Preface

Author cocitation analysis is a subfield of informetrics. Informetrics is a broader term that encompasses electronic communication of media including the Internet and World Wide Web, books, and journals. Informetrics is defined 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" (Tague-Sutcliffe, 1992). The development of the Internet has expanded the scope of bibliometrics into electronic communication media. These new areas are often called Webometrics, cybermetrics, technometrics, or scientometrics.

The terms bibliometrics, librametry scientometrics, and informetrics are frequently used interchangeably. Even in the late 1980s, all three terms were not clearly distinguishable from one another. The chaotic state of terminologies existed until the late 1980s. Now, the library and information science area seems to have accepted "informetrics" (Björneborn & Ingwersen, 2004; Wormell, 1998) as the umbrella term enveloping all subfields to study all the quantitative aspects of various information resources including journals, books, and information resources on the Web and the Internet.

This book focuses on a small spot regarding the study of informetrics and author cocitation analysis. The huge body of knowledge that exists today is the result of a cumulative research tradition. Researchers build on each other and their own previous work. Definitions, topics, and concepts are shared and interesting lines of inquiry need to be continuously followed up. In this process of knowledge creation, it is necessary to identify, examine, and trace the intellectual linkage to each other in 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.

Over the past 80 years, the way we count and analyze the citation frequency has dramatically changed from the early manual transcribing and statistical computation of citation data to computer-based citation data creation and its manipulation. The term statistical bibliography was coined by Hulme (1923) as a research tool

for examining the intellectual development and structure of an academic discipline. Since then, we have seen continuous development in the field of bibliometrics. The principal method of bibliometrics is citation analysis through counting and analyzing the citation frequencies. The most important milestone in the development of citation analysis was established by Garfield. He presented an idea for the management of scientific information using a comprehensive citation index in 1955 and three years later founded the Institute for Scientific Information (ISI) (Garfield, 1955). For a detailed description of theory and application of citation indexing, see (Garfield, 1979). A citation index is a listing of all referenced or cited source items published in a given time span associated with the citing articles. The Web version of citation index appeared in 1997 is Web of Science®. The Web of Science provides access to multidisciplinary citation index information from approximately 8,700 high impact research journals in the world.

Due to the rich information resources available today such as Web of Science®, bibliometric analysis researchers can easily access rich bibliographic information using the World Wide Web. There are two important recent developments in author cocitation analysis: The use of Pearson correlations coefficients, r, as a similarity measure and several new developments in ACA visualization tools such as Pathfinder networks (White, 2003), AuthorLink (Lin, White, & Buzydlowski, 2003), and VxInsight (Boyack, Wylie, & Davidson, 2002). Although there are some developments in applying common bibliometric methods to Web co-link analysis (Zuccala, 2006), Chapter I briefly discusses only two streams of developments in the ACA area.

THE AUDIENCE OF THIS BOOK

This book is for graduate students and researchers in any academic discipline who want to learn 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. Author cocitation analysis (ACA) is one of research methodologies that transcends 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. For example, in the area of management information systems, there are a total of 2,744 individuals listed in the database of MIS faculty directory. This service was developed and is operated by the Information and Decision Sciences Department and the MIS Research Center of the Carlson School of Management at the University of Minnesota. Of these 2,744, less than 10 researchers have conducted and published ACA research over the past four decades.

This book aims to open the vast expanse of wasteland. Considering the limited exposure of this research methodology to our area, this book covers all essential ACA topics for graduate students and researchers who want to learn the basics of ACA as well as recent developments in ACA such as controversial debates on proximity measurer, diagonal values in cocitation frequency matrix, visualization tools and techniques. The basics of ACA include how to retrieve cocitation frequency counts from online commercial bibliographic databases and how to build custom databases using spreadsheet and database management systems. The basics also include the step-by-step procedures of ACA using the factor, cluster, and multi-dimensional scaling procedures.

THE OBJECTIVE AND CONTRIBUTIONS OF THIS BOOK

This book introduces an alternative approach to conducting author cocitation analysis (ACA) without relying on commercial citation databases such as index ISI citation index. It is based on a custom bibliographic database and cocitation matrix generation systems specifically developed to use the custom database. The alternative approach can be an effective research tool overcoming several weaknesses of the commercial online data-based ACA research. The custom data-based ACA is not a replacement for the commercial data-based ACA. These two approaches are complementary to each other. Our approach clearly has advantages but its critical drawback is the time and effort needed to build the database.

First, the approach we are introducing here has the capability to access the non-primary authors of cited references. The non-primary authors refer to all authors other than the first author. The inability to access non-primary authors is a critical shortcoming of ACA research utilizing the commercial databases. Theoretically, the contributions made by non-primary authors must be counted when examining the intellectual structure of an academic discipline.

Second, strict criteria can be applied to the selection of citing articles. A researcher does not always write articles in a specialized field throughout his/her lifetime. Research interests can shift from one subspecialty area to other areas within an academic discipline. Custom bibliographic databases can be built to include only writings in a specific field. Custom database requires hard labor and a time-consuming process from selecting citing articles, entering cited references from the citing articles, and to maintaining the databases.

Third, the alternative approach effectively identifies the intellectual structure of an academic field and its reference disciplines more accurately as well. All previous ACA studies, except the ones conducted by Eom and his colleagues (Eom, 1996, 1998a, 1998b, 2002; Eom & Farris, 1996; Eom, Lee, & Kim, 1993), failed

to identify the reference disciplines of an academic field. The reason for the failure was the method used to select authors to use for ACA. The method starts with a predetermined list of authors selected by the subjective judgments of researchers. It is impractical for ACA researchers to include all authors in the reference disciplines of an academic field prior to conducting ACA analysis. If ACA researchers somehow managed to include authors in the reference disciplines of an academic field, ACA would produce empirical maps of prominent authors selected by the researchers. However, with the approach introduced in this book, ACA becomes an exploratory tool. It can dig up the roots (reference disciplines), locate the trunk (foundations of an academic discipline), and sift through 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 entire databases to finalize the author set for ACA analysis. This can be called the bottom-up approach. The majority of , if not all, ACA studies using commercial databases are based on the top-down approach – selecting authors applying the subjective judgments prior to ACA analysis. The end result of the top-down approach is simply clustering the subjective author set into several subgroups. With this approach, ACA is inherently a limited tool for identifying the changing structure of an academic field and tracing emerging/fading scholars.

Fourth, the custom databases can be built to include only writings in a specific domain/subspecialty. For example, if anyone wants to study the intellectual structure and main themes and reference disciplines used by the researchers who attended only the International Conference on Information Systems (ICIS), the existing commercial database cannot be used. The only way is to build a custom database from the proceedings of the ICIS. Building custom databases requires hard labor, making it a time-consuming process. However, there are important advantages in using custom databases. The ISI social science citation index includes bibliographic information, author abstracts, and cited references found in more than 1,700 scholarly social science journals covering more than 50 disciplines. To identify the intellectual structure of the decision support systems area, social science citation index-based research could possibly reach inaccurate results due to the technical limitations of ISI citation index files. In this case, building custom databases could be an effective approach.

Fifth, this book describes step-by-step ACA procedures for novice SAS users as well as SPSS. The SAS® system is an integrated system of software that provides complete control over data access, management, analysis, and presentation. The SPSS® is a statistical and data management package for analysts and researchers. This book provides explicit instructions to build bibliographic databases, compile a cocitation matrix, prepare SAS input files, and interpret the results. This book provides the reader with a useful, instructional guideline to conduct ACA research regardless of the bibliographic databases used; in-house databases or commercial

citation databases. With commercial citation databases, the cocitation matrix can be easily retrieved to create the bibliographic databases. After the retrieval of author cocitation counts, many steps and procedures must still be followed to accomplish the goals of ACA as shown in Figures 1 and 2 (Chapter VII). Each and every step is an unstructured process for those inexperienced researchers. This book is intended to help them conduct ACA research.

This book can also be useful for those who are not familiar with the three multivariate statistical techniques (factor analysis, cluster analysis, and multidimensional scaling). The book shows the entire procedure to prepare SAS data files, process them, and analyze the outputs. Some of the chore activities must be learned from trial and error, which is often time-consuming and frustrating. Even to those who are not ACA researchers, the book provides useful tips on each process of research using multivariate techniques. Although I have included the basics of SAS and SPSS programs for three multivariate statistical analysis techniques (factor analysis, cluster analysis, and multidimensional scaling), this introduction is not intended to give a comprehensive one-step guideline for ACA students. It is an introduction of multivariate statistical techniques using the SAS and SPSS systems to analyze cocited author counts. With this introduction, ACA students are in a better position to study SAS and SPSS language and procedures; SAS graph software, and SAS/STAT users' guide. The sample SAS and SPSS programs in the book are working programs that can be used with different data sets.

THE STRUCTURE OF THE BOOK

The book consists of five sections: Foundations, Fundamental Issues in ACA Online Data Retrieval, Alternative Approaches of Building Custom Databases, ACA Procedures, and ACA Applications.

Foundations

The first section, which includes one chapter, is concerned with the foundation of ACA. Chapter I provides readers with a big picture of bibliometrics and introduces ACA as a subfield of bibliometrics. Author cocitation analysis (ACA) is a branch of bibliometrics. Bibliometrics/informetrics is one of the older areas of information science research. This chapter briefly overviews the bibliometrics, including the basic concepts, scopes, and study area of bibliometrics. The area of study covers bibliometric distribution, citation, and cocitation analyses, and library use studies. The study of bibliometric distribution led to the invention of Lotka's law of scientific productivity, Bradford's law of core scatter in journals, and Zipf's law of word oc-

currence. The researchers in the citation and co-citation areas identify the pattern of how published documents are cited over time using many different approaches such as bibliometric coupling, document cocitation analysis, author cocitation analysis, and co-ward analysis. The last section briefly discusses the assumptions, purposes, benefits, limitations, and criticisms of ACA.

Fundamental Issues in ACA Online Data Retrieval

Section II consists of 3 chapters. Chapter II introduces the basics of the Institute for Scientific Information online data retrieval, using the Web of Science and Dialog Classic. The Web of Science provides access to multidisciplinary citation index information from approximately 8,700 high impact research journals in the world. Users can navigate to electronic full-text journal articles with complete bibliographic data, cited reference data, and direct links to the full text. A citation index, developed by ISI, is an alphabetical listing by author, of all the references found in footnotes and bibliographies of the journals covered in the index. This chapter overviews three search options: general search, cited reference search, and advanced search. The following section provides some useful information about the entire procedure to retrieve cocitation frequency counts using Dialog Classic and the free ONTAP® (ONline Training And Practice) site. This chapter points out several technical limitations of the ISI online citation index databases including multiple authorship: all citation index files permit retrieving records only by the last name and initials of the first author only. Another limitation is name-homographs: SSCI indexes only 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 author. Another limitation is synonyms: the same author's initials are recorded in many different ways. Some examples of synonym are Keen, P., or Keen, P. G. W., Lee, S. or Lee, S. M.

Chapter II introduces the first of the two fundamental and long standing issues in ACA using ISI online databases. The majority of ACA research has relied on the Institute for Scientific Information (ISI) citation databases. ISI convention allows only the retrieval of papers citing works of which the author is the first or sole author. Non-primary authors (authors whose name appear in second or a later position) will not be counted when assembling a cocitation frequency matrix. This chapter empirically examines the impact of the ISI convention on the results of ACA. Virtually all ACA studies use Thomson's ISI citation indexes that use the first author to retrieve the cocitation counts. Therefore, this has been a methodological issue in ACA study. First, literature survey is conducted to review what has been done to deal with this issue. Second, based on the survey of literature, we further argue that previous research has addressed and shed light on some parts of method-

ological issues. However, it had failed to address issues such as to what extent the use of a different approach has resulted in different outcomes in terms of an actual intellectual structure of a given academic discipline. Using our data and cocitation matrix generation systems, we compare the differences in the process and outcomes of using different cocitation matrices. Three conclusions can be reached based on our study. First, an all author-based ACA is better than first author-based ACA to capture all influential researchers in a field. Second, it identifies more subspecialties. Finally, an all author-based ACA and first author-based ACA produce little differences in stress values.

Chapter IV investigates the second of the two fundamental and long standing issues in ACA using ISI online databases. Diagonal values in the cocitation frequency counts matrix have been considered a fundamental issue in ACA study. Diagonal values are the cocitation frequency counts between the author and himself/herself. Finding the exact values of diagonal values in the co-citation matrix requires the manual procedure of examining the total number of contributions including journal articles, books, proceedings, and so forth. For that reason, ACA researchers suggested many different approaches to fill the diagonal cells in the cocitation matrix. They include the mean cocitation count, missing values, zeroes, highest off-diagonal counts, adjusted off-diagonal values, and the number of times cocited with himself/herself. The majority of ACA researchers prefer to use either the adjusted value approach by adding the three highest off-diagonal values and dividing by two or the missing value approach. This chapter empirically examines the impact of these different approaches on the ACA outcomes. Based on the results of this study, if the pure cocitation counts are not used, the next best alternatives are as follows. They are the missing value approach, mean cocitation value approach, and the highest off-diagonal value approach in the order of the highest total variance explained.

Alternative Approaches: Building Custom Databases

The third section of the book presents two other alternative approaches in Chapters V and VI that can be used to retrieve cocitation counts in lieu of using the ISI citation index files and Dialog Classic. Chapter V introduces the first, using a popular database management system, of the two alternative approaches to overcome the technical limitations associated with online cocitation counts retrieval using Dialog Classic and citation index files. Certainly Dialog Classic is an attractive alternative because the user is using the readily available bibliographic databases and retrieval software. The majority of ACA researchers have used ISI databases and Dialog Classic to retrieve cocitation counts. However, this approach has some technical limitations as discussed earlier. They include the issue of *Multiple Authorship, Name-Homographs, and Synonyms.* This chapter introduces an alternative approach

to retrieving co-citation counts from the custom databases through the system we have designed and implemented. Custom database and retrieval systems need time and investment for development, but they can manage most of the technical limitations discussed. This chapter introduces the fox-base approach, the first of the two, in developing custom databases and the cocitation matrix generation system. The first part is concerned with the design of databases. The second part describes the cocitation retrieval system. We also discuss how our system can eliminate or minimize the technical limitations of the Thomson ISI database and Dialog Classic Software system.

Chapter VI introduces the second alternative approach using a spreadsheet program, Microsoft Excel. McIntire (2007) invented this approach as part of his Master's thesis at the University of Columbia. His thesis is based on the International Textile and Apparel Association (ITAA) publication database. The motive for the design of database and cocitation counts system was simply that the ISI citation index files do not include the specific journal in the textile and apparel area. The chapter shows the design of databases and retrieval of cocitation counts using the spreadsheet based cocitation counts generation system.

ACA Procedures

Section IV deals with the procedures of ACA analysis and consists of 5 chapters. Chapter VII overviews several important steps in author cocitation analysis. ACA consists of the six major steps beginning with the selection of author sets for further analysis, collection and statistical analysis of the cocitation frequency counts, and the validation and interpretation of statistical outputs.

The remaining 4 chapters (VIII through XI) focus on statistical procedures using the SAS and SPSS systems. Chapter VIII describes principal component analysis using the factor procedure of the SAS system. The first section of the chapter begins with the definition of factor analysis. It is the statistical techniques whose common objective is to represent a set of variables in terms of a smaller number of hypothetical variables (factor). We also present many different approaches of preparing datasets including importing from external sources, manual data inputs, and in-file statements. We discuss each of the key SAS statements including DATA, INPUT, CARDS, PROC, and RUN. In addition, we examine several option statements to specify the following: method for extracting factors, number of factors, rotation method, and displaying output options.

Chapter IX describes the distance and cluster procedures of the SAS system. Cluster analysis is a data reduction technique for grouping various entities (e.g. individuals, variables, objects) into clusters so that the entities in the same cluster have more similarities with each other with respect to some predetermined selection

criteria. The first section of this chapter explains the creation of a distance matrix, which is the input to the cluster procedure. The second part of this chapter focuses on the PROC CLUSTER statement which sets out the CLUSTER procedure steps. This chapter includes the discussions of generations of a distance matrix, the PROC CLUSTER Statement, and interpreting results of cluster analysis.

Chapter X presents multidimensional scaling (MDS) procedures in the SAS system. MDS is a class of multivariate statistical techniques/procedures to produce two or three dimensional pictures of data (geometric configuration of points) using proximities among any kind of object as input. Three SAS procedures (MDS, PLOT, and G3D) are necessary to convert the author cocitation frequency matrix to two or three dimensional pictures of data. The distance matrix produced earlier by using xmacro.sas and distnew.sas programs in SAS version 8 or the DISTANCE procedure in version 9 is converted to a coordinate matrix, to produce two-dimensional plots and annotated three-dimensional scatter diagrams. This chapter also discusses how to label data points on a plot. The annotate facility in the SAS system produces figures with the name of the author on each data point. The PROC MDS procedure includes many of the features of the ALSCAL procedure.

Chapter XI briefly introduces the use of SPSS version 15.0 to conduct ACA analysis. The SPSS accepts data files in many different formats including spreadsheets, database files, tab-delimited, and other types of ASCII text files. Assuming that cocitation frequency counts are stored in a spreadsheet file in Excel, we demonstrate each step of ACA analysis to produce outputs using factor, cluster, and multi-dimensional scaling analyses.

ACA Applications in the MIS Area

Section V introduces an ACA study in the management information systems area to demonstrate some concepts that cannot be adequately explained with the smaller dataset used in prior chapters. Throughout this book, we use a small data set to demonstrate the step-by-step procedures of converting the dataset to the final ACA outputs. Advantages of using such a small number of variables include a clearer understanding of data preparation steps and an easier interpretation of outputs. On the other hand, a smaller data set may make it difficult to fully demonstrate the problems that can arise with a large number of variables such as scree plot, finding the optimal number of factors based on the factor interpretation, and so forth.

Section V has 2 chapters. Chapter XII infers the intellectual structure of the decision support systems (DSS) field by means of an empirical assessment of the DSS literature from 1969 to 1989. Three multivariate data analysis tools (e.g. factor analysis, multidimensional scaling, and cluster analysis) are applied to an author cocitation frequency matrix derived from a large database file of comprehensive DSS

literature over the same period. Seven informal clusters of DSS research subspecialties and reference disciplines were uncovered. Four of them represent DSS research subspecialties—foundations, group DSS, model/data management, and individual differences. Three other conceptual groupings define the reference disciplines of DSS—organizational science, multiple criteria decision making, and artificial intelligence. DSS is a very young academic field that is still growing. DSS has entered the era of growth after 20 years of research. During the 1990s, DSS research was further grounded in a diverse set of reference disciplines. Furthermore, it is in the active process of solidifying its domain and demarcating its reference disciplines.

The last chapter of the book, Chapter XIII, extends an earlier benchmark study (Eom, 1995), which examined the intellectual structure, major themes, and reference disciplines of decision support systems (DSS) over the last two decades (1969-1990). Factor analysis of an author cocitation matrix over the period of 1990 through 1999 extracted 10 factors, representing 6 major areas of DSS research: group support systems, DSS design, model management, implementation, and multiple criteria decision support systems, and 5 contributing disciplines: cognitive science, computer supported cooperative work, multiple criteria decision making, organizational science, and social psychology. We have highlighted several notable trends and developments in the DSS research areas over the 1990s.

REFERENCES

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

Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain Visualization Using Vxinsight for Science and Technology Management. *Journal of the American Society for Information Science and Technology, 53*(9), 764-774.

Eom, S. B. (1995). Decision Support Systems Research: Reference Disciplines and a Cumulative Tradition. *Omega: The International Journal of Management Science, 23*(5), 511-523.

Eom, S. B. (1996). Mapping the Intellectual Structure of Research in Decision Support Systems through Author Cocitation Analysis (1971-1993). *Decision Support Systems, 16*(4), 315-338.

Eom, S. B. (1998a). The Intellectual Development and Structure of Decision Support Systems (1991-1995). *Omega, 26*(5), 639-658.

Eom, S. B. (1998b). Relationships between the Decision Support System Subspecialties and Reference Disciplines: An Empirical Investigation. *European Journal of Operational Research, 104*(1), 31-45.

Eom, S. B. (2002). *Decision Support Systems Research (1970-1999): A Cumulative Tradition and Reference Disciplines*. Lewiston, New York: Edwin Mellen Press.

Eom, S. B., & Farris, R. (1996). The Contributions of Organizational Science to the Development of Decision Support Systems Research Subspecialties. *Journal of the American Society for Information Science, 47*(12), 941-952.

Eom, S. B., Lee, S. M., & Kim, J. K. (1993). The Intellectual Structure of Decision Support Systems (1971-1989). *Decision Support Systems, 10*(1), 19-35.

Garfield, E. (1955). Citation Indexes for Science. *Science, 122*, 108-111.

Garfield, E. (1979). *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. New York: Wiley.

Hulme, E. W. (1923). *Statistical Bibliography in Relation to the Growth of Modern Civilization*. London: Grafton.

Lin, X., White, H. D., & Buzydlowski, J. (2003). Real-Time Author Co-Citation Mapping for Online Searching. *Information Processing & Management, 39*(5), 689.

McIntire, J. S. (2007). *The Clothing and Textile Research Base: An Author Cocitation Study.* Unpublished Master's Thesis, University of Missouri, Columbia, Columbia, Missouri.

Tague-Sutcliffe, J. (1992). An Introduction to Informetrics. *Information Processing & Management, 28*(1), 1-3.

White, H. D. (2003). Pathfinder Networks and Author Cocitation Analysis: A Remapping of Paradigmatic Information Scientists. *Journal of the American Society for Information Science and Technology, 54*(5), 423-434.

Wormell, I. (1998). Informetrics: An Emerging Subdiscipline in Information Science. *Asian Libraries, 7*(10), 257-268.

Zuccala, A. (2006). Author Cocitation Analysis Is to Intellectual Structure as Web Colink Analysis Is To...? *Journal of the American Society for Information Science and Technology, 57*(11), 1487-1502.