analysis

ACA is a subarea of bibliometrics. ACA is a research tool whose idea originated in the late 1960s (Rosengren 1968). A series of papers from researchers at the College of Information Studies at Drexel University have made ACA a popular research tool in the area of library science (White and Griffith 1981, White 1983, White and Griffith 1982, White 1981). ACA, introduced in 1981, is a more general approach to identify, examine, and trace the intellectual structure of an academic discipline. This is done by counting the frequency with which any work of an author is cited by another author in the references of citing documents (Bayer, Smart, and McLaughlin 1990).

ACA, a major area of bibliometrics, is a technique that applies quantitative methods to various media such as books, journals, conference proceedings, and so on. ACA is "a set of data gathering, analytical, and graphical display techniques that can be used to produce empirical maps of prominent authors in various areas of scholarship" (McCain 1990). The cocitation of authors occurs when a citing paper cites any work of authors in reference lists. Many information scientists and author cocitation analysis researchers define an author as "a body of writings by a person" or "a body of contributions by a person." The term "contributions" may be better since it can include any type of contribution that can be cited as a reference, such as speeches delivered at professional meetings, personal communications including conversation and letters, and other media. The term "person" refers to a single author or one of multiple authors. These different uses of terms are related to the citation databases used in the study.

Most commercial citation databases and software access only the first author, regardless of the number of multiple authors, when retrieving author cocitation counts. This has been the critical weakness of using the commercial citation databases and software. However, this book is based on the bibliographic database I have created, which includes all contributions such as speeches delivered at various meetings and software we have developed that can access all multiple authors. With custom-built bibliographic databases, and the bottom-up approach of the selection of author sets, ACA becomes an exploratory tool for digging up the roots (reference disciplines), locating the trunk (foundations), and sifting through the branches (subspecialties) of a tree (an academic discipline). The critical element that makes ACA an exploratory tool is the custom bibliographic databases and the author selection method of screening the entire databases to finalize the author set for ACA analysis. For an overview and discussion of the continuing relevance of ACA to the study of the intellectual structure of literatures, see a special issue of *Journal of the American Society for Information Science*, vol. 41, no. 6, 1990. The issue contains a brief introduction by Howard D. White (Guest Editor) and a technical overview of the steps in ACA (McCain 1990).

Assumptions of Author Cocitation Analysis

Author cocitation analysis is based on the assumptions that "bibliographic citations are an acceptable surrogate for the actual influence of various information sources" (McCain 1986) and that the cocitation analysis of a field yields a valid representation of the intellectual structure of the field (Bellardo 1980, McCain 1984, 1990, Smith 1981). According to Bellardo (1980), the fundamental premise of cocitation analysis is that the greater the frequency a pair of documents/authors are cited together, the more likely it is that they are related in content. The cocitation frequency of authors represents relationships among authors. Authors whose works are cited together frequently are interpreted as having close relationships with one another. ACA is based on the assumptions that "cocitation is a measure of the perceived similarity, conceptual linkage, or cognitive relationship between two cocited items (documents or authors)" and "cocitation studies of specialties and fields yield valid representations of intellectual structure" (McCain 1986).

Purposes and Benefits of Author Cocitation Analysis

Citation analysis is often used to determine the most influential scholars, publications, or universities in a particular discipline by counting the frequency of citations received by *individual* units of analysis (authors, publications, etc.) over a period of time from a particular set of citing documents. However, citation analysis cannot establish relationships among units of analysis. ACA is the principal bibliometric tool to establish *relationships* among authors in an academic field. It can thus identify subspecialties of a field and how closely each subgroup is related to each of the other subgroups. By establishing relationships among authors, ACA provides a basis for revealing the intellectual structure of literature and defining the principal subject (major area of subspecialties in an academic discipline and their contributing disciplines) through the empirical consensus of numerous authors in an academic discipline.

In her landmark ACA-based research, which examined the intellectual evolution and development of the MIS area, Culnan (1986, p.156) discusses the importance of the study of the intellectual development of a field of study:

> Researchers in any academic discipline tend to cluster into informal networks, or "invisible colleges," which focus on common problems in common ways (Price 1963). Within these networks, one researcher's concepts and findings are soon picked up by another to be extended, tested and refined, and in this way, each person's work builds on that of another. The history of exchanges between members of these subgroups in a discipline describes the intellectual history of the field. .....
>
> Researchers can benefit by understanding this process and its outcomes because it reveals the vitality and the evolution of thought in a discipline and because it gives a sense of its future. In a relatively new field such as MIS, this understanding is even more beneficial because it identifies the basic commitments that will serve as the foundations of the field as it matures....

Limitations /Criticisms of Author Cocitation Analysis

ACA is a quantitative tool that cannot be used by itself to determine the intellectual structure of academic disciplines. This is a supporting quantitative tool that must be used with further qualitative analysis of bibliographic data. In regard to citation behavior of authors, Smith (1981, p. 84) enumerated fifteen reasons for citation based on the work of Garfield (1965).

1. Paying homage to pioneers
2. Giving credit for related works (homage to peers)
3. Identifying methodology, equipment, etc.
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticizing previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of fact-- physical constants, etc.
12. Identifying original publications in which an idea or concept was discussed
13. Identifying original publications or other works describing an eponymic concept or term

14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage).

While citation analysis can be a useful research tool due to its unobtrusive, precise, and objective characteristics, there are limitations of ACA stemming from the citation behavior of authors and bibliographic databases. Many problems can also arise in relation to the sources of citation data and mechanics of deriving citations from existing ISI citation indexes. Table 1.1 summarizes the problems of citation analysis. The table is taken from (MacRoberts and MacRoberts 1989).

Table 1.1: Event-Data Problem of Citation Analysis

1. Formal influences not cited.
2. Biased citing.
3. Informal influences not cited.
4. Self-citing.
5. Different types of citations.
6. Variations in citation rate related to type of publication, nationality, time period, and size and type of specialty.
7. Technical limitations of citation indices and bibliographies.
    a. Multiple authorship
    b. Synonyms
    c. Homonyms
    d. Clerical errors
    e. Coverage of literature

The technical problems consist of multiple authorship, self-citations, homographs, synonyms, unification problems, etc. (Lindsey 1980, Long, McGinnis, and Allison 1980, Smith 1981). The use of Social Science Citation Index (SSCI) and the Science Citation Index (SCI) can raise a potential problem since these sources can exhibit English language bias (Baker 1990). The use of custom databases and the cocitation matrix generation system we have developed can eliminate many of the problems discussed above.

MULTIPLE AUTHORSHIP

SSCI lists records by first author only. All authors except first authors will not be counted when compiling the cocitation frequency matrix. This has been a fundamental issue in ACA. Chapter 3 of Eom(2009) fully addresses this issue and concludes that the ISI convention of relying on only the name of the first author in assembling the cocitation matrix in the investigation of the intellectual structure of academic disciplines may often fail to identify all possible underlying factors.

NAME-HOMOGRAPHS

SSCI indexes only the author’s last name and initials. Consequently, citation records by an author of the same last name and initials may not be authored by the same person. In the case of common English surnames, such as Smith, Davis, and Williams, indexing only initials creates significant problems of name-homographs such as Smith G., Smith GA., Smith GN., Smith GD., Smith GR.

SYNONYMS

To further complicate matters, the same author’s initials can be recorded in many different ways. Some examples are Keen, P. or Keen, PGW., Lee, S. or Lee, SM. Furthermore, many individuals have the same names. Another problem is when the name changes, there is no easy way to handle this situation. The name of women authors can change when they marry, or some change their names for many different reasons.

UNIFICATION PROBLEMS

This problem of unification is concerned with the way each author’s name and the journal title in each cited record is entered into the citation index. In other words, journal titles are entered in many un-standardized ways. For example, *MIS Quarterly* can be entered into Citation Index as MIS Q, MISQ, MIS Quart., etc. due to the fact that some journals (e.g., *Omega*) use their own abbreviated journal name in the references.

COVERAGE OF LITERATURE

Due to the simple fact that SCI and SSCI do not cover all science and social science literatures, the use of these citation indices are undoubtedly problematic. See (MacRoberts and MacRoberts 1989) for a more detailed discussion of this topic.

Recent Development in Informetrics

Since the late 1990’s, a new subset of informetrics called webometrics/cybermetrics, has become part of the mainstream library and information science research area. In 2004, *Journal of the American*

Society of for Information Science and Technology published a special issue to discuss the emerging area (Thelwall and Vaughn 2004)

Björneborn and Ingwersen introduced a basic framework for webmatrics and provided a broad picture of relationships between the library and information science fields of informetrics, bibliometrics, cybermetrics, webometrics, and scientometrics as shown in figure 1 (Björneborn and Ingwersen 2004). The conceptual and terminological confusions of the emerging phenomena seem to be settling down. According to Björneborn and Ingwersen (2004, pp. 1216-1217),

> A range of new terms for the emerging research field were rapidly proposed from mid-1990s, for example, netometrics (Bossy 1995); webometry (Abraham 1996); internetometrics (Almind and Ingwersen 1996); webometrics (Almind and Ingwersen 1997); cybermetrics (journal started 1997 by Isidro Aguillo); Web bibliometry (Chakrabarti et al. 2002). This and similar more specific conceptual diversity and development often made (and make) it difficult to understand what actually is analyzed in the contributions. The transformation over a year from internetometrics to webmetrics by the same authors, Almind and Ingwersen (1996, 1997), is typical of conceptual confusion.

Cybermetrics

The term cybermetrics refers to the quantitative studies of the nature of scientific communication over the Internet and its impact on diffusion of ideas and formation, whereas bibliometrics aims to understand the communication process of authors using the analysis of journal articles to infer the intellectual structure of an academic discipline, and to assess the journal impact factor. Often webometrics and cybermetrics are used as synonyms. But Informatricians tend to agree that cybermetrics is a broader area that encompasses webometrics (Björneborn and Ingwersen 2004).

Cybermetrics is proposed as a generic term for the quantitative study of all Internet applications.

> The study of the quantitative aspects of the construction and use of information resources, structures, and technologies on the whole Internet drawing on bibliometric and informetric approaches (Björneborn 2004).

The coverage of cybermetrics includes the following:

- Statistical study of the World Wide Web and computer-mediated communication on the Internet (Herring 2002) such as discussion groups(Matzat 1998), mailing lists (Hernández-Borges, Pareras, and Jiménez 1997), usenet newsgroup (Bar-Ilan 1997), etc.

- Quantitative measure and analysis of the Internet backbone technology, topology, and traffic (Molyneux and Williams 1999).
- Analysis of Web contents, link structure, web-usage in information systems or computer science, etc. such as
    - Web ecology (Chi et al. 1998).
    - Cybergeography and cyber cartograph (Dodge 1999, Dodge and Kitchin 2001, 2002)
    - Web mining (Etzioni 1996, Cooley, Mobasher, and Srivastava 1997).
    - Web graph analysis (Broader et al. 2000).
    - Web dynamics (Levene and Poulovassilis 2001), and
    - Web intelligence (Yao et al. 2001).

Webometrics

Webometrics is proposed as a generic term for the quantitative study of the World Wide Web phenomena.

> The study of the quantitative aspects of the construction and use of information resources, structures, and technologies on the Web drawing on bibliometric and informetric approaches (Björneborn 2004)

The coverage of webometrics includes the following four main areas (Björneborn and Ingwersen 2004):

- Web link structure analysis/web colink analysis
- Web page content analysis
- Web usage analysis
- Web technology analysis

As figure 1.1 shows, cybermetrics refers to the quantitative studies of the Internet-related phenomena including discussion groups, mailing lists, computer-mediated communication, etc. Webometics is a subarea of cybermetrics which focuses on only the World Wide Web-related phenomena. Especially, the hyper-link is the core of webomerics. *Journal of American Society of for Information Science and Technology* published a special issue on Webometrics. Most of the articles in that issue investigated the issues surrounding hyperlinks (Thelwall and Vaughn 2004).

Web colink analysis (WCA) is an emerging field in webometrics. In webometrics, colink is established when two web pages "both have inlinks from a third pages" (Thelwall 2004, p.5). The link analysis is concerned with the analysis of inlinks and outlinks. Inlinks are defined as follows (anonymous 2008):

> Backlinks (or back-links (UK)) are incoming links to a website or web page. In the search engine optimization (SEO) world, the number of backlinks is one indication of the popularity or importance of that website or page (though other measures, such as PageRank, are likely to be more important). Outside of SEO, the backlinks of a webpage may be of significant personal, cultural or semantic interest: they indicate who is paying attention to that page.
>
> In basic link terminology, a backlink is any link received by a web node (web page, directory, website, or top level domain) from another web node (Björneborn and Ingwersen, 2004). Backlinks are also known as incoming links, inbound links, inlinks, and inward links.

The inlinks are classified into two types: internal inlinks and external inlinks. These two types of inlinks make up total inlinks. In ACA, the inclusion of self-citation has been an issue in analyzing and interpreting the ACA results. WCA may include links within the site itself or examine only external links. For the comparison of ACA and WCA, authors are referred to Zuccala (2006). Detailed comparisons are made between ACA and WCA in term of the following:

- Selecting author names and Web pages
- Retrieving cocitation frequency matrix and Web colink frequency matrix
- Mapping and Interpretation of ACA and WCA

Table 1.2 is constructed based on Zuccala (2006) and it highlights some differences between ACA and WCA. As the table indicates, these two tools are different from each other and do not share many things in common. One exception is that the data matrix is constructed by similar procedures and processed by the same multivariate statistical techniques such as factor analysis, cluster analysis, and multidimensional scaling techniques. Although differences in data sources, data currency, data selection, and data stability exist, the most critical difference occurs when interpreting the results of ACA and WCA. To interpret what each factor/cluster/dimension means, ACA and WCA need interpretation based on citation theory and link theory, respectively.

Table 1.2: Comparison of ACA and WCA

	Author Cocitation Analysis (ACA)	**Web Colink Analysis (WCA)**
Data source	SciSearch or Social ScieSearch Custom databases	World Wide Web data
Data selection	Highly cocited authors in a discipline or research area	Well-linked Web pages in the area of common themes such as business or academic pages
Currency of data	Historical	Up-to-the minute
Stability of cocitation/colink	Stable and reliable	Fluctuate daily
Retrieval of homonymous data	Homonymous data	No homonymous data
Inputs	Cocitation frequency matrix	Colink frequency matrix
Interpretation of outcomes (Maps)	Intellectual structures (cognitive linkages) of an academic field	Some types of Web structure (geography, mission, subject-area orientation, etc.)
Theory to interpret the results	Citation theory	Hyperlink theory

Conclusion

This chapter briefly introduced the concept of bibliometrics. The terms bibliometrics, scientometrics, and informetrics are frequently used interchangeably. Bibliometrics is the application of quantitative methods to communication media. Scientometrics is the application of quantitative tools to the study of scientific communications (Leydesdorff 2001). Informetrics is a subfield of information science. Nowadays, informetrics is a broader term that encompasses electronic communication media, books, and journals. We also briefly discussed the scope of bibliometric studies including bibliometric distribution, citation analysis, library use studies, cocitation analysis, coword analysis, and bibliographic coupling.

The second part of this chapter also discussed the origin, assumptions, purposes, benefits, limitations and criticisms of ACA. Technical limitations of ACA stem from the use of ISI citation indices. Of these various issues raised by MacRoberts & MacRoberts, most technical limitations can be effectively managed by developing custom databases and cocitation counts generation systems.

The third part of this chapter briefly overviewed the recent developments in ACA (2003).

References

Abraham, R.H. . 2004. *Webometry: Measuring the complexity of the World Wide Web.* Visual Math Institute, University of California at Santa Cruz 1996 [cited July 9 2004]. Available from http://www.ralph-abraham.org/vita/redwood/vienna.html.

Almind, T.C., and P. Ingwersen. 1997. "Informetric analyses on the World Wide Web: methodological approaches to “webometrics”." *Journal of Documentation* no. 53 (4):404-426.

Almind, T.C., and P. Ingwersen. 1996. Informetric analysis on the World Wide Web: A methodological approach to internetometrics. Centre for Informetric Studies: Royal School of Library and Information Science, Copenhagen, Denmark

Anonymous. 2008. *Backlink* [Wikipedia, the free encyclopedia]. Wikimedia Foundation Inc., 19 June 2008, at 19:05. 2008 [cited June 23 2008]. Available from http://en.wikipedia.org/wiki/Inlinks.

Baker, Donald R. 1990. "Citation Analysis: A methodological Review." *Social Work Research & Abstracts* no. 26 (3):3-10.

Bar-Ilan, J. 1997. "The "mad cow disease," usenet newsgroups and bibliometric laws." *Scientometrics* no. 39 (1):29-55.

Bayer, Alan E., John C. Smart, and Gerald W. McLaughlin. 1990. "Mapping Intellectual Structure of a Scientific Subfield Through Author Cocitations." *Journal of the American Society for Information Science* no. 41 (6):444-452.

Bellardo, T. 1980. "The Use of Co-Citations to Study Science." *Library Research* no. 2:231-237.

Björneborn, Lennart. 2004. *Small-world link structures across an academic Webspace: A library and information science approach.* Doctoral dissertation, Royal School of Library and Information Science Copenhagen, Denmark.

Björneborn, Lennart, and Peter Ingwersen. 2004. "Toward a basic framework for webometrics." *Journal of the American Society for Information Science and Technology* no. 55 (14):1216-1227.