A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions

A Thesis
Presented to
The Academic Faculty

By

Cherie R. Courseault

In Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy in
Industrial and Systems Engineering

Georgia Institute of Technology
May 2004

A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions

Approved by:

Professor Alan L. Porter, Advisor

Professor William B. Rouse

Professor Jye-Chyi Lu

Professor Donghua Zhu

Professor Milena Mihail

Date Approved March 7, 2004

EPIGRAPH

"Information in itself is silent; it is the use to which it is put, in terms of inferring, interpreting, projecting, analyzing, manipulating, computing and decision-making, that is important." Oskar Morgenstern (Jantsch, 1967, p95)

DEDICATION

This work is dedicated to my family and to my Lord and Savior, Jesus Christ. Thank you, mom and dad, for your ongoing support, in both natural and spiritual ways. Thank you, Gregory Trumbach, Jr., my husband, for your patience. Thank you, Anya, my daughter, for giving me the motivation to finish. Thank you, Lord, for giving me the wisdom, courage, and strength to overcome every obstacle.

Heavenly Father,

I pray that in finishing this degree, I am readily equipped to accomplish all the plans and purposes that you have for me.

In Jesus Name,

Amen

ACKNOWLEDGEMENTS

At this point, I am thoroughly amazed at how many people it actually takes to complete a dissertation. There are so many people who, without their assistance, I would not have been able to complete this document.

First and foremost, I would like to thank my advisor, Dr. Alan. Porter. It is because I had such a great experience working for him as a Master's GA that I even thought of pursuing a PhD. I would like to also thank others working on TOA: Webb and Doug for the hours of algorithm programming; Nils, Paul, and Bob for the opportunities provided along the way; Alisa, for all the errands, but mostly for hanging around the States until I also finished ☺; and Buddy, for so many things that I cannot name; however, the late night and last minute assistance sit at the top of the list.

I cannot fail to mention all my Atlanta friends, who have been so supportive throughout the years. I am afraid to make a list for fear of forgetting someone.

Thank you to all of those individuals at The Agency for providing a topical learning experience that I will never forget.

Thank you to all of those who gave me their time for interviews, questionnaires, and focus groups, especially USACERL, the National Finance Center, and the companies located in the UNO Research and Technology Park. Thank you to the faculty of the University of New Orleans, especially Olie for your support and patience and Sandy for being such a great sounding board and proofreader. There are many others that I could thank, but that might take a book by itself. As I said, "At this point, I am thoroughly amazed at how many people it actually takes to complete a dissertation."

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **v**

LIST OF TABLES **ix**

LIST OF FIGURES **xi**

LIST OF ABBREVIATIONS **xiii**

SUMMARY **xiv**

CHAPTER 1: INTRODUCTION **1**

CHAPTER 2: LITERATURE REVIEW **6**

2.1 Technical Intelligence **7**

2.1.1 The Benefits of Technical Intelligence *9*

2.1.2 Technical Intelligence Viewpoints *12*

2.1.3 Technical Intelligence Methodology *15*

2.2 Text Mining **30**

2.2.1 Retrieval *33*

2.2.2 Extraction *34*

2.2.3 Data Cleansing *35*

2.2.4 Data Mining *37*

2.2.5 Visualization *42*

2.3 Literature Review Conclusion **48**

CHAPTER 3: DESCRIPTION OF RESEARCH **50**

3.1 Step One: Determine the Technologies/Functions to be Monitored **53**

3.2 Step Two: Determine the Information Needs of the Technology Decision-Makers **56**

3.3 Step Three: Develop a Concept-Clumping Algorithm **57**

3.4 Step Four: Compare Keywords and Abstract Phrases Clusters **65**

3.5 Step Five: Determine Metrics for an Example Technology **68**

3.6 Step Six: Evaluate the Framework **70**

CHAPTER 4: FRAMING THE NEEDS OF THE TARGET USERS **72**

4.1 The Technologies/Functions to be Monitored **72**

4.1.1 Five Technology Cases *72*

4.1.2 The Target Audience *73*

4.2 The Information Needs of the Technology Decision-Makers **80**

CHAPTER 5: DATA PREPARATION **86**

5.1 The Concept-Clumping Algorithm **86**

5.1.1 Preparation *86*

5.1.2 Calculation Precision *88*

5.1.3 The Effect of the Algorithm *91*

5.2 Keywords and Abstract Phrases Clusters Comparison **98**

CHAPTER 6: METRIC FINDINGS AND EVALUATION **107**

6.1 The Metrics for an Example Technology **107**

6.1.1 General Organizational Monitoring *109*

6.1.2 Global Organizational Monitoring *111*

6.1.3 Hiring for Cutting Edge *113*

6.1.4 The Progress of the Technology *113*

6.2 Framework Evaluation **116**

CHAPTER 7: CONCLUSIONS AND FUTURE RESEARCH **123**

7.1 Information Products 123

7.2 Text Data Mining Methods 124

7.3 Future Research 125

APPENDIX A: FUNCTIONS AFFECTED BY CTI 129

APPENDIX B: INTELLIGENCE QUESTIONS IN THE PRODUCT LIFE CYCLE 132

APPENDIX C: INTERVIEW QUESTIONS 135

APPENDIX D: INFORMATION REQUIREMENTS QUESTIONNAIRE 137

APPENDIX E: MAP OF QUESTIONNAIRE STATEMENTS TO PUBLICATION METRICS 142

APPENDIX F: TRANSCRIPT OF FOCUS GROUP INTRODUCTION 144

APPENDIX G: QUESTIONNAIRE STATEMENT SUMMARIES 146

APPENDIX H: TECHNOLOGY CASES- ZIPF DISTRIBUTION GRAPHS 153

APPENDIX I: TECHNOLOGY CASES- CLUSTER MAPS 160

APPENDIX J: CLUSTER DATA CORRELATION MATRIX 175

REFERENCES 180

LIST OF TABLES

Table 3.1	Questionnaire Score Categories	56
Table 3.2	Sonochemistry Keywords vs. Abstract Phrases	59
Table 3.3	Comparison of Cohesion Measures	68
Table 4.1	Technology Cases: Record Counts	73
Table 4.2	Categorized Early Warning Topics	78
Table 4.3	Summary of Information Evaluation Criteria	79
Table 4.4	Participant Industry Groups	81
Table 4.5	Participant Profiles	81
Table 4.6	Questionnaire Score Categories	82
Table 4.7	Strong Agreement Survey Questions	82
Table 5.1	Hard Disk Drive Matches	88
Table 5.2	High Density Recording Matches	89
Table 5.3	Technology Cases: Clumping Algorithm Precision Calculations	91
Table 5.4	Fuel Cell Top 20 Abstract Phrases	92
Table 5.5	Remote Sensing Top 20 Abstract Phrases	94
Table 5.6	Magnetic Storage Top 20 Abstract Phrases	95
Table 5.7	GIS Top 20 Abstract Phrases	96
Table 5.8	Pollution Monitoring Top 20 Abstract Phrases	97
Table 5.9	Dataset Sample Sizes	98
Table 5.10	Quantitative Cluster Measures	101

Table 5.11	Quantitative Measures of Clusters: Comparison of Means	102
Table 6.1	Size of Research Teams: US vs. Foreign	112
Table 6.2	Analysis of Non-Technical Terms	115
Table 6.3	Focus Group Evaluation Results	119
Table 6.4	Focus Group Consensus Opinions	120
Table B.1	Intelligence Questions for Life Cycle Stages	132
Table E.1	Questionnaire Statements Mapped to Publication Database Metrics Calculated in VantagePoint	142
Table G.1	Questionnaire Statement Responses- Descriptive Summary Statistics	146
Table H.1	Fuel Cell Clumped Abstract Phrases Ranks and Frequencies	153
Table J.1	Quantitative Cluster Comparison Data: Correlation Matrix	175

LIST OF FIGURES

Figure 2.1	**Traditional Intelligence Cycle**	**16**
Figure 2.2	**The Herring Protocol**	**17**
Figure 2.3	**Technology Delivery System (TDS)**	**29**
Figure 2.4	**The Text Mining Process**	**31**
Figure 2.5	**Semantic Depth of Field**	**47**
Figure 3.1	**Research Information Flow**	**52**
Figure 3.2	**Participant Evaluation Email**	**54**
Figure 3.3	**Screenshot of VantagePoint**	**58**
Figure 4.1	**Questionnaire Statement Clusters**	**83**
Figure 5.1	**Remote Sensing Clumped Abstract Phrases Map**	**99**
Figure 6.1	**Magnetic Storage Home Page**	**108**
Figure 6.2	**General Organizational Monitoring**	**108**
Figure 6.3	**Magnetic Storage Conferences and Journals Web Page**	**109**
Figure 6.4	**Cross-Correlation Map**	**110**
Figure 6.5	**Global Monitoring**	**111**
Figure 6.6	**Global Activity Over The Years**	**112**
Figure 6.7	**The Leading Universities in Magnetic Storage Research**	**113**
Figure 6.8	**Cumulative Magnetic Storage Records**	**114**
Figure H.1	**Fuel Cell Zipf Distribution Graphs**	**155**
Figure H.2	**Magnetic Storage Zipf Distribution Graphs**	**156**

Figure H.3	**Remote Sensing Zipf Distribution Graphs**	**157**
Figure H.4	**Geographical Information Systems Zipf Distribution Graphs**	**158**
Figure H.5	**Pollution Monitoring Zipf Distribution Graphs**	**159**
Figure I.1	**Fuel Cell Keywords Cluster Maps**	**160**
Figure I.2	**Fuel Cell Cleaned Abstract Phrases Cluster Maps**	**161**
Figure I.3	**Fuel Cell Clumped Abstract Phrases Cluster Maps**	**162**
Figure I.4	**Magnetic Storage Keywords Cluster Maps**	**163**
Figure I.5	**Magnetic Storage Cleaned Abstract Phrases Cluster Maps**	**164**
Figure I.6	**Magnetic Storage Clumped Abstract Phrases Cluster Maps**	**165**
Figure I.7	**Remote Sensing Keywords Cluster Maps**	**166**
Figure I.8	**Remote Sensing Cleaned Abstract Phrases Cluster Maps**	**167**
Figure I.9	**Remote Sensing Clumped Abstract Phrases Cluster Maps**	**168**
Figure I.10	**Geographical Information Systems Keywords Cluster Maps**	**169**
Figure I.11	**Geographical Information Systems Cleaned Abstract Phrases Cluster Maps**	**170**
Figure I.12	**Geographical Information Systems Clumped Abstract Phrases Cluster Maps**	**171**
Figure I.13	**Pollution Monitoring Keywords Cluster Maps**	**172**
Figure I.14	**Pollution Monitoring Cleaned Abstract Phrases Cluster Maps**	**173**
Figure I.15	**Pollution Monitoring Clumped Abstract Phrases Cluster Maps**	**174**

LIST OF ABBREVIATIONS

AMA	American Marketing Association
CEO	Chief Executive Officer
CIO	Chief Information Officer
CTI	Competitive Technical Intelligence
IR	Information Retrieval
KDD	Knowledge Discovery in Databases
NLP	Natural Language Processing
NTIS	National Technical Information Service
R&D	Research and Development
TLC	Technology Life Cycle
TOA	Technology Opportunities Analysis
TPAC	Technology Policy and Assessment Center

SUMMARY

This research developed a comprehensive methodology to quickly monitor key technical intelligence areas, provided a method that cleanses and consolidates information into an understandable, concise picture of topics of interest, thus bridging issues of managing technology and text mining. This research evaluated and altered some existing analysis methods, and developed an overall framework for answering technical intelligence questions. A six-step approach worked through the various stages of the Intelligence and Text Data Mining Processes to address issues that hindered the use of Text Data Mining in the Intelligence Cycle and the actual use of that intelligence in making technology decisions. A questionnaire given to 34 respondents from four different industries identified the information most important to decision-makers as well as clusters of common interests. A bibliometric/text mining tool applied to journal publication databases, profiled technology trends and presented that information in the context of the stated needs from the questionnaire.

In addition to identifying the information that is important to decision-makers, this research improved the methods for analyzing information. An algorithm was developed that removed common non-technical terms and delivered at least an 89% precision rate in identifying synonymous terms. Such identifications are important to improving accuracy when mining free text, thus enabling the provision of the more specific information desired by the decision-makers. This level of precision was consistent across five different technology areas and three different databases. The result is the ability to use abstract phrases in analysis, which allows the more detailed nature of abstracts to be captured in clustering, while portraying the broad relationships as well.

CHAPTER 1

INTRODUCTION

Each year billions of dollars are spent on Research and Development projects. Over $275 billion was spent on R&D in 2001 alone (R&D Magazine, 2002). At a time when firms faced a few well-known competitors, informal intelligence gathering practices were sufficient. However, frequent changes in technology and increased competition, among other factors, means technology has been difficult to predict, while at the same time companies must now act quickly upon new innovations (Ashton and Klavans, 1997). Technology managers are faced with the challenge of identifying emerging technologies with the greatest economic potential. The project decisions faced by these decision-makers may involve basic science research, specific product development, or purchasing decisions. Which project or product should they support? In answer to these challenges, competitive intelligence efforts have arisen in recent years. It was not until 1986 that The Society of Competitive Intelligence Professionals was formed and the associated journal was initiated in 1990.

Competitive Intelligence (CI) is the organizational process for systemically collecting, processing, analyzing and distributing information about an organization's external environment to the people who need it (Hohhof, 1997). Many decisions affected by external forces, such as entering new markets and businesses, investing in and acquiring new technologies; making major capital investments; selecting strategic partners, forging alliances; and implementing trade and public policy initiatives, require intelligence support- (Herring, 1998). A CI system may track: a competitor's capabilities and strategies; the industry's structure and trends; the market and customer behavior;

political, economic, and social forces; and technological developments (Hohhof, 1997). The essence of any CI system is its contribution to better and more timely decisions, which can have strategic implications for the survival of the organization (Gilad, 2000)

This particular research focuses on the technological aspects of CI affecting all of the decision areas mentioned above, including investments, acquisitions, partnerships, policy, and new business ventures. The issues related to technology and decision-making are not new. In 1967 Erich Jantsch recollects that, before 1960, technological forecasting was deemed "a purely exploratory exercise." And in the same book he states, "Today, all leading forecasting institutes and consulting firms producing technological forecasts-… - regard their forecasting function as closely related to their consulting function in corporate planning." In 1971, Marvin Cetron and Christine Ralph were writing about methods to help research and development planners to identify "what appears to be the most fruitful areas for the investment of funds… to provide a basis for decisions to initiate, increase, cut back, or terminate particular research projects… [and to] discern where and when he will be in jeopardy if no action is mounted to effect a timely response to the hazards unseen." Their research concluded that 95% of companies conducted formal long-range planning activities, that among the most relevant external factors were competition, technological change, diminishing product lifetime, and lengthening market lead-times; that technological forecasts were an important input into the planning process; and that management information systems could be employed in future planning systems (Cetron and Ralph, 1971). Over the years, there have been a number of names and method variations to provide substantive information to technology decision-makers. Some of them are:

- Technological Forecasting: The prediction of innovations and advances in a particular technology (AMA, 2004)
- Scientometrics: The statistical analysis of research patterns in the physical and life sciences (Wolfram, 1994)
- Competitive Technical Intelligence: The analytical process that transforms disaggregated competitor technology data into relevant and usable strategic technology knowledge about competitors' positions, size of efforts, and trends (Coburn, 1999).
- Bibliometrics: counts of publications, patents, or citations to measure or interpret technological advances (Watts and Porter,1997)
- Innovation Forecasting: is the attempt to scientifically predict which technologies will successfully evolve through the development cycle into application (Watts and Porter, 1997).

This research will use the term "Technical Intelligence." Technical Intelligence implies information packaged to provide insight beyond mere information. The word competitive is not used because, although even research labs can be in competition for funding and internal technology offices, such as IT departments, may be in competition with potential outsourcing companies, most individuals tend to think only of private enterprises in product competition and not simply the "competitive" position of the firm or department itself. The focus may end up only on the direct "competitor," whereas technical intelligence encompasses industry knowledge, competitor knowledge, and knowledge of suppliers and enabling technologies.

As mentioned previously, the need for technical intelligence is not new. However, the environment is new. In 1967, Jantsch wrote "There is even one additional obstacle in the present case: there exists nothing which resembles a systematic abstract service." In modern times, there are numerous abstract services. Vast amounts of information are easily accessible; however, it is so vast as to be impossible for a manager to stay current on technology advancements. Dialog, the leader in providing online-based information services to organizations, has more than 800 million unique records of key information contained in 900 different databases (Dialog, 2004) The volume of scientific writing is growing exponentially (Cunningham, 1998) For instance, a search for all 1970 records in EI Compendex, a database of Engineering Research, indexed 44,677 abstracts. In 1996, that same database indexed over 436,000 abstracts. And in the information age, effectively managing knowledge, the most important company asset, is critical to business (Tjaden, 1998). As a result, technology managers are in need of additional tools to support the decision-making process.

Bibliometrics and "text mining" have arisen offering approaches to analyze and contextualize large amounts of information (Watts et al, 1999). Bibliometrics is counting publication, patent, or citation activity (Watts et al. 1998). Text mining involves extracting information from text and mining the text for discovering rules and patterns (Nasukawa and Nagano, 2001).

One community, that bridges the issues of managing technology and text mining, focuses on using published information to aid in the technology decision-making process. The Technology Policy and Assessment Center at Georgia Tech (TPAC) formulated an area entitled "Technology Opportunity Analysis (TOA), which uses a bibliometric/text

mining tool with publication databases in order to profile trends in research and development (R&D) and can help identify emerging or unfamiliar research that may intersect the functional interest of the client. This research will investigate the development of these methods for further utilization in Technical Intelligence gathering activities.

While there are many sources for information and methods to analyze information, this project seeks to develop a comprehensive methodology to quickly monitor key technical intelligence areas, in order to streamline aspects of the technical intelligence gathering effort, providing a method that quickly cleanses and consolidates information into an understandable, concise picture of topics of interest. This approach caters to the technology manager's need for timeliness and completeness. This research will determine which journal publication database metrics may be useful to technology managers. It will evaluate and alter some existing analysis methods for more accurate and effective results, and develop an overall framework to obtain expedited answers in a user-friendly, consistent format (McDonald and Richardson, 1997). This research is important because it has the opportunity to provide a powerful tool that will help technology decision-makers quickly get a picture of not only the general landscape of the technology that they are reviewing but also identify important changes in trends that might not be achievable in any other way. Monitoring technologies can take a whole new direction. The goal of this research is to provide a methodology that can be integrated into larger monitoring efforts and provide analysis that can be adjusted with feedback from the target audience.

CHAPTER 2

LITERATURE REVIEW

This research is investigating both improving the tools utilized for text analysis and bridging analysis tools with the intelligence gathering processes. Therefore, this literature review deals with several topic areas. It seeks to explore the relationships among three different stakeholder groups: Competitive Technical Intelligence Professionals, Technology Decision-Makers, and Text Data Mining Researchers. The competitive technical intelligence professionals gather and analyze a variety of information related to the industry, market situation, and competitors' activities for the second group, the technology decision-makers. These individuals must make decisions about strategic efforts in researching and developing technologies or purchasing technologies. While the first two sets of stakeholders have a clear overlap in interests, the third group, the text-mining professionals, work in the distinct area of analyzing text, an analysis method that has only minimally been applied to the competitive or technical intelligence domains. The core of this research is the formulation of technical intelligence for the technology decision-maker incorporating text mining approaches.

In order to fully explore the literature in these areas, this review will be separated into two sections: the first on technical intelligence and the second on text mining. The first section will include a general discussion of technical intelligence and an in-depth review of the research related to each step of the TI process. Those steps include Planning, Collection, Processing, Analysis, and Dissemination. The interaction between decision-makers and TI professionals will be discussed in the Planning step, since the

purpose of the planning stage is to scope the intelligence that will be provided to those decision-makers. The most in-depth research coverage will take place at the analysis step, which will include the foundations of TI analysis leading up to current methods of TI analysis. Included in this step will be a discussion of the current methods that have begun to incorporate basic text mining methods. Finally, this first section will conclude with a discussion on measuring the effectiveness of TI.

The second section of this literature review is an overall discussion of the steps in the text mining process and will be organized around the steps in text mining, in the same way as the first section revolves around the established TI process. While this research is primarily focused on cleansing text data and mining text for analysis purposes, the overall text mining process includes Retrieval, Extraction, Cleansing, Mining, and Visualization.

In discussing the literature in technical intelligence and text mining in separate sections, this review reflects the current situation where there is very little purposeful overlap between these two groups of researchers. However, this research seeks to increase the intersection between these areas.

2.1 Technical Intelligence

Technical Intelligence (TI) is part of the overall Competitive Intelligence (CI) process. TI is that component of the CI system that supports project and scientific funding decisions and helps decision-makers calculate the relative strength of other organizations (Hohhof, 1997). It emphasizes the R&D function of an organization, but can also encompass other technology-driven activities such as strategic planning, technology acquisition, and process equipment investments. Since TI is a subset of CI, much of the research concerning general CI covers the TI aspect. There are only two

major books written on the topic of Technical Intelligence, in the modern context, each offering a definition of TI. Ashton and Klavans define TI as

> "…the business-sensitive information on external scientific or technological threats, opportunities or developments that have the potential to affect the company's competitive situation."

They go on to emphasize that TI is externally focused, business-sensitive, and action-oriented.

Coburn defines TI as

> "The analytical process that transforms disaggregated competitor technology data into relevant and usable strategic technology knowledge about competitors' positions, size of efforts, and trends."

There are two noticeable differences in these definitions. The first is that Ashton-Klavans defines TI as "information" and the Coburn definition defines TI as a "process." Secondly, the Coburn definition identifies TI as information about the competitor, whereas the Ashton-Klavans definition leaves the definition open to any relevant external information, recognizing that competitive advantage can be affected by government, customers, suppliers, and general scientific developments, in addition to competitor actions. While both books cover generally the same areas, the foundational difference in viewpoint is evident in the fact that the Ashton-Klavans book covers broader issues of understanding the technology and science of an industry. On the other hand, while both authors make it clear that TI should be usable for strategic action and discuss areas where TI can aid in the strategic decision making process, the Coburn book deals more with the process of implementing those actions. For example, the last five chapters of the Coburn book discuss Strategic Alliances. In any case, TI starts

with the gathering of information and ends with the actions performed based on that information.

2.1.1 The Benefits of Technical Intelligence

Coburn and Ashton-Klavans essentially state the same set of objectives for TI. Coburn lists the purpose of TI as:

1) Avoidance of being blindsided,
2) Increasing external awareness and focus
3) Acquisition of Technology
4) Input for a component of a Tactical Plan of Action
5) Input for a component of a Strategic Plan

Ashton-Klavans list the objectives of TI as:

1) To provide early warning of external technical developments or company moves,
2) To evaluate new product, process, or collaboration prospects created by external technical activities,
3) To anticipate and understand S&T related shifts or trends in the competitive environment for organizational planning. (Ashton and Klavans, 1997)

These objectives also reveal the difference in definitions by the two books. Coburn's objectives reflect more process orientation, while Ashton-Klavans work relates more to the information itself and the need for a broad reach of useful information.

Ashton and Klavans (1997) state that TI is not for every company. The companies that benefit the most from TI –

1) Operate in technologically dynamic industry environments where the pace of change is rapid or new technologies are likely to surface
2) Emphasize technology-intensive products where technology is a differentiating factor, product introduction rate is fast, market entry timing is important, regulatory approval of new products is complex
3) Manage a significant R&D portfolio
4) Expect a high share of near-term business revenue growth from new products

Some of the industries that fit into these categories are computers, telecommunications, pharmaceuticals, and energy. For such companies, technology is a basic determinant of a company's competitive position and is the source of future growth. A company, of course, does not have to fit into all of the aforementioned categories in order for technology to be a basic determinant of position and growth. The participants in this study were selected based on their meeting at least two of the characteristics. Dou (1999) asserts that TI focused on the development of new products or new services is strongly linked to innovation, which can better benefit the "Contender" companies rather than "Native" or "World Class" ones. And while large corporations may be more equipped to support full CI department feeding R&D efforts, the lack of finances may be the best incentive for small companies to engage in TI. Raymond goes a step further. He states that it is essential for SMEs to use intelligence to detect trends and understand strategic issues that stem from the global knowledge economy (Raymond, 2003) Brandau and Young (2000) discuss how CI in start-ups can occur. The main success factors for start-up efforts are speed, simplicity, organization, efficiency, and effectiveness. If a company has incorporated intelligence gathering and utilization at an early stage, these

activities will become part of the culture. They list many resources that small firms can utilize. Resources for intelligence are becoming more available for SMEs. Dialog, who offers access to a a variety of databases, has begun offering a flat-fee set of services aimed at small businesses (Anonymous, 2002). In addition, Smith asserts that it is a myth that competitive analysis is only necessary in highly competitive environments or that it is costly and therefore only necessary for major decisions (Smith, D. C. and Prescott, J. E., 1987). Tibbetts (1997) agrees. He asserts that

> "Any of these companies can benefit from TI. Technology monitoring can lead to the discovery of new technologies, trends, research in progress, emerging or advanced technologies, improved processes and production methods, outsourcing possibilities and expertise, technology breakthroughs, proven products, potential strategic alliances, organizations doing research in targeted fields, potential mergers or acquisitions, new markets, new applications; regulations, standards, and laws; solutions to technological problems."

CTI can benefit a company by eliminating negative surprises, improving portfolio management, improving the selection of projects, and identifying competitive threats (Ashton and Klavans, 1997). For example, Davison (2001) mentions the example of Dr. Eger, whose pharmaceutical company saved $16million dollars when the CI department discovered that they were too far behind the competition in development to catch up. Another example, provided by Herring (1993), concerns a company Vice President who discovered at a talk that their company's next big innovation was already being researched at a small Far Eastern company. A publication search revealed that the company had an 18 month advantage over their company. The small firm was acquired and with the combined resources, the introduction date was accelerated by one year. A study by Prescott and Smith (1989), whose sample consisted of 172 corporate CI practitioners from the membership of the Society of Competitive Intelligence

Professionals (SCIP) found that the most beneficial outcomes of general CI efforts are the "Identification of New Business Opportunities" and the "Sharing of Ideas". Other benefits named included "Improving their ability to Anticipate Surprises" and "Improving the Managers' Analytical Skills." Ten years later, another study, involving 137 CEOs and CIOs showed that these individuals valued CI as important for identifying both threats and opportunities and that CI was useful in deciding Business Strategies (Vedder et. al., 1999.) Appendix A identifies numerous areas in which TI can be applied to benefit a company.

2.1.2 Technical Intelligence Viewpoints

There are different categorizations and breakouts that affect the approach to TI. For example, the most common breakout of TI is Coburn's discussion of TI purposes: strategic vs. tactical (Davison, 2001; Parker, 2000; Prescott and Smith 1989). Davison defines strategic CI output as being "forward looking to accommodate the need for successful long-term planning. He defines tactical CI as achieving "a particular short term aim." An example of tactical a TI decision would be whether or not to invest in certain equipment. Davison discusses tactical vs. strategic decision in the context of measuring results, an issue which will be discussed later. While Prescott and Smith (1989) also find that CI is equally applicable to strategic and tactical decisions, they also categorize CI activities based on the mission. They define three types of missions: information, offensive, and defensive. Informational missions are for general understanding purposes. Offensive missions are geared at understanding the competitor's organization, such as competitor vulnerabilities, while defensive missions are focused on the internal organization and actions that competitors may take against the organization,

such as recognizing gaps in the organization's capabilities and how the competitor may obtain competitive advantage based on the organization's internal weaknesses.

Comstock and Sjolseth (1999), in their case study of Weyerhaeuser Co., demonstrate how the categories of TI affect the structure of the R&D organization. Weyerhaeuser supports three categories of R&D activities: Core research Programs, Strategic Programs, and Development Projects (tactical). They use Technology Assessment to understand how external events affect the business needs. The Technology Assessment activities have a direct impact on the activities of the three types of programs. One divergent view of CI is presented by Burkhart (2001). She states that TI should be considered in terms of the company's or industry's location on the product life cycle. The product life cycle describes the developmental evolution of products, companies, and industries, consisting of introductory, growth, maturity, and decline stages. Burkhart states that if your industry is at an early stage, recognizing "surprise" competitors should be the focus of intelligence gathering. On the other hand, if the industry is moving toward maturity, intelligence that helps the company maintain market share is most important. Linking intelligence with life cycle information can help management better prepare for the future. Some of the intelligence may be used to help define the location of the product, industry or company in the Life Cycle. Appendix B is a table demonstrating the particular intelligence questions that are most appropriate for different Life Cycle Stages.

Another categorization of TI divides technical intelligence activities by the target: technology surveillance and organizational surveillance. Technology surveillance is the systemic, continuous watching or searching external environment for relevant technology

developments or trends, includes monitoring and analysis. Technology surveillance can be used to:

1) Provide technical descriptions of existing or emerging technology systems, developments, events, trends, or capabilities
2) Identify or predict significant shifts in the rate of progress in an area or the occurrence of technical breakthroughs that will make a new capability technically or economically feasible
3) Identify when substitute of competing technologies are becoming available.
4) Assess the impact on the directions and rate of technology development from new market-influencing technology forces such as government regulations or shifts in consumer preferences.

Organizational Surveillance entails efforts to:

1) Recognize patterns of activity by other organizations that can have consequences for a firm's market relationships
2) Identify emerging capabilities or external organizations strengths and weaknesses
3) Compare products and methods of the state of the art to others
4) Compare a competitor's product or process technology performance or cost data with past records to discern trends (Ashton & Klavans, 1997).

The metrics evaluated in this research aid in both technology and organizational surveillance.

Taken as a whole, these findings suggest that many organizations would benefit from CI, if the benefits of CI are provided in forms that meet the needs of the users – the technology managers themselves – and if they are available in cost-effective, user-

friendly formats. This review next considers the literature support for the six- step methodology for gathering technical intelligence.

2.1.3 Technical Intelligence Methodology

There seems to be a substantial consensus concerning the general process model for gathering intelligence. Coburn (1999) provides the most simplistic model, breaking activities into three steps: Data Collection, Data Analysis, and Action. At the collection stage, Coburn takes the perspective that good design will dictate that no more information should be gathered than necessary. At the Data Analysis stage, Coburn emphasizes the need to bring the data together into a cohesive story using good judgment. Finally, three criteria must be met in order to perform the Action stage: the data must be believable, the decision maker must trust the analysis, and the conclusions must represent a logical fallout from the analysis. In other words, what to do must be obvious. Stacey (1998) concurs with these steps but adds Planning to the front end of the steps and Evaluation at the tail end. Ashton and Klavans (1997) provide a more complex model. They separate Delivery from the Application Step and demonstrate the relationships among the steps in the system, including the input of the intelligence needs and the output of the impacts of the actions taken. More important than an academic model of the process which is generally agreed upon is to consider examples of the CI process in operation. This section next examines several of the steps in greater detail, in the context of the model (Figure 2.1), another example of a more complex model. This model, presented by Herring (1999), is the traditional intelligence cycle proposed by the Central Intelligence Agency. It will be used as the basis for the discussion of the TI process.

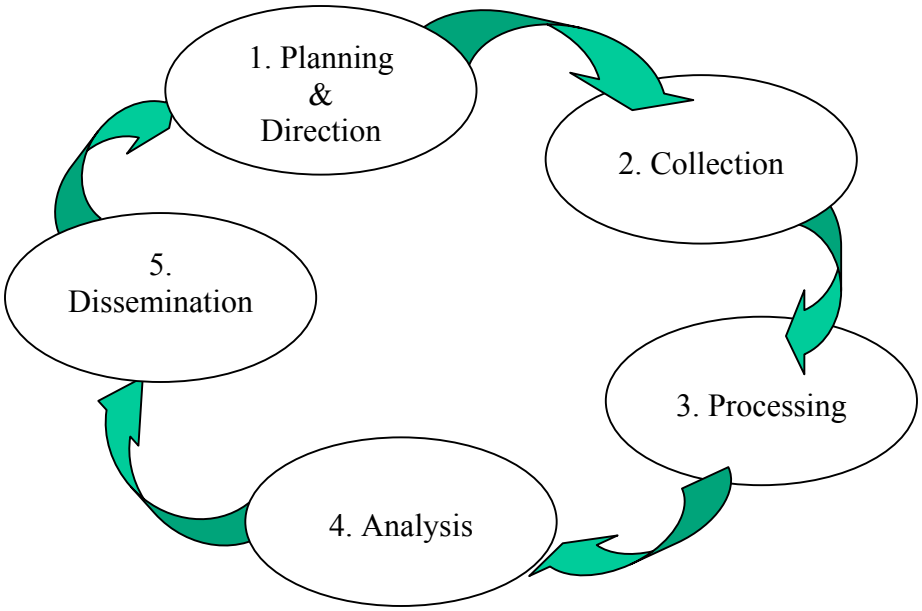


Figure 2.1 Traditional Intelligence Cycle.

2.1.3.1 Step One: Planning and Direction- Identifying Key Intelligence Needs

The traditional intelligence cycle model actually begins with the information needs. Herring has been at the forefront of researchers who are concerned with determining the key intelligence needs of the decision-makers. However, generally very little research has been done to determine the information that decision-makers actually *want*. Herring notes the lack of research on determining the information needs of decision-makers and points out that a mismatch between the needs of decision-makers and the information provided is cited often as the reason decision-makers do not use information provided by competitive intelligence professionals. Herring first embarked on a study to determine national level needs for Science and Technology intelligence. His protocol for determining these needs was then adapted to corporate interests. Herring determined that a company's intelligence needs fall into three functional areas: Strategic,

Early Warning, and Key Players. He then developed a two-level protocol that can be utilized in order to determine the information needs of the decision-maker. This protocol was adapted into the first stage of this study to determine the needs and expectations of the decision-makers (Figure 2.2)

HERRING PROTOCOL

1. **Business Decisions and Strategic/Tactical Topics**

   What decision and/or actions will you/your team be facing in the next ___ months, where CI could make a significant difference?

   a. How will you use that CI?
   b. When will it be needed?

2. **Early-Warning Topics**
   (Begin by identifying/discussing a past "surprise" in your industry, business, or company.)
   Identify several potential surprise topics that you do not want to be surprised by.
   For example, new competitors, technology introductions, alliances & acquisitions, regulatory changes, etc.

3. **Key Players in Our Marketplace**: Competitors, Customers, Suppliers, Regulators, etc.
   Identify those players you believe the company needs to better understand.
   a. Who are they?
   b. What specifically do we need to know?

Figure 2.2 The Herring Protocol

Despite the process presented by Herring, other research reveals that there is a general disconnect between the decision-maker's needs and the work of intelligence analysts. Breeding (2000), in an article based on a case study of Shell Services International (SSI), and especially the entity that provided IT services for the Shell

operating companies, provides insight into the concerns of users of CI systems. Breeding identifies the problems described by users of CI: information is too shallow, the credibility of the information, the timeliness, and managerial need for some say on how the report is assembled.

Research and subsequent papers (Porter et al 2000, Porter et. al, n.d.) by the Georgia Tech Technology Policy and Assessment Center attempts to bridge the communication gap between Information Analysts and Information Product users. In one project, 26 technology professionals and managers were interviewed in order to determine the factors that affect the utilization of information products in their decision-making. The follow-up research resulted in a checklist of ten action items for managers to improve the likelihood of receiving usable Information from their Information Analysts.

2.1.3.2 Step Two: Collection- Intelligence Gathering and Reporting

Gathering Information involves two elements: the sources and the attack plan to obtain information from those sources. In order to perform these activities, a wide variety of sources may be used to find financial, market, legislative, competitor, and technological information. The possibilities and techniques are endless, especially when considering the methods to obtain expert opinion. Mockler (1992) names daily employee reporting systems, including that of Kodak, which developed an approach to gather oral communications from its employees worldwide. Similarly, as mentioned in Teo (2000), a number of companies use the internet to review their competitors' advertising, and targeting strategies. However, this discussion will address only text sources that can be mined to provide information relevant to technology decisions.

Gathering can either be assigned to an individual or department, as is often the case in large companies or, as is more typical in R&D laboratories, each expert is responsible for keeping up in a specific field. In either case, the internet has become an important source for gathering intelligence. A survey by Teo (2000) found that 600 companies in Singapore showed above average agreement that the internet was useful to review articles in industry/ trade websites, monitor government information, and check on competitors' products/services. By using the internet, a company can monitor competitor activities, track customer viewpoints, seek new ideas, and gain international expertise. The Internet can also be a source for obtaining customer feedback, which can be stored and mined for trends (Teo, 2000).

While a general Internet search on a search engine, such as Google, may provide interesting information or lead to a new useful database, the most comprehensive and useful approach is to search external databases. A number of companies make use of services such as NEXIS, Dow Jones News/Retrieval, and Dialog (Mockler, 1992). Cambridge Scientific Abstracts is another portal for access to a large number of databases. Research reveals numerous sources that can be utilized for various purposes. Business Information can be found using ABI/Inform at Proquest, Hoovers on-line, EBSO business index, among other databases of business articles. Press Releases can be found at LEXIS-NEXIS, and NTIS provides or a service to find government publications. There are a number of other government databases available, also. Other services are more specialized. Some of the leading databases are:

- Science Citation Index- SCI® provides access to bibliographic information, author abstracts, and cited references found in 3,700 of the world's leading scholarly science and technical journals.
- Chemical Abstracts Service (CAS) has indexed and summarized 23 million chemistry-related articles from more than 40,000 scientific journals, patents, conference proceedings and other documents.
- MEDLINE is a database of abstracts maintained by the National Library of Medicine containing over 8.4 million abstracts from 3,800+ medical journals.
- EI Compendex covers almost seven million records referencing 5,000 engineering journals and conference materials dating from 1970.
- INSPEC, produced by the Institution of Electrical Engineers, contains 5.8 million records from over 4,000 technical journals, 2,000 conference proceedings plus books and reports annually from over 60 countries in physics, electrical engineering, electronics, computing, control and information technology.
- Derwent World Patents Index (DWPI) provides access to information from more than 22.9 million patent documents.
- Pollution Abstracts contains almost 300,000 records on scientific research and government policies on pollution, including coverage of journal literature, conference proceedings, and hard-to-find documents.

Some of these sources are fee-based while others, such as Medline, provide access to abstracts for free.

2.1.3.3 Step Three: Processing and Storage-

This step includes the transformation of information into a form usable by analysts through decryption, language translations, and data reduction. This area is pertinent to this research in that unstructured text must go through a cleaning process in order to retain relevant phrases and reduce the noise that is a normal function of language. Extensive discussion of this topic will be included in the text mining section of this literature review. Decryption, translation, and storage are outside of the scope of this research.

2.1.3.4 Step Four: Analysis

With all of these sources, it is important to have a systematic approach to analysis of the new data for emerging trends and new direction, in research, and there are a number of techniques that have been developed for such a task. However, before discussing the modern techniques, it is important to first understand more about the history of analysis. Around 1953-54, corporate long range planning became heavily discussed and recognized. Technological forecasting began to be used around 1959-1960. However, it wasn't until 1965-66 that the integration of the two areas started to occur (Jantsch, 1967). By 1969, a book was written entitled "Technology Forecasting and Corporate Strategy" (Wills et al., 1969) that demonstrated maturing of the idea that technology was strongly related to strategy. The "Technological Forecasting" literature represented the first attempts to apply analytical methods to data in order to provide intelligence to technology decision-makers. The challenge to early forecasters was the availability of information. Therefore, the early methods were developed in the environment of few data points. The Delphi method was the main approach to obtain

industry-wide information. The early methods were not based on systemic literature evaluation, but were applied primarily to specific elements of a technology, such as price or speed (Jantsch, 1967).

In the Information Science world, terms like "content analysis" had already begun to appear, although the articles were not focused particularly on science and technology content. By 1972, in Martino's 700 page book entitled "Technological Forecasting for Decision-Making," content analysis received a one paragraph description. Around 1972, bibliometrics and citation analysis appeared in the literature. Next, and representing an interesting trend in the literature, the discussion of breakthrough technologies and substitution became more popular. An analysis of the "technological forecasting" literature revealed that, prior to 1975, very little was written about technological breakthroughs, at most one article per year. Then, in 1975 three articles were written and the upward trend continued. In 1978, 10 articles were written mentioning words like breakthrough and substitution theory. Harold Linstone and Devendra Sahal (1976) edited a book that was a collection of articles concerned with technological substitution. The book discusses the shortened Technology life cycle and the impact on management. During this same time period, patent analysis and scientometrics appeared as methods to quantify science and technology.

Patent Analysis and Scientometrics are basically categories of Bibliometrics, and involve analyzing text information in text databases by counting. These numbers are indicators of technological activity. An "indicator" is a statistic used to measure something intangible. Technology indicators are statistics that measure technology indirectly. They include, 1) R&D expenditures, 2) number of scientists and engineers,

and 3) number of scientific and technical publications (Ashton, Klavans, 1997). Studies have shown that when large numbers of patents are analyzed, the number of patents is positively associated with the level of technological activity (Ashton, Klavans, 1997). Moreover, there are numerous bits of information in patents that can be counted and analyzed. For example, data can be obtained by comparing entities by classification, calculating proportion of patents in classes divided by the company's proportion of all patents, in order to measure strengths and weaknesses; monitoring inventors; and analyzing changes in levels of patenting over time. There are all analysis that can be conducted on patents (Ashton, Klavans, 1997). Other metrics include the size and composition of a patent family, the number and timing of subsequent patent citations and the timing of the decision to let a patent lapse (Ashton and Klavans, 1997). Work by Rajman and Besanon (1998) uses textual and statistical methods with correspondence and cluster analysis to identify technology interactions and trends. Their research aims to profile the research of various countries. Additionally, studies of the scientific literature have found that in rapidly changing fields, the references tend to be made to more recent articles. The metric representing the speed of the Technology Life Cycle is the median age in years of the references on a company's recent patents. Analyzing patent indicators is a good start for quantifying technological activity because patents represent perceived economic potential, contain a good level of detail, and contain information not found in other places. However, the analyst must be careful because, while counting patents reveals the amount of activity. Notice that the information can be misleading because the importance of the patent varies. Also, not all discoveries are patentable, and others are

patented just to prevent others from doing so; and evidence from one country doesn't represent global technological activity (Ashton, Klavans, 1997).

Similar approaches can also be applied to publication databases, as will be done in this research. The analysis of publication databases as a means of quantifying science and technology was initially referred to as Scientometrics. Other names that refer to like concepts are Informetrics, Bibliometrics, and Text Data Mining (TDM). Publication databases, which hold publication information primarily from entities with cultures of publishing, such as universities, represent research in its earliest stages (Tibbetts, 1997). Therefore, technological terms tend first to appear in technical publication databases years before patents, and up to a decade before the same terms appear in business periodicals (Courseault, 2001). For example, microelectromechanical systems first appeared in the INSPEC database as early as 1993 and did not appear in Business Index until 1998. Since much of the work in refereed journals and conferences is early stage basic research, it has a long shelf life, allowing for trend tracking over time. Some ways publication databases are utilized include the evaluation of national scientific capabilities, policy evaluation, and compiling intelligence. Publication databases can be analyzed in order to: enhance information retrieval, identify the technology infrastructure (which Kostoff defines as the authors, journals, and organizations), identify main themes in the literature and discover relationships between different fields of research and elements of the infrastructure.

The BESST Project is an example of using publication databases to evaluate national capabilities. The project involved a study in the UK in which publication indicators were used to determine the amount of scientific output in different fields, map

changes in the collaboration efforts of scientist within the UK and internationally, and to explore policy-relevant questions (Katz and Hicks, 1997).

Georgia Tech Technology Policy and Assessment Center (TPAC) provides other examples. The TPAC website contains a recent paper that analyzes Iraqi engineering over a fifteen year period (Porter, 2003). Other TPAC examples include the Technology Opportunities Analysis (TOA) method that the TPAC research team led by Alan Porter has been developing since the early 1990s. They include an early study for the Malaysian government profiling the correlation between their R&D and actual industrial activity, and the study that initiated the TOA process which was a project to "identify Japanese companies actively developing electronic packaging to contact in arranging a US study mission (What is TOA?, (n.d.)). The TOA process for analyzing publication databases has moved from simple counting of basic information located in fields to a wide range of applications. Indicators such as Technology Life Cycle Status, Innovation Context Indicators which represent topic groupings and relationships, and Product Value Chain and Market Prospects Indicators, based on gap analysis of actual research versus needed research for a technology to become operational, have been formulated. Studies have been performed on ceramic engines, fuel cells and a host of other technologies (Watts and Porter, 1997 & 1998). The Hot Tech project is in progress and represents an attempt to automate the determination of the aforementioned categories of indicators (Hot Technologies, (n.d.)).

Moreover, the methodology is expanding to include Text Data Mining. Whereas bibliometrics refers to simple counting, Text Data Mining refers to the processes used to extract meaningful relationships from a large corpus. A study on Interoperability reveals

a process where a clustering technique is utilized in a process to identify unique information in text. The methodology captures terms that are related to the central themes of the text but are found at the outskirts of those themes. The study finds that those terms are likely to fall into one of three categories: a new application of an existing concept, an emerging capability, or noise (Watts et al, 2000). Alan Porter contributes to this body of knowledge in a Futures Research Methodology book chapter in which he identifies numerous applications of text data mining for technology foresight purposes. One area of research that he references is that of Kostoff and Swanson in which text data mining is used to identify cross-relationships between topics in a body of research not evident to researchers in individual domains (Porter, 2003).

Kostoff, in the Office of Naval Research, publishes heavily in applying text data mining to science and technology questions. Like TPAC, Kostoff also utilizes examples of technology studies in order to discuss methodology. One important difference in these two research teams is that while the Georgia Tech team is working toward automated technologies through the Hot Tech initiative (Porter, 2003), Kostoff rejects the notion of automated processes and stresses strategies that include expert opinion more extensively.

Kostoff presents a general methodology for conducting Science and Technology Text Data Mining which is framed in three major steps: Information Retrieval, Information Processing, and Information Integration. Retrieval is the selection of text to analyze, Processing is the application of the computational approaches, and Integration combines the expert interpretation to the computational results. In this paper, Kostoff uses the Biomedical domain to present the value of Science and Technology Text Mining along with his generic approach and discusses roadblocks to high quality results (Kostoff,

2001). The roadblocks discussed include the limitations created by using a limited number of databases, the variation in the quality of information contained in abstract records between different databases, and the difficulty in retrieving a corpus that is both comprehensive and contains a high percentage of relevant documents. In another paper, Kostoff and DeMarco (2001) analyze a corpus of documents published in the journal "Analytical Chemistry," demonstrating the vast array of analysis that can be conducted on a body of technology abstracts. The four main categories of application are improving further document retrieval, identifying infrastructure elements involved in publishing on the specified topic and their relationships to each other - identifying themes and relationships within the corpus, and the discovery of links, relationships, and opportunities not discernable from reading each record separately (Kostoff, DeMarco, 2001).

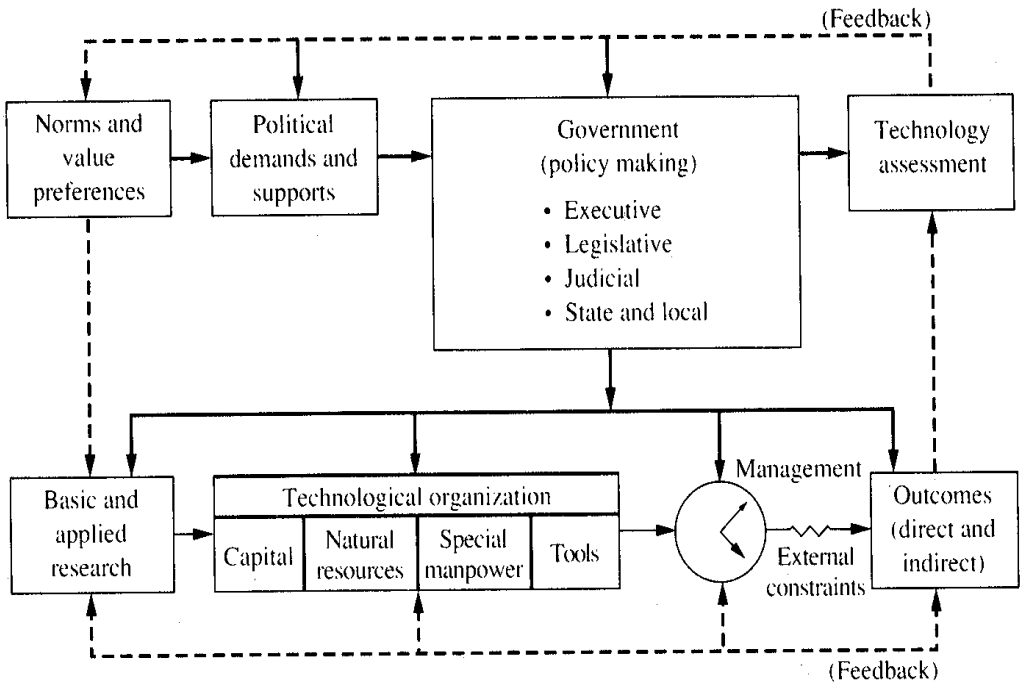
Publication databases can provide a plethora of intelligence to Science and Technology Managers at the earliest stages of development. However, just as patents have drawbacks, so do publication databases. One drawback is that publishing has a time delay in months or even years in addition to the lag in time before the article is indexed (Hohhof, 1997). Other potential pitfalls are that articles from less reputable sources are given the same weight as other articles, translation of foreign languages, excessive extraneous information, or the same information reported in many articles (Kinzey and Johnson, 1997). Publication practices also vary across fields, making interpretation more problematic for areas that cross fields. Choosing the appropriate database may also pose a challenge (Wining, 2003). With the benefits of all these approaches, one major weakness

in using any of these indicators is the lack of research validating them, an issue that must be addressed in the future.

2.1.3.5 Step Five: Dissemination

The issues and problems facing effective information transfer in the competitive technological environment still focus on the basic problem of getting the right information to the right person at the right time (Hohhof, 1997). A planned method of distribution or communication is essential (Bryant, 1997). A study of 95 Competitor Intelligence programs revealed the challenges of dissemination. Different users prefer different formats. The study reveals the challenges faced by CI practitioners in getting their message into the hands of decision-makers. Some of the challenges noted were deciding who should get the information, the lack of feedback on the information needed in order to produce a useful product, and the challenge of reaching a large audience in a timely manner using the method of communication to which each of those individuals will respond. Feedback is essential to clarify the user needs, identify missing information, and identify new areas to research (Prescott and Smith, 1989). Unfortunately, dissemination is often thought of only as getting information into the hands of the decision-maker. When information was limited, such a definition perhaps was sufficient. Now, however, the presentation of intelligence has become an essential part of dissemination. For example, Wenk and Kuehn (1977) developed a method known as the Technology Delivery System (Figure 2.3). It is a systems model used in understanding the development and delivery of technologies. It includes the organizations and factors that either hinder or enable the development of a technology. The issue is that many individuals interpret the TDS and the development process in different ways. The

two researchers have experimented with changing the name and format of the TDS to clarify the purpose of the map to no avail at this point in time.



Reprinted from "Forecasting and Management of Technology" (Banks et al, 1981)

Figure 2.3 Technology Delivery System (TDS)

Visualization is an important part of dissemination of information because, as the Prescott study demonstrated (1989), CI managers found personal communication to be the most effective method for obtaining information. However, it is not the most practical. Visualization will be discussed in more detail in the Text Data Mining section of this review.

2.1.3.6 Step Six: Measuring

When all is said and done in implementing technology monitoring, the system is useless unless it adds value to the company in some manner. In general, there are two ways in

which CTI organizations can measure their effectiveness. One way is to generate metrics to measure such hard-to-quantify benefits as lack of being blind sighted, lack of product development failure, success stories, and demand for CTI services (suggesting CTI credibility). These methods potentially measure the direct impacts of CTI. The other way of evaluating a CTI program is based on the overall company result of CTI, including beating competitors to market, faster development, changes in product or service attributes, better marketing position, more competitive cost, or better design (Ashton and Klavans, 1997). These are long-term measures of intelligence and not a focal point in this research. However, it is important to note even as late as 2001, the idea of measuring the impact of competitive intelligence was still at the stage of "investigating the need and ability of competitive intelligence (CI) departments to become accountable" (Davison, 2001)

2.2. Text Mining

To this point, the literature review has emphasized the need for developing Competitive Intelligence (CI) and Technical Intelligence (TI) as competitive tools in many organizations, and has recognized the need to provide TI/CI in ways which meet the needs of decision makers and which are efficient and cost effective. Next, consideration was given to the development of a model to guide the process. In this research, text mining, using VantagePoint, a commercial text data mining tool, is the TI tool under study. Note, however, that there are number of available text mining tools that have various strengths and weakness. In comparing these tools, two components must be considered: the domain and the technique, which account for the primary difference in functionality. The primary domains in which text mining approaches are being applied

are the web, research publication databases such as Medline and Engineering Index, patent databases, and news source databases. VantagePoint was selected because it is designed to analyze text gathered from large databases in order to scan the records, identify trends, profile, map, and decompose technologies, needs which are the focus of this study. In this section of the literature review, emphasis is placed upon examining how the text mining tool itself fits into the stages of the process developed in the previous section, and to examine needs for improvement at two of the stages.

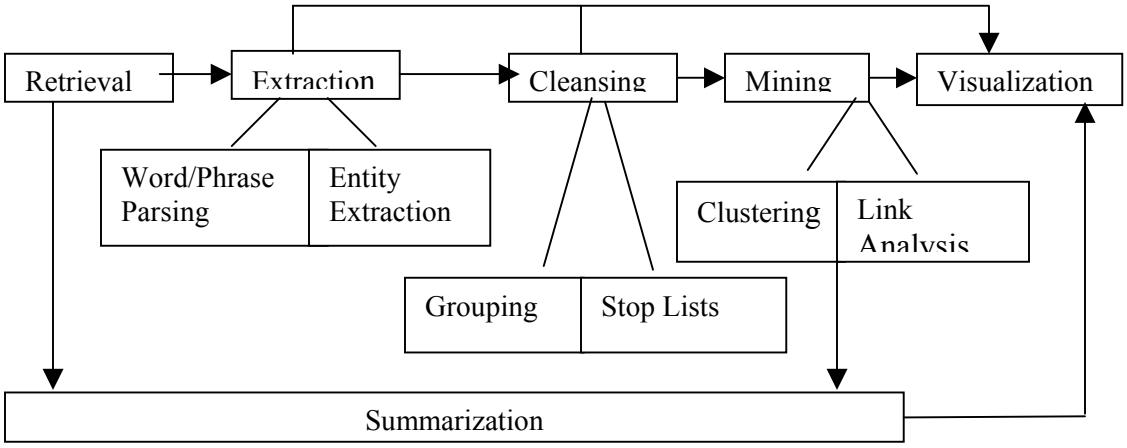


Figure 2.4 The Text Mining Process

There are approximately five major technique categories in the overall text mining process: Document Retrieval, Data Extraction, Data Cleansing, Mining, and Visualization. As part of the text mining process, there are a number of technique categories that are subcategories of, or supplements to, these major categories, such as Clustering, Visualization or Summarization (Figure 2.4)

In technology assessment, one of the basic assumptions about the nature of technology development is that, initially, research is focused in individual areas and as the technology matures, research domains begin to overlap. This assumption holds true for text mining. Early research focused on a particular domain, which included: information retrieval, entity extraction, clustering, summarization, and link analysis. Recent research combines methods. For example, a search in Engineering Index was done on the term "information retrieval and document." The abstract words were analyzed and words related to other methods such as extraction, clustering, grouping, visualization etc. were put into a group. It was found that the percentage of records containing words related to other methods increased gradually over time. From 1981 to 1986, the percentage of records containing words related to other methods averaged 10%, from 1987 to 1991-12% from 1991 to 1997 the percentage had increased to 30%. In 2001, almost half of all IR abstracts contain words addressing other methods. For example, entity extraction and clustering are being used to improve information retrieval, to determine what records are selected and the relevance of the documents selected (McCabe, 2000; Muller and Hamp, 1999). Note that the trend in overlapping topics can make the linear process depicted in the model difficult to discuss at a variety of points.

This section discusses examples from current research in the above core areas: Retrieval, Extraction, Cleansing, Mining and Visualization, with a focus on the latter three categories. In the course of the discussion, the tool VantagePoint from Search Technology will be referenced frequently as an example. As noted, VantagePoint is the tool that is utilized in the conduct of this research project.

2.2.1 Retrieval

"Information Retrieval" is the term often used to describe the act of retrieving information from documents or retrieving documents from a document collection. In this case, what will be described is document retrieval. Much of the research in document retrieval is being done in support of Web search engines. In 2001, 47% of the documents in our information retrieval documents contained either the word "Internet," "web," or "w w w". 11% contained "TREC." The size of the web and subsequently the large number of returns from a search is causing frustration in the user population as they attempt to identify the most relevant pages in a sea of thousands of hits. For example, an "information retrieval" search returned 63,600 documents. As a result, researchers are attempting to find new ways of scoring and presenting the search results. The general trend in document retrieval is to incorporate methods that were initially considered post-retrieval techniques and incorporating these techniques into the retrieval process. For example, Google uses a type of link analysis to identify relevant documents, and has been quite successful. Other methods include clustering, both static and dynamic, as a method to present the documents to users and allow them to focus on the area of interest (Kaji 1999; Tsuda, 1999; Muller & Hamp, 1999). Clustering may also be used as a type of query expansion. Documents not containing the search term, but clustered with the search term are also returned. Other researchers are looking to improve the document set by returning extracted entities into the search (McCabe, 2000). The results, although they overall have demonstrated some improvement, can also greatly deteriorate the resulting set, depending on the entity that is returned into the search. Further research into the best manner in which the entities can be used may lead to better results. Overall, the trend in

retrieval is combining methods and including user interaction. In this research, retrieval is not a primary issue. The search string will be crafted in such a way as to be as broad as possible while limiting the noise in the dataset.

2.2.2 Extraction

Extraction can take two forms; one is to identify the parts of speech and the other is to identify the specific type of entity extracted, such as whether a particular entity is a person, organization, phone number, date, address, or geographic location. There are numerous companies working in this domain. Many users require knowing the difference between an organization and a person, or want to be able to associate certain activities with the appropriate proper noun. For these purposes, in a very large corpus, recall is much more important than precision because, in general, it would not be challenging to remove names from the "Proper Name" group, because the number of proper names would not be too large to accomplish this type of revision. However, missing proper names requires that an individual sort through the entire word list to identify proper names missed by the tool. Some of the tools that perform entity extraction are SRA's NetOwl, Lockheed Martin's NL Toolset, Inxight's Thingfinder, and IBM's Intelligent Miner. These extraction tools boast wonderful results in the 90% area for recall and precision in MUC and TREC data, which includes primarily newspaper sources. However, for less predictable formats, such as publication abstracts or web sites, the effectiveness drops considerably. In these areas, the drop in proper noun recall may cause problems.

Another type of extraction is "parts of speech," specifically noun phrases, which are important for capturing domain specific concepts (Kaji, 1999). The problem with

parts of speech extraction is that the same word can be used as different parts of speech such as “census” as in “take a census” or “census the population.” A powerful, applicable tool could take advantage of capabilities of both types of extraction to at least be able to identify accurately verbs and complex proper nouns. VantagePoint uses Natural Language Processing to extract phrases from abstracts.

2.2.3 Data Cleansing

The Data Cleansing literature is primarily discussed in relation to cluster analysis, but has impacts on all forms of data mining. It consists of the algorithms and methods that determine the final information that feeds the clustering algorithm or link analysis. Data Cleansing impacts the quality of other text mining techniques and determines the quality of the information that is fed into the clustering algorithms a as well as how it is structured. For the databases currently used by VantagePoint, and others as well, the first decision is what part of the record to use: keywords, title words, abstract phrases/words, for full text words/phrases. In this research, data mining will be applied to abstract phrases.

Other issues are related to the selection and compression of the words that are used. Selection is the way words from text are determined to be candidate keywords for analysis. Selection issues relate to identifying a word as a potential keyword for analysis and determining the significance of that word in the document. The first step in the selection process is the defining of the word. For instance, words can be determined by every space or determined by Natural Language Processing algorithms to identify actual phrases (i.e. “Information Retrieval”). Another approach is simply to use windows of adjacent words. Selection also involves narrowing the number of words for analysis once

they have been identified. For example, VantagePoint only uses words that meet a minimum frequency for clustering. Ahonen-Myka et al. (1999) break words into sequences, and only use maximal frequent sequences, which are sequences of words that are frequent in the document collection and that are not contained in any other longer frequent sequence. A frequency threshold is defined for the document set. In order to bolster the frequency of terms in abstracts or full text documents, compression is used. Compression is grouping together words that are different, but have the same meaning. The most basic type of compression involves the variations of the same phrases such as "management of technology" and "technology management. " VantagePoint's List Cleanup function uses a stemming algorithm and shared words in reverse order to improve the compression. At a more sophisticated level is the compression of words that are different but have the same meaning. For example, in literature, Internet commerce and web commerce mean the same thing. Ahonen-Myka et al. (1999) described using the concept of equivalence class, which they defined as sets of phrases that occur together in the same documents frequently enough. Phrases belonging to some equivalence class are replaced by the name of the class. Most software products currently on the market, however, only view data cleansing as a task within a document as a component of entity extraction. For example, NetOwl will link a last name listed in a document with a full name in the same document. The same is true for company acronyms and company full names. However, if the acronym or last name is in a different document, then the association is missed.

The final issue is the determination of strength between keywords based on location of the words. This information is not exactly data cleansing, but the method of

capturing this information has an impact on the type of analysis that can be performed on the text data. Some tools may identify that two words are in the same paragraph or the proximity of two words to each other in the document.

2.2.4 Data Mining

2.2.4.1 Link Analysis

Link Analysis is the linking of information within documents. The most basic type of link analysis shows networks of word relationships, usually involving co-occurrence of some sort. Depending on the number of links, these networks can get very large and complex. The more powerful type of Link Analysis tools involve linking particular types of verbs with the doer and the object(s) of that action. SRA international incorporates this type of link analysis in order to identify links between entities in text and to identify key events in text. Hearst (1999) is also doing some work in this area, Her efforts, using computational linguistics, although the most powerful, require a significant amount of training for individual domains. Her focus is finding knowledge, such as developing a disease hypothesis or uncovering a social impact that is not contained in any one document. Kostoff (1997) is pursuing a similar, but his approach is a more statistically based effort.

2.2.4.2 Clustering

Since any type of text clustering is based on co-occurrence of words, whether some type of keyword or words contained in a document or abstract, it would seem that the actual clustering algorithm chosen will not bring about large differences in the actual clusters developed. This hypothesis is supported by two separate investigations on

clustering performed by TPAC graduate research assistants. Both projects investigated term clustering which allowed for multiple locations in space of the same term; terms were not required to fit in any one cluster, but terms that did not fit in a cluster were not included in the mapping. The first project investigated using two different methods, Principal Component Analysis and Maximum Likelihood Estimation, for determining the initial factors in factor analysis (Courseault, n.d.). The second project compared Principal Component Analysis and a Probabilistic approach to clustering. In both cases, the clusters had very little deviation between methods (Parasarathy, n.d.). Further support comes from research conducted by Chris Ding at Lawrence Berkeley National Laboratory. In his research, he finds that the partitioning indicator vectors found when clustering using a Hopfield network results in LSI index vectors, and that PCA is equivalent to the MinCut in graph theory. He also identifies a connection between Hopfield, PCA, and K-means. The basis of these similarities is the fact that the objective of all clustering is to minimize associations between clusters and maximize the relationships within clusters (Ding, 2003). Different algorithms simply have different starting points. This statement does not necessarily mean that the results are not somewhat different. The details of the chosen clustering algorithm are important to the end result and must be determined by the end goal. The difference, however, is primarily based on factors such as the following: whether the clusters are term clusters or document clusters, whether the clusters are distinct groups or whether certain items can be excluded from any cluster, whether the location of certain words or documents have a distinct location in the space or can have multiple locations, whether the clusters remain the same each time the algorithm is run

or, as in the case of probabilistic methods, the clusters may change each time the algorithm is run.

The cluster research contains a plethora of clustering techniques and additions to well known methods designed to improve the ability to find either documents or bits of information, as well as to provide a general landscape of the documents. These techniques fall into a number of categories. Hierarchical methods group items in a treelike structure. The methods can start with small groups and aggregate those clusters into larger clusters or start with one or more larger clusters and break those into smaller ones. In contrast, non-hierarchical methods simply break the corpus into subsets (Leouski & Croft, 1996). Partitioning clustering divides the data into disjoint sets. Density-based clustering groups neighboring objects into cluster, based on density criteria. A cluster is defined by a given density threshold. Statistical clustering method, such as factor analysis, use similarity measures to partition documents (Halkidi and Vazirgiannis, 2001). While factor analysis is a more linear statistical approach, there are other statistical approaches, such as the probabilistic approach offered in Vinkourov and Girolami (2000). Bayesian Clustering is another probabilistic approach which uses Bayesian probability theory to calculate the probability that a certain object belongs in a certain group (Rauber et al, 2000). Kohonen Self-Organizing Maps is an artificial intelligence approach based on unsupervised neural networks. In general, each of these methods is based on term frequency of co-occurrence. One unique method is offered by Shah. In this method, the semantic relationships between words in the document are captured. The Kohonen Self Organizing Map is used to cluster documents that have the most similar semantic maps (Shah, 2002).

In conducting text mining, clustering can be utilized in a number of different ways for a variety of purposes. Clustering may also serve as the basis for other types of analysis, such as those presented by Watts, Courseault, and Kapplin (2000). In this paper, an algorithm based on combining various clustering techniques is used to find emerging technologies that accomplish a particular function in a corpus containing over 10,000 publication records. Clustering may be used to discover topic hierarchies giving structure to a corpus and allowing an individual to explore the corpus in a more organized fashion (Larsen & Aone, 1999). Merkl and Rauber use the Self Organizing Map as the basis for an approach designed to uncover associations between documents. Their approach is intended to make explicit the associations between clusters (Merkyl & Rauber, 1999). Clustering can also be reapplied to the original document set in order to improve information retrieval. Ding applies a probabilistic model for dimensionality reduction to a corpus as a means of conducting word sense disambiguation and thus permitting the filtering of information and improving information retrieval. Therefore, if the user types in the word "capital," articles related to a city vs. venture capital can be separated and the user can then focus their search on the type of capital that is their interest (Ding, 2000). Similarly, Kaji et. al. (1999) present a method for generating a thesaurus using term clustering as a means to traverse a domain-specific corpus. The thesaurus is designed to cluster generic terms first. Then, allow the user to "zoom-in" to a cluster and identify more specific terms in that cluster by analyzing the statistical correlation between terms (Kaji et. al, 1999). Beil et al (2002) also present a method for term-based text clustering with the intent of offering a method that better handles very large corpuses and improves the retrieval process. However, this method includes the

added affect of cluster descriptions based on the frequent terms in the cluster. A hierarchical and a non-hierarchical approach is presented (Beil et al, 2002).

Most clustering methods use document clustering as a way to maneuver through documents, especially as clustering is being promoted as a visualization method for document retrieval (Lowden and Robinson, 2002). The increased number of internet sites have sparked a greater interest in this area (Zamir and Etzioni, 1998). Therefore, much of the most recent research in this area is based on web pages. Broder et al (1997) offers a method for determining the syntactic similarity of web documents for the purpose of filtering search results, updating web pages and identifying copyright violations. Zamir and Etzioni (1998) evaluate clustering algorithms used on web documents and offer an algorithm called Suffix Tree Clustering, which analyzes phrases shared by multiple documents.

There are as many methods for evaluating clusters as there are for actually clustering. Evaluation techniques fall into four major categories: separateness, cohesion,, precision, and recall. Separateness and Cohesion are both based on the similarity. Separateness measures the distinctiveness of each cluster. The object is to minimize the similarity between clusters. Separateness can be measured either based on the cluster members closest to the next cluster or by the distance between the centroids of the clusters. Cohesion is a measure of the "tightness" of the clusters. Cohesion is a bit more difficult because it looks at the relationship between terms in the cluster for every cluster. Dunn's indices, which incorporate both separateness and cohesion, interprets cohesion as the maximum distance between any two elements of a cluster. However, this approach is subject to a high degree of influence by noise. Basically, noise would define the cohesion

of the cluster (Halkidi et al, 2002). An alternative, yet more complex, approach is to calculate the average pairwise similarity between each term in a cluster and calculate a map value by averaging those values (Watts et. al, 2002).

Essentially all of the research on text clustering is based the need for document retrieval. This research differs from other research in that it utilizes term clustering as a means of understanding concepts within documents and not a representation of documents. VantagePoint is one of the few software packages that have this feature. Some of the methods mentioned throughout the literature review are based on frequent terms but do not display the relationships among term concepts. Therefore, some of the approaches taken must be adjusted to consider the representation of terms and not documents.

2.2.5 Visualization

A review of information visualization literature can be a complex process. The visualization of text data mining results may be about visualizing text but not in this area exclusively. It may include the visualizing of data about text. Therefore, valuable information can be found in general data visualization literature, as well as literature concerned with the visualization of textual concepts.

The big challenge that hinders the effectiveness of text data mining techniques is the visualization of results. An effective interface should allow the user to review, manipulate, search, explore, filter, and understand large volumes of data (Gershon 1997). The challenge is to integrate human perceptual abilities to large datasets (Kein 2001). However, the melding together of the two powerful elements, the human mind and the computer, is limited by the fact that the communication must pass through some form of

video display, severely reducing fast, complex communication (Cawkell 2001). Information visualization has emerged as a field merging human computer interaction and information retrieval in large databases (Hawkins 1999). However, there is no comprehensive theory of information visualization to handle the increasing scale of datasets, nor is there a general methodology to measure the effectiveness of large data set representations (Fabrikant 2001) . General understanding of the dataset is generally not assessed in research (Fabrikant 2001).

The visualization literature covers three main areas: a presentation of general visualization principles, results from task completion testing of a selection of representations and descriptions of tools under development. The articles generally overlap at least two of these categories. The most advanced work in tool development and task completion appears in conference proceedings. Very little work, even in journals, is focused on the theoretical elements of data representation, especially in relation to the representation of text.

An early work by Robertson, Card, and Mackinlay (1993), addressed the issue of increasing the speed of information access to complete work processes. They suggested various types of visualizations depending on the type of data. They suggested using a cone tree to represent hierarchical structures, a perspective wall for linear structures, a data sculpture for continuous data, and an office floor plan for spatial data. (Robertson et al. 1993) Text analysis could result in any of these data types.

It may seem odd that an article in 1993 is considered an early work in visualization. However, Cawkwell (2001), who provides an overview of the history of visualization research, notes the limited inclusion of visualization as a research

component of Information Science. He concludes that only seven relevant books have been written on the topic and four of the seven had been in the prior three years. Early work in visualization work conducted by Tuftes was based on one simple premise "there are right and wrong ways to show data; there are displays which reveal the truth and displays which do not" (Cawkell 2001). The article shows the difference between a simple flow diagram for representing citation analysis results and a more complex representation of literature completed in Pathfinder with zoom in capabilities and providing a "mind's eye" representation of a large corpus. He also mentions one very important work in Information visualization. Card (1999) published a book which is the most cited information visualization book. It is a collection of classic visualization papers. Card includes articles that deal with visualization space and dimension, user interaction with visualizations, focus + content methods, and visualization tools. Document Visualization is one of eight chapters in this 650 page book. The chapter looks at document visualization in 1D, 2D, 3D, and 3D+ Time, including perspective walls, network diagrams, and 3D geographical representations.

3D is becoming a popular element in more complex visualization tools. However, the value of 3D or the optimal use of 3D is yet to be determined. Sebrechts et al (1999) compared 3D, 2D, and text versions of the visualization tool NIRVE on a corpus of documents that had been clustered. The participants were given some information and asked to complete tasks requiring them to locate, compare and describe documents or clusters. The structure of the documents was hierarchical. The 3D condition was presented as cluster boxes on the surface of a sphere and the 2D condition was simply a flattened sphere. The speed of task completion was the measure used to evaluate the

results. The text presentation had the fastest average completion time. However, as the participants gained experience, the 3D representation showed significant improvement. The use of color was another way to add dimensions to the visualization used in this study. Color itself was not tested as a variable. However, users were allowed to utilize a color feature that associated different concepts with colors. Users consistently made use of this feature. However, the effectiveness of color was found to decrease once there were more than five concepts shown (Sebrechts et al. 1999). Many studies reveal the value of color as an added dimension. The most valuable use of color appears to be when it is used to represent concepts (Sebrechts et al. 1999) (Stasko et al. 2000).

Baker and Bushell (1995) experimented with adding dimensions and clarity to data visualization by revisiting and improving upon a classic visualization video, "Study of a Numerically Modeled Severe Storm". This article provides a number of instructive points on good visualization, such as the concept of "just noticeable difference," a parameter referring to the minimal variation in color required for a human to perceive change. Colors should reflect some natural order to enable the viewer to make a mental link between the image and what the image is representing. The linkage may differ depending on the field of study. The article mentions that if color is being used to represent quantitative data, then the hues should be clearly discriminant and matched to a color bar. They also advise using no more than seven colors in order to keep the colors distinguishable. This idea is similar to those perpetuated in the Sebrechts paper, which found difficulty with distinguishing after five colors were used. The article demonstrates the benefits of coordinate axes, labels, and visual cues among other helpful tips for visualization. Animation is one of the visual cues discussed. It is used to show the

formation of a cloud. The actual steps to the formation are shown using a timeline. In the same way, the formation of a cluster could also be represented. In general, however, the article does advise limited animation activity.

Animation is one way the cluster development could be visualized. Yet, how will the actual clusters be represented? An article by Henry Small on visualizing scientific relationships determined by mapping citation links provides one method. He presents a directed graph with chronologically ordered vertices and nested maps.

Fabrikant (2001) published three papers related to mapping text information. She is a geographer who applies principles of cartography to mapping Reuters documents. A hierarchy is created using Latent Semantic Indexing to create layers of granularity in the topic clusters. The topics are then depicted in the same way that a state consists of multiple counties and counties consist of multiple cities.

In text data mining clustering is an important concept that requires visualization, especially for large cluster maps. One of the challenges is providing the ability to maneuver through a cluster map, focusing on the details of the cluster, while maintaining perspective in relation to the entire cluster. Also, finding specific details in the midst of a large cluster map remains a challenge. Kosara et al (2002) present an interesting concept to aid in viewing detailed information while maintaining an accurate sense of location in the data. They take existing Focus + content methods like hyperbolic trees, and have created a concept called Semantic Depth of Field (Figure 2.5.) In the text application, the lines surrounding a keyword are displayed yet slightly blurred. The keyword sentence is sharp with the keyword highlighted.

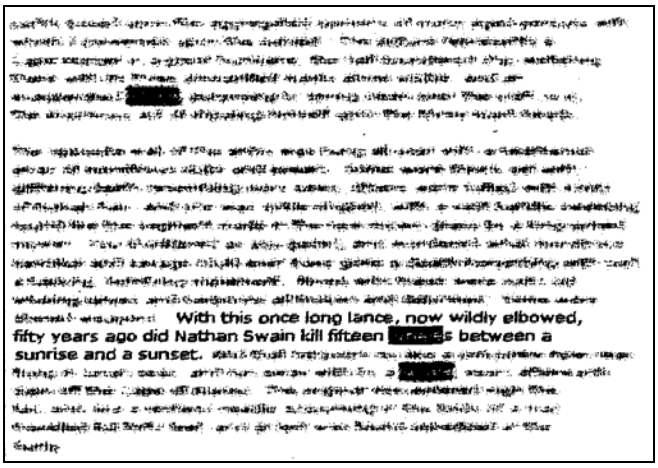


Figure 2.5 Semantic Depth of Field.

The idea of using blur and cues can also be applied to finding documents in a tree structure or identifying an additional dimension of trends in a scatter plot. This same idea can be used with other methods. The authors apply the same idea to a geographical map made in layers. Textual concepts first put into a geographical metaphor may also apply these concepts to improve search capabilities (Kosara et al. 2002).

Throughout the literature on visualization are visualization principles. However, Mirel (1998) provides the most extensive review of the literature that summarizes important principles. The article provides a plethora of principles for usable visualizations. It also addresses problems in usable visualizations and areas where research is weak. The principles provide important guidelines that can be applied to cluster maps, reiterating much of the other literature such as avoiding over-labeling, and noting that icon appearances should represent their function. Mirel also summarizes what is known in the research regarding three areas: perceptual sophistication of users, completing complex tasks, and visual querying. Mirel cites studies that stress the need for simplicity in representation and the proper use of color. The most important aspect of this work is the array of areas that she determines have not been studied effectively. Those

areas include the effect of users' prior knowledge and expectations on their interpretation of what they see (a fact very relevant to the analysis of text data mining results), the trade-off between the need for labels and the limits of screen real estate, the limitations in users' perceptual processing capabilities, users' actual questions, user patterns of action and points of comprehension, designs that best aid problem-solving and recall, and the affect of devices for querying (Mirel 1998). The visualization in this research is dependent on the capabilities of VantagePoint

2.3. Literature Review Conclusion

For the purposes of this research, it is important to note the parallels between the technical intelligence and text data mining processes. Any business effort must begin with planning as portrayed in the intelligence cycle. In the text data mining that we are offering, the planning stage consists of the development of the search strings, the appropriate databases, and identifying the needs of the target user. Collection is the next step, which is simply done by retrieving the documents from the appropriate databases. Processing entails the steps of Extraction and Cleansing. This research will add to this body of knowledge by developing an algorithm to improve the accuracy of the representation of abstract phrases. Analysis is the main focus of intelligence efforts. This research project intends to demonstrate that the analysis of technical publication abstracts provides viable technical intelligence. Dissemination and Visualization both relate to the actual interaction between the user and the information. There are many opportunities for future work in this domain. Measuring the impact of the analysis is challenging. There is also significant space for future research in this domain.

CHAPTER 3

DESCRIPTION OF RESEARCH

The context of this research centers on the problems encountered by technology managers and the utilization of a text-mining tool called VantagePoint to address some of those problems. It is a commercial text data-mining tool designed to analyze text gathered from large databases in order to scan the records, identify trends, profile, map, and decompose technologies. The basic elements of VantagePoint include list creation and grouping capabilities, Natural Language Processing (NLP) parsing of abstract and title phrases, stemming-based list cleanup capabilities, a versatile thesaurus creation and editing ability, matrices, and three variations of Principal Component Analysis. VantagePoint is configured to input datasets consisting of detailed abstract records from a variety of technical databases. Each record contains several fields including the title of article, the authors, the first author's affiliation, the year that the article was published, the source, country, keywords, and an abstract of the work. Frequency lists of the items in these fields can be created as well as matrices demonstrating co-occurrence relationships. The lists can be cleaned in two ways: through the list cleanup function or through the thesaurus function. The thesaurus can also add additional dimensions to the analysis. For instance, there is a thesaurus that brings together universities, corporate entities, and government organizations into three simple terms named for the affiliation type. Similarly, these different types of entities may simply be grouped together. Analysis functions such as matrices or maps allow these groups to be analyzed as single entities and allows for analysis of a single group individually. VantagePoint offers three type of mapping functions: the standard factor map which may, for example, map the

relationship among selected terms based on their co-occurrence with each other. This type of map shows clusters of terms and the relationship among clusters; the autocorrelation map shows individual items such as authors with links showing the strength of the relationship among them. The autocorrelation map is equivalent to taking the individual terms in a cluster from a factor map and showing the links among those particular terms. The cross-correlation map shows the relationship among terms based on the co-occurrence of items in another field. For example, the map may show the strength of relationships between authors based on the keywords that they publish under rather than the fact that individuals simply publish with each other. With these basic functions a number of metrics can be developed for analysis.

This research project has two main focal areas. First is the development and evaluation of useful technology monitoring metrics from publication databases as determined by the needs of technology decision-makers. Second are individual projects intended to improve the text data mining process that creates the technology metrics. The approach taken in this research is a six-step process that loosely follows the chronology of conducting both intelligence and text data mining activities. The six steps in this research project are:

1) Determine the Technologies/Functions to be Monitored
2) Determine the Information Needs of Technology Decision-Makers
3) Develop a Concept-Clumping Algorithm
4) Compare Keywords and Abstract Phrases Clusters
5) Determine Metrics for an Example Technology
6) Evaluate Framework

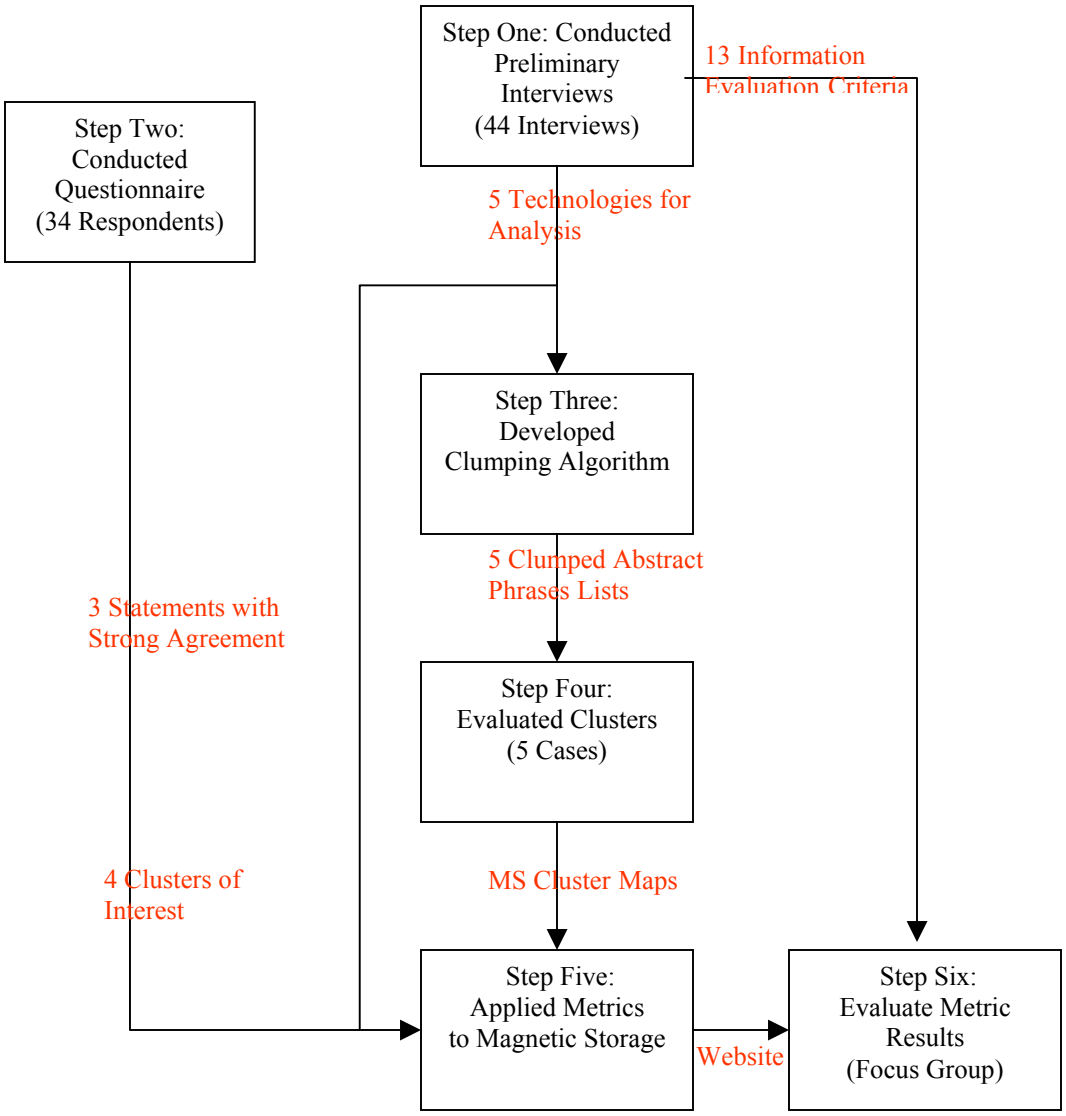


Figure 3.1 Research Information Flow

Figure 3.1 shows the relationship among these steps. The steps can be grouped into three parts. The first two steps frame the needs of the target users. They correspond to the "Planning and Direction" step in the Intelligence Cycle. These two steps determine the technologies of interest to the target audience, their technology challenges, and the information needs of these technology decision makers. The second part, which includes

Steps Three and Four, is concerned entirely with data preparation. In this section, the research is concerned with the effectiveness of the text-mining tool. Step Three introduces an algorithm for clumping together terms that are conceptually the same, but have not been reduced to one term in the stemming-based list cleanup function included in VantagePoint. This step clearly corresponds with the "Processing" step in the Intelligence Cycle. In Step Four, the effect of using this algorithm on multiword abstract phrases in concept clustering is evaluated and using abstract phrases is compared to the traditional method of clustering keywords. The final section, consisting of Steps Five and Six, is focused on evaluating the results of using this methodology. At Step Five, the information needs determined in Step Two are linked with a VantagePoint capability. Those metrics are determined for one of the technologies of interest found in the initial interviews, corresponding with the "Analysis" step in the Intelligence Cycle. Then, users can view the results on a website. The website corresponds to the "Dissemination" step. Finally, in Step Six, a subset of the decision makers initially surveyed evaluated the website. The methods utilized at each step will be described in detail in the sections below.

3.1. Step One: Determine the Technologies/Functions to be Monitored.

"Scoping" is bounding the investigation in order to identify its purpose. The project must bound technological functions, systems, applications and institutions (Rossini et al. 1998). In order to bound the project, a group of accessible technology decision-makers were interviewed to determine the technologies and challenges that they work with every day. First, a group of 25 Principal Investigators from the United States Army Construction Engineering Research Lab (USACERL) were interviewed as part of a

project sponsored by CERL and the Army Environmental Policy Institute (AEPI). Although these individuals are not in commercial competition with anyone, they do compete for research funding. They must also have knowledge of technical issues and the technical landscape in order to make the decisions that are best for the objectives that they are trying to accomplish. A rolling interview process starting with a list of corporate and government organizations obtained through personal contacts added an additional 19 participants. These individuals were sent an introductory e-mail naming the person who referred them, along with a profile of an appropriate interview subject (Figure 3.2.) Some of the initial contacts were only useful for the purposes of obtaining new names. The interview consisted of 11 questions (Appendix C) based on Herring's Key Intelligence Topics (KIT) Process (Herring, 1999).

Hi,

I am conducting research on the utilization of information products in technology decision-making. I received your name from _____, because _____. I was wondering if I could have 30 minutes of your time to speak with you about my research and to interview you about issues related to technology decision-making. I would also be interested in speaking with anyone that you know who is a technology decision-maker or researcher in a business/government entity that meets at least two of the following criteria:

1) Technologically dynamic environment where the pace of change is rapid?
2) New technologies are likely to surface?
3) Technology is a differentiating competitive factor?
4) Product introduction rate is fast?
5) Market entry timing is important?
6) Regulatory approval of new products is complex?
7) Manages a significant R&D portfolio?

If you are able to assist me in anyway, please reply to this email with your availability for an interview and/or additional contacts. Thank you for your time.

Figure 3.2 Participant Evaluation Email

The interviews determined the particular technologies and/or functions that are of interest to the target audience and to bound the analysis (Rossini et al 1998). While not used in this research, note that analysts may use other techniques to scope a project including brainstorming, expert opinion techniques like the Delphi method, research through initial monitoring, and initial bibliometrics. In order to determine the set of topics for this research, the participants answered questions about the technology challenges and decisions that they face and the terms they use in order to search the Internet or library databases for their technology interests. Five technologies relevant to the target audience became the basis of the analyses conducted. An initial list of search terms was developed and a search conducted. VantagePoint produced an initial keywords list, which was used to identify documents unrelated to the technology in the record set. Adding additional terms to the search string, using "AND" or "NOT," reduced noise in the dataset. "Noise" entails records that are not related to the technology. Another method is to force the terms to be adjacent (ADJ) or near (NEAR#) each other in the record. In this research, datasets for five technologies were retrieved for analysis.

In addition to the primary purpose of scoping the research studies, the interviews obtained demographic and industry information, determined technology issues, determined information sources and how the sources were utilized, and determined how the decision makers preferred to obtain information. The interviews were coded according to relevant answer characteristics. For example, the participants were asked which terms they used in searching the internet or library databases for their technical area. If they gave an answer such as "xml", it was coded as "specific technologies." An answer such as "security" was coded "broad technology area." The interview results were

then summarized and this additional information obtained in the interviews helped gauge the expectations of the decision-makers, contributed to framing the questionnaire in the next step, and provided the criteria for evaluating the final information product.

3.2. Step Two: Determine the Information Needs of Technology Decision-Makers.

While the interviews provided the technologies to be monitored, a questionnaire found the common information about the technologies that most interest decision-makers. The CTI and Management of Technology (MOT) literature provided the general basis for the questionnaire (Appendix D). The participants rated their level or agreement on various statements about information that may aid in their decision-making. Thirty-four technology decision-makers from a variety of Corporate, Academic, and Government organizations participated. First, the mean score and standard deviation for each statement was calculated. The statements were then ranked and categorized according to the level of agreement with those statements (Table 3.1). The statements with the highest level of agreement have the lowest average score.

Table 3.1 Questionnaire Score Categories.

Category	Category Range of Score
Strong Agreement	1.0 - 1.6
Agreement	1.6 – 2.2
Neutral	2.2 – 2.8
Disagreement	2.8 – 3.4
Strong Disagreement	3.4 – 4.0

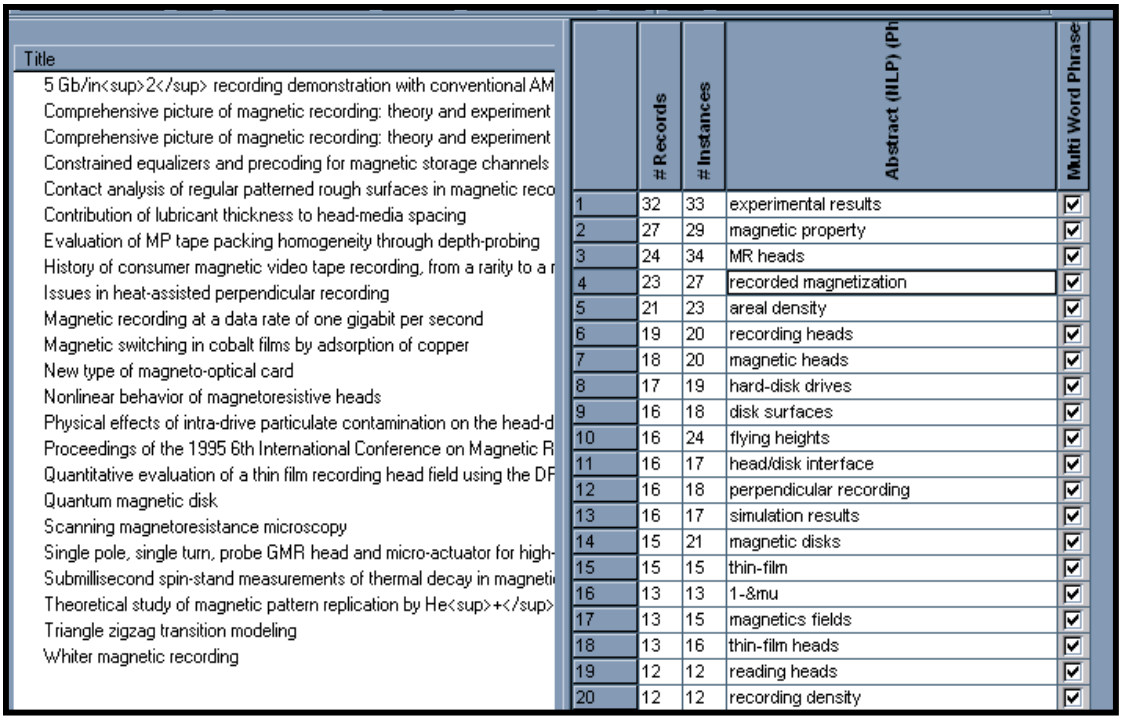
Further analysis identified patterns in the responses. Factor analysis determined groups of statements that decision-makers might be interested in obtaining as a package.

This method for evaluation provides an opportunity to determine the information that technology decision-makers deem important, and to determine if managers agree with the statements made in the literature. The pattern analysis using correlation and clustering allows the opportunity to identify any linkages between the information needs and the view of information, the technology challenges, or any demographic information. The results from this section drive the monitoring metrics that will be described in Step Four.

3.3. Step Three: Develop a Concept-Clumping Algorithm

In order to more effectively discuss the data cleansing process, several definitions are needed. For the purposes of this research, a "word" is a string set apart by spaces, a "phrase" is one or more words, and a "term" is a phrase that is identified by VantagePoint to be a unique phrase from the abstract. A "phrase" consists of one or more *words* and every *phrase* belongs to a set of *phrases* that is a subset of *words* in a *term*. Each line in a VantagePoint abstract phrases list is considered a "term." Figure 3.3 depicts a screen shot from an Abstract Phrase List in VantagePoint. "Recorded magnetization" has a highlight box around it. This term is also a two-word phrase.

This research investigates the use of abstract phrases in clustering; however significant clean up is required to adequately utilize abstract phrases. While a number of methods seek to determine the appropriate words for use in clustering, most do not address a particular challenge unique to text, which is that there are words that are essentially the same in meaning but which are written or extracted in slight variations. Words such as "engineering science" and "general engineering science" should be



Title		# Records	# Instances	Abstract (IILP) (Ph	Multi Word Phrase
5 Gb/in² recording demonstration with conventional AM	1	32	33	experimental results	☑
Comprehensive picture of magnetic recording: theory and experiment	2	27	29	magnetic property	☑
Comprehensive picture of magnetic recording: theory and experiment	3	24	34	MR heads	☑
Constrained equalizers and precoding for magnetic storage channels	4	23	27	recorded magnetization	☑
Contact analysis of regular patterned rough surfaces in magnetic reco	5	21	23	areal density	☑
Contribution of lubricant thickness to head-media spacing	6	19	20	recording heads	☑
Evaluation of MP tape packing homogeneity through depth-probing	7	18	20	magnetic heads	☑
History of consumer magnetic video tape recording, from a rarity to a r	8	17	19	hard-disk drives	☑
Issues in heat-assisted perpendicular recording	9	16	18	disk surfaces	☑
Magnetic recording at a data rate of one gigabit per second	10	16	24	flying heights	☑
Magnetic switching in cobalt films by adsorption of copper	11	16	17	head/disk interface	☑
New type of magneto-optical card	12	16	18	perpendicular recording	☑
Nonlinear behavior of magnetoresistive heads	13	16	17	simulation results	☑
Physical effects of intra-drive particulate contamination on the head-d	14	15	21	magnetic disks	☑
Proceedings of the 1995 6th International Conference on Magnetic R	15	15	15	thin-film	☑
Quantitative evaluation of a thin film recording head field using the DF	16	13	13	1-&mu	☑
Quantum magnetic disk	17	13	15	magnetics fields	☑
Scanning magnetoresistance microscopy	18	13	16	thin-film heads	☑
Single pole, single turn, probe GMR head and micro-actuator for high-	19	12	12	reading heads	☑
Submillisecond spin-stand measurements of thermal decay in magneti	20	12	12	recording density	☑
Theoretical study of magnetic pattern replication by He⁺					
Triangle zigzag transition modeling					
Whiter magnetic recording					

Figure 3.3 Screenshot of VantagePoint

common words such as “the” and “of.” VantagePoint adds a step to the cleaning process by using a stemming algorithm to group two terms into one. With the stemming algorithm, “computer” and “computers” become one term. However, this method is insufficient. Since the purpose of this research is to make technology linkages, more accurate concept representations means more accurate end-results. The discussion that follows highlights need for a *concept-clumping algorithm* when working with abstract phrases.

While using abstract phrases may be desirable, doing so can potentially introduce problems. One of the problems with abstracts is the variation in the words that are used. In an abstract, there are words that provide no conceptual insight into the content of the paper, such as “novel means” shown in the example in Table 3.2. Additionally, there are

occasions where the same concepts may be discussed in a variety of ways, even within the same abstract.

Therefore, in order to effectively analyze the information, the data must be cleansed and clumped to accurately portray the prevalence of the concepts in the dataset. The idea is to remove as much "junk" as possible and to combine words that represent the same concept. Many such methods were discussed in the literature review. The data clumping algorithm developed for this research first identifies a list of relevant noun phrases and then applies a rule-based algorithm for identifying synonymous words based on shared words in each phrase. The algorithm does not claim to be generalizable to all

Table 3.2 Sonochemistry Keywords vs. Abstract Phrases.

List of Keywords	List of Abstract Phrases
• Pollution control • Sonochemistry • Mass Transfer • Ultrasonic applications • Reaction Kinetics • Sonochemical Reacting Systems	• Environmental Sonochemistry • Environmental remediation • Ultrasonic waves • Kinetic analysis • Sonochemical engineering • Chemical analysis • Mass transfer • Aqueous solutions • Chemical processing • Cheaper reagents • **Novel means** • Shorter reaction cycles • Smaller plants • Large-scale applications • Growing area • Existing knowledge • Outline directions • Exciting field

The table above comes from the article "Sonochemistry: Environmental Science and Engineering Applications ". It demonstrates the difference in terms listed in the keywords list versus those listed in the abstract phrases list.

text sets, but is intended for use with technical periodical abstracts. Further research is necessary to determine the generalizability of results to other types of text document sets.

In performing additional phrase clumping, the intention is to increase the analytical validity of using abstract phrases to perform additional analysis. The basic outline of the algorithm is as follows:

1) Remove hyphens, numbers, punctuation
2) Remove common words
3) Clump phrases with four or more words in common into a new phrase.
4) Name the new phrase the shortest phrase name
5) Calculate the importance of the remaining words
6) Clump phrases with three words in common into a new phrase
7) When a conflict arises, use a similarity measure to determine with which group of phrases that the conflicted phrase will clump.
8) Name the new phrase the phrase name with the highest prominence
9) Repeat steps 5) – 7) for two word matches.

The basic starting point for the algorithm is a cleaned list of abstract phrases as determined by VantagePoint. Non-alphanumeric characters are removed. This step combines terms such as "high-density" and "high density." Then, the algorithm removes common single words from the list. Common single words are removed from the list using a published list from White (1999) of the most frequently used words from two to ten letters. The goal of the research is to use Abstract Phrases to map technical terms. Common words, such as "study," "uses," or "results," may occur frequently in the dataset, yet do not represent technology and may falsely show relationships more related

to the type of article, rather than a technical relationship. The list was revised to remove potential technical terms. Any terms left on this list are eliminated. Finally, with only multiword noun phrases and uncommon single word noun phrases remaining, the list is ready for analysis. VantagePoint abstract phrases list consists of only noun phrases. Noun phrases are used because language research finds that text is understood by understanding noun phrases (Chen and Chen, 1994).

Once the list of relevant words has been determined, the clumping algorithm is applied. The basis of the remaining portion of the algorithm is the existence of shared words. Shared words are the words that exist together in more than one term. For example, engineering science and "general engineering science" share two words. Identifying equivalent concepts is a difficult process; by starting with shared words, a high level of precision can be achieved and the number terms compared to one another is limited.

The algorithm searches for phrases with four words in common. If a phrase has four words in common, these words will be combined together and named for the shortest phrase. In the rare occasion that a conflict arises, VantagePoint chooses the first grouping that occurs in the thesaurus. This approach is somewhat random, however, some initial analysis revealed that these terms are likely *all* conceptually the same and should be grouped together in the three-shared words step in the algorithm .

Secondly, phrases sharing three words in common will be grouped together and given a prevalence rating. The formula for the prevalence rating is:

$$P(b) = \sum_{\substack{\forall \text{ Docs where} \\ (b) \in D(i)}} \frac{\text{Instances of (b) in D(i)}}{\text{\# of relevant phrases in Doc (i)}} \qquad (3.1)$$

where:

P(b) = prevalence rating for term (b)

(b) = a term in the abstract phrase list

D(i) = the set of terms contained in Document (i) in the record set

In VantagePoint, the prevalence ranking will be determined by first constructing the matrix of "records" by "*group* of relevant abstract phrases" using the *instances* option for the matrix cells. For each relevant abstract phrase, the cell value will be divided by the row total. Each of these values will be summed together. This method is used because it gives a higher rating to both words that appear in many documents and words that appear more frequently in one document. Words are also given a higher prevalence if they appear in shorter abstracts.

Once the prevalence rating is determined, the algorithm searches for groups of terms that share a three-word phrase. These terms are clumped into one term using the *thesaurus* feature in VantagePoint. If a term shares phrases with multiple groups, a similarity measure will determine the group to which the term belongs. The basis of the similarity measure is the standard approach to similarity used in Information Retrieval where similarity of terms has been researched most frequently. The premise is that two terms are semantically similar if they occur in the same context (Crestani, 2000). Other approaches to similarity are taxonomy-based. The similarity between two items depends on the relationship or distance of the terms in a hierarchically structured lexical resource, such as WordNet (Basu et. al, 2001). Taxonomy-based approaches would require incorporating a lexical resource such as WordNet into Vantage Point. Such a resource would have to map technical terms. A problem with such an approach, for the purposes

of this research, is that the terms that are most likely represented differently in the record sets occur in newer technical areas. These areas would less likely appear in a lexical resource. Therefore, a contextual similarity approach is more suitable for technical publications. This algorithm from Cutting et. al (1992) asserts that a term is most similar to the group of terms that co-occur with words most like the terms with which the conflict term co-occurs.

Therefore, for each document α in a corpus C, let $c(\alpha)$ be each word in the document and its frequency. Let V be the set of unique terms occurring in C. Then $c(\alpha)$ can be represented a vector of length $|V|$;

$$c(\alpha) = \{f(w_i, \alpha)\}_{i=1}^{|V|} \tag{3.2}$$

w_i = ith word in V

$f(w_i,\alpha)$ = the frequency of w_i in α.

Using the cosine between monotone element-wise functions of $c(\alpha)$ and $c(\beta)$, the similarity measure between two documents can be determined by

$$s(\alpha, \beta) = \frac{\langle g(c(\alpha)), g(c(\beta)) \rangle}{\|g(c(\alpha))\| \|g(c(\beta))\|} \tag{3.3.}$$

where g is a monotone damping function using a component-wise square-root, "(,)" denotes inner product, and "|| ||" denotes vector norm.

If similarity is considered to be a function of document *profiles* $p(\alpha)$, then

$$p(\alpha) = \frac{g(c(\alpha))}{\|g(c(\alpha))\|}, \tag{3.4}$$

in which case

$$s(\alpha,\beta) = \langle p(\alpha), p(\beta) \rangle = \sum_{i=1}^{|T|} p(\alpha)_i \, p(\beta)_i. \tag{3.5}$$

Applying the aforementioned equations to match the similarity between the group of documents in which the group of terms that share a phrase appear (Γ) and the documents in which the term that shares phrases with multiple groups appears (x). Then, let Γ have a profile defined as the normalized sum of profiles of the contained individuals. Therefore,

$$\tilde{p}(\Gamma) = \sum_{\alpha \in \Gamma} p(\alpha) \tag{3.6}$$

is the unnormalized sum profile, and the normalized profile is

$$p(\Gamma) = \frac{\tilde{p}(\Gamma)}{\|\tilde{p}(\Gamma)\|}. \tag{3.7}$$

By employing this definition, the cosine measure can be extended to Γ and the similarity between a document x and the document set Γ can be found by the following equation:

$$s(\Gamma, x) = \langle p(\Gamma), p(x) \rangle. \tag{3.8}$$

Once all of the three common phrase matches have been made, the "two common word" clumping process will take place. The same process utilized in matching terms that share three common words is utilized to match terms that share two common words. The starting point is the prevalence ranking for the appropriate terms. This research stops at two shared-words in common. Future research may look at improving the algorithm to

effectively handle terms that only share one word in common. The assumption is that as the number of shared words decreases, the less likely it is that the shared words indicate a similarity and, therefore, different approaches will be necessary.

"Precision" tests the ability of the algorithm to accurately identify that two words are synonymous. The overall precision was evaluated by running the algorithm against an abstract record corpus. Each term was manually compared to the term that the algorithm named the group for determination as to whether it is actually similar in concept.. The naming algorithm is important because it ultimately determines the term that is chosen to represent all of the terms in the group.

3.4. Step Four: Compare Keywords and Abstract Phrases Clusters

Prior analysis conducted using VantagePoint has taken advantage of the Keywords contained in abstract records provided by the database company (Watts, 1998 and 2000). However, these Keywords are more generalized than Abstract Phrases and, as mentioned, often come from the database provider and not the author of the paper. This research hypothesizes that using Abstract Phrases provides a more informative set of clusters than Keywords. The abstracts provide a richer source of information. For example, compare the keywords and Abstract Phrases found in the pollution prevention articled entitled Sonochemistry: Environmental science and engineering applications (Table 3.2). The Abstract Phrases are more specific. This research hypothesizes that the more specific term creates more meaningful clusters.

Before actually engaging in clustering, the method for clustering must be determined. For scientific publication database clustering, three scenarios were evaluated

1. Keywords

2. Cleaned Abstract hrases
3. Clumped Abstract Phrases

Clusters were created for Keywords, Cleaned Abstract Phrases, and Clumped Abstract Phrases for a sample from each of the five datasets crafted from the search terms found in Step One. These clusters were compared to each other.

In general, existing research has avoided declaring a standard measure to operationalize "good" or "better" for clusters, primarily because different clustering algorithms have different guidelines. For example, one traditional approach to testing precision and recall is to craft datasets to determine if the algorithm places the documents in the same sets as would be performed manually. This approach is not feasible when using PCA in term clustering, because not every term is placed in a cluster, terms can be placed in multiple clusters, and many terms may be found in the same document. However, numerous quantitative measures are available and this research will investigate the clustering alternatives from various viewpoints -- both qualitative and quantitative -- to evaluate the clusters. First, a profile of the clusters will be created using basic qualitative or pseudo-quantitative information.

- A Description of the clusters
- The number of clusters and links
- The Strength/number of links between clusters
- The number of words per cluster (average and distribution)

As stated in the literature review, quantitative methods to evaluate clustering center around four issues: the separateness of the clusters, the tightness of each cluster, precision and recall. Precision and recall are used when every document is placed in a

cluster. There are numerous methods for determining separateness because many of the clustering algorithms are based on separateness. This evaluation used entropy as defined in the paper by Watts (2002). Entropy for each cluster is calculated by

$$Entropy_j = -\sum_{i=1}^{m} P_{ij} \log(P_{ji}). \tag{3.9}$$

Pij represents the probability that a member of cluster j also belongs to cluster i. It is calculated in VantagePoint as (the number of co-occurances of terms in group j)/(the number of records in which terms from group i appear), m represents the number of clusters in the cluster map. In order to calculate the entropy for the entire cluster map, we define

$$TotalEntropy = \sum_{j=1}^{m} \frac{n_j * Entropy_j}{n} \tag{3.10}$$

where n_j equals the number of abstracts in cluster j, and n equals the total number of abstracts in the dataset.

Cohesion is measured by the average similarity of the terms in each cluster of the record set. As stated in Chapter Two, Watts (2002) proposed a pair-wise similarity approach between the terms in a cluster. However, this pair-wise similarity approach defines the cohesion of the clusters along the dimensions of inclusion or exclusion in documents as opposed to co-occurrence with similar terms. This difference may seem inconsequential except that there actually can be a difference in the calculation, when compared with the approach used in this paper (Table 3.3). Table 3.3 shows the difference in Cohesion measures for three approaches on a sample of clusters. This research is more interested in cohesion based on terms. Therefore, a different approach to cohesion is taken. Since, cohesion is a measure of the similarity of the terms in a cluster

and similarity of terms has been used with other terms in a group for our clumping algorithm, the same approach is utilized for cohesion. Cohesion is defined as the average similarity between each term and the other terms in the group as defined in equation (3.7). A "tighter" cluster has a larger value.

Table 3.3 Comparison of Cohesion Measures

Fuel Cell Clusters	Cohesion Document-based (duplicated pairwise)	Cohesion Document-based (exclusive pairwise)	Cohesion Term-based
carbonates	0.54	0.31	0.739
cathodes	0.41	0.21	0.783
synthesis	0.38	0.18	0.703
polyelectrolytes	0.69	0.38	0.752
methane	0.22	0.14	0.736

Finally, an overall assessment of the trends and differences among the clusters was assessed.

3.5. Step Five: Determine Metrics for an Example Technology

The questionnaire identified information important to decision-makers as well as groups of important related information. For each of these statements, the associated metric in VantagePoint was found. (See Appendix E for the association between each question and the publication metrics.) Those metrics were applied to a random sample taken from the dataset of one of the selected technologies. Sampling kept the size of the dataset manageable. A macro in VantagePoint tagged every nth record and created a new dataset. The goal was to keep the most records possible while allowing for reasonable run time for the algorithms in VantagePoint. The metrics calculated were:

1) Affiliations appearing/disappearing in the Dataset: This metric identifies organizations that are just starting to publish on the technology or have potentially abandoned publishing. It is an indicator of either burgeoning or waning interest in the technology.
2) Links between affiliations and
    a. Terms: Indicates the topics that an organization has expertise
    b. Other affiliations: indicates collaboration among various organizations
    c. Proper noun phrases in the abstract: indicates the specific interests for an organization
3) Abstract phrases appearing/disappearing: indicates specific topics within the technology domain where research interest is either burgeoning or waning.
4) Keywords appearing/disappearing: Indicates broad topic areas where interest in research in relation to the technology is either burgeoning or waning.
5) Relationships among keywords and abstract phrases: Indicates the research relationships that exist among topics within the technology domain.
6) Comparisons between the United States and Other Countries: Identifies differences in research behavior between the United States and the International Community
7) Journals and the topics discussed: Identifies the latest research topics

8) Conferences and topics discussed: Indicates the latest research topics
9) Non-technical terminology: Provides insight into the non-technical aspects of the discussion of a technology. Potential information may include hints into the Life Cycle progress of the technology or social impact issues
10) Cumulative Number of Records/Year: Indicates position on the Technology Life Cycle)

These metrics were organized on a website according to results from the questionnaire.

3.6. Step Six: Evaluate Framework

In software engineering, an empirical evaluation of a system may test the usability of the system design. In such cases, the user is required to make choices about an interface. In this case, the evaluation is not of the interface design but of the information provided by the system. Future research investigating the visualization aspects of the system could improve the utility of the information products provided.

Five technology decision-makers, who were familiar with Magnetic Storage technology participated in a focus group evaluating the website information along the dimensions found in the initial interviews. Two additional evaluators, who were not able to attend the focus group, provided their input online. The evaluators in the focus group were a mix of practitioners and researchers. The participants evaluated the information primarily based on the type of information and not the information itself.

The session began with a briefing on the background of the research and instructions. A transcript of the opening statement can be found in Appendix F. Each evaluator sat in front of a computer with the Welcome Screen for the website. The

evaluators were also given a sheet listing the evaluation criteria. They spent 15-20 minutes reviewing the information before the discussion began. Following the review period, the leader went through each dimension and encouraged discussion based on each point. The Results were then summarized in Chapter 6.

CHAPTER 4

FRAMING THE NEEDS OF THE TARGET USER

4.1. The Technologies/Functions to be Monitored

4.1.1 Five Technology Cases

The interviews conducted had three components: Target Audience Profile, Monitoring System, and Early Warning Topics. "Target Audience Profile" consists of seven questions designed to better understand the participants' technology challenges., the sources of information that they use, and how that information is used in the decision-making process. "Monitoring System" consists of two questions related to the ways that the participants would like the findings to be organized and how they evaluate the usefulness of such a system. "Early-Warning Topics" gathers the type of information that the Investigators seek. This section provides the information to support the main goal of the interviews, to determine the technologies of interest to the target audience, from which a subset is used in the analysis steps. From the array of answers supplied, five topic areas were selected to use for analysis. Search strings were developed for each of these areas. The search strings and the total number of abstract records retrieved are as follows:

1. "Fuel cells" or "fuel cell"
2. "Remote sensing" or "remote sensor"
3. "geographical information system(s)" or "geographic information system(s)"
4. "pollution monitoring"

5. “magnetic disk storage” or “magnetic storage” or “magnetic bubble memories” or “magnetic data storage” or “magnetic film storage” or “magnetic tape storage” or “magnetic core storage” or “magnetic heads”

Table 4.1 Technology Cases: Record Counts.

Technology	**Database**	**# Records**
FC	Compendex	7495
RS	Compendex	11,105
GIS	Inspec	7,556
PM	Pollution Abstracts	7500
MS	Inspec	6985

*Information regarding these databases can be found in section 2.1.3.2

4.1.1 The Target Audience

Those interviewed come from a range of areas. Most are from IT or environmental industries. They include representatives from government, corporate, and academic organizations. Their job functions include CEOs, CIOs, Consultants, Head Researchers, and School Deans. The individuals interviewed are responsible either for obtaining funding, deciding on the research that will be conducted, selling their products or services, or implementing a technology. In the sections below are summaries of the responses. These responses came from at least six individuals spread across at least two types of respondents (i.e. different industries or types of organizations.) Throughout the sections are statements that were made by at least a few of the respondents but were not

consistent across profile boundaries. These statements are noted with terminology such as "For some…" or "Researchers…" At the end of this section is a table that captures the overall themes that were dominant throughout the responses.

4.1.1.1 Target Audience Profile

The primary technology challenges faced by these decision-makers are finding and adjusting existing tools to meet their needs. Obtaining long-term, flexible, cost-effective technology planning; and remaining aware of the most advanced options in their field. Other challenges include finding validation for claims for existing products, acquiring funding, managing expectations of technology, and security. In the next year, they will make decisions primarily about new products or methodologies to be applied. These projects may be ongoing from this year or the start of new multiyear projects. A surprising finding was that a number of individuals did not know what they would be working on in the next year. For some, new laws, additions to the endangered species list, or new weapons being developed would be factors, which would drive those decisions. For others, customer demand or the weaknesses revealed by 9/11 would drive that work. Investigating and developing various forms of modeling technologies will also be popular in the coming year, as well as information brokering issues.

These individuals unanimously felt that keeping up with new information is crucial to their success. In order to find technologies that meet their needs, these decision-makers seek information from a number of sources. The most frequently mentioned sources were the internet as a source for research, product information, and even demonstrations; and conferences/conference proceedings. Peer-reviewed journals are the most important source for any type of researcher while many corporate decision makers

did not mention refereed journals. Other sources include professional organizations, government documents, vendors, and other agencies and experts. While these investigators prefer to receive information from abstracting services which they can use to sort through and then access the full documents, newsletters, or conference proceedings, they are willing to receive valuable information in any form. However, many of the subjects do not like e-mail alerts because they were too numerous and usually not useful.

Once obtained, the sources are used in a number of ways. They are used to search for new ideas, to search for products that can solve a particular problem, or to validate the claims of found products. Frequently, the information found is used in proposals or to justify a particular decision. Other times, the information is used simply to increase the decision-maker's overall knowledge and view of the problem area.

The participants were asked the value of and need for intelligence-type information in their decision-making. One of the most common areas mentioned was the desire to better understand the upcoming needs of funding agencies and other decision-makers. The individuals who do not know what projects they would be researching in the upcoming year especially mentioned this need. Government and academic researchers both stated that they would also like to know what other agencies/researchers are doing so as not to duplicate efforts and also to know additional technical options that are available to them. The aforementioned information would save time and money, prevent the decision makers from having a focus that is too narrow, help them to understand importance of issues in the long term, and, in general, help maintain the competitiveness of the organization.

4.1.1.2 Monitoring System

Participants were asked their preferences for the organization of an information analysis system. The responses indicate the features that are important to the participants. However, not all requests are possible to incorporate. One important observation is that the participants wanted to see brief information that connects to full articles, in line with current preferences. This desire implies that the results of a search would be presented as a list of documents. Many participants had not even considered other multidimensional or flexible forms of presentation. This project introduces new forms of presentation to the users while attempting to incorporate the users' preferences. However, the ability to access full articles is limited by the database to which the users subscribe.

The most frequently mentioned features indicate the importance of customization in the searches. The ability to customize searches, set up search criteria, and rate those search criteria were the most frequently mentioned organizational elements. Next, users would like to see information rated according to their search criteria; two methods mentioned were a familiar relevance score or providing an explanation as to how the returned document matched the criteria. Some of the information that the users would like to see about the documents is the value that others have placed on the documents, which can be judged by the citation frequency of the document. A number of organizational sorting methods were also named. The most commonly mentioned were the following: by technology, by geography, by technique, by chronology, by affiliation or affiliation type, by author, by source or source type, and by keyword. Almost all of these arrangements will be possible, by using VantagePoint to analyze the records returned from the search. A more advanced feature recommended by the users is the

capability to learn their behavior from previous searches as the computer receives updates from listed search preferences.

Other information or capabilities mentioned were the ability: to associate certain methods with an author, an overview of the area searched, to mark abstracts already read so that they are not reread, to provide links between technologies, to identify where the overlaps in search criteria take place, to determine partnerships, to find other works by the same author, to determine if studies or methods are duplicated in the research, to have the ability to deal with the variety of jargon used for the same area in the area, to find population trends and distribution and track that as new information comes in—graphical maps, trend charts, the ability to determine a technology's maturity and applications of the technology. One other recommendation worth mentioning is that one participant mentioned Ebay website as a model of how he would like to search for information.

Researchers were also asked how they would judge the credibility of information received. The most popular responses were credibility of researcher, credibility of the journal, and applicability to the problem. Other frequently mentioned methods are the fit with the search criteria, the accuracy of the information, and usability of the information in reporting. Other interesting criteria were if a new bit of information was found, and if a point of contact was included. Only two individuals didn't think it was possible to measure the usefulness either because you couldn't know what was missed or because the information was absorbed into overall thinking and not incorporated in a measurable way. Many participants were focused on finding specific studies and methods to evaluate those studies, which is outside of the scope of this research project. The methods in this

project can only identify factors that may indicate the validity of the study such as the source, affiliation, or author.

4.1.1.3 Early Warning Topics

Participants were asked more specifically what type of information they were interested in regarding the technologies, organization, and experts in their research area. Some of those answers are listed in Table 4.2. The questionnaire incorporated the areas from this list that VantagePoint is capable of capturing.

Table 4.2 Categorized Early Warning Topics.

About the Technology	About Organizations	About Experts	Other
Applications	Who are they	Contact Information	Source type
New developments	What are they doing	What are they doing	Reference sources
Basic information	Motivation	Who are they	Latest political action
How it works	Future plans	Motivation	New guidelines
What exists	Funding available		Language of journal
Latest research	Funding sources		
Health risks	EPA activities		
Patents	Breakdown of barriers between organizations		
Problems			
Technology Parameters			
Associated equipment			
Military applications			
Maturity			
Cross uses of the technology			
Modeling of the technology			
Human or animal research			

The interviews provide an opportunity to become familiar with the work of the decicion-makers and their use of information in that work. One thing that stands out in these interviews is the similarity across a wide range of industry and environments. The participants overwhelmingly view information as important to their success and want new and better ways to obtain applicable information. In general, they are not already performing "intelligence" gathering activities such as monitoring the activities of experts, other organizations, or the technology itself in an organized fashion. The decision-makers brought up many interesting ways in which they would like to monitor information that they are currently not doing, a finding which demonstrates that they, for the most part, have considered these needs. In general, the most important part of mining information for the decision-makers is "not missing anything" and "being able to find new information quickly."

Throughout these interviews, the participants identified how they use information in making technology decisions, the goals that the information helps them to achieve, and how they evaluate the usefulness of information (Table 4.3.) These comments became the evaluation criteria for the information developed by this research framework.

Table 4.3: Summary of Information Evaluation Criteria

1. Uses Peer-reviewed journals. 2. Provides results by category 3. Provides access to new ideas. 4. Used to solve problems. 5. Used in proposals 6. Provides overall knowledge 7. Provides insight into funding sources. 8. Leads to less duplication of research efforts	9. Provides Additional options. 10. Leads to broader research focus. 11. Provides understanding of long-term issues. 12. Leads to organizational competitiveness. 13 Used in strategic technology decision-making. 14. Allows for Customization

4.2 The Information Needs of the Target Users

While the Early Warning Topic interview questions provide some insight into the type of information that decision-makers desire, an online questionnaire provided more specifics information corresponding with the capabilities of VantagePoint.

Individuals who filled out the questionnaire were obtained in a number of ways. Interview participants received an e-mail inviting them to participate in the survey and encouraging them to send the e-mail to others who fit the criteria of a technology decision maker. This method was generally unsuccessful in obtaining a satisfactory sample size. The second and final method involved the internet. Technology decision makers involved on Advisory Boards were contacted and asked to answer the questions by telephone. In the end, 34 individuals completed the questionnaire. The questionnaire consisted of 27 questions and participants rated each statement referring to the importance of particular types of information as either Strongly Agree, Agree, Disagree, or Strongly Disagree. The statements reflected the interview answers and information found in Competitive Technical Intelligence literature. Additionally, the participants recorded their company name, position, and job description. Inferences about the organization type, size of the organization, industry, and level in the organization were made from these responses.

In the sample, 21 participants are corporate and 13 are from either universities or government entities. Nine of the participants are executives and 25 are at the level of Head Researcher, Lead Engineer, or Manager. The organizations are labeled small if they contain fewer than 100 people, medium if there are fewer than 10,000 and large otherwise. Government agencies are placed in the medium category because agencies are

considered individually. The participants come from six different industry groups (Table 4.4). They are also identified by their organization type, the size of the organization, and their position level (Table 4.5).

Table 4.4 Participant Industry Groups.

Industry	**# Participants**
Science/Engineering	3
Computer Technology	13
Energy	4
Environmental	14

Table 4.5 Participant Profile

Position Level		**Organization Type**		**Organization Size**	
9	Executives	21	Corporate	9	Small
25	Non-Execs	13	Not Corporate	19	Medium
				6	Large

Participants generally responded to every question. Two statements are missing two responses and eight statements are missing one response. One participant did not answer four questions.

The responses are assigned numeric values. Strongly agree =1, Agree = 2, Disagree = 3, and Strongly Disagree = 4. Appendix G contains the summary information for each statement, including the mean rating, standard deviation, minimum, and

maximum ratings. The statements fall into one of six categories. The scores are categorized based on the following scale (Table 4.6).

Table 4.6 Questionnaire Score Categories

Category	**Category Range of Score**
Strong Agreement	1.0 - 1.6
Agreement	1.6 – 2.2
Neutral	2.2 – 2.8
Disagreement	2.8 – 3.4
Strong Disagreement	3.4 – 4.0

Table 4.7 Strong Agreement Survey Statements

Question	**Avg Score**
R19: I would like to know what periodicals are publishing in my technical domain.	1.55
R1: I would like to see an overview of the research conducted in my technical domain.	1.56
R21: It is important to know which conferences cover my technical domain and the specific topics covered.	1.59

Table 4.7 lists the three questions that have an average score falling into the "Strong Agreement" category. It is, therefore, important to include the answers to these questions in any information provided to decision makers. Additionally, Factor Analysis was run on the variables using Principal Component Analysis with Varimax Rotation. Factor analysis discovers clusters of information that a decision-maker may want

together. Therefore, the three statements with the strongest level of agreement were removed from the analysis assuming that these would be offered to all decision-makers. A scree plot was used to determine the number of factors to accept. Using .5 as the minimum factor loading, a conservative approach, four factors were accepted. A loading of .45 was used as the breakpoint for high loading terms. Four clusters were identified (Figure 4.1.)

Questions included in each cluster:

Organizational Monitoring- (R: 5,7,9,10,11,12,13,14,16,20)
Technology Progress- Journals- (R: 20, 25,26)
Global Organizational Monitoring- (R: 10,11,13,17,18)
Hiring for Cutting Research- (R: 8,22)

Figure 4.1 Questionnaire Statement Clusters

Cluster 1 includes all questions that are related to knowing what other organizations are doing and it is labeled *Organizational Monitoring.*

Cluster 2 includes statements related to the life cycle of the technology and knowing the topics published in periodicals. It suggests a search of periodicals to determine the progress a technology is making, and it is labeled *Technology Progress – Journals.*.

Cluster 3 includes statements about concern with global activity in a technical domain and general interest in what other organizations are doing. It is labeled *Global Organizational Monitoring.*

Cluster 4 includes statements about interest in early stage research and interest in recruiting from universities conducting research in their technical domain. This activity may indicate that these respondents are particularly interested in working on cutting edge technologies, and the factor is labeled *Hiring for Cutting Edge*.

In addition to the clusters, there are other sets of metrics that a decision-maker may want to choose. For example, as can be expected, individuals whose technical decisions were affected by external human factors were also interested in the impact of their technology.

Also, the only significant correlation in relation to the position of the respondent is that non-executives are more inclined to believe that commercial readiness could be determined by the information in publication databases and are more inclined to believe that the organization would be impacted by changes by suppliers. This finding may alter the information packets made available for executives versus non-executives.

Most questions fell into the "agreement" category. The remaining responses fell into the "neutral" category. No question had a mean score falling into any of the "disagreement categories. However, all of the questions falling into the neutral category had greater than ten responses indicating disagreement. R3, R17, and R18 had the most number of "strong disagreement responses, with at least three. R17 and R18 are related to the importance of information related to behavior in other countries. R3 states, "The organization has been slow in detecting emerging technological breakthroughs in our domain." It is interesting to note that this statement has the largest standard deviation of all of the responses, with the highest mean disagreement from the participants in the energy industry. This response also has a significant correlation with R17, R21, and R25, a finding which suggests that the individuals who most felt as though their organization was good at detecting emerging technologies, also considered global activity as an important part of the decision process, thought that it was important to know which conferences covered their technical area, and believed it is important to know the position of their technology in its life cycle.

CHAPTER 5

DATA PREPARATION

5.1 The Concept-Clumping Algorithm

5.1.1 Preparation

The Concept-Clumping Algorithm is intended to identify and combine terms that are conceptually the same, improving the conceptual accuracy of analyses using Abstract Phrases. Implementing the Algorithm proved to be challenging. First, the algorithm is written as a script in VantagePoint. Therefore, it required a programmer knowledgeable of scripting in VantagePoint and able to program in Visual Basic. Four individuals, all with other commitments, fit that profile. Additionally, the programming is more complicated then initially expected, further limiting the programmers capable of finishing the task. Finally, the second programmer found a memory leak in VantagePoint that was preventing the algorithm from running properly. The Algorithm ran on the Cleaned Abstract Phrases from samples of the five record sets from the selected topic areas (Table 4.1). Each sample consists of between 176-263 records taken from one year out of the entire record set. The manual requirement to evaluate the precision of the algorithm prohibits evaluating an entire record set at this time.

One major adjustment was made to the algorithm. In some cases, because the algorithm forces the term to choose between groupings starting at the level of the greatest number of shared words, the multiword search terms create some inaccurate groupings, if that term appears in numerous separate concepts. The reason is that the different variations in spelling of the search term would be considered at the same time as different

categories of the search term. “Carbonate fuel cell systems” has as many shared words with “solid oxide fuel cell” as it does with “carbonate fuel cells.” The algorithm ran at sufficient accuracy for the “geographic information system” and the “pollution monitoring” record sets. However, the problem became evident after running the algorithm on the “remote sensing” and “fuel cell” record sets. At the two shared-words iteration, “carbonate fuel cell system” would have to choose between “solid oxide fuel cell” and “carbonate fuel cell”. Since the terms cell(s) very rarely appear without fuel, ignoring “cell(s)” improves the accuracy of the algorithm. “Carbonate fuel cell system” would not have to consider “solid oxide fuel cell” as a partner. In the remaining record sets, the noun part of the search term which may appear in a variety of forms was ignored, meaning “sensing,” “sensor,” “cell,” and “cells,” by the algorithm. Ignoring the search term word that rarely appears without the other is a way of forcing additional strength between concepts that contain the search term. It requires an additional shared word, allowing different categories of the search term to be considered before variations in spelling of the search term itself.

The algorithm macro now gives the user the option of ignoring a string or set of strings from consideration. In the future, something like “sub” might be ignored. “Sub” is used in abstracts to indicate a subscript. So, in scientific abstracts “O_2” would be written as O(sub)2. Further research is required to determine what terms should be added to a list of terms to ignore. If there are terms that should be ignored across all record sets, the algorithm should be programmed to read these words from a stopwords list. The goal is to create such as list that is not domain specific.

5.1.2. Calculating Precision

The output file produced is a set of VantagePoint thesaurus files, which combined together provide the entire clumped group and the term that is ultimately chosen as the representative term for the group of terms deemed similar. . For example, the output file contained the following segment:

**hard disk drives
100 1 ^hard disk drives$
100 1 ^double prime hard disk drives$
100 1 ^hard drives$

The "**" indicates the name that the terms in the lines below it will be given.

Each term was evaluated to determine if the representative term provides an accurate portrayal of the term under consideration. The file was opened as an Excel Spreadsheet and each term in the group was evaluated to determine if "hard disk drives" is a conceptually accurate representation of the term. For this segment, all of the terms are "Good Matches." Therefore, the spreadsheet was marked as in Table 5.1.

Table 5.1: Hard Disk Drive Matches

Bad Matches	Good Matches	**hard disk drives
	1	100 1 ^hard disk drives$
	1	100 1 ^double prime hard disk drives$
	1	100 1 ^hard drives$

The column totals were tabulated in order to determine the precision of the algorithm in that record set. Due to conflict resolutions, some of the output groups contained only one term. These are not calculated into the precision, because clearly a

group of one term named for that term would be a correct representation. Only output combining terms are considered. So, consider the following output in Table 5.2.

Table 5.2: High Density Recording Matches

B	G	**high density recording
	1	100 1 ^high density recording$
	1	100 1 ^high density magnetic recording$
	1	100 1 ^high density magnetic recording applications$
	1	100 1 ^high density magnetic recording materials$
	1	100 1 ^excellent high density recording capability$
	1	100 1 ^good high density recording performance$
	1	100 1 ^high density magneto optical recording$
	1	100 1 ^high density overwrite recording$
	1	100 1 ^high density recording disks$
	1	100 1 ^high density recording media$
	1	100 1 ^high frequency high density tape recording$
	1	100 1 ^high linear density recording$
	1	100 1 ^recording density$
	1	100 1 ^predicted recording density$
	1	100 1 ^recording areal density$
	1	100 1 ^recording density curves$
	1	100 1 ^recording density D sub sub$
1		100 1 ^recording linear density$
		**high density television
1		100 1 ^high density$
1		100 1 ^high bit density$
1		100 1 ^high density partial response channels$
	1	100 1 ^high density television$
1		100 1 ^high superficial density$
		**magnetic property
		100 1 ^magnetic property$
		**thin film head elements
	1	100 1 ^thin film$
	1	100 1 ^polished thin film disk$
	1	100 1 ^thin film head on disk wear tests$
	1	100 1 ^thin film rigid disk$
	1	100 1 ^thin film disks$
	1	100 1 ^isotropic longitudinal CoCrTa Cr thin film head$
	1	100 1 ^thin film head elements$

Table 5.2 (cont.)

	1	100 1 ^Co Pt thin film patterns$
	1	100 1 ^conventional thin film head sliders$
	1	100 1 ^thin film corrosion$
	1	100 1 ^thin film corrosion model$
	1	100 1 ^thin film discs$
	1	100 1 ^thin film magnetism$
	1	100 1 ^thin film optics$
	1	100 1 ^thin film type recording head$
		**magnetic heads
	1	100 1 ^magnetic heads$
	1	100 1 ^small magnetic heads$
		**thin films heads
	1	100 1 ^thin film inductive heads$
	1	100 1 ^conventional thin film inductive heads$
	1	100 1 ^inductive thin film magnetic recording heads$
	1	100 1 ^thin film inductive recording heads$
	1	100 1 ^thin film magnetic recording heads$
	1	100 1 ^thin film recording heads$
	1	100 1 ^CoTaZr amorphous thin film disk heads$
	1	100 1 ^thin film inductive disk drive heads$
	1	100 1 ^thin film magnetic heads$
	1	100 1 ^thin film read write magnetic heads$
	1	100 1 ^conventional thin film heads$
	1	100 1 ^modified thin film heads$
	1	100 1 ^similar thin film heads$
	1	100 1 ^thin film heads TFHs$
	1	100 1 ^thin films heads$
		**amorphous magnetic film
		100 1 ^amorphous magnetic film$

The “B” column is a marker for “Bad Matches” and the “G” column is a marker for “Good Matches”. Notice that the group member, “amorphous magnetic film” does not have a “1” in either column. This term is the only term in its group and therefore, was not included in the calculation. There are 50 terms that are considered Good Matches and

5 that are considered "Bad Matches." In some cases, judgments were made by reviewing individual abstracts to determine the context of the term in the record set.

Where precision = (Good Matches)/ (Good Matches + Bad Matches), the above sample had a precision of 50/55 or 90.9%

The precision of the algorithm was above 89% for all five record sets (Table 5.3).

Table 5.3 Technology Cases: Clumping Algorithm Precision Calculations

File	# Recs	Precision
Fuel Cells (1995)	197	91.1%
Remote Sensing (2002)	263	89.7%
Magnetic Storage (1992)	220	92.2%
GIS (1992)	176	90.7%
Pollution Monitoring (2003)	181	91.4%

5.1.3 The Effect of the Algorithm

The effect of the algorithm is apparent in the "Top 20" term list for each of the record sets. The Clumped Abstract Phrases list is shown alongside the Cleaned Abstract Phrases list, where cleaning refers to the stemming algorithm already in VantagePoint. Cleaning is maintained in the clumping algorithm as a preparation step. Individual points are discussed below each Top 20 list (Tables 5.4 – 5.8).

Consider the lists in Table 5.4. The cleaned abstract phrases list only contains two multiword phrases containing "fuel cells" (the search term itself) and "solid oxide fuel cells." However, clumping allows for many of the multiword concepts to increase in prominence on the list. Three additional terms containing the phrase "fuel cells" are now on the list and the concept "solid oxide fuel cells" increases from 11 records to 30

records. The combined “solid oxide fuel cells” entry consists of the following original terms:

solid oxide fuel cells
solid oxide fuel cells SOFCs
reduced temperature solid oxide fuel cells SOFCs
novel solid oxide fuel cell SOFC system
SOFC Solid Oxide Fuel Cells interconnector material
solid oxide fuel cell SOFC cells
solid oxide fuel cell SOFC performance
chemical cogenerative solid oxide fuel cell
solid oxide fuel cell electrolytes
solid oxide fuel cell systems

Table 5.4: Fuel Cell Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Clumped
1	50	Fuels cells	50	fuels cells
2	33	C	33	C
3	24	developments	31	deg
4	24	results	30	solid oxide fuel cells SOFCs
5	20	effects	24	developments
6	14	study	15	direct methanol polymer electrolyte membrane fuel cells
7	14	temperatures	15	molten carbonate fuel cells
8	14	uses	14	temperatures
9	13	operator	12	current density
10	12	cells	12	electrodes
11	12	electrodes	12	electrolytic
12	12	electrolytic	12	hydrogenation
13	12	hydrogenation	12	increasing
14	12	increasing	12	oxygen
15	12	oxygen	12	yttria stabilized zirconia YSZ
16	12	systems	11	applications
17	11	applications	10	high efficiency
18	11	solid-oxide fuel cells	9	cathodically
19	10	activity	9	phosphoric acid fuel cells
20	10	catalysts	9	proton exchange membrane fuel cells

The simple ability to combine “solid oxide fuel cells” and “solid oxide fuel cells SOFCs” would increase the representation of the this type of fuel cell from 11 records to 18 records.

Some other important terms not on the list originally were: direct methanol polymer electrolyte membrane fuel cells, molten carbonate fuel cells, phosphoric acid fuel cells, yttria stabilized zirconia YSZ, and proton exchange membrane fuel cells.

Using the concept-clumping algorithm, “yttria stabilized zirconia YSZ” is counted in 12 records. Without the algorithm, the most frequent variation of this term only appears in 2 records. Therefore, without the algorithm it would not be used in the mapping function at all. Phosphoric acid fuel cells is another term that makes the Top 20 list only after clumping. It consists of the following terms.

four phosphoric acid fuel cell monocells
kilowatt phosphoric acid fuel cell
phosphoric acid fuel cell cathodes
phosphoric acid fuel cell technology
phosphoric acid fuel cells
pressurized phosphoric acid fuel cell
phosphoric acid electrolyte
platinum bearing phosphoric acid
pyro phosphoric acid

Two phosphoric acid fuel cell terms that are not included in this grouping are “phosphoric acid fuel cell power plants” and “PAFC power plants”, which the algorithm determined were more similar to a fuel cell power plants grouping.

After numerical and punctuation characters are removed from the list, common words with up to ten letters are removed. Notice the impact that this has on the Abstract Phrase list for Remote Sensing (Table 5.5.) The five most frequent terms (results, data, study, methods, used) are removed from the list. Terms are removed that would be

included in a wide array of records but do not uniquely distinguish the scientific concepts in the record.

Table 5.5: Remote Sensing Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Clumped
1	72	results	79	remote sensing
2	40	data	26	applications
3	35	study	24	estimators
4	34	methods	22	development
5	32	used	19	approaches
6	26	applications	15	Synthetic Aperture Radar SAR images
7	26	presented	14	experimental results
8	25	remote sensing	14	techniques
9	24	effects	12	Atmosphere
10	24	estimators	12	information
11	22	accuracy	12	potentiality
12	22	analysis	11	land cover classification
13	22	development	11	ms
14	21	surfacing	11	relationships
15	21	systems	10	classifications
16	20	measures	10	combinations
17	19	approaches	10	km
18	18	problems	10	vegetation
19	17	images	9	conditions
20	16	regions	9	Gaussian maximum likelihood GML classification

Notice the Magnetic Storage Cleaned Abstract Phrases contain a number of generic single terms (Table 5.5). In the Clumped Abstract Phrases list, there are a few "thin film" entries, such as "thin film heads," that were not in the Top 20 in the Cleaned Abstract Phrases list. The output file looks as follows:

**thin films heads
100 1 ^thin film inductive heads$

100 1 ^conventional thin film inductive heads$
100 1 ^inductive thin film magnetic recording heads$
100 1 ^thin film inductive recording heads$
100 1 ^thin film magnetic recording heads$
100 1 ^thin film recording heads$
100 1 ^CoTaZr amorphous thin film disk heads$
100 1 ^thin film inductive disk drive heads$
100 1 ^thin film magnetic heads$
100 1 ^thin film read write magnetic heads$
100 1 ^conventional thin film heads$
100 1 ^modified thin film heads$
100 1 ^similar thin film heads$
100 1 ^thin film heads TFHs$
100 1 ^thin films heads$

Table 5.6: Magnetic Storage Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Clumped
1	34	Results	32	Mu
2	29	Heads	20	High density recording
3	28	Uses	20	Ms
4	27	Effects	20	Thin film recording media
5	21	Presents	17	Thin film heads
6	20	Ms	16	Developments
7	19	Disks	16	Thin film magnetic recording disks
8	18	Measures	15	Techniques
9	17	Methods	15	Thin film head elements
10	16	Described	14	Magnetic property
11	16	Developments	12	Deg
12	15	Techniques	11	Applications
13	14	Magnetic property	11	Experimental results
14	13	Functions	11	MIG heads
15	13	Systems	11	Recording heads
16	12	Magnets	10	Finite element method FIM
17	12	Taping	10	Intermittent head disk contacts
18	11	Applications	9	Air bearing surfaces
19	11	C	9	Directions
20	11	problems	9	Disk drives

Table 5.7: GIS Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Clumped
1	54	GIS-Geographic Information System	83	GIS Geographic Information System
2	43	GIS	63	geographical information systems
3	36	data	43	GIS
4	32	geographical information systems	24	applications
5	32	results	24	developments
6	31	systems	21	spatial data
7	30	uses	13	U S
8	24	applications	12	multiple remote sensing images
9	24	developments	12	researches
10	20	analysis	11	land use category
11	20	informing	11	relationships
12	20	study	10	ground water
13	18	maps	10	processing
14	16	timing	10	remotely sensed
15	15	spatial data	9	data sets
16	14	areas	9	land uses
17	14	numbers	8	United States
18	14	plans	8	water resources
19	14	tools	7	approaches
19	14	users	7	Extensive water quality data

The GIS list reveals the limitation of the clumping algorithm. The first three terms on the list are “GIS Geographic Information System” “Geographical Information Systems” and “GIS.” These terms are clearly the same concept, but share at most only one word in common. The algorithm only reviews terms that share at least two words in common, because while this case is clear, imagine the number of terms that may include the word “information” in a record set. Reapplying the concepts of ignoring common words, stemming, and similarity could result in a more powerful algorithm that could address these issues.

Table 5.8: Pollution Monitoring Top 20 Abstract Phrases

	# Recs	Abstract Phrases Cleaned	# Recs	Abstract Phrases Clumped
1	61	results	42	concentrations
2	51	study	21	Zn
3	42	concentrations	19	contamination
4	36	data	19	Pb
5	29	sites	17	Cu
6	21	Zn	16	Cd
7	20	effects	13	heavy metals
8	19	contamination	13	pollutants
9	19	Pb	12	air pollution
10	18	soils	12	air quality
11	18	used	12	Co
12	17	Cu	11	contributions
13	16	Cd	11	distributions
14	15	impacts	11	environmental heavy metal ions
15	15	low	11	Ni
16	15	sampling	10	PM sub
17	14	analysis	10	study area
18	14	area	9	high concentrations
19	14	increases	9	indicators
20	14	sediments	9	polycyclic aromatic hydrocarbons

In the case of Pollution Monitoring, some terms rose in prominence on the list, while terms such as “heavy metals,” “ environmental heavy metal ions,” and “polycyclic aromatic hydrocarbons” were included on the list. The group for “polycyclic aromatic hydrocarbons” consist of the following terms:

**polycyclic aromatic hydrocarbons
1 100 1 ^polycyclic aromatic hydrocarbons PAHs$
1 100 1 ^polycyclic aromatic hydrocarbons$
1 100 1 ^low molecular weight polycyclic aromatic hydrocarbons PAH$
1 100 1 ^particle bound polycyclic aromatic hydrocarbons$
1 100 1 ^polycyclic aromatic hydrocarbon PAH exposure$

The most frequent occurrence of any one of these terms is the title term, which appears in two records. A term that clearly conceptually belongs with this group is "PAHs", which occurs in 7 records. An improvement in the algorithm should attempt to match such a term with like concepts.

The algorithm has a high level of precision. However, clearly from the Top 20 lists we also see terms that have the same meaning that are still not identified as being conceptually the same. Therefore, additional work should be done to improve the recall of the algorithm without reducing the precision. The lists also reveal additional opportunities for improvement. If VantagePoint is to be used on files with chemical elements discussed, a thesaurus for the elements in the periodic table may be useful.

5.2. Keywords and Abstract Phrases Cluster Comparison

The impact of the clumping algorithm is evident in the Step Three results. However, can clumping create more accurate clusters? In this step, the five datasets from the search terms in Step One were randomly sampled to create workable size datasets Table 5.9.

Table 5.9: Dataset Sample Sizes

Fuel Cells	880
GIS	520
Magnetic Storage	693
Pollution Monitoring	434
Remote Sensing	445

The first step in creating a cluster map based on PCA is determining which terms will be included in the clustering. There are a number of ways to make this determination; however, in any case, the term must occur in at least two documents in order for any co- occurrence based method to work. Using all terms with at least two occurrences, is one method and another is to take a percentage of the terms. This research

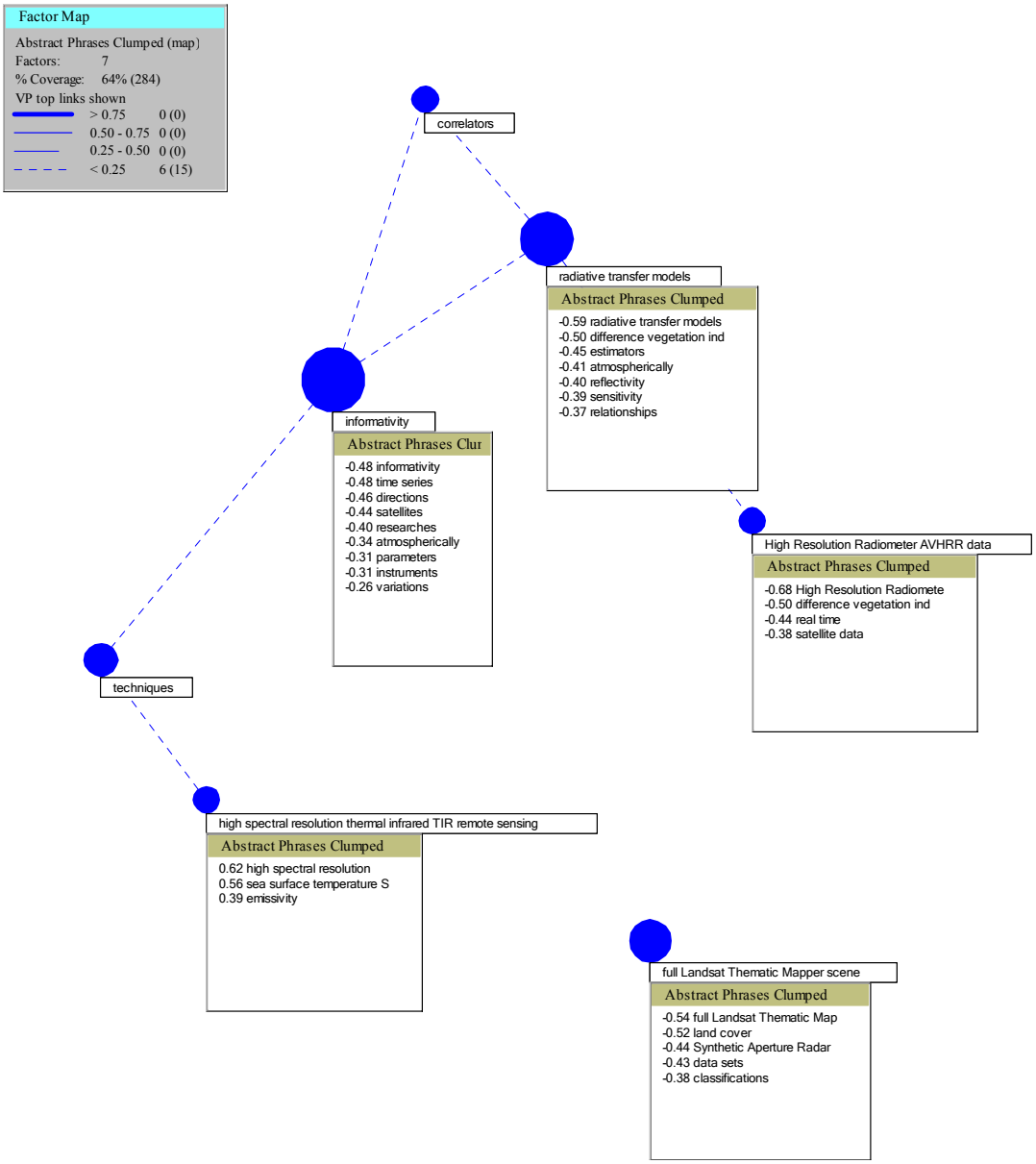


Figure 5.1 Remote Sensing Clumped Abstract Phrases Map

uses a method based on the Zipf distribution. The Zipf distribution asserts the log of the rank vs. the log of the frequency of a term is linear. The method used finds that line and the terms with the highest and lowest rank that fall below the line are eliminated. This method was chosen because of the variation between the number of terms in keywords versus abstract phrases. Appendix H contains the graphs of the Zipf distributions for each of the three lists in each of the five datasets. After the terms for the cluster map were determined, three maps were created for each of the five datasets, a Keyword map, a Cleaned abstract phrases map, and a Clumped abstract phrases map. The Remote Sensing Clumped Abstract Phrases Map shown in Figure 5.1 is an example of one of the maps. Appendix I contains the entire set of maps.

Table 5.10 captures the numerical data regarding the links, clusters, and terms for each cluster. Consider the data on the Remote Sensing Clumped Abstract Phrases Map (Figure 5.1), The Remote Sensing clusters were created starting with the original list of 8004 terms. The Zipf Distribution graph kept only 46 terms, or 0.6% of the original terms. Then the clustering algorithm created seven clusters. There are seven blue circles on the map, corresponding to the seven clusters. There are 6 links between the clusters all with a link strength of less than 0.25. Therefore, there are only weak relationships between the clusters. The clustering algorithm used 33 of the 46 terms, or 71.70% of the terms. There were an average of 5.14 terms included in each cluster and least one term that is included in a cluster occurs in 64% of the documents. The last two columns in Table 5.10 represent the tightness and separateness of the clusters. The Remote Sensing clusters have an Entropy of .74, and a Cohesion of .45.

Using SPSS, a correlation matrix was developed in order to determine if

Table 5.10 Quantitative Cluster Measures

					Strength of Links										
	Total # Terms	# Terms used	% term	# of links	<.25	.25-.5	.5-.75	>.75	# of Clstrs	Avg # terms per Clstr	# terms in Clstrs	% terms	% docs Used	Entropy	Cohes.
Fuel Cells															
Key	1210	81	6.70%	8	8	0	0	0	9	5.56	48	59.30%	62	1.62	0.75
Clean	14564	298	2.00%	9	9	0	0	0	11	4.64	47	15.80%	45	9284.8	0.49
Clum	9799	184	1.90%	13	13	0	0	0	15	5	74	40.20%	74	11536.2	0.48
GIS															
Key	773	146	18.9%	10	9	1	0	0	10	6.6	66	45.20%	71	7732.5	0.69
Clean	10668	31	0.30%	8	8	0	0	0	9	3.44	29	93.50%	69	1828.2	0.42
Clum	7614	224	2.90%	10	10	0	0	0	12	5.08	59	26.30%	49	2789.2	0.42
Mag Stor															
Key	1120	230	20.5%	13	13	0	0	0	15	4.93	74	32.20%	69	40794.5	0.69
Clean	10832	139	1.30%	9	9	0	0	0	10	5.2	52	37.40%	62	9284.8	0.44
Clum	7207	149	2.10%	9	9	0	0	0	10	5.4	64	43.00%	52	3810.1	0.41
Poll Mon															
Key	2072	272	13.1%	14	13	1	0	0	16	6.88	84	30.90%	87	3091.2	0.73
Clean	12330	167	1.40%	11	10	0	1	0	13	4.62	54	32.30%	70	4149.2	0.56
Clum	8702	67	0.80%	8	8	0	0	0	9	4.89	36	53.70%	54	0.84	0.48
Rem Sens															
Key	2175	215	9.90%	13	13	0	0	0	15	5.27	79	36.70%	68	23214.6	0.62
Clean	11153	90	0.80%	7	6	1	0	0	8	5.75	43	47.80%	71	1.28	0.44
Clum	8004	46	0.60%	6	6	0	0	0	7	5.14	33	71.70%	64	0.74	0.45

Note: Entropy requires the log(Pij) However, if there is no overlap between clusters Pij=0 and the log0 is negative infinity. In such a case if there was one set of clusters with no overlap the Entropy of the entire map would be negative infinity. In order to create some basis for comparison the log0 was calculated as –10,000. A sufficiently large number.

correlations in the variables existed regardless of the method used to construct the cluster maps (Appendix J). Some notable correlations that existed at a .01 significance level include a positive correlation between the number of terms actually included in the clusters, the number of links, and the entropy. These three variables have a negative correlation with the percentage of terms actually used from the initial terms used to determine the clusters. Entropy is also positively correlated with the number of clusters. Therefore, when there are more clusters on the map, the clusters are more distinct.

Table 5.11 Cluster Quantitative Measure Comparison of Means

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)
Kywd Mean	3	594	1470	189	13.8	12	13	5.85	70	41	71	14967	0.7
Std. Dev.	2	190	620	75	5.9	3	3	0.85	14	12	9	16974	0.1
Clnd Mean	21	594	11909	145	1.2	9	10	4.73	45	45	63	4910	0.5
Std. Dev.	5	190	1620	100	0.7	2	2	0.86	10	29	11	4256	0.1
Clmp Mean	15	594	8265	134	1.6	9	11	5.10	53	47	59	3627	0.5
Std. Dev.	4	190	1019	76	1.0	3	3	0.19	18	17	10	4733	0.1

A) terms per document
B) Number of documents
C) Total number of terms
D) Number of terms used in clustering
E) Percentage of terms considered for clustering
F) Number of links on cluster map
G) Number of clusters on cluster map
H) Average Number of terms per cluster
I) Number of terms assigned to a cluster
J) Percentage of terms assigned to a cluster
K) Percentage of documents covered by the clusters
L) Entropy
M) Cohesion

Cohesion does not have a strong correlation with any other variable at the .01 level. However, at the .05 level, cohesion is positively correlated with the percent of all the term that are selected to use in clustering. The higher the percentage of terms used in the clustering results in "tighter" clusters. A comparison of means was calculated using keywords, cleaned phrases, and clumped phrases as the grouping variables. Table 5.11 contains the results.

This paper first compares Clumped Abstract Phrases to Cleaned Abstract Phrases and then compares the Clumped Abstract Phrases to Keywords. In comparing Cleaned Abstract Phrases to Clumped Abstract Phrases, this table reveals that clumping reduced the Abstract Phrase terms by 30% and reduced the number of terms included in the clustering. With those reductions came an increase in the number of clusters and the average number of terms per cluster. So what was the impact on the factors that are used to compare the quality of clusters? The percentage of documents covered was decreased, as was the cohesion. Are these numbers a surprise? Somewhat. While the practice of combining terms means that the same terms represent more documents and therefore a greater spread of terminology, it is a surprise that the elimination of the common words did not result in greater cohesion. Does this mean that clumping phrases is bad? Certainly not! As stated earlier, these methods do not necessarily reflect good or better clusters. The purpose of the clusters is to show the relationships that exist in the data. The precision and impact of the clumping algorithm reveal that clumping conceptually represents the dataset well. The more important evaluation of the value of clumping in clustering is revealed in the actual clusters themselves. The biggest difference between

the two types of Abstract Phrase maps is the technical specificity of the terms included. Cleaned Abstract Phrases are dominated by the common generic terms. This circumstance exists for two reasons: the most common words are not removed and secondly because the more technical terms are included in phrases that are not gathered together as in the Clumped Phrases. For example, the "friction" cluster in Cleaned Abstract Phrases includes the terms: " friction", "surfaces", "lubrication", "coefficients", "wearing", and "tribology". A similar cluster in the Clumped Phrase map contains phrases like " head disk interface" "surface roughness" "slider disk spacing" "Contact Start Stop durability" and "stiction". Cleaned Abstract Phrases contains more clusters that have little meaning because of the broad terminology included. Clusters such as these in the Remote Sensing dataset:

Accounts: used, limits, accounts, interpreting, selection, important

Presents: presents, ones, techniques, atmospherically, described, viewing, experimental results, improvements

In contrast, some of the Clumped Abstract Phrases clusters are:

AVHRR data: difference vegetation index NDVI, real time, satellite data, High Resolution Radiometer AVHRR data

TIR remote sensing: high spectral resolution thermal infrared TIR remote sensing, sea surface temperature SST, emissivity

Clearly, clumping provides richer details in the clusters.

In reviewing the clusters, Clumped Abstract Phrases sits in-between Cleaned Phrases and Keywords, having some clusters that match each side and not the other. The conceptually equivalent terms that are included in a keyword cluster, but are missing

from the Clumped Abstract Phrases clusters, fall into two categories: either the concept is defined by a single word, such as “pervoskite” in the fuel cell case, that occurs frequently with different words, or one of the words has different variations such as “fault tolerant” and “fault tolerance.” In order to capture those concepts found in the keyword clusters in the clumped abstract phrase clusters, and adjustment has to be made in VantagePoint. Currently, the number of clusters is determined by an algorithm based on the number of terms tagged for clustering. If the default number of factors is increased, then more of the keyword cluster concepts are captured in the clumping clusters. For example, in the fuel cell case, X-ray diffraction analysis is contained in the keyword clusters obtained using the VantagePoint default number of factors, but not in the clumping clusters. If the number of factors is increased, the term appears in the clumping clusters, as are some other interesting clusters.

Keyword clustering has been the method used in conducting conceptual clustering for a number of reasons. This research initially set out to determine if abstract phrases could be used in place of keywords for science and technology mapping. In comparison to abstract phrases, keywords used a higher percentage of the total terms which resulted in a larger number of clusters and links. The average number of terms per cluster was higher, as was the percentage of documents covered; the entropy and cohesion were both higher. Keywords do an excellent job of covering the dataset. However, it has been determined that the two complement each other and may be used for different purposes. Keywords use more general terms and therefore capture relationships with broader fields outside the specifics of the topic area. For example, magnetic storage has a cluster with the terms computer system recovery, security of data, computer software, computer

operating system, computer networks, fault tolerant computer systems. redundancy, computer hardware. However, clumped abstract phrases provide the specific details of the technology areas. The word "computer" only appears in the abstracts a few times in different terms. Abstract Phrases capture more specific concepts such as perpendicular recording, single pole type head, head medium spacing fluctuation, medium noise, transition shift distortion, and writing head field.

In Science and Technology Analysis, clustering is used to “discover new concepts or new relationships from literature to identify promising research or technological opportunities, to identify themes and sub-themes in a large body of technical literature, allows technical taxonomies to be generated, and to link major themes” (Kostoff, 2001). Keywords can identify where the technology area sits among broad categories and can identify promising research at the crossroads of these broad areas. Larger themes can be identified using Keywords. Clumped Abstract Phrases on the other hand can be used to identify sub-themes within those broad areas. Since keywords come from a taxonomy that already exists, Clumped Abstract Phrases is better suited to identify new themes or concepts and developing new taxonomies. More specific identification can also take place using abstract clumping in clustering.

CHAPTER 6

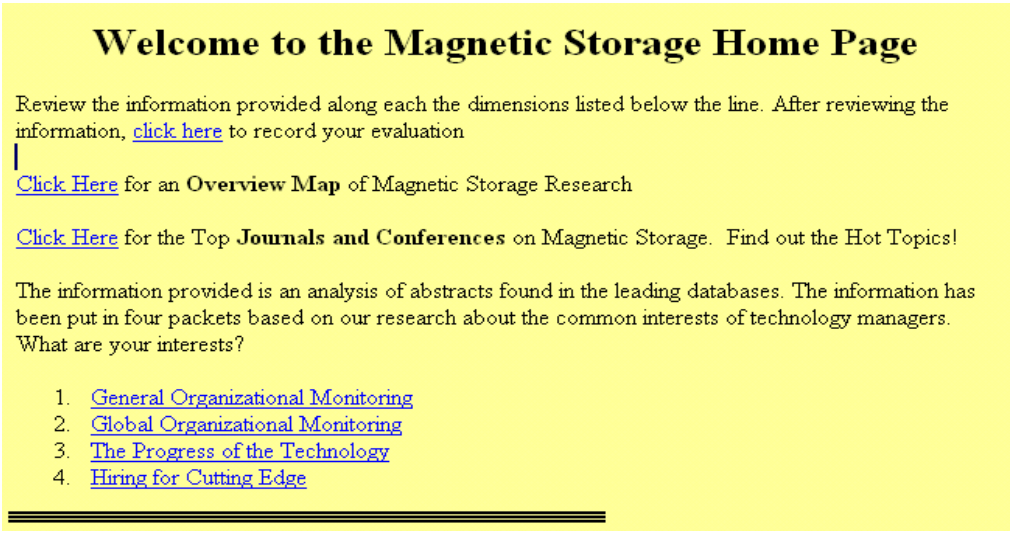
METRIC FINDINGS AND EVALUATION

6.1 The Metrics for an Example Technology

The analysis of the Questionnaire results reveals that participants strongly agreed with the importance of three of the information statements (Table 4.4) and indicates that there are four clusters of information interests. From these results, information packets on Magnetic Storage packaged on a website. The website opened to a Welcome Screen which provided background information. Three situations provided the evaluators with the perspective of a technology decision-maker who would use the information.

> **"Situation 1)** You manage research in a magnetic storage technology. You are trying to determine if you should make any changes in your research strategy.
> **Situation 2)** You are a developing a technology that relies on storage technology as a component of your design. You are trying to determine if you should invest research dollars investigating magnetic storage or look into some other option. If you decide to go with a magnetic storage technology, you must decide if you should do the work in-house or partner with another organization.
> **Situation 3)** You are a company that must maintain millions of customer records per year. Should you use a magnetic storage technology or investigate another route?"

The Welcome Page is linked to the "Magnetic Storage Home Page" (Figure 6.1). From the Home Page, the reviewer could link to both a Keyword and Abstract Phrases overview map. There is also a link to a list of conferences and journals with the topics covered (Figure 6.2). In addition to these links, which provide a general overview of the information contained in the dataset, the Home Page contains links to each of the clusters of interest.



Welcome to the Magnetic Storage Home Page

Review the information provided along each the dimensions listed below the line. After reviewing the information, click here to record your evaluation

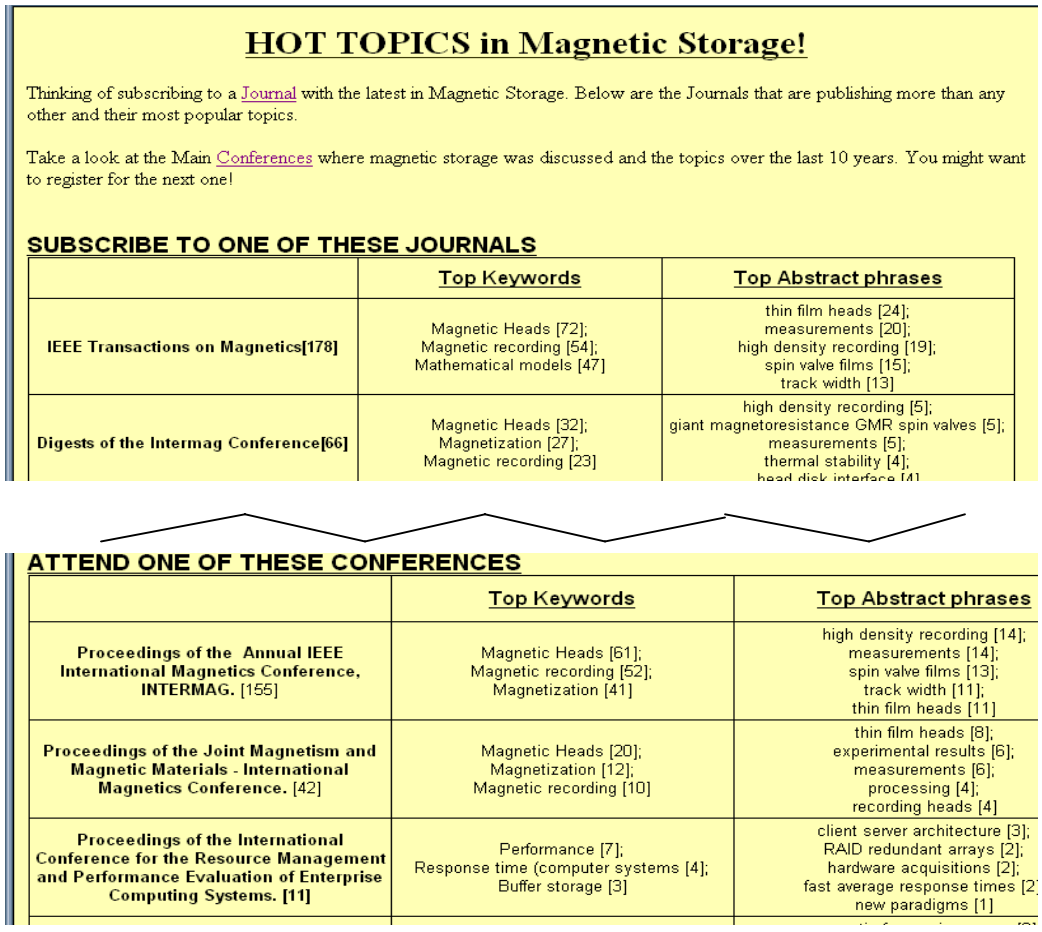
Click Here for an **Overview Map** of Magnetic Storage Research

Click Here for the Top **Journals and Conferences** on Magnetic Storage. Find out the Hot Topics!

The information provided is an analysis of abstracts found in the leading databases. The information has been put in four packets based on our research about the common interests of technology managers. What are your interests?

1. General Organizational Monitoring
2. Global Organizational Monitoring
3. The Progress of the Technology
4. Hiring for Cutting Edge

Figure 6.1 The Magnetic Storage Home Page



HOT TOPICS in Magnetic Storage!

Thinking of subscribing to a Journal with the latest in Magnetic Storage. Below are the Journals that are publishing more than any other and their most popular topics.

Take a look at the Main Conferences where magnetic storage was discussed and the topics over the last 10 years. You might want to register for the next one!

SUBSCRIBE TO ONE OF THESE JOURNALS

	Top Keywords	Top Abstract phrases
IEEE Transactions on Magnetics[178]	Magnetic Heads [72]; Magnetic recording [54]; Mathematical models [47]	thin film heads [24]; measurements [20]; high density recording [19]; spin valve films [15]; track width [13]
Digests of the Intermag Conference[66]	Magnetic Heads [32]; Magnetization [27]; Magnetic recording [23]	high density recording [5]; giant magnetoresistance GMR spin valves [5]; measurements [5]; thermal stability [4]; head disk interface [4]

ATTEND ONE OF THESE CONFERENCES

	Top Keywords	Top Abstract phrases
Proceedings of the Annual IEEE International Magnetics Conference, INTERMAG. [155]	Magnetic Heads [61]; Magnetic recording [52]; Magnetization [41]	high density recording [14]; measurements [14]; spin valve films [13]; track width [11]; thin film heads [11]
Proceedings of the Joint Magnetism and Magnetic Materials - International Magnetics Conference. [42]	Magnetic Heads [20]; Magnetization [12]; Magnetic recording [10]	thin film heads [8]; experimental results [6]; measurements [6]; processing [4]; recording heads [4]
Proceedings of the International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems. [11]	Performance [7]; Response time (computer systems [4]; Buffer storage [3]	client server architecture [3]; RAID redundant arrays [2]; hardware acquisitions [2]; fast average response times [2]; new paradigms [1]

Figure 6.2 Magnetic Storage Conferences and Journals Web Page

6.1.1 General Organizational Monitoring

Figure 6.3 is an example of an information page based on a cluster from the questionnaire results. From this page, a decision-maker can discover the activity of leading organizations and identify the leading individual publishers.

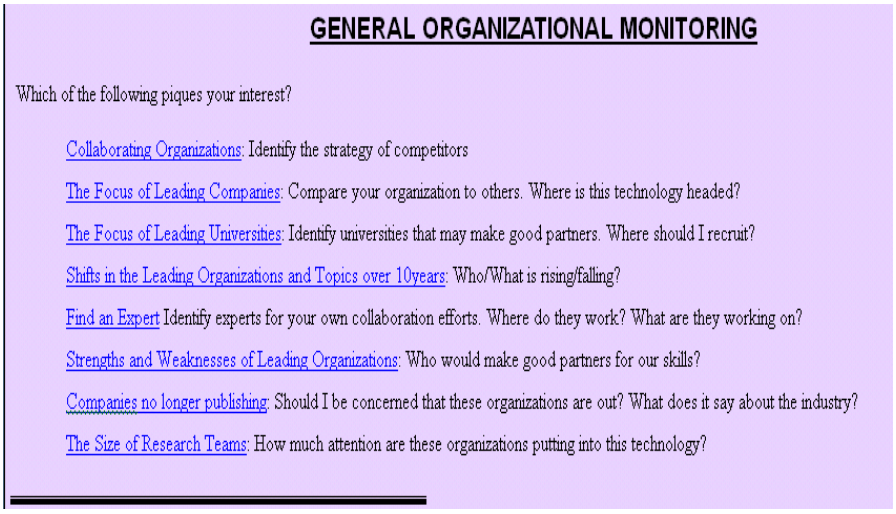


Figure 6.3

A user can discover which organizations have researchers who are publishing together using the Collaboration Map. This map shows the links between the top 27 organizations based on the authors who published together (Figure 6.4).

Figure 6.4 shows which organizations have authors who are collaborating in Magnetic Storage. Note, for example, that Seagate Technology authors have collaborated with a number of organizations. They have collaborated, not only with universities such as UC San Diego, Carnegie Mellon, and the University of Minnesota, but also with other corporations such as Read-Rite Corporation and Western Digital Corporation.

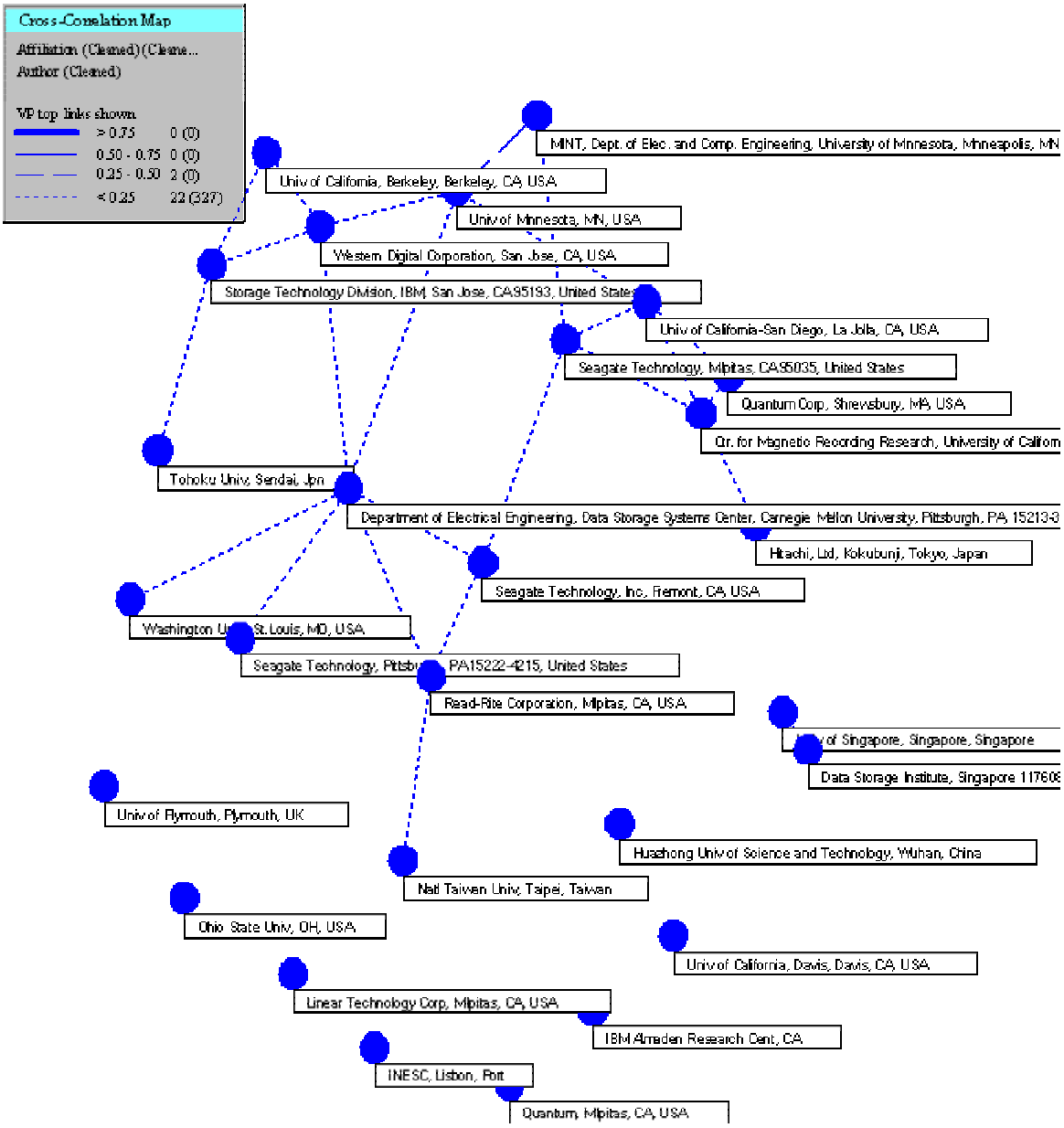


Figure 6.4 Magnetic Storage Collaboration Map

One interesting finding is the number of organizations that are no longer publishing. Four companies (Western Digital Corporation, Fujitsu, Quantum Corp, and Phillips Research) had published at least nine articles before 2001 and no articles

afterwards. This fact is an indication that the technology may be more mature and the research discussion is changing. The information in the "Technology Progress" area provides more information on the maturity of the technology.

6.1.2 Global Organizational Monitoring

On the Global Monitoring page, the user views a graph of the total publishing activity in the United States compared to the rest of the world. Users can also see the

Figure 6.5 Global Monitoring Web Page

main research topic areas of different countries and foreign organizations, and compare the size of foreign research teams to those in the United States (Figure 6.5). The graph of Global Activity (Figure 6.6) shows very similar year-to-year behavior between the United States and other Countries. It also shows a significant drop in the number of articles published in recent years; yet another sign of the maturity of this technology. Another interesting discovery about Global Activity is the difference between the size of research teams in the US compared to foreign organizations (Table 6.1). There are fewer

publications for the size of the research teams in the foreign companies than in the US companies, indicating greater collaboration.

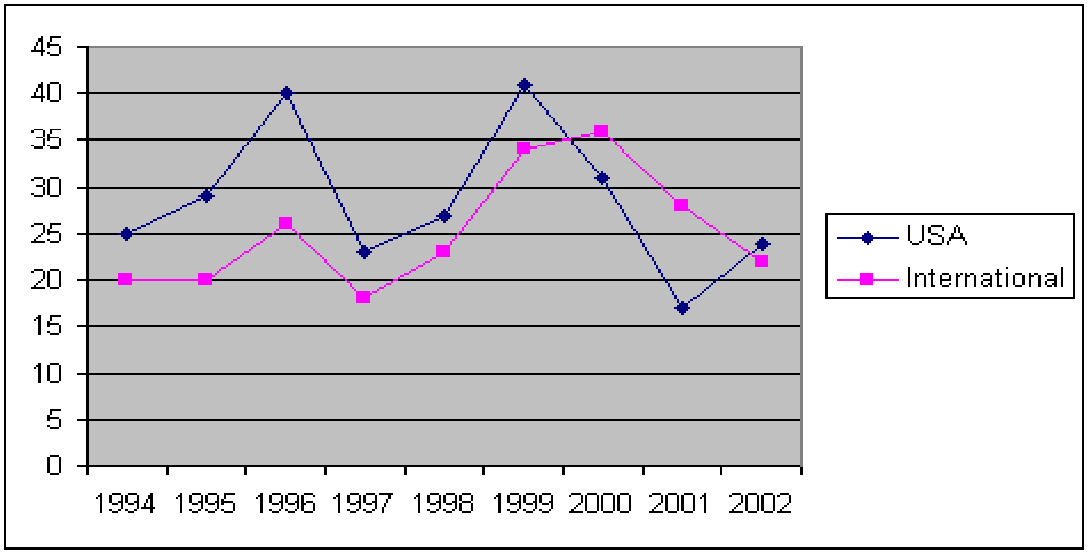


Figure 6.6 Global Activity Over The Years

Table 6.1 Size of Research Teams: US vs. Foreign

Foreign	Recs	Size	US	Recs	Size
Hitachi, Ltd	23	66	Seagate Technology	36	100
Fujitsu	7	35	Western Digital Corp	20	68
Sony	10	34	Storage Tech Div., IBM	18	65
Data Storage Inst., Singapore	12	31	Read-Rite Corp.	13	38
Toshiba, Japan	7	21	Quantum	14	25
Philips	4	18	Linear Technology	3	11
Samsung	3	9	Hewlett Packard	4	6

6.1.3 Hiring for Cutting Edge

TOP UNIVERSITIES

RECRUIT AT ONE OF THESE SCHOOLS!

These are the Top Universities Publishing in Magnetic Storage: Is their focus an area where you have needs? Maybe you need to recruit one of their graduates.

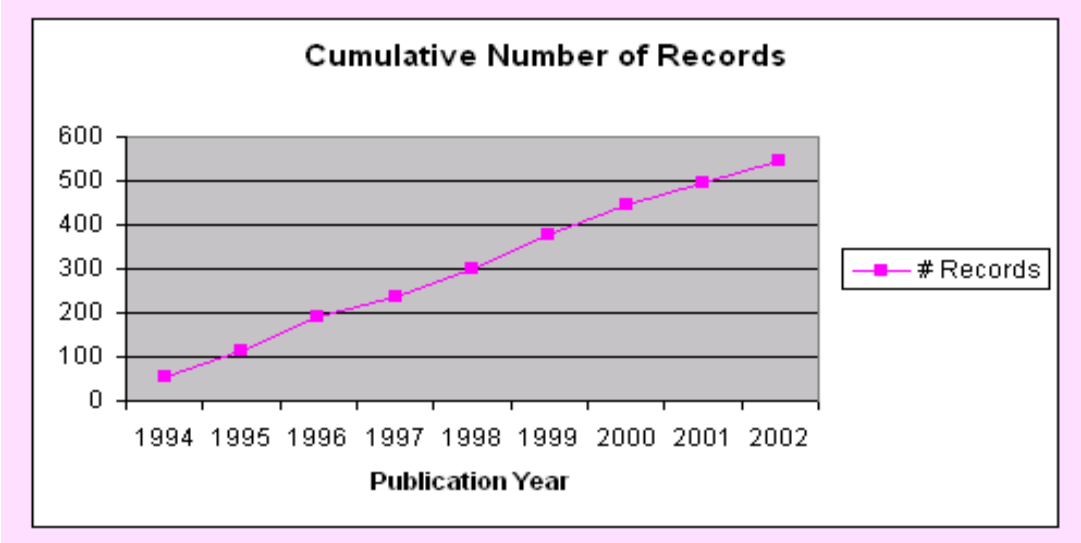
Top Universities	Keywords	Abstract phrases
Top 19	**Top Terms**	**Top Terms**
Data Storage Syst Center, Carnegie Mellon Univ, Pittsburgh, PA, USA[13]	Magnetic recording [5]; Mathematical models [4]; Magnetoresistance [4]; Annealing [2]; Multilayers [2]; Nickel alloys [2]; Ferrites [2]	measurements [3]; thicknesses [2]; magnitudes [2]; scaling equation [1]; NiFe [1]
Univ of California-San Diego, La Jolla, CA, USA[13]	Magnetic Heads [6]; Mathematical models [5]; Friction [3]; Interfaces (materials [3]; Interferometry [3]; Tribology [3]; Magnetic disk storage [3]	developments [3]; measurements [3]; contact recording [2]; Contact Start Stop CSS durability [2]; head medium spacing fluctuation [2]
Univ of Minnesota, MN, USA[13]	Magnetic recording [9]; Magnetic disk storage [3]; Magnetic Heads [3]; Magnetization [3]; Mathematical models [2]; Microscopic examination [2]; Nanotechnology [2]	thin film heads [6]; high density recording [3]; spin stand tester [3]; transitions [3]; NiFe [2]

Figure 6.7 The Leading Universities in Magnetic Storage Research

A number of universities are conducting research in Magnetic Storage. On this cluster web page, a technology manager can find a university where research in a particular area of Magnetic Storage is being conducted (Figure 6.7). These universities are potential partners or may have students graduating with the skills that an organization needs.

6.1.4 The Progress of the Technology

Overall, the results show that Magnetic Storage is a mature technology. The indications of maturity are the progression along the S-curve (Figure 6.8) and the language progression (Table 6.2). Expert evaluators also supported these findings (Ortego, 2004; Domingue, 2004).



The y-axis value is the number of articles published in the given year.
Figure 6.8

The maturity of Magnetic Storage seems to be confirmed by the non-technical terms found in the record set (Table 6.2). The later years terms indicate a further point along the Innovation Cycle. In the earlier years, terms such as —"theory, functions, commercial" appear. Midway, in 1997, "standards" first appears in the literature, characteristic of the Transitional Phase of the Innovation Cycle. In later years, the terms "customers, affordable, and reproducibility" first appear, indicating improvements or more process innovations, which are indicative of the last stage of the Innovation Cycle (Abernathy and Utterback, 1978). The information in this section combined with corresponding information in other sections indicates that someone making a decision about Magnetic Storage may need to expand their search to more innovative forms of this technology. This technology, as it has traditionally been discussed, is towards the end of its life cycle. These facts do not necessarily mean that a technology is obsolete; however, the discussion of the technology is changing and an innovation that may change the

nature of the technology is likely to appear on the horizon. In this case, emerging topics center on networks, instead of devices.

Table 6.2: Analysis of Non-Technical Terms

FIRST YEAR OF APPEARANCE	NON-TECHNICAL TERMS	
1995	**functions**	efficient
	theory	environments
	commercial	copy
	techniques	constraints
	obsolescence	policy
	issues	predictions
	costs	configurations
	devices	components
1996	interfaces	demonstration
	behaviors	instability
	products	repeatability
	models	capabilities
	relationships	fundamentals
	stable	obsolete
	established	start-up
1997	challenges	
	applied	
	standards	
	usefulness	
1998	interactions	
	domains	
	barriers	
	basics	
	quality	
	success	
	economies	
	infrastructure	
	utilized	
1999	appliances	
2000	**reproducibility**	
2001	**customers**	
2002	**affordable**	

6.2 Framework Evaluation

The purpose of a focus group is not to be representative but to gain the greatest possible understanding of a topic from the point of view of participants. In contrast to standard survey and interview techniques, the goal of focus groups is to understand the "why" of participant responses. Focus groups have been used to understand user interaction with information products (Eysenbach and Kohler, 2002). In this research, a focus group served as outside evaluators of the information produced, along previously defined dimensions. The focus group provided insight not only as to whether the information products met the requirements stated by decision-makers, but as to what improvements could be made in order to further attain those goals.

In the interviews conducted at Step One, the initial participants conveyed the role that information plays in their decision-making. The participants named 14 expectations that they have of such information. The dimensions are:

1. Emphasizes significant use of peer reviewed journals.
2. Provides an ability to obtain results by category (chronology, geography, etc.).
3. Provides access to new ideas.
4. Produces results which can be used to solve problems.
5. Provides factual data for support, as in proposals and decision justification.
6. Permits increase in overall knowledge of a problem area.
7. Provides a better understanding of the needs of funding sources.
8. Leads to less duplication of research efforts (i.e., through seeing what others are doing and sharing information).
9. Permits identification of additional approaches/techniques/options.

10. Leads to a broader research focus.
11. Leads to understanding of long-term issues.
12. Leads to improvement of organizational competitiveness.
13. Can be used in making strategic technology decisions.
14. Can be Customized

The last dimension was clearly outside of the scope of this particular project and, therefore, the first 13 dimensions became the basis for analyzing the Information on the website developed in Step Six.

Five technology decision-makers who were familiar with Magnetic Storage technology participated in a focus group evaluating the Website information along the above dimensions. Three additional evaluators, not able to attend the focus group, provided input online. Two online evaluators were provided information on the consensus opinions of the focus group. The evaluators consisted of two consultant practitioners who make recommendations to clients, two organizational implementers, who make decisions on what technology to implement in their own organization, two organizational innovators, who developed technology solutions, and two academic researchers. The instructions emphasized that they were to evaluate the type of information provided and not the information itself. The evaluators who offered their input online received background information in an email, along with the consensus results from the focus group.

The focus group session began with a briefing on the background of the research and instructions. A transcript of the opening statement can be found in Appendix G. Each evaluator sat in front of a computer with the Welcome Screen for the website. The

evaluators were also given a sheet listing the evaluation criteria. They spent 15-20 minutes reviewing the information before the discussion began. Following the review period, the focus group engaged in discussion along each of the evaluation criteria. In addition to the Magnetic Storage sample set, a dataset from the search string "storage area network" was also provided. This dataset, which includes only 76 records, was shown on a projector. This additional information provides a comparison, as magnetic storage in general is an older technology and storage area networks are considered emerging.

The participants evaluated the information in comparison to the current forms of information that they receive. The results of the focus group are tabulated in Table 6.3. The table records categories of statements made by evaluators and notes which evaluator expressed agreement with the statement. A missing check does not indicate disagreement with the statement, but simply means that agreement was not overtly expressed. In some cases, the difference is simply based on whether the evaluator was involved in the focus group or not. The first five columns in the table represent individuals who participated in the focus group. The last three columns represent those individuals who responded online. Only those sentiments expressed by at least two individuals are recorded.

The results from the focus group show that the information product provides a good broad overview of the topic area and that it is good for R&D decision makers and technology implementers. There are many areas where the evaluators had a consensus opinion. Table 6.4 summarizes these opinions. As can be expected, the information was deemed less useful for consultant practitioners who are recommending technologies that are already developed by other organizations. While organizational implementers are also

Table 6.3: Focus Group Evaluation Results

Evaluation Statements	CP-1	CP-2	AR-1	AR-2	IMP-F	IMP-O	INN-1	INN-2
Uses peer-reviewed journals	√	√	√	√	√	√	√	√
Wants breakout between applied and theory	√	√	√	√		√		
Good for researchers		√	√	√			√	√
Like the categories and presentation of the categories	√	√	√	√	√	√	√	√
Want to see topic branches/compare technologies	√	√	√	√	√			
Little direct access to new ideas/more history	√	√	√	√				
Want to see links to emerging technologies	√	√	√	√	√			
Better than current methods for accessing new ideas						√	√	√
Somewhat better than current methods for solving problems					√	√	√	√
Technologies need to be defined more specifically	√	√						
Significantly/Somewhat better for proposals				√	√	√	√	√
Good source for references (whose doing what with whom)	√	√	√	√	√			
Comparison information needed for proposals	√	√	√	√	√			
Significantly/Somewhat support for overall knowledge					√	√	√	√
Categories would help develop overall knowledge	√	√	√	√	√			
Poor source for information on funding sources	√	√	√	√	√	√		
Somewhat better for information on funding sources							√	√
Whose doing what- Abstract better than Keywords(Specific)	√	√	√	√	√			
Whose doing what- want drilldown to titles	√	√	√	√	√			
Somewhat better for avoiding duplicating effort					√	√	√	√
Good for seeing different applications of the technology							√	√
Broadens research focus, but needs more specifics	√	√	√	√	√	√	√	√
Life Cycle is helpful, but needs comparison information	√	√	√	√	√	√	√	√
Not good for a corporate sell	√	√						
Not better than current methods for competitiveness					√	√		
Significantly better for competitiveness, but not guaranteed							√	√
Somewhat/Significantly better for strategic decisions					√	√	√	√
Drill down and comparison would be better for strategy	√	√	√	√	√			
Visualization of some areas needs improvement	√	√	√	√	√	√	√	√

only implementing a technology developed by others, this group may differ in their analysis of the type of information in the example output because of the difference in the level of knowledge about the technology. The organizational implementers may use the broad-based information provided to better understand a technology about which they are

not intimately familiar, but decide on its feasibility. To this end, the approach of starting with the stated needs of decision-makers and framing the information product in those contexts was successful. The participants expressed appreciation for that format, as well as the numerous categorical perspectives. However, there is still significant room for improvement in the presentation of many components of the website pages.

The participants found that both the relationships between organizations portrayed in the collaboration map and the ranking of the organizations according to their publication numbers provided relevant competitive information. Lastly, the participants deemed the Life Cycle information both accurate and useful in decision-making. However, the lack of comparison information made that information more difficult to interpret.

Table 6.4: Focus Group Consensus Opinions

Likes	Wants
• Broad Overview • Information for Researchers • Categories • Overall Presentation (Answer Approach) • Collaboaration Map • Organization Rankings • Life Cycle Information • Abstract Phrases Over Keywords	• Customizability • Drill Down Capabilities • Comparison Information • Better Visualization • Explanation of Concept Maps • Separation between Applied and Theoretical • Access to More Specifics

It is noteworthy that the participants want more specific information. On the website, both keywords and Abstract Phrases were listed as the research topics for organizations and authors. The participants favored the Abstract Phrases because they were more specific than Keywords. They also expressed a desire for more specific

information and greater detail in the relationship between those phrases. Using Abstract Phrases as the basis for additional methods, would make it possible for decision-makers to have access to information at a level of detail not possible when using Keywords. However, the participants also appreciated the broader information. "Drill-down Capability" was the main principle expressed. Along with the drill-down capability, the focus group reiterated the earlier stated desire for customizability. One participant stated that the information was overwhelming and primarily wanted the ability to control how many members of a list were shown at once.

While some sections of the website were easy to comprehend, others needed further explanation. In most cases, the lack of clarity was expelled by providing some form of context, either comparison information or greater explanation of the results. The online evaluators each had a particular item that they found to be unclear as well. In some cases, the method of visualization was the hindering factor. The focus group participants understood the collaboration map, but not the concept map. Those maps have an almost identical visualization method. It appears that the visual form seems more suited to the cross-correlation maps in VantagePoint, than to the conceptual factor maps. Some information, such as the gap analysis of organizations and keywords, became more clear when shown in VantagePoint, underscoring the need for customizability even in the presentation of the information.

The evaluators were given the 13 dimensions discussed in Chapter Four along which to evaluate the information product. The focus group interaction provided rich details into how the participants interacted with the information. For example, when participants didn't understand a certain presentation of the information, they requested

that an explanation be made available. However, in some cases, where a visual relied upon a written explanation, the participants did not read the information. This fact underscores the need for visuals that expressly state the intent of the information presented. Also, this research took an approach of starting with the statements of information needs and crafting the information to address those statements. While this seems to have been an effective method, the problems experienced when providing VantagePoint lists, matrices, and maps with *no* context can only be mitigated to the degree to which an exhaustive context is provided. For example, the website presents conference information under the following heading: "Attend one of these Conferences." One of the consulting practitioners stated that this was not useful because conferences are too far ahead into the future. It never occurred to him that perhaps he could review the proceedings from the prior conference, because that was not expressly stated. These types of issues should be kept in mind when crafting the visualization and contextual representation of these information products.

It is also important to note that while the evaluators expressed interest in a customizable solution with drill-down capabilities and comparison information included, they also stated that the information was overwhelming. A solution must be addressed in the context of visualization. This research indicates that presenting information in packets geared toward stated information needs is meaningful for technology decision-makers. However, additional research must be done in order to determine the appropriate balance between the desire for access to more information and information overload. Improved visualization and a drill-down approach, which is currently utilized in Executive Information Systems, may mitigate some of those issues.

CHAPTER 7

CONCLUSIONS AND FUTURE RESEARCH

The purpose of this research was to provide a method of producing technical intelligence that provides an advantage over methods currently used by decision-makers. Information Analysts and Text Data Mining Professionals have had the ability to apply advanced techniques to provide information to Technology Decision-Makers. However, there has been a disconnect between the information provided and the presentation of that information in a manner that decision-makers find useful. This research used a six-step approach to work through the various stages of the Intelligence and Text Data Mining Processes in order to address issues that may hinder the use of Text Data Mining in the Intelligence Cycle and the actual use of that Intelligence in making technology decisions. Figure 3.1 demonstrates the flow of the research and the contributions of one step to another. The evaluation efforts from this research found that presenting mined information in packets based on the stated information needs, in the context of an action for the decision-makers, offers an improvement over the current methods used by these decision-makers along a number of dimensions. However, refinements are needed and should be addressed by future research.

7.1 Information Products

The similarity of expressed needs among a wide range of technology managers reveals that certain information and methods can be used to address common needs. Herring (1998) introduced a protocol that he recommended be used each time an analyst initiated an intelligence gathering effort. This research started with that protocol as a

basis to identify common interests among technology managers regardless of industry and size of the organization.

Interviews of technology managers produced a set of technologies to serve as example cases throughout the analysis of the methods and overall framework. Those interviews also provided insight into how decision makers use information, identifying dimensions along which to evaluate the information product (Table 4.3). A questionnaire given to 34 respondents from four different industries identified the information most important to decision-makers as well as clusters of common interests. This information was used to create a website for Magnetic Storage, our example technology. Each cluster of interest had its own web page on the site. The bibliometric and text data mining results were presented in the context of the stated needs from the questionnaire. A group of 8 decision-makers from a variety of technology perspectives evaluated the information products along 13 dimensions identified from the initial interviews. These dimensions represent how decision-makers may use the information. The evaluation results are found in Tables 6.3 and 6.4. The evaluators found that the information provided an improvement along most of the dimensions. However, they wanted more customizability and drill-down capabilities. They also wanted more detailed information, better visualization, and the ability to compare technologies along the given metrics. Implementing the information from the evaluations can further enhance the usability of these types of information products in the decision-making process

7.2 Text Data Mining Methods

In addition to identifying the information that is important to decision makers, this research made improvements to the methods for analyzing information. Identifying

terms that are synonymous is important to improving accuracy when mining free text, thus enabling the provision of the more specific information desired by the evaluators. An algorithm was developed that has delivered at least an 89% precision rate in making such identifications. This level of precision was achieved across five different technology areas (pollution monitoring, remote sensing, magnetic storage, fuel cells, and geographic information systems) and was used in three different databases (Compendex, Inspec, and Pollution Abstracts), all with about the same level of precision. These results indicate that the algorithm may be used with other types of free text such as Patents and the Internet. The impact of this algorithm can be seen in Tables 5.4 – 5.8. Terms that are conceptually important to the dataset (solid oxide fuel cells) have replaced very generic common words (study, results) at the top of the list. Also, the viability of using Abstract Phrases improves because the concept-clumping algorithm reduces the number or terms to consider for clustering by 30%. The terms left are the more technical terms. The comparison of cluster terms in Chapter Five found more technically meaningful clusters. The result is the ability to use abstract phrases in analysis, which allows the more detailed nature of abstracts to be captured in a clustering format. Clumped Abstract Phrases capture the broad relationships as well.

7.3 Future Research

There are a plethora of future research opportunities sparked from this research, both to improve the presentation of information products and in the techniques used to develop those products.

- There are significant future research opportunities in the presentation of the information. More can be done to make use of visualization research to improve

the clarity of the information provided. For example, the focus group participants did not immediately understand the nature of the overview maps and clearly did not see a difference between the overview factor maps and the cross-correlation maps, because they look very similar. The use of metaphors, color, and dimensions may improve the clarity and usefulness of those maps. Visualization methods should also aim to reduce the "overwhelming" elements of the information provided.

- It was very clear that although packetizing the information into categories was helpful to the users, the users want more flexibility in the packets and a drill-down component. To this end, additional research using expert system approaches may improve the usability of the information. Research on combining text mining with Executive Information Systems may also provide an answer for how to provide drill-down capabilities in a manner that decision-makers want.
- In the process of conducting this research, achieving a high level of recall in clumping like concepts together was difficult. However, some foundational methods (i.e. shared words, similarity, stop words lists, and the ignore feature) have been established that can improve the algorithm further. In particular, by applying the methods to single word matches and incorporating stemming into the algorithm warrant investigation. Other research to improve the clumping results includes examining the number of default factors when clustering terms.
- In the evaluation of the different cluster types, there were indications that the number of factors traditionally used for keywords was insufficient when using the clumped abstract phrases. The same research that was performed in this study to

make the determination of number of factors for keywords should be replicated for clumped abstract phrases.

- Additional research can be conducted to determine the effect of clumping when dealing with websites or patents. Clumping may also be useful for improving patent searches, especially in identifying emerging technologies in the patent dataset.
- In the process of evaluating the clusters, it became clear that the quantitative methods for evaluating clusters that have traditionally been used in document clustering do not seem to be the best approach for term clustering. In term clusters, the accuracy of the conceptual linkages is more important than any "physical" characteristic of the cluster. However, the traditional way in which "accuracy" is measured becomes increasingly complex when making term-concept clusters. This issue occurs primarily because terms can be in multiple clusters and some terms selected for clustering, when using PCA, are not actually included in any cluster.
- Finally, in combining the methods with the information products, this research used some of the more basic methods to produce information. Additional research should seek to apply more advanced methods of text mining to address the information needs of decision-makers. Link Analysis, linking more specific information in research, may address the decision-makers' need for more specific information. Sequential analysis, or identifying metrics over time that occur frequently in a sequential pattern, may do the same, as well as provide some forecasting insights. The techniques should also be applied to different types of

databases for comparison, such as patent databases and business databases. More research could also be done to further study the lexical nature of technology publication databases; a key example is further identifying terminology that indicates a technology's position in its life cycle.

APPENDIX A: FUNCTIONS AFFECTED BY CTI

Business and Technology Strategy

- Strategic Technology Roles and Directions
  - What role advanced technology will play in business, vision, goals, and strategy
  - Which technology-based business and market directions to pursue
  - Which core technical competencies to create and/or nurture
- Technology Needs and Opportunity Evaluation
  - What priority to assign current product, process, and operations technology needs
  - Which new technology applications (product, process, service, or operations developments to pursue
  - Whether to enter a technology-based product line with strong competitors
- Technical Information and Property Security
  - How to protect intellectual property
  - How to protect sensitive company information

Technology Acquisition

- Technology Acquisition Planning
  - How best to acquire a new technology: internal R&D, external purchase. Licensing-in, hiring or partnering
  - How much to invest in technology acquisition and R&D budgets
- Technology Collaboration Choices
  - Whether to enter into joint technology development venture with another organization and what gains to expect from it

- Which technology partners to consider and which terms of agreement to establish

- Technology Acquisition Implementation
  - What external technology sources to pursue
  - How to get the best bargain from external acquisitions

R&D Program and Portfolio Management

- R&D Investment Portfolio Decisions
  - Which new product development initiatives or improvements to make
  - What allocation of R&D funds near-versus long-term projects
  - Whether to terminate or delay work on a project or in an S&T area
- Technical Research, Product or Process Development Strategies
  - Which technical approach to take in developing new product or process technologies
  - What technical objectives to set for R&D programs

Technology Deployment Investments or Divestiture Actions

- Product and Process Investment Decisions
  - Which new product options to select for investment
  - Which capital expenditures to make for facility or process technology needs
- Technology Transfer Mechanisms
  - How to transition new know-how from R&D to manufacturing operations
  - Whether to permit external disposition of technology and how to transfer or limit distribution of rights or results

Production and Delivery Operations

- Manufacturing and Distribution Operations
    - How to qualify suppliers, customers or partners
    - What kind of technology training and operational procedures to establish
- Technology Maintenance and Replacement
    - What technology maintenance, repair, and replacement policies to use
    - How to trouble-shoot product or manufacturing technology problem
        - Tracking competitor activities
        - Identifying emerging technologies that can aid or hurt the company

APPENDIX B: INTELLIGENCE QUESTIONS IN THE PRODUCT LIFE CYCLE

Table B.1 Intelligence Questions for Life Cycle Stages

	Introduction	Growth	Maturity	Decline
Intelligence Questions	Which of our competitors will wish to compete in this new product market? What are their strengths and/or weaknesses? How long will it take to enter the market? How do we need to prepare for this competitive market? (**SWOT analysis**)	Is the form, quantity, or intensity of competition going to change during this growth stage? How? (**market scanning**)	What measures have our strongest competitors used to position themselves in this established market? What new competitive strategies could be most profitable for us during this established period (most surprising to our competitors)? Why? (**market monitoring, SWOT analysis, and strategy development**)	How long can this market be projected to remain profitable? Are we positioned to continue generating profits or should we exit this market? (**market projections profitability ratios, and comparative assessments**)
	Could competition come from another industry? From overseas? In what form? (**market forecasting**)	How is industry and individual corporate growth going to affect competition? What competitive strategies are our competitors going to use to gain market share? (**growth measurement & competitive analysis**)	What competitive insight has our firm acquired about the market environment? How can we use this in future markets or other product divisions? (**internal corporate learning assessment**)	How can we develop foresight to identify future markets? Are we positioning ourselves to compete in tomorrow's market? Likewise, are our competitors moving out of the current market? Why or why not? Are our competitors moving into the next market? Why? Which firm(s) are taking a leadership role in defining future markets? How? (**market foresight, positioning, and market scanning**)

(continued)

37

Table B.1 (continued)

Introduction	Growth	Maturity	Decline
If our firm introduced a new product, how can we maintain our competitive market edge? Can we also be the first company to introduce the next product? Why? If we didn't introduce this product, how can we compete in this market? What market advantage can we employ? How can we be first to market with the next product? (**competitive market & extra-market analysis**)	Are alternate products going to appear on the market—possibly taking market share? (**extra-market assessments**)	What is the projected length of this established and comparatively stable maturity stage? How can our firm and/or industry protect itself against premature market decline? (**market projections and assessments of market positioning**)	What advantage can be drawn from this declining market and applied to future markets? Is there a component of our current product that can apply strategically to other products or markets? Can our full product be revitalized via introduction to other markets and/or by finding alternate application(s) for the product? (**learning curves, internal corporate review/assessment**)
What types of product modifications might help us to gain market share against our competitors? (**strategic product improvements**)	How quickly is the market going to grow? For how long and how far? (**market forecasting**)	How can we position our firm for dominance during the decline stage (i.e., how can we secure remaining market sales)? (**strategic positioning**)	

(*continued*)

Table B.1 (continued)

Introduction	Growth	Maturity	Decline
Have we developed a clear market forecast? (**market forecasting**)	Which additional competitors are quietly positioning themselves as real or potential competitive threats? How can we identify competitors who may enter from alternate markets? Which alternate markets may have players capable of entering this growing market? How long might it take them to enter this market? (**SWOT analysis and market scanning**)	Are we preparing for future now by seeking the next market (looking for market indicators)? (**market forecasting & strategic planning**)	
Has this market been clearly defined? Have we developed a long-term market view? (**market definitions**)			

APPENDIX C: INTERVIEW QUESTIONS

Name

Organization

Phone Email

Position

Is the person a

Scientist/Engineer Marketing Personnel

Executive Policy Maker/Regulator

Other

I. Target Audience Profile

1. What are some technology challenges that you face?
2. Who are other organizations operating in your same field?
3. What decisions do you make concerning technology?
4. What information sources (magazines, vendors, internet, consulting research etc) do you utilize in order to make those decisions? How is this information incorporated into the decision process?
5. What types of technical monitoring information would you prefer to receive? (e.g., analytical alerts, competitor assessments, short briefings, etc.)

6. What technology decisions and/or actions will your group face in the next year, where early technical information could make a significant difference?

In what ways would information/intelligence or monitoring affect the overall success of your organization?

II. A Monitoring System

Suppose a system was developed that would automatically seek out information on your technology, analyze the information, and present the results on your desktop, How would you suggest the system be organized?

1. How will you evaluate the value of the information that you receive?

III. Early-Warning Topics

1. If you were searching on the internet for technologies/functions that you must regularly monitor, what search terms would you use?
2. In regards to the topic that you searched on, what would you like to know about?
    - About the technology?
    - About other organizations and experts?

APPENDIX D: INFORMATION REQUIREMENTS QUESTIONNAIRE

The following questionnaire is intended to determine the type of information that you consider important to the long-term strategic technology/research decisions that your organization must make. Please select the circle that indicates your level of agreement with the statement above the scale.

Top of Form

Name and/or E-mail address_____

Organization: _____

Position & Job Description: _____

I. Technology Topic Profile

1. I would like to see an overview of the research conducted in my technical domain.

Strongly Agree Agree Disagree Strongly Disagree

2. I want to know the gaps in my organization's activities in comparison to the full scheme of research in our technical domain.

Strongly Agree Agree Disagree Strongly Disagree

3. The organization has been slow in detecting emerging technological breakthroughs in our domain.

Strongly Agree Agree Disagree Strongly Disagree

4. We want to be aware of constraints/difficult issues faced in developing a particular technology.

Strongly Agree Agree Disagree Strongly Disagree

II. Organizations

5. It is important to know the names of universities, companies, agencies, and labs that are publishing in my technical domain.

Strongly Agree Agree Disagree Strongly Disagree

6. Sometimes emerging competitors/organizations come from an entirely different industry. I need to identify those organizations and their research interests.

Strongly Agree Agree Disagree Strongly Disagree

7. It is important for me to know if an organization is no longer researching in my technical domain and perhaps why they are no longer doing so.

Strongly Agree Agree Disagree Strongly Disagree

8. I would like to know which universities are researching in my technical domain in order to strategically recruit for my organizational needs.

Strongly Agree Agree Disagree Strongly Disagree

9. It is important for me to know the technical strengths and weaknesses of other organizations in my technical domain.

Strongly Agree Agree Disagree Strongly Disagree

10. I would like profiles of the work being done by other organizations working in my area. I am interested in

a. their current research activity,

Strongly Agree Agree Disagree Strongly Disagree

b. how their activities have changed over time,

Strongly Agree Agree Disagree Strongly Disagree

c. their partnerships,.

Strongly Agree Agree Disagree Strongly Disagree

d. the size of research teams

Strongly Agree Agree Disagree Strongly Disagree

11. If I knew which organizations in my technical domain were publishing together, then I would have greater insight into their strategic direction.

Strongly Agree Agree Disagree Strongly Disagree

III. Suppliers

12. My organization would be adversely affected by major technical changes by suppliers.

Strongly Agree Agree Disagree Strongly Disagree

IV. Experts

13. Publication databases contain the names of individuals publishing from around the world. I want to track the activities of relevant external subject matter experts. .

Strongly Agree Agree Disagree Strongly Disagree

V. Global Activity

14. My company is evaluating our future opportunities in a technical domain. Increases or decreases in global activity regarding this technology may determine my organization's interest in this technology.

Strongly Agree Agree Disagree Strongly Disagree

15. A profile of the expertise located in other countries would be helpful to my decision-making.

Strongly Agree Agree Disagree Strongly Disagree

VI. Periodicals

16. I would like to know what periodicals are publishing in my technical domain.

Strongly Agree Agree Disagree Strongly Disagree

17. It is important to know the spread of topics discussed in my domain's most important periodicals.

Strongly Agree Agree Disagree Strongly Disagree

18. It is important to know which conferences cover my technical domain and the specific topics covered.

Strongly Agree Agree Disagree Strongly Disagree

19. Publication databases contain conference proceedings. Conferences represent research in its earliest stages. It is important for me to be aware of this research.

Strongly Agree Agree Disagree Strongly Disagree

VII. Regulations and Standards

20. My technical decisions can be affected by changes in international, political, social, economic or regulatory situations.

Strongly Agree Agree Disagree Strongly Disagree

VIII. Miscellaneous

21. It is important that I know about the potential social, economic, environmental, or cultural impact of technology developments.

Strongly Agree Agree Disagree Strongly Disagree

22. It is important that I know the position of specific technologies in their life cycle.

Strongly Agree Agree Disagree Strongly Disagree

23. The commercial readiness of a technology can be determined by the topics of discussion in technology publications.

Strongly Agree Agree Disagree Strongly Disagree

24. Our organization markets the ability of our products/services to accomplish certain functions. Part of my responsibility is to determine the appropriate technology to fulfill each function.

Strongly Agree Agree Disagree Strongly Disagree

APPENDIX E: MAP OF QUESTIONNAIRE STATEMENTS TO PUBLICATION METRICS

Table E.1 Questionnaire Statements Mapped to Publication Database Metrics in VantagePoint

Statement	Metric
R1. I would like to see an overview of the research conducted in my technical domain.	• Concept map • Top keywords
R2. I want to know the gaps in my organization's activities in comparison to the full scheme of research in our technical domain.	• Keywords X Affiliation (grouped)
R4. We want to be aware of constraints/difficult issues faced in developing a particular technology.	• Abstract Phrases (Topics Discussed)
R5. It is important to know the names of universities, companies, agencies, and labs that are publishing in my technical domain.	• Affiliations list
R6. Sometimes emerging competitors/organizations come from an entirely different industry. I need to identify those organizations and their research interests.	• Affiliations X Keywords or Abstract Phrases
R7. It is important for me to know if an organization is no longer researching in my technical domain and perhaps why they are no longer doing so.	• Affiliations list X Year Matrix
R8. I would like to know which universities are researching in my technical domain in order to strategically recruit for my organizational needs.	• Affiliations list (Group Universities)
R9. It is important for me to know the technical strengths and weaknesses of other organizations in my technical domain.	• Affiliations X Keywords matrix or Abstract Phrases
R10. I would like profiles of the work being done by other organizations working in my area. I am interested in a. their current research activity,	• Affiliations X Keywords
R11. how their activities have changed over time,	• Affiliations X Keywords X Time
R12. their partnerships,.	• Cross correlation of Affiliations by Authors
R13. the size of research teams	• Affiliation X Author X

Table E.1 (continued)

R14 If I knew which organizations in my technical domain were publishing together, then I would have greater insight into their strategic direction.	• Cross correlation of Affiliations X Authors
R15. My organization would be adversely affected by major technical changes by suppliers.	• Affiliations X Keywords or Abstract Phrases
R16. Publication databases contain the names of individuals publishing from around the world. I want to track the activities of relevant external subject matter experts. .	• Authors List
R17. My company is evaluating our future opportunities in a technical domain. Increases or decreases in global activity regarding this technology may determine my organization's interest in this technology.	• Countries X KeywordsX Year
R18. A profile of the expertise located in other countries would be helpful to my decision-making.	• Countries X Keyword • Countries X Abstract Phrases
R19. I would like to know what periodicals are publishing in my technical domain.	• Journals List
R20. It is important to know the spread of topics discussed in my domain's most important periodicals.	• Journals X Keywords • Journals X Abstract Phrases
R21. It is important to know which conferences cover my technical domain and the specific topics covered.	• Sources-- Conferences
R22. Publication databases contain conference proceedings. Conferences represent research in its earliest stages. It is important for me to be aware of this research.	• Conferences X Abstract Phrases
R23. My technical decisions can be affected by changes in international, political, social, economic or regulatory situations.	• Abstract Phrases
R24. It is important that I know about the potential social, economic, environmental, or cultural impact of technology developments.	• Abstract Phrases
R25. It is important that I know the position of technologies in their life cycle.	• Cumulative Records X Year • Abstract Phrases
R26. The commercial readiness of a technology can be determined by the topics of discussion in technology publications.	• Abstract Phrases

APPENDIX F: TRANSCRIPT OF FOCUS GROUP INTRODUCTION

Thank you for participating in this focus group. Let me first give you some background. You all know the challenge of making long-term technology decisions in today's information intense competitive environment. Perhaps you are a researcher looking to avoid duplicated efforts, or you are looking for that company with whom to partner or merge, perhaps you are looking to hire someone, sufficiently knowledgeable. Perhaps, you need to know if the technology is at its maturity peak or not. Will your investment be obsolete next year?

At the beginning of this research, I interviewed 40+ technology decision-makers, people who have to make long-term technology decisions either as a researcher, developer, or implementer. I asked them about the information that they use in order to make those decisions, the sources that they use, and the criteria that they use in order to judge the value of the information. In front of you are the criteria that the interviews revealed are used to evaluate information. You will use those dimensions to evaluate the information that I will show you today.

So, what is this information and how did we choose what will be included? I have found that there is a 5-7 year gap between what is published in business magazines and what appears in journal/conference publication databases. If a decision maker waits until they read it in their favorite business magazine to make a decision, it is too late. Monitoring of the technology landscape must be an ongoing process and should start at the publication database level. However, the results can be numerous. The information in front of you is based on Magnetic Storage. There were 693 abstract records over a ten

year period. This information product analyzes that information to provide answers to questions that technology decision makers deemed important. We provided our interview subjects with a list of 33 statements concerning technology, other countries, and other organizations; and asked them to determine how important the information is to their decision-making. These information products reflect the information that was deemed the most important or clusters of interest by participants.

APPENDIX G: QUESTIONNAIRE STATEMENT SUMMARIES

Table G.1 contains the mean, maximum, minimum and standard deviation for each of the statements on the questionnaire. Following the summary table is the frequency information for each question, R1- R27.

Table G.1 Questionnaire Statement Responses-Descriptive Summary Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
R1	34	1	2	1.56	.504
R2	34	1	3	1.65	.544
R3	34	1	4	2.44	.927
R4	34	1	3	1.62	.551
R5	34	1	3	1.74	.751
R6	34	1	3	1.82	.673
R7	33	1	3	2.30	.728
R8	34	1	3	1.97	.627
R9	34	1	4	1.71	.676
R10	34	1	3	1.68	.589
R11	33	1	4	2.09	.723
R12	34	1	3	1.88	.591
R13	34	1	4	2.21	.687
R14	33	1	3	2.00	.433
R15	31	1	4	2.32	.791
R16	33	1	3	2.24	.614
R17	34	1	4	2.35	.849
R18	33	1	4	2.45	.754
R19	33	1	3	1.55	.564
R20	34	1	3	1.91	.668
R21	34	1	3	1.59	.557
R22	34	1	3	1.85	.610
R23	32	1	4	1.91	.818
R24	34	1	4	1.94	.736
R25	33	1	3	1.88	.696
R26	32	1	4	2.19	.821
R27	33	1	3	1.94	.659
Valid N (listwise)	27				

R1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	44.1	44.1	44.1
	2	19	55.9	55.9	100.0
	Total	34	100.0	100.0	

R2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	13	38.2	38.2	38.2
	2	20	58.8	58.8	97.1
	3	1	2.9	2.9	100.0
	Total	34	100.0	100.0	

R3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	6	17.6	17.6	17.6
	2	11	32.4	32.4	50.0
	3	13	38.2	38.2	88.2
	4	4	11.8	11.8	100.0
	Total	34	100.0	100.0	

R4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	14	41.2	41.2	41.2
	2	19	55.9	55.9	97.1
	3	1	2.9	2.9	100.0
	Total	34	100.0	100.0	

R5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	44.1	44.1	44.1
	2	13	38.2	38.2	82.4
	3	6	17.6	17.6	100.0
	Total	34	100.0	100.0	

R6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	11	32.4	32.4	32.4
	2	18	52.9	52.9	85.3
	3	5	14.7	14.7	100.0
	Total	34	100.0	100.0	

R7

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	14.7	15.2	15.2
	2	13	38.2	39.4	54.5
	3	15	44.1	45.5	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	20.6	20.6	20.6
	2	21	61.8	61.8	82.4
	3	6	17.6	17.6	100.0
	Total	34	100.0	100.0	

R9

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	13	38.2	38.2	38.2
	2	19	55.9	55.9	94.1
	3	1	2.9	2.9	97.1
	4	1	2.9	2.9	100.0
	Total	34	100.0	100.0	

R10

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	13	38.2	38.2	38.2
	2	19	55.9	55.9	94.1
	3	2	5.9	5.9	100.0
	Total	34	100.0	100.0	

R11

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	6	17.6	18.2	18.2
	2	19	55.9	57.6	75.8
	3	7	20.6	21.2	97.0
	4	1	2.9	3.0	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R12

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	8	23.5	23.5	23.5
	2	22	64.7	64.7	88.2
	3	4	11.8	11.8	100.0
	Total	34	100.0	100.0	

R13

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	4	11.8	11.8	11.8
	2	20	58.8	58.8	70.6
	3	9	26.5	26.5	97.1
	4	1	2.9	2.9	100.0
	Total	34	100.0	100.0	

R14

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	8.8	9.1	9.1
	2	27	79.4	81.8	90.9
	3	3	8.8	9.1	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R15

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	14.7	16.1	16.1
	2	12	35.3	38.7	54.8
	3	13	38.2	41.9	96.8
	4	1	2.9	3.2	100.0
	Total	31	91.2	100.0	
Missing	System	3	8.8		
Total		34	100.0		

R16

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	8.8	9.1	9.1
	2	19	55.9	57.6	66.7
	3	11	32.4	33.3	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R17

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	14.7	14.7	14.7
	2	15	44.1	44.1	58.8
	3	11	32.4	32.4	91.2
	4	3	8.8	8.8	100.0
	Total	34	100.0	100.0	

R18

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	5.9	6.1	6.1
	2	17	50.0	51.5	57.6
	3	11	32.4	33.3	90.9
	4	3	8.8	9.1	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R19

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	16	47.1	48.5	48.5
	2	16	47.1	48.5	97.0
	3	1	2.9	3.0	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R20

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	26.5	26.5	26.5
	2	19	55.9	55.9	82.4
	3	6	17.6	17.6	100.0
	Total	34	100.0	100.0	

R21

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	44.1	44.1	44.1
	2	18	52.9	52.9	97.1
	3	1	2.9	2.9	100.0
	Total	34	100.0	100.0	

R22

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	26.5	26.5	26.5
	2	21	61.8	61.8	88.2
	3	4	11.8	11.8	100.0
	Total	34	100.0	100.0	

R23

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	11	32.4	34.4	34.4
	2	14	41.2	43.8	78.1
	3	6	17.6	18.8	96.9
	4	1	2.9	3.1	100.0
	Total	32	94.1	100.0	
Missing	System	2	5.9		
Total		34	100.0		

R24

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	26.5	26.5	26.5
	2	19	55.9	55.9	82.4
	3	5	14.7	14.7	97.1
	4	1	2.9	2.9	100.0
	Total	34	100.0	100.0	

R25

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	10	29.4	30.3	30.3
	2	17	50.0	51.5	81.8
	3	6	17.6	18.2	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

R26

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	6	17.6	18.8	18.8
	2	16	47.1	50.0	68.8
	3	8	23.5	25.0	93.8
	4	2	5.9	6.3	100.0
	Total	32	94.1	100.0	
Missing	System	2	5.9		
Total		34	100.0		

R27

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	8	23.5	24.2	24.2
	2	19	55.9	57.6	81.8
	3	6	17.6	18.2	100.0
	Total	33	97.1	100.0	
Missing	System	1	2.9		
Total		34	100.0		

APPENDIX H: TECHNOLOGY CASES- ZIPF DISTRIBUTION GRAPHS

The Zipf distribution asserts the log of the rank vs. the log of the frequency of a term is linear. The method finds that line, and eliminates the terms with the highest and lowest ranks that fall below the line. On the following pages are the Zipf distribution graphs for each type of term list (Keywords, Cleaned Abstract Phrases, Clumped Abstract Phrases) for each of the five technology cases. VantagePoint contains a script that opens Excel and records a single entry for each frequency and the last rank at which that frequency occurs. Then, it plots the log (rank) vs. log (frequency.) The Excel output for Figure H.1 [c] Fuel Cell Zipf Distribution Graphs: Clumped Abstract Phrases is below:

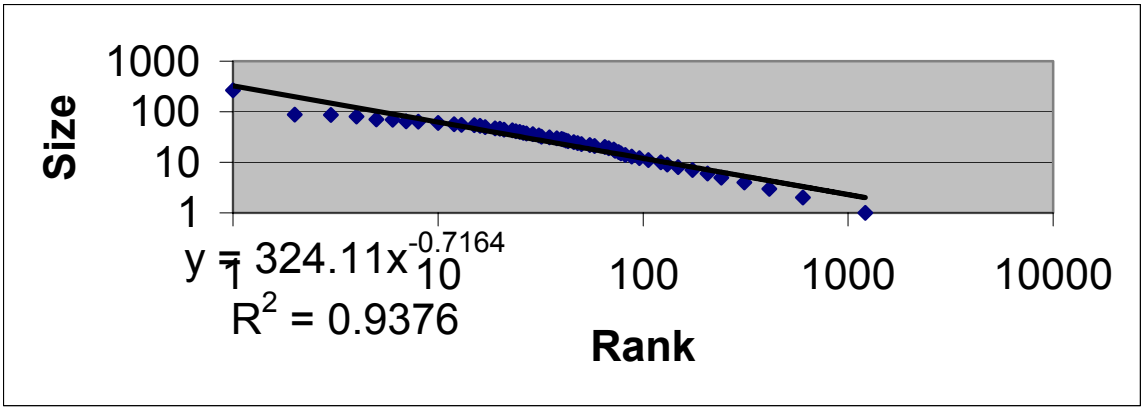
Table H.1: Fuel Cell Clumped Abstract Phrases Ranks and Frequencies

NumTerms	Rank	Size
1	1	212
1	2	163
1	3	93
1	4	88
1	5	71
1	6	69
1	7	66
1	8	58
1	9	57
1	10	56
2	12	54
2	14	53
1	15	47
2	17	45
1	18	43
1	19	39
1	20	36
4	24	34
3	27	32
1	28	30
1	29	29
3	32	28

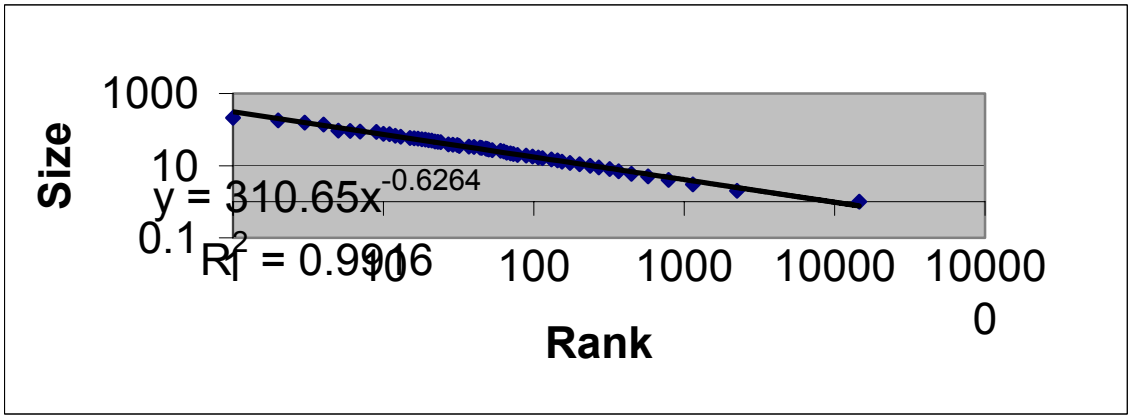
Table H.1: (continued)

1	33	27
3	36	26
1	37	25
4	41	24
4	45	23
3	48	22
9	57	21
2	59	20
3	62	19
6	68	18
8	76	17
5	81	16
13	94	15
8	102	14
8	110	13
15	125	12
17	142	11
27	169	10
25	194	9
35	229	8
47	276	7
74	350	6
112	462	5
171	633	4
368	1001	3
1107	2108	2
7691	9800	1

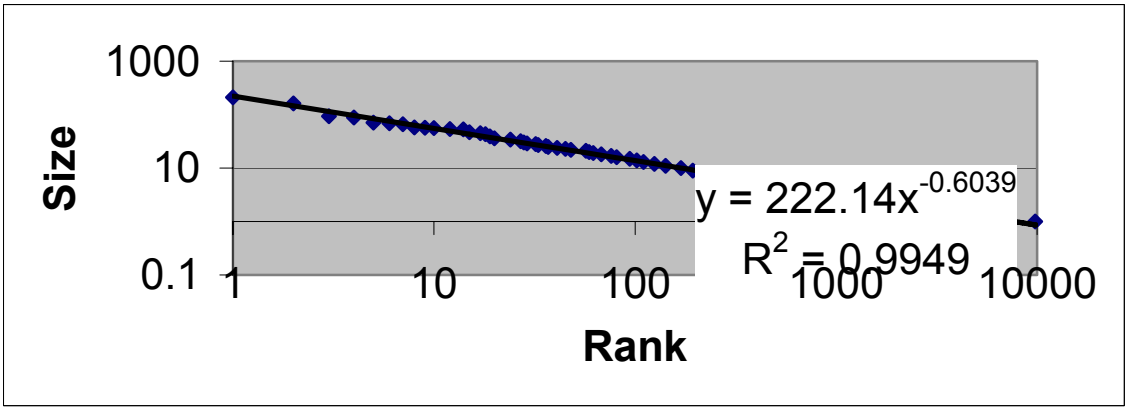
The Highlighted rows correspond to the cutoff points for inclusion in the clustering maps. These are the points above the line in Figure H.1[c]. 25 terms occur nine times, ending with the 194th term. The first term “fuel cells” is the highest ranking term, occurs in 212 records and will be excluded from clustering. “Fuel cells” is the search term, reflecting the intent of this method. This method eliminates such terms because all of the record in the dataset have some relationship with fuel cells. By removing that term , the relationship between the other term becomes evident. At the other end, the terms have frequencies too low to have significant influence.



[a]

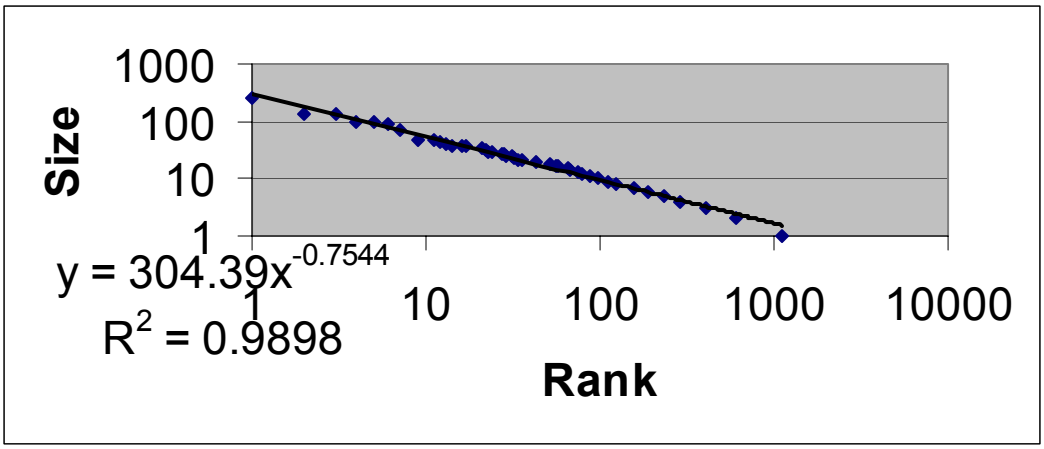


[b]

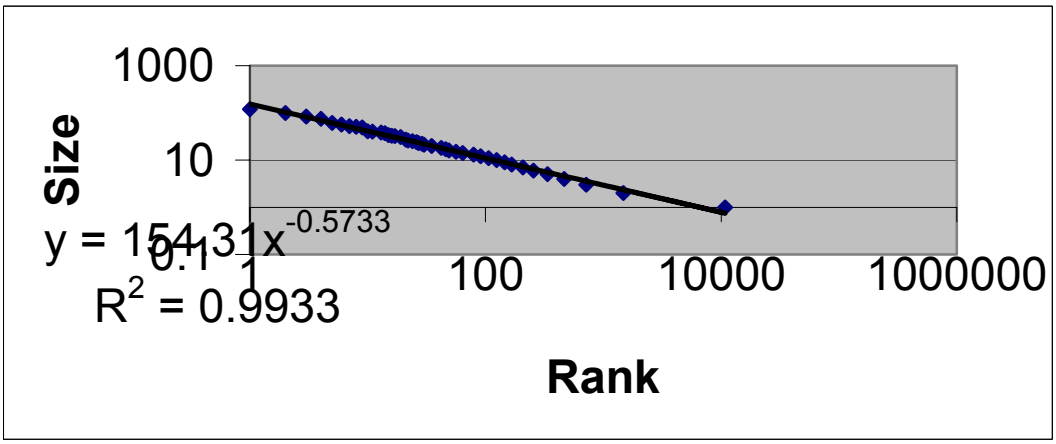


[c]

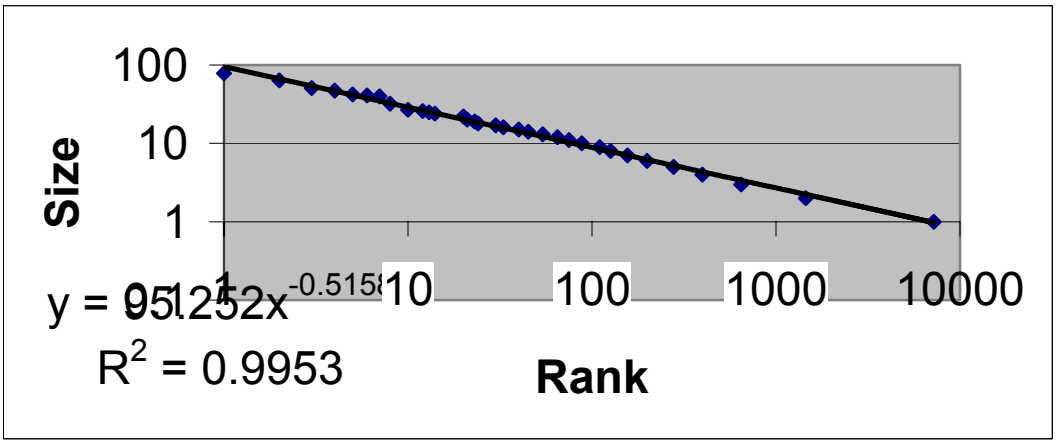
Figure H.1 Fuel Cell Zipf Distribution Graphs: [a] Keywords [b] Cleaned Abstract Phrases [c] Clumped Abstract Phrases



[a]

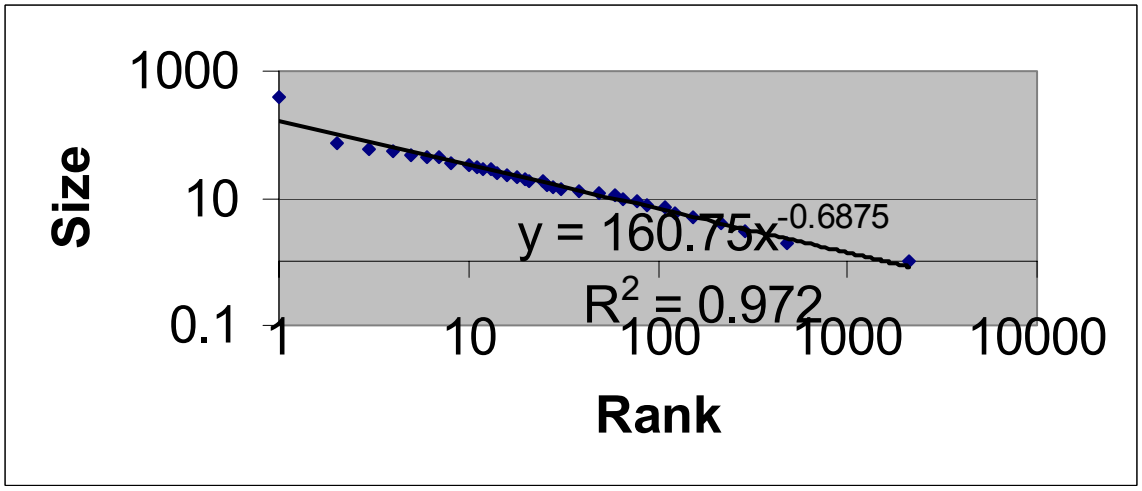


[b]

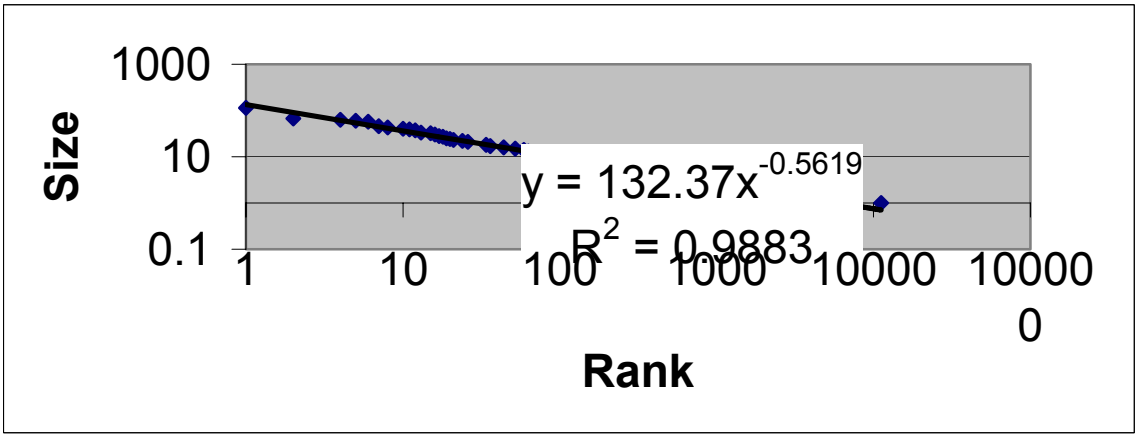


[c]

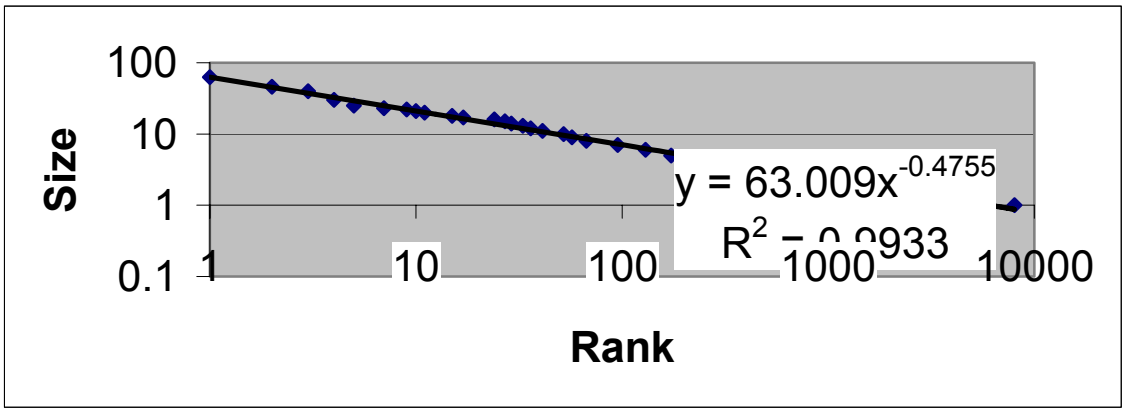
Figure H.2 Magnetic Storage Zipf Distribution Graphs: [a] Keywords [b] Cleaned Abstract Phrases [c] Clumped Abstract Phrases



[a]

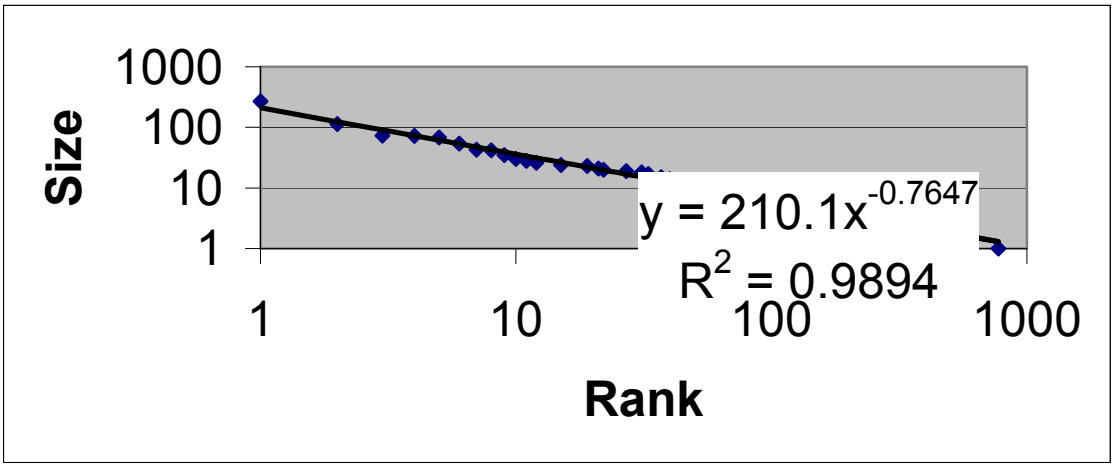


[b]

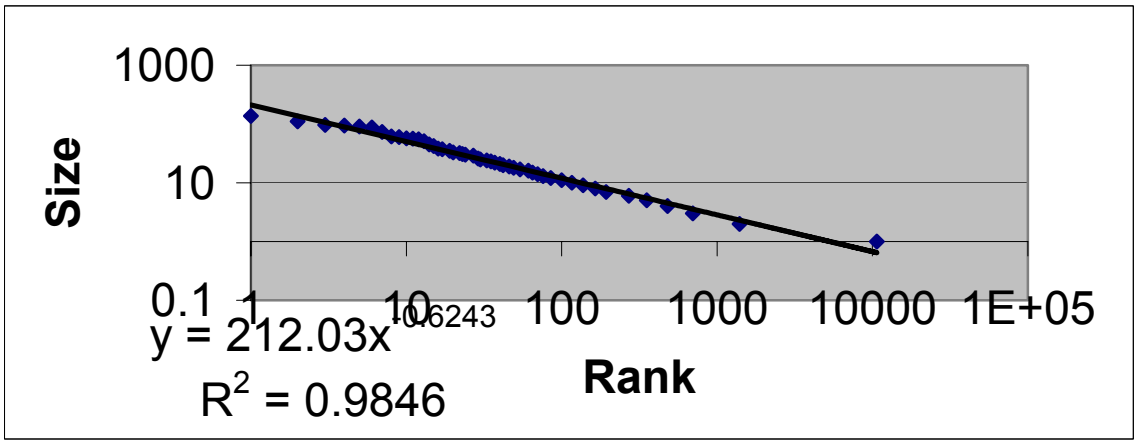


[c]

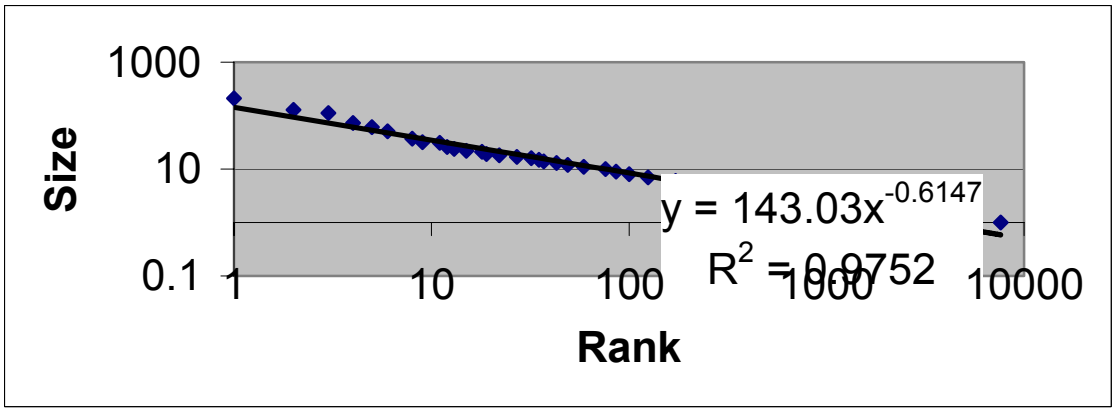
Figure H.3 Remote Sensing Zipf Distribution Graphs: [a] Keywords [b] Cleaned Abstract Phrases [c] Clumped Abstract Phrases



[a]

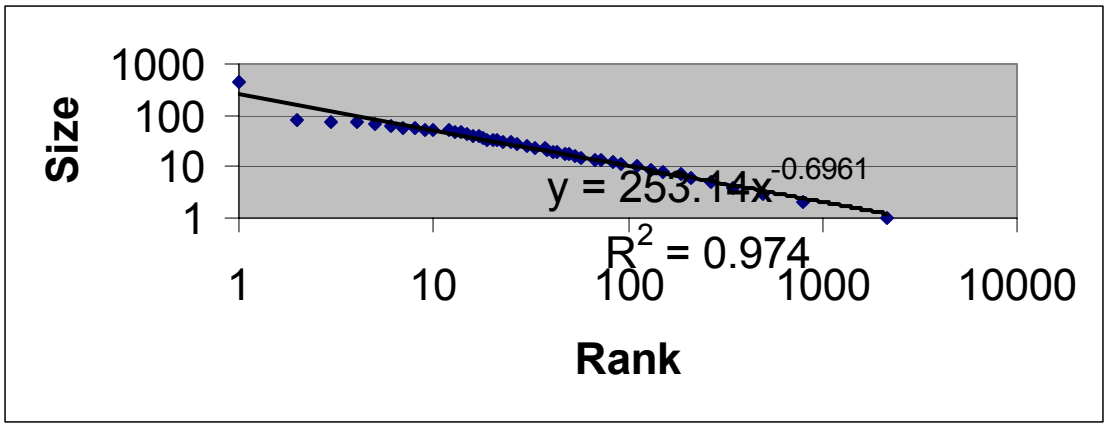


[b]

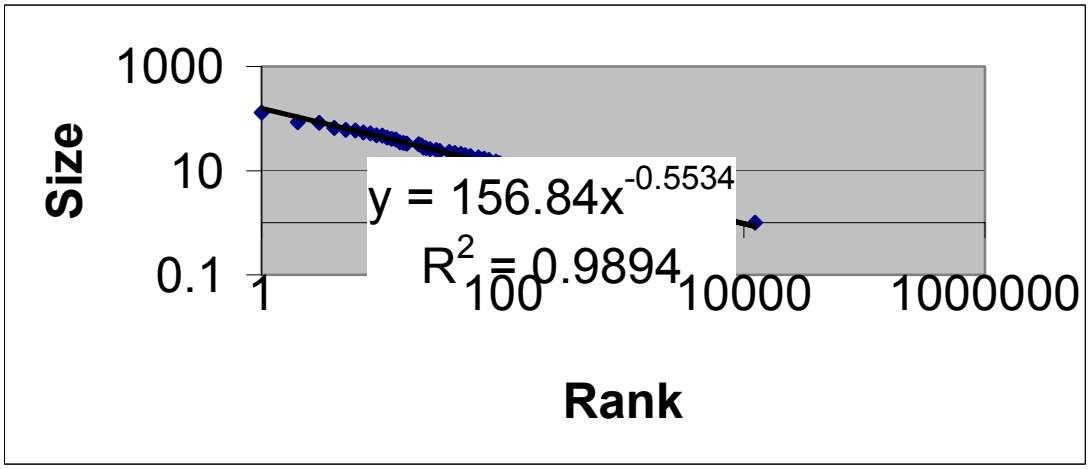


[c]

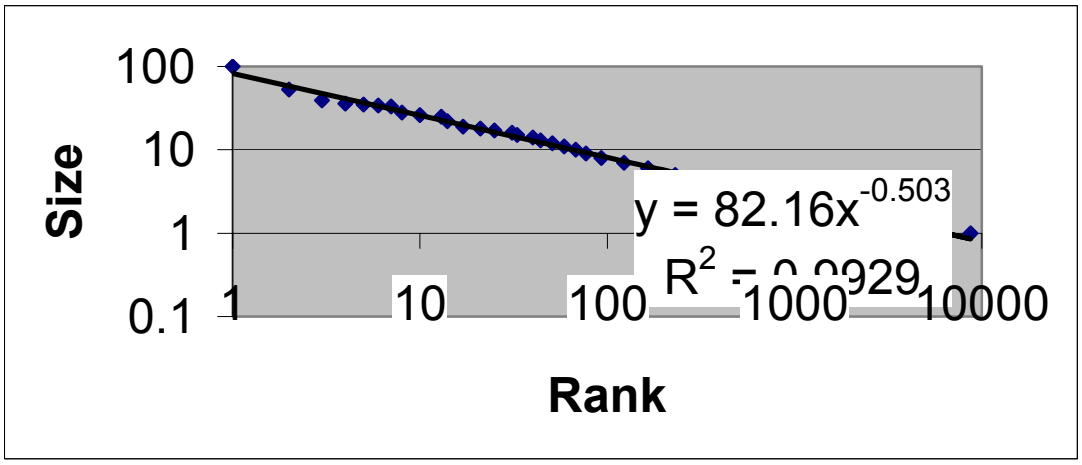
Figure H.4 Geographical Information Systems Zipf Distribution Graphs: [a] Keywords [b] Cleaned Abstract Phrases [c] Clumped Abstract Phrases



[a]



[b]



[c]

Figure H.5 Pollution Monitoring Zipf Distribution Graphs: [a] Keywords [b] Cleaned Abstract Phrases [c] Clumped Abstract Phrases

APPENIDIX I: TECHNOLOGY CASES- CLUSTER MAPS

Appendix I contains Figures I.1 – I.15, which are all of the Cluster Maps for Keywords, Cleaned Abstract Phrases, and Clumped Abstract Phrases for each of the five technology cases.

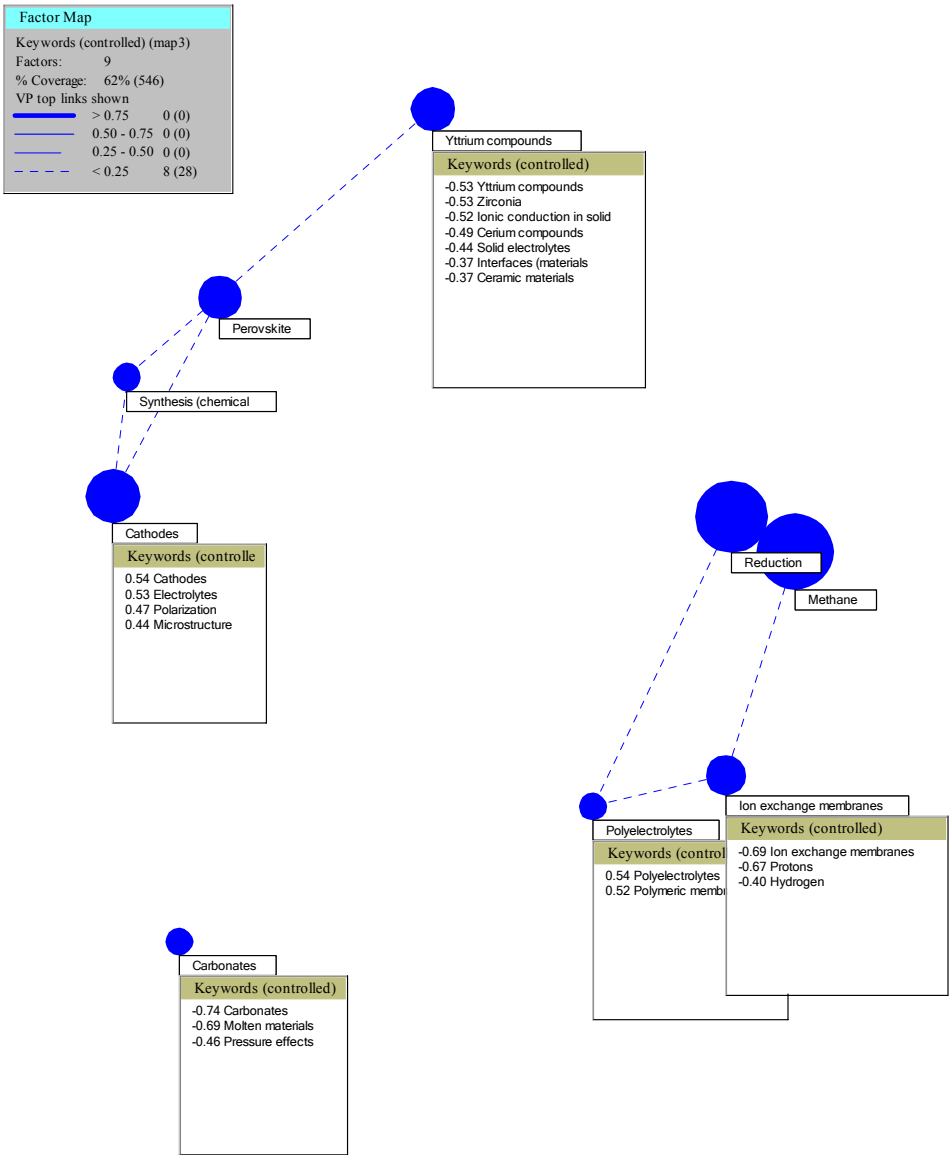


Figure I.1 Fuel Cell Keywords Cluster Maps

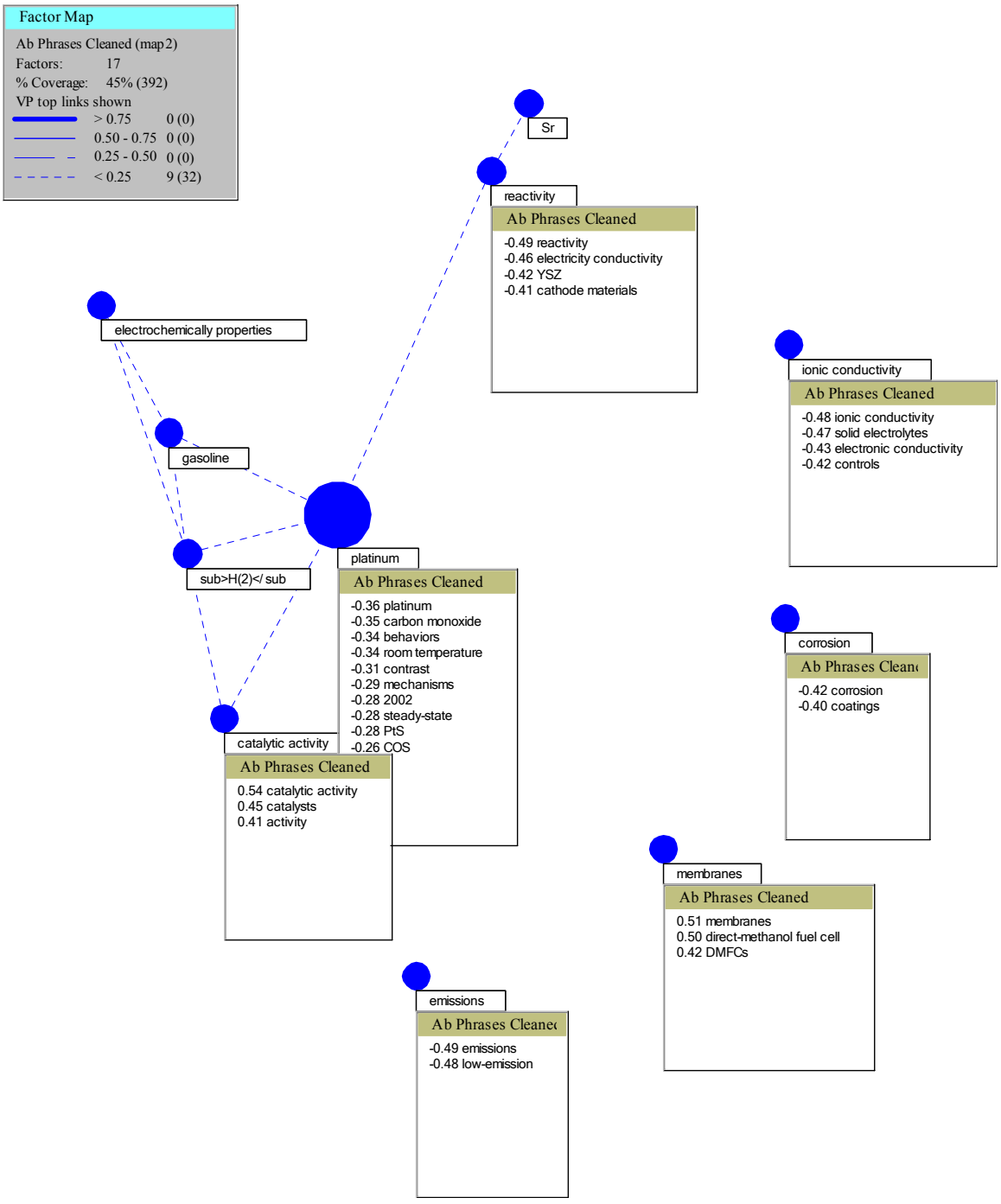


Figure I.2 Fuel Cell Cleaned Abstract Phrases Cluster Maps

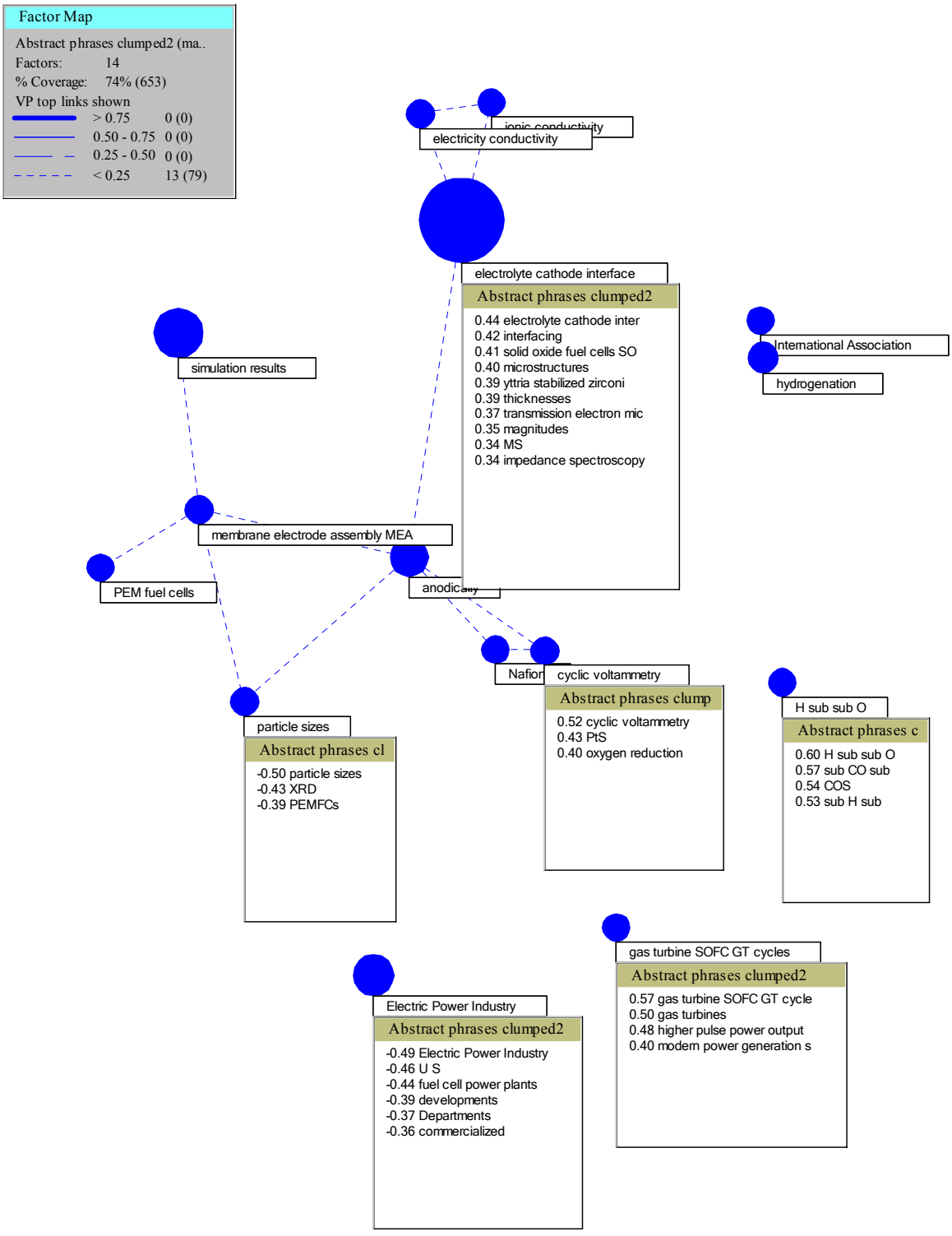


Figure I.3 Fuel Cell Clumped Abstract Phrases Cluster Maps

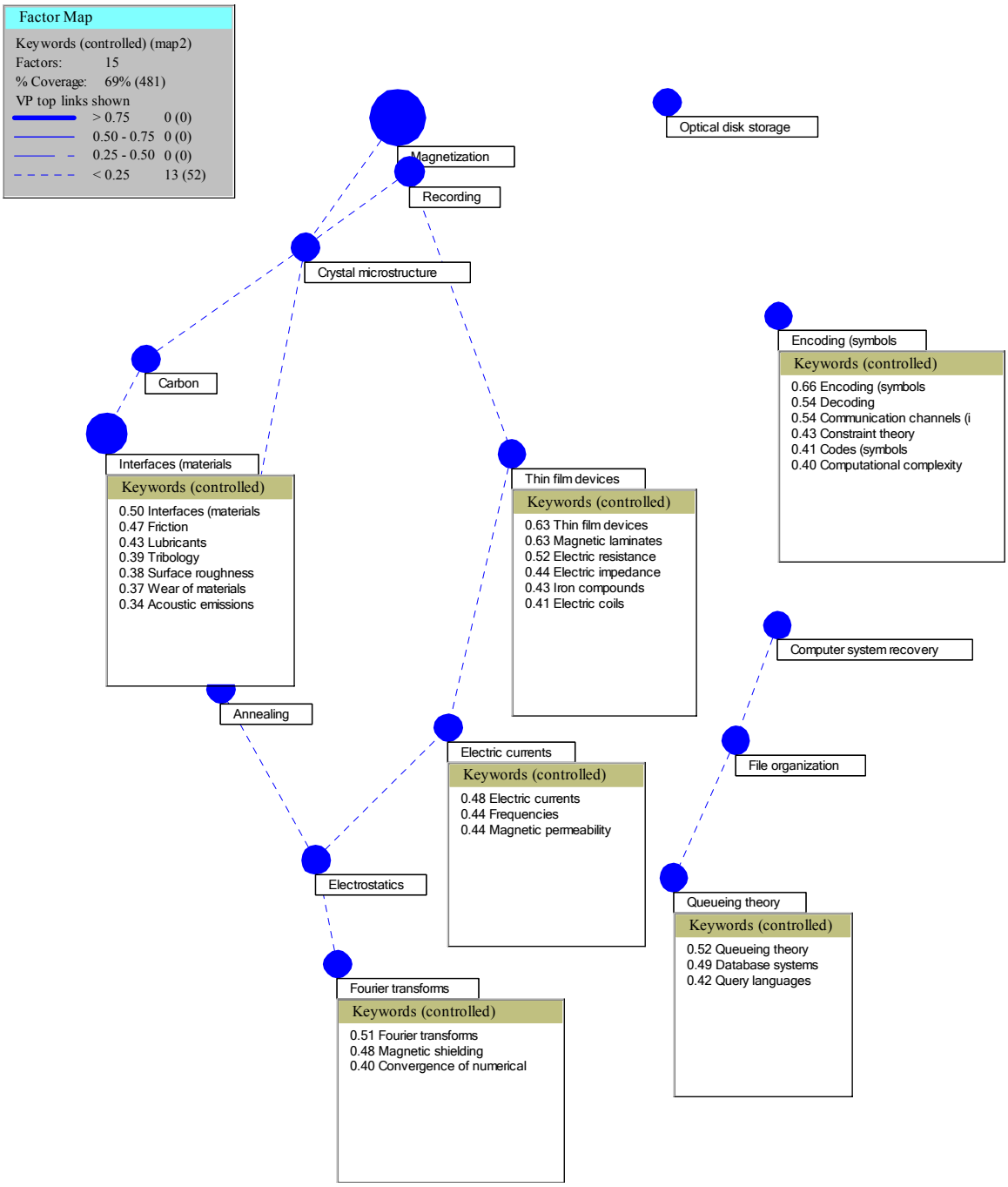


Figure I.4 Magnetic Storage Keywords Cluster Maps

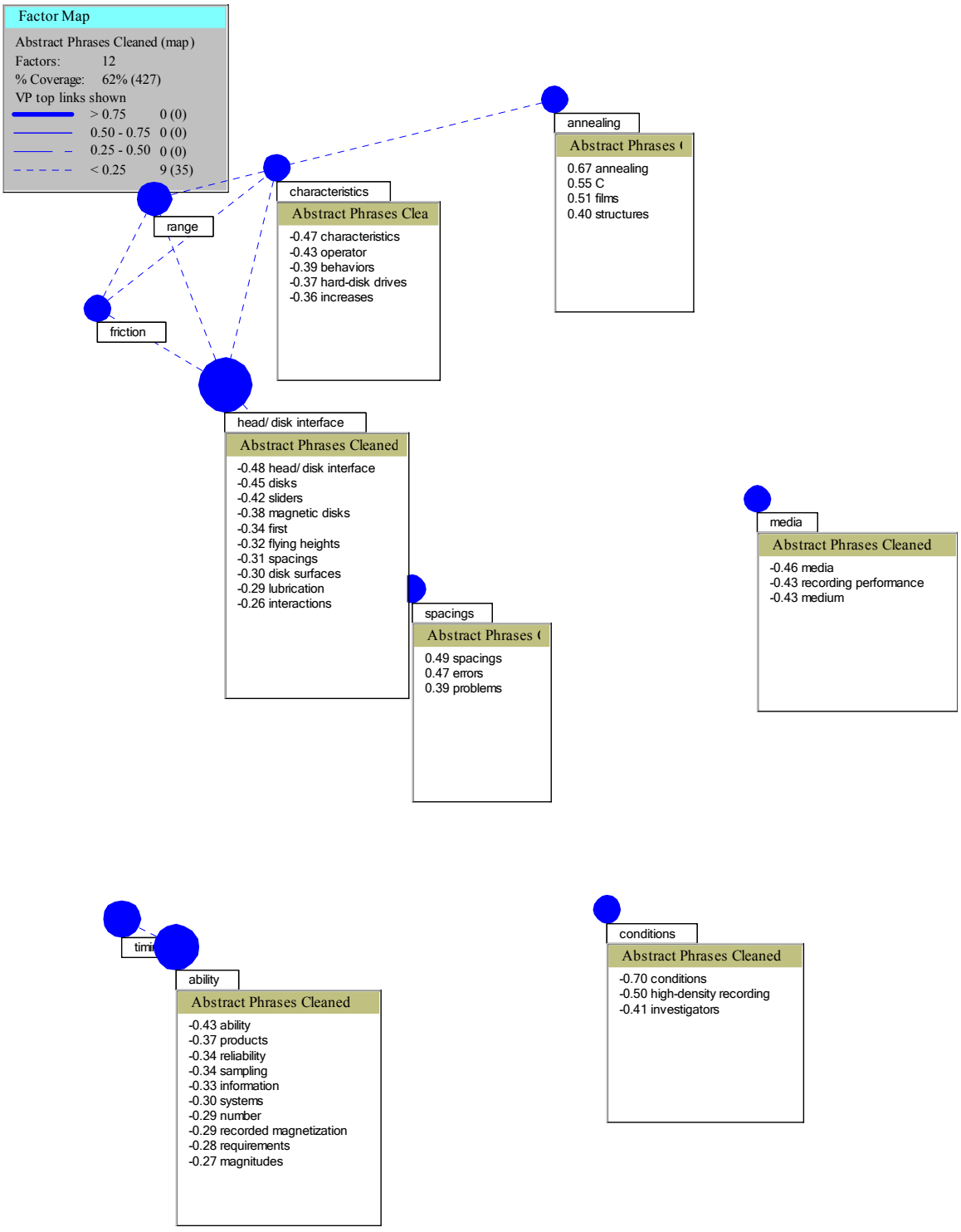


Figure I.5 Magnetic Storage Cleaned Abstract Phrases Cluster Maps

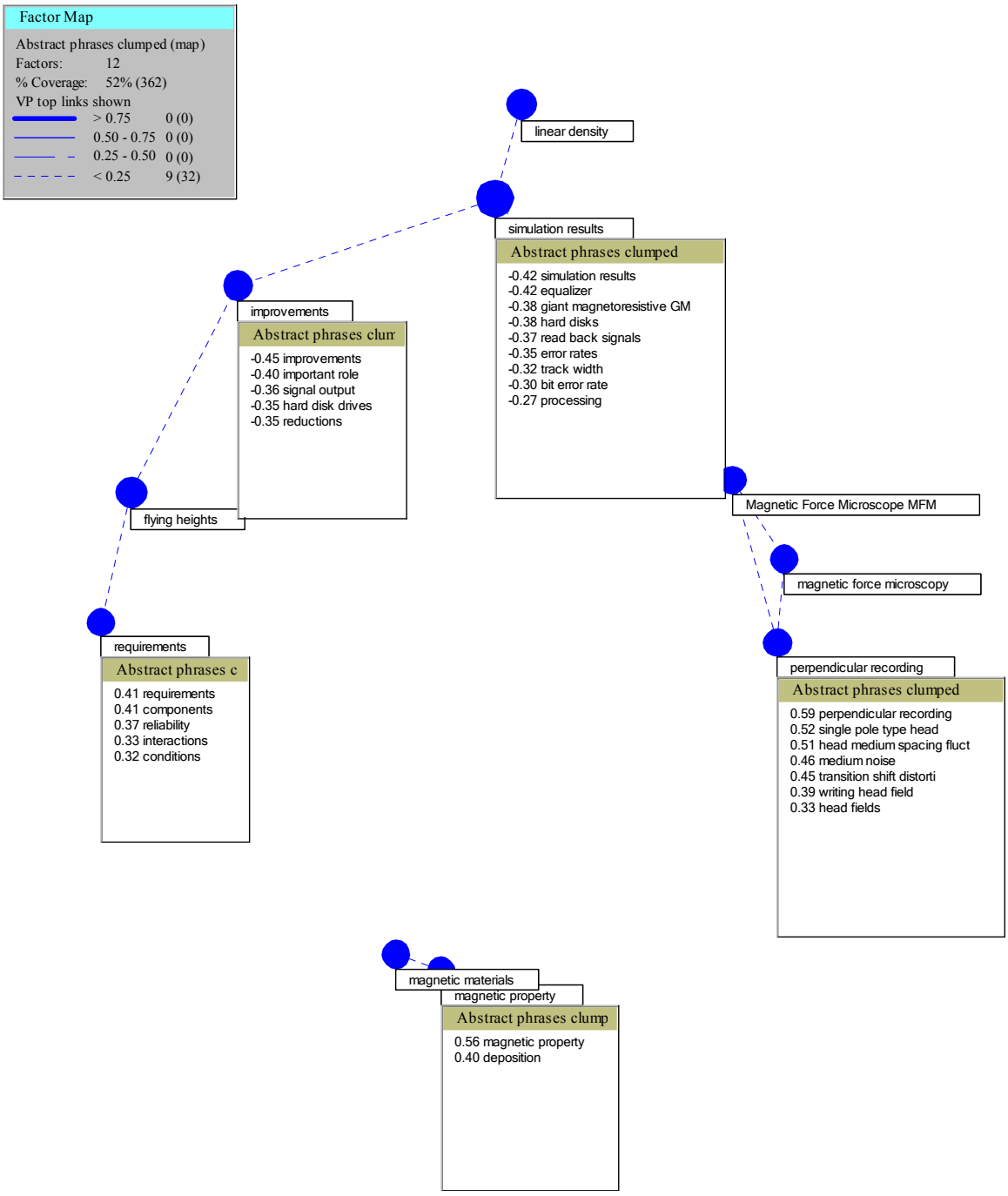


Figure I.6 Magnetic Storage Clumped Abstract Phrases Cluster Maps

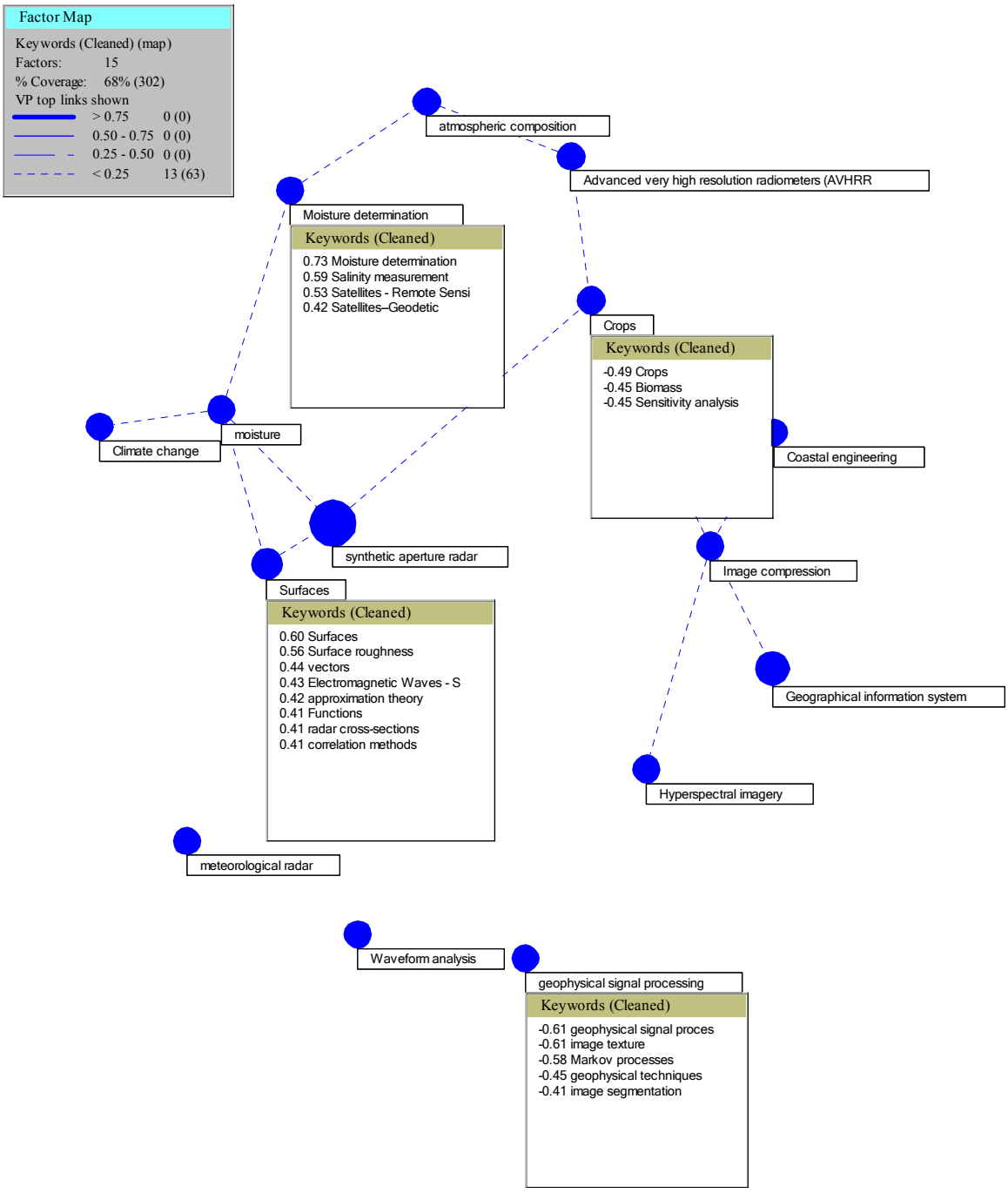


Figure I.7 Remote Sensing Keywords Cluster Maps

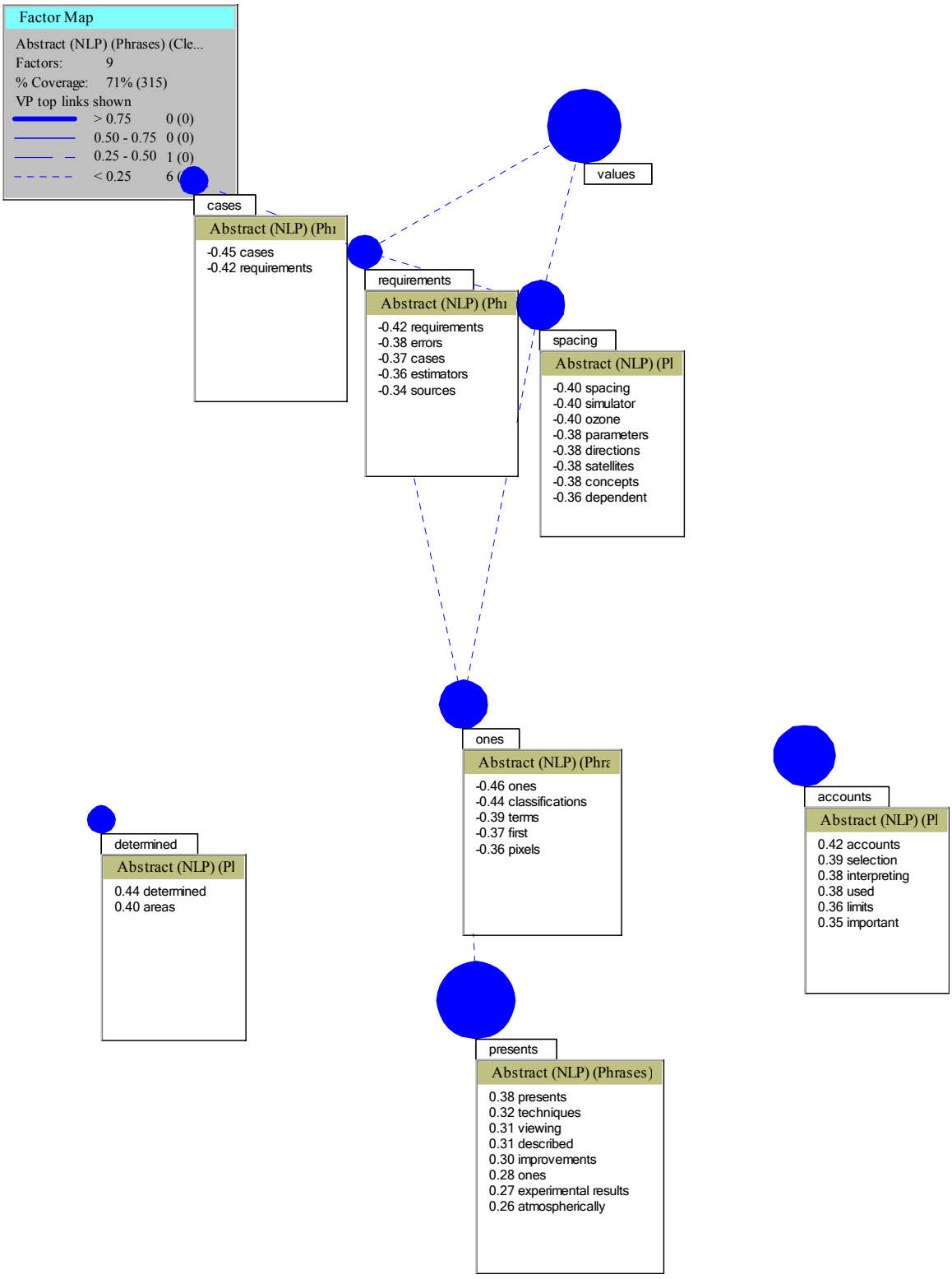


Figure I.8 Remote Sensing Cleaned Abstract Phrases Cluster Maps

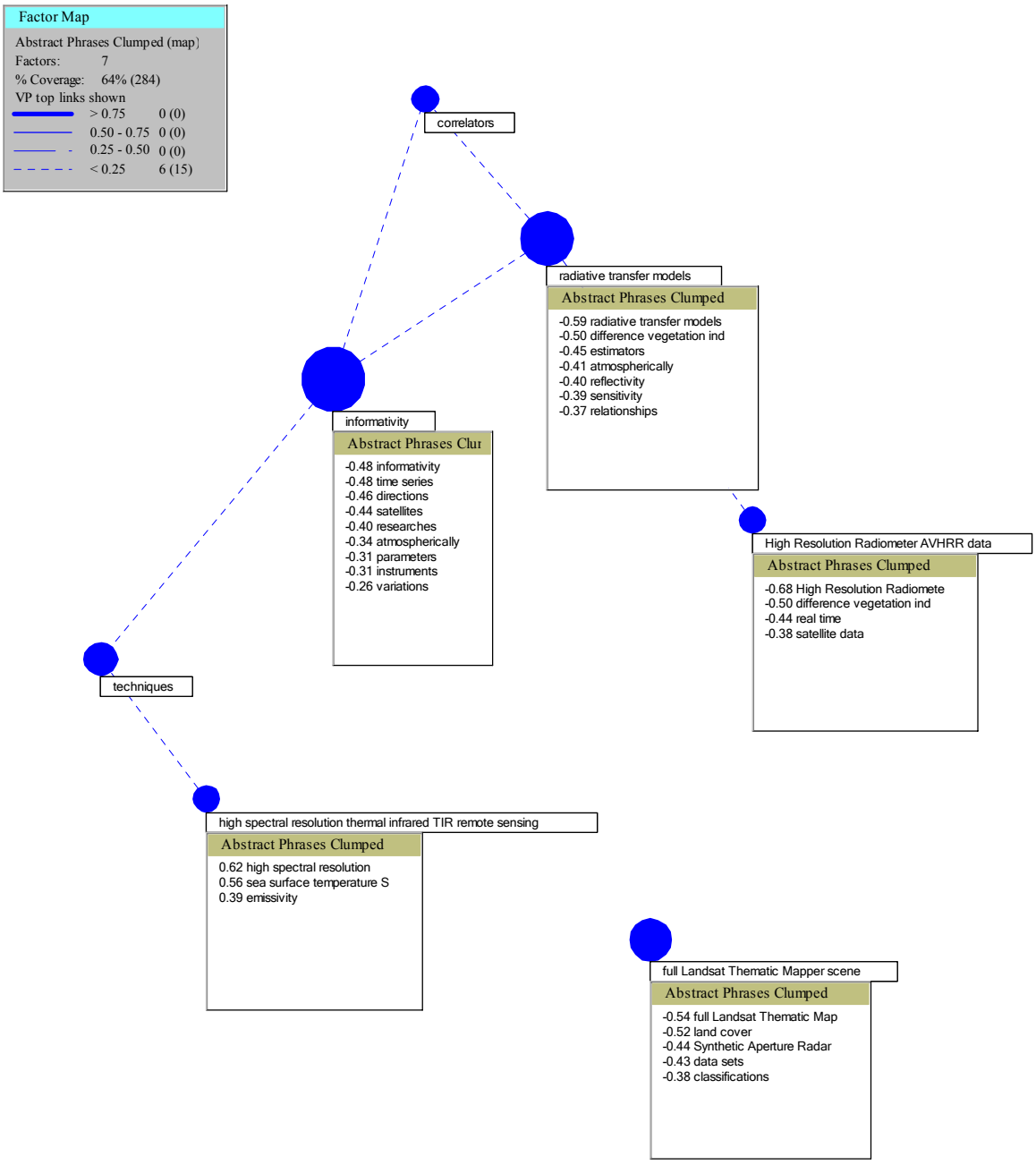


Figure I.9 Remote Sensing Clumped Abstract Phrases Cluster Maps

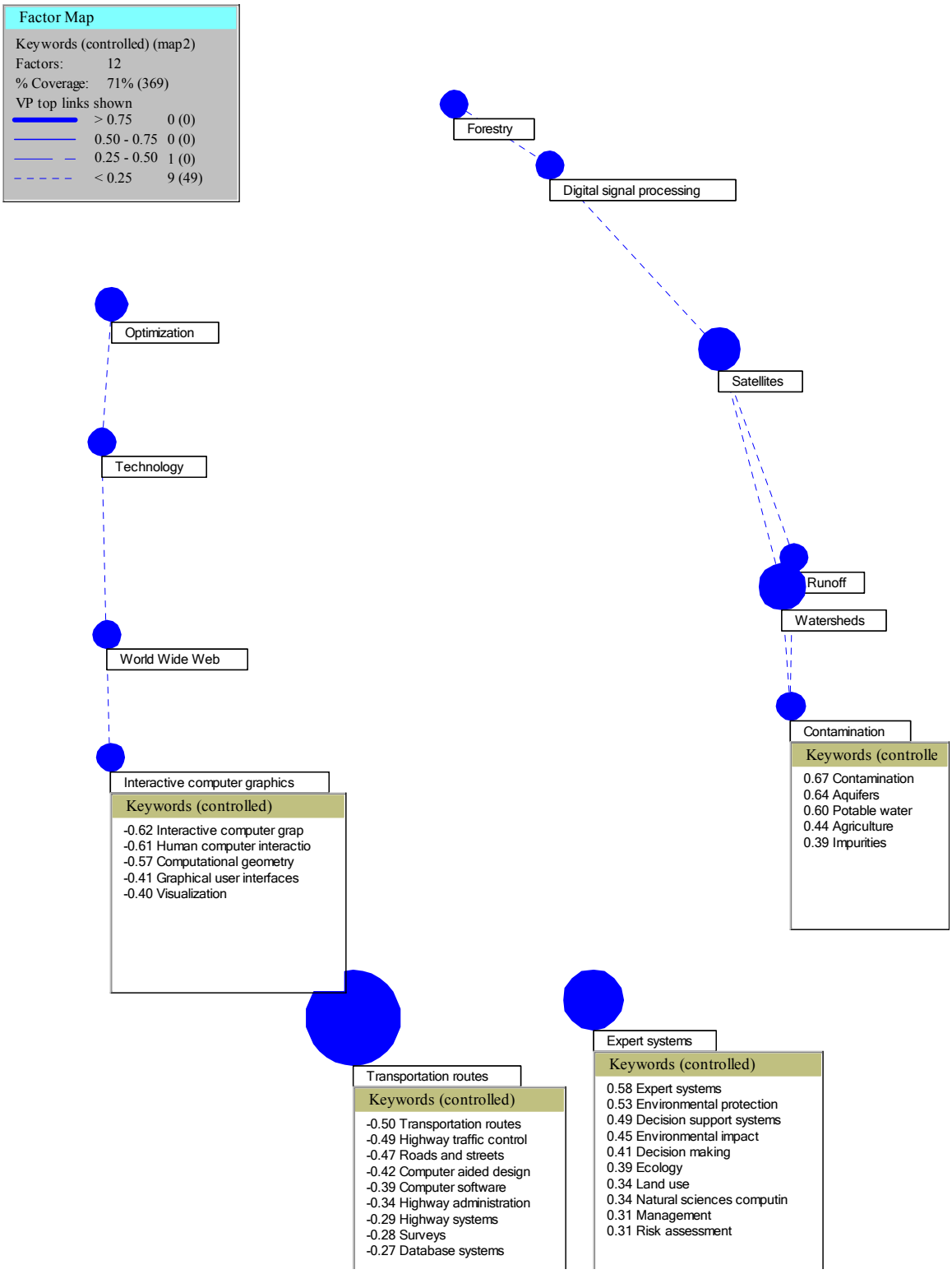


Figure I.10 Geographical Information Systems Keywords Cluster Maps

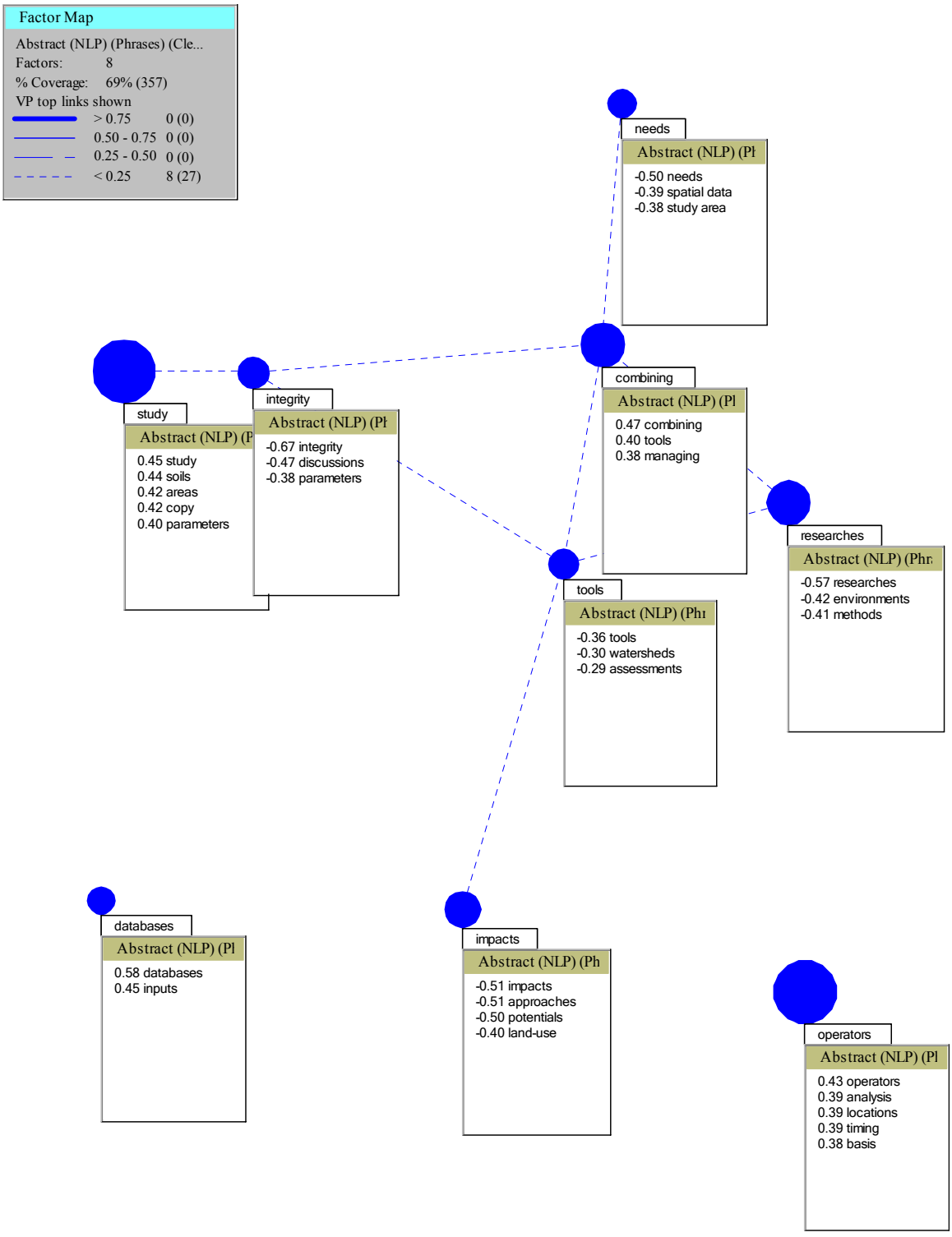


Figure I.11 Geographical Information Systems Cleaned Abstract Phrases Cluster Maps

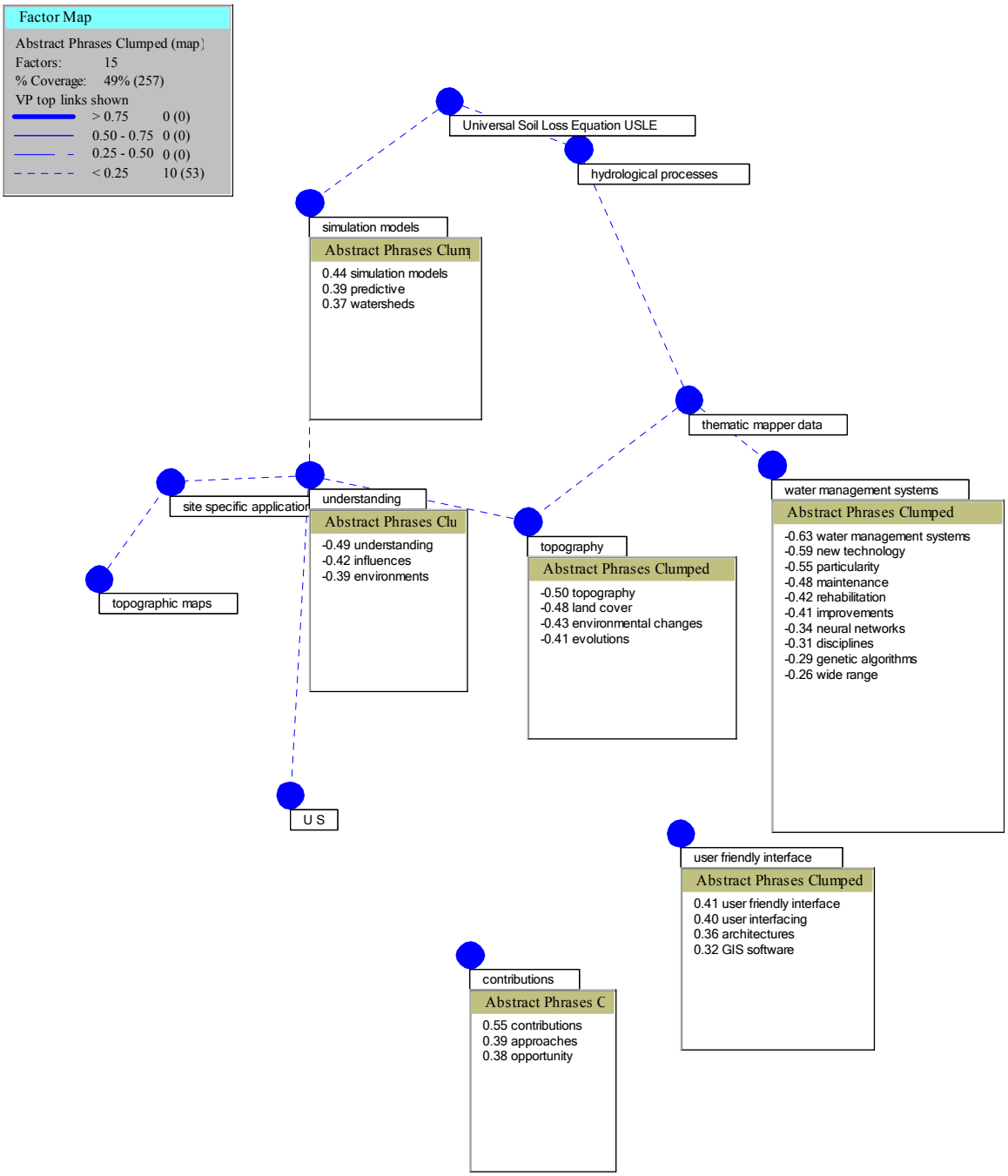


Figure I.12 Geographical Information Systems Clumped Abstract Phrases Cluster Maps

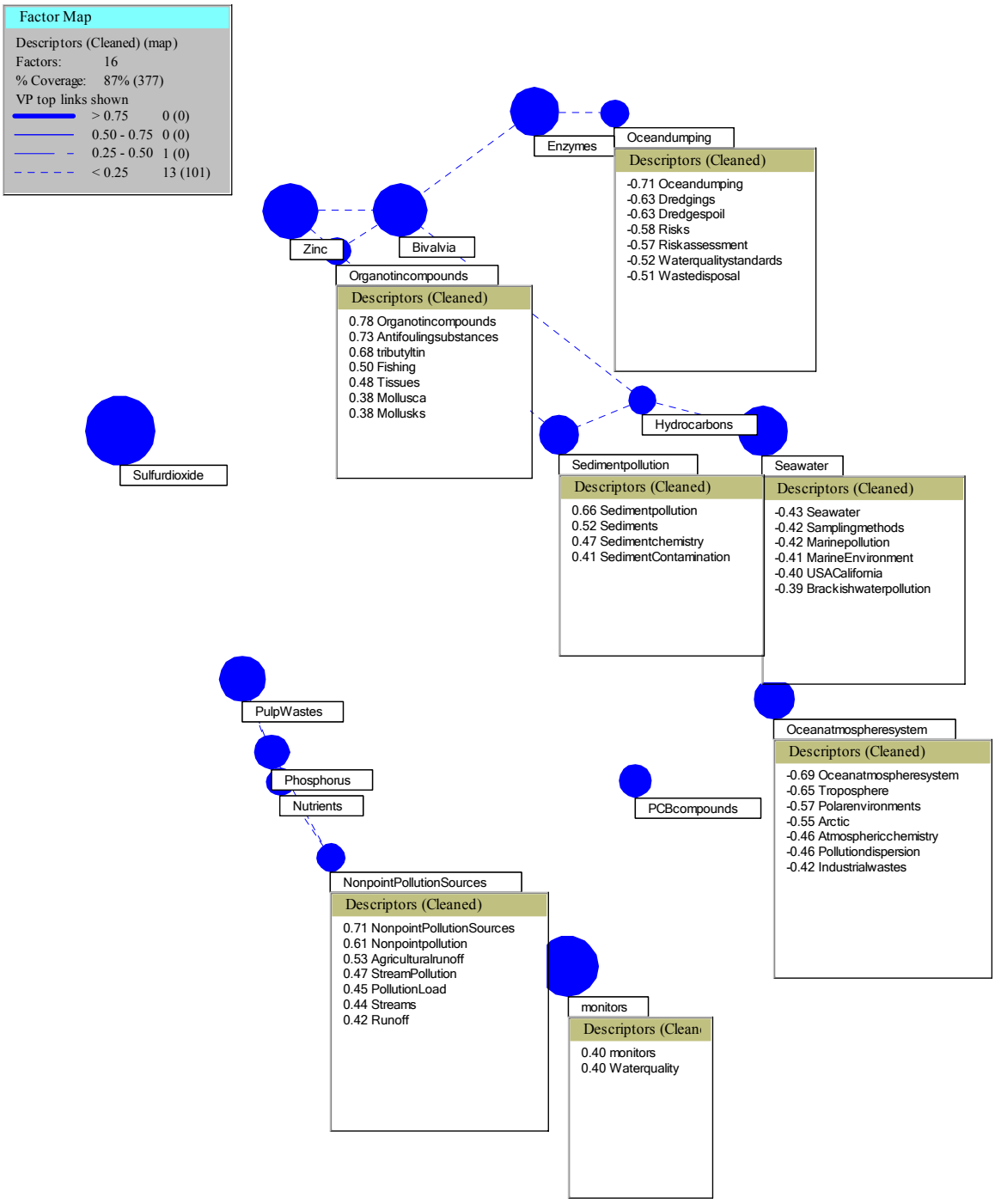


Figure I.13 Pollution Monitoring Keywords Cluster Maps

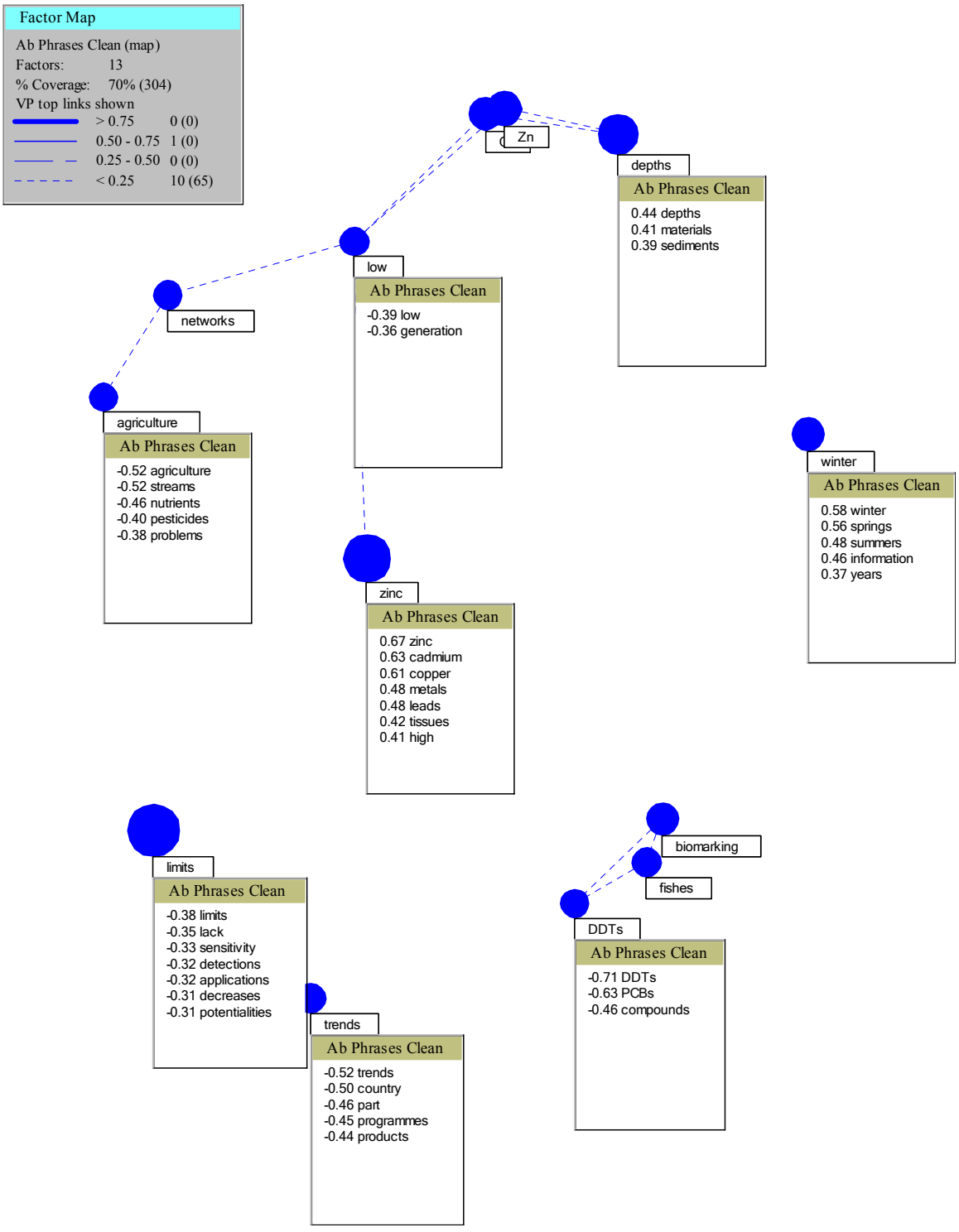


Figure I.14 Pollution Monitoring Cleaned Abstract Phrases Cluster Maps

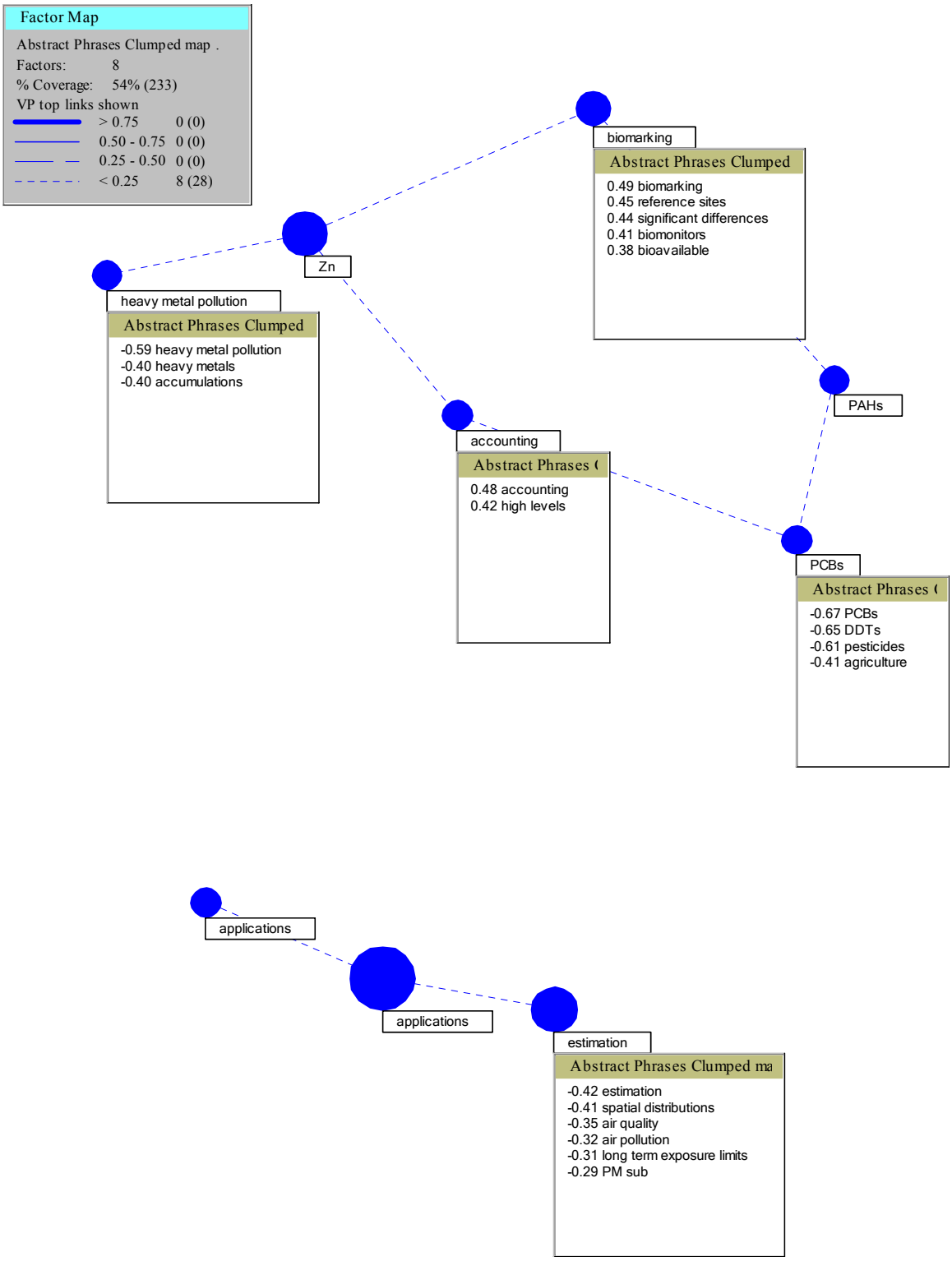


Figure I.15 Pollution Monitoring Clumped Abstract Phrases Cluster Maps

APPENIDIX J: CLUSTER DATA CORRELATION MATRIX

Table J.1: Quantitative Cluster Comparison Data: Correlation Matrix

	terms /doc	#Docs	Total# Terms
Spearman's rho terms /doc Correlation Coefficient	1.000	-.415	.868*'
Sig. (2-tailed)		.124	.000
N	15	15	15
#Docs Correlation Coefficient	-.415	1.000	-.044
Sig. (2-tailed)	.124		.877
N	15	15	15
Total# Terms Correlation Coefficient	.868**	-.044	1.000
Sig. (2-tailed)	.000	.877	
N	15	15	15
# Terms used Correlation Coefficient	-.350	.164	-.107
Sig. (2-tailed)	.201	.560	.704
N	15	15	15
% term Correlation Coefficient	-.854**	107	-.739*'
Sig. (2-tailed)	.000	.458	.002
N	15	15	15
# of links Correlation Coefficient	-.459	-.011	-.332
Sig. (2-tailed)	.085	.969	226
N	15	15	15
# of Clusters Correlation Coefficient	-.390	.033	-.220
Sig. (2-tailed)	.151	.907	.430
N	15	15	15
Avg # terms per Cluster Correlation Coefficient	-.568*	-.065	-.550*
Sig. (2-tailed)	.027	.817	.034
N	15	15	15
# terms(in clusters) Correlation Coefficient	-.661"	.049	-.540*
Sig. (2-tailed)	.007	.862	.038
N	15	15	15
• terms Correlation Coefficient	.171	-.076	-.057
Sig. (2-tailed)	.541	.787	.840
N	15	15	15
• docs Coverage Correlation Coefficient	-.075	-.312	-.165
Sig. (2-tailed)	.790	.258	.557
N	15	15	15
ENTROPY Correlation Coefficient	-.395	.410	-.120
Sig. (2-tailed)	.145	.130	.671
N	15	15	15
COHESION Correlation Coefficient	-.550*	-.076	-486
Sig. (2-tailed)	.034	.787	.066
N	15	15	15

Table J.1 (continued)

	# Terms used	% term	# of links
Spearman's rho terms /doc Correlation Coefficient	-.350	-.854**	-.459
Sig. (2-tailed)	.201	.000	.085
N	15	15	15
#Docs Correlation Coefficient	.164	.207	-.011
Sig. (2-tailed)	.560	.458	.969
N	15	15	15
Total# Terms Correlation Coefficient	-.107	-.739"	-.332
Sig. (2-tailed)	704	.002	.226
N	15	15	15
# Terms used Correlation Coefficient	1.000	.682"	.790
Sig. (2-tailed)		.005	.000
N	15	15	15
% term Correlation Coefficient	.682**	1.000	.705
Sig. (2-tailed)	.005		.003
N	15	15	15
# of links Correlation Coefficient	.790"	.705**	1.000
Sig. (2-tailed)	.000	.003	
N	15	15	15
# of Clusters Correlation Coefficient	.856**	.653**	.981
Sig. (2-tailed)	.000	.008	.000
N	15	15	15
Avg # terms per Cluster Correlation Coefficient	.050	.468	.081
Sig. (2-tailed)	.860	.079	.773
N	15	15	15
# terms(in clusters) Correlation Coefficient	.727*'	.826**	.919
Sig. (2-tailed)	.002	.000	.000
N	15	15	15
• terms Correlation Coefficient	-.950"	-.546*	-.712
Sig. (2-tailed)	.000	.035	.003
N	15	15	15
• docs Coverage Correlation Coefficient	.014	.147	.390
Sig. (2-tailed)	.960	.602	.150
N	15	15	15
ENTROPY Correlation Coefficient	.677"	.570*	.760
Sig. (2-tailed)	.006	.026	.001
N	15	15	15
COHESION Correlation Coefficient	.282	.607*	.399
Sig. (2-tailed)	.308	.016	.140
N	15	15	15

Table J.1 (continued)

	# of Clusters	Avg # terms per Cluster	# terms(in clusters)
Spearman's rho terms /doc Correlation Coefficient	-.390	-.568*	-.661
Sig. (2-tailed)	.151	.027	.007
N	15	15	15
#Docs Correlation Coefficient	.033	-.065	.049
Sig. (2-tailed)	.907	.817	.862
N	15	15	15
Total# Terms Correlation Coefficient	-.220	-.550*	-.540
Sig. (2-tailed)	.430	.034	.038
N	15	15	15
# Terms used Correlation Coefficient	.856"	.050	.727
Sig. (2-tailed)	.000	.860	.002
N	15	15	15
% term Correlation Coefficient	.653"	.468	.826
Sig. (2-tailed)	.008	.079	.000
N	15	15	15
# of links Correlation Coefficient	.981	.081	.919
Sig. (2-tailed)	.000	.773	.000
N	15	15	15
# of Clusters Correlation Coefficient	1.000	-.011	.871
Sig. (2-tailed)		.969	.000
N	15	15	15
Avg # terms per Cluster Correlation Coefficient	-.011	1.000	.416
Sig. (2-tailed)	.969		.123
N	15	15	15
# terms(in clusters) Correlation Coefficient	.871	.416	1.000
Sig. (2-tailed)	.000	.123	
N	15	15	15
• terms Correlation Coefficient	-.791"	.057	-.601
Sig. (2-tailed)	.000	.840	.018
N	15	15	15
• docs Coverage Correlation Coefficient	.288	.297	.381
Sig. (2-tailed)	.298	.282	.161
N	15	15	15
ENTROPY Correlation Coefficient	.762"	-.100	.694
Sig. (2-tailed)	.001	.723	.004
N	15	15	15
COHESION Correlation Coefficient	.350	.286	.409
Sig. (2-tailed)	.201	.302	.130
N	15	15	1

Table J.1 (continued)

	% terms	% docs Coverage
Spearman's rho terms /doc Correlation Coefficient	.171	-.075
Sig. (2-tailed)	.541	.790
N	15	15
#Docs Correlation Coefficient	-.076	-.312
Sig. (2-tailed)	.787	.258
N	15	15
Total# Terms Correlation Coefficient	-.057	-.165
Sig. (2-tailed)	.840	.557
N	15	15
# Terms used Correlation Coefficient	-.950**	.014
Sig. (2-tailed)	.000	.960
N	15	15
% term Correlation Coefficient	-.546*	.147
Sig. (2-tailed)	.035	.602
N	15	15
# of links Correlation Coefficient	-.712"	.390
Sig. (2-tailed)	.003	.150
N	15	15
# of Clusters Correlation Coefficient	-.791"	.288
Sig. (2-tailed)	.000	.298
N	15	15
Avg # terms per Cluster Correlation Coefficient	.057	.297
Sig. (2-tailed)	.840	.282
N	15	15
# terms(in clusters) Correlation Coefficient	-.601*	.381
Sig. (2-tailed)	.018	.161
N	15	15
• terms Correlation Coefficient	1.000	.111
Sig. (2-tailed)		.694
N	15	15
• docs Coverage Correlation Coefficient	.111	1.000
Sig. (2-tailed)	.694	
N	15	15
ENTROPY Correlation Coefficient	-.615*	.137
Sig. (2-tailed)	.015	.626
N	15	15
COHESION Correlation Coefficient	-.154	.387
Sig. (2-tailed)	.585	.154
N	15	15

Table J.1 (continued)

	ENTROPY	COHESION
Spearman's rho terms /doc Correlation Coefficient	-.395	-.550*
Sig. (2-tailed)	.145	.034
N	15	15
#Docs Correlation Coefficient	.410	-.076
Sig. (2-tailed)	.130	.787
N	15	15
Total# Terms Correlation Coefficient	-.120	-.486
Sig. (2-tailed)	.671	.066
N	15	15
# Terms used Correlation Coefficient	.677"	.282
Sig. (2-tailed)	.006	.308
N	15	15
% term Correlation Coefficient	.570*	.607*
Sig. (2-tailed)	.026	.016
N	15	15
# of links Correlation Coefficient	.760**	.399
Sig. (2-tailed)	.001	.140
N	15	15
# of Clusters Correlation Coefficient	.762**	.350
Sig. (2-tailed)	.001	.201
N	15	15
Avg # terms per Cluster Correlation Coefficient	-.100	.286
Sig. (2-tailed)	.723	.302
N	15	15
# terms(in clusters) Correlation Coefficient	.694*'	.409
Sig. (2-tailed)	.004	.130
N	15	15
• terms Correlation Coefficient	-.615*	-.154
Sig. (2-tailed)	.015	.585
N	15	15
• docs Coverage Correlation Coefficient	.137	.387
Sig. (2-tailed)	.626	.154
N	15	15
ENTROPY Correlation Coefficient	1.000	.231
Sig. (2-tailed)		.408
N	15	15
COHESION Correlation Coefficient	.231	1.000
Sig. (2-tailed)	.408	
N	15	15

**Correlation is significant at the .01 level (2-tailed).

*Correlation is significant at the .05 level (2-tailed).

REFERENCES

American Marketing Association (2004) Retrieved on April 5, 2004 from http://www.marketingpower.com/live/mg-dictionary-view3155.php

Abernathy, William J. and James M. Utterback (1978), "Patterns of Industrial Innovation", Technology Review. Cambridge (MA): MIT Press, pp. 40-47. reprinted in Strategic Management of Technology and Innovation. eds. Burgelman, Christensen and Wheelwright (2004) 4th ed. McGraw-Hill

Ahonen-Myka, Helena; Hienonen, Oskari; Klemettinen, Mika. (1999). Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery. Proceedings of the Sixteenth Joint Conference on Artificial Intelligence-IJCAI99. Workshop IRF-3: Text Mining: Foundations, Techniques, and Applications

Anderberg, Michael R. (1973). Cluster Analysis for Applications. Academic Press, New York

Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. 2nd Edition. John Wiley & Sons, New York

Anonymous (2002, April). Dialog Targets Small Firms. Information World Review, Oxford, Iss 179, pg 4

Ashton, W. Bradford, Klavans, Richard A. (eds) (1997). Keeping Abreast of Science and Technology. Battelle Press, Columbus, Richland

Baker, M. P. & Bushell, C. (1995, May), After the storm: considerations for information visualization. Computer Graphics and Applications, IEEE, Vol 15, Issue 3, pp. 12-15

Banks, J.; Mason, T.; Porter, Alan L.; Roper, A.T.; Rossini, F.; Weiderholt.(1991) Forecasting and Management of Technology. Wiley-Interscience

Basu, S.; Mooney, R. J.; Pasupuleti, K. V.; and Ghosh, J. (2001). Evaluating the novelty of text-mined rules using lexical knowledge. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

Beil, F.; Ester, M.; Xu, X. (2002). Frequent Term-Based Text Clustering. Proceedings of the 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD)'2002

Berson, Alex; Smith, Stephen; Thearling, Kurt (2000). Building Data Mining Applications for CRM. McGraw Hill, New York

Bramer, MA (ed.) (1999). Knowledge Discovery and Data Mining. The Institution of Electrical Engineers, London, United Kingdom

Brandau, J. & Young, A. (2000). Small Business CI: Competitive Intelligence in Entrepreneurial and Start Up Businesses. Competitive Intelligence Review, 11(1): 74–84

Breeding, B. (2000). CI and KM Convergence: A Case Study at Shell Services International. Competitive Intelligence Review, 11(4): 12-24

Broder, Andrei; Glassman, Steven C.; Manasse, Mark S.; Zweig, Geoffrey. (1997). Syntactice Clustering of the Web. Selected Papers from the Sixth international conference on the World Wide Web, pp. 1157-1166, Elsevier Science Publishers Ltd, Santa Clara

Bryant, Patrick J.; Coleman, James C.; Krol, Thomas F. (1997) Organizing a Competitive Technical Intelligence Group. Keeping Abreast of Science and Technology; Ashton, W. Bradford, Klavans, Richard A. (eds) Battelle Press, Columbus, Richland

Burkhardt, K. E. Competitive Intelligence and the Product Life Cycle. (2001). Competitive Intelligence Review, Vol. 12(3). 35-43.

Card Stuart K., Mackinlay, Jock D., and Shneiderman, Ben (1999). *Readings in Information Visualization: Using Vision To Think*. Morgan Kaufmann, San Fransisco.

Cawkell, Tony (2001). Progress in Visualization. Journal of Information Science, 27 (6), pp. 427-438

Cetron, Marvin J. and Ralph, Christine A. (1971). Industrial Applications of Technological Forecasting: Its Utilization in R&D Management. Wiley-Interscience, New York

Chen, Kuang-hua and Hisn-Hsi Chen. (1994) Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, p 234--241.

Cheng, Chun Hung; Kumar, Ashkok; Jaideep, Motwani G.; Reisman, Arnold; Madan, Manu S. (1999, February). A Citation Analysis of the Technology Innovation Management Journals. IEEE Transactions on Engineering Management, Vol 46:No1.

Cios, Krzysztof, Pedrycz, Witold, Swiniarski (1998). Roman Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers

Coburn, Mathias M. (1999). Competitive Technical Intelligence: A Guide to Design, Analysis, and Action. Oxford University Press

Comstock, G. L., Sjolseth, D. E. (1999). Aligning and Prioritizing Corporate R&D. Research Technology Management; Washington.

Cook, Thomas D., Campbell Donald T. (1979). Quasi-Experimentation: Design & Analysis Issues for Field Settings Chicago: Rand McNally College Publishing Company

Courseault, Cherie (n.d.) TPAC internal report. Retrieved from www.tpac.gatech.edu

Crestani F. (2000). Exploiting the Similarity of Non-Matching Terms at Retrieval Time, Information Retrieval, vol. 2, nr 1, Kluwer Academic Publishers

Cunningham, Scott (1998). Revolutionary Change in the Electronic Publication of Science. The Information Revolution: Current and Future Consequences, Alan L. Porter and William H. Read (eds) Ablex Publishing Corporation Greenwhich, Connecticut pgs 149-160

Cutting, Douglass R.; Karger, David R.; Pederson, Jan O.; Tukey, John W. (1992) Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. Proceedings of the Fifteenth Annual International ACM SIGIR Conference. Also available as Xerox PARC technical report SSL-92-02

Davison, Leigh. (2001). Measuring Competitive Intelligence Effectiveness: Insights from the Advertising Industry. Competitive Intelligence Review, 12(4):25-38.

Dialog, A Thomson Business (2004) Retrieved March 11, 2004 from www.dialog.com/about

Ding, C. H. (2000, October). A probabilistic model for dimensionality reduction in information retrieval and filtering. In Proc. of 1st SIAM Computational Information Retrieval Workshop

Ding, Chris H. Q. (2003). Document Retrieval and Clustering from Principal Component Analysis to Self-Aggregation. Proceedings from the 9th International Workshop on Artificial Intelligence and Statistics. Key West

Dou, H., Dou, J.-M., Jr. (1999). Innovation Management Technology: Experimental Approach for Small Firms in a Deprived Environment. International Journal of Information Management, 19, 401-412.

Ebecken, Nelson F.F. (1998). Data Mining WIT Press Computational Mechanics Publication, Boston

Eysenbach, G. and Kohler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. British Medical Journal, 324(7337),

Fabrikant, S. I. (2001). Evaluating the Usability of the Scale Metaphor for Querying Semantic Spaces, Spatial Information Theory: Foundations of Geographic Information Science. Conference on Spatial Information Theory, Berlin, Germany, pp. 156-171.

Fabrikant, S. I. (2001). Visualizing Region and Scale in Information Spaces. Proceedings, The 20th International Cartographic Conference, ICC 2001, Beijing, China, Aug. 6-10, pp. 2522-2529

Gershon, Nahum; Eick, Stephen G. (August/September 1997). Information Visualization, Guest Editors Introduction, pp. 29-31.

Gilad, Ben (2000). What Should You Tell Management When They Ask What is the Value of CI? Conference Proceedings: 2000 Annual International Conference & Exhibit for the Society of Competitive Intelligence Professionals, Atlanta, GA

Gottman, John M., Notarius, Cliff. (1978). Sequential Analysis of Observation Data Using Markov Chains. In T.R. Kratochwill (Ed.), Single Subject Research: Strategies for Evaluating Change (pp. 237-285). New York: Academic Press

Halkidi, Maria; Batistakis, Yannis; Vazirgiannis, Michalis, (2002). Clustering validity checking methods: part II. ACM SIGMOD Record, Vol 31, Issue 3, pp. 19-27. ACM Press, New York

Halkidi, Maria; Vazirgiannis, Michalis. (2001). Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. Athens University of Economics & Business, Department of Informatics

Hawkins, Donald (May/June 1999). Information visualization product development, Wilton, Vol 23, Issue 3, pp. 96-98.

Hearst, Marti (1999, June 20-26). Untangling Data Mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, (invited paper).

Herring, Jan. P. (1993). Scientific and Technical Intelligence: The Key to R&D. Journal of Business Strategy 14(3): 10-12

Herring, Jan P. (1998). Creating Successful Scientific and Technical Intelligence Programs. Keeping Abreast of Science and Technology; Ashton, W. Bradford, Klavans, Richard A. (eds) Battelle Press, Columbus, Richland

Hohhoff, Bonnie (1997). Computer Support Systems for Scientific and Technical Intelligence. Keeping Abreast of Science and Technology; Ashton, W. Bradford, Klavans, Richard A. (eds) Battelle Press Columbus Richland

Hot Technologies (n.d.) Retrieved July 20, 2004 from Georgia Tech Technology Policy and Assessment Center website tpac.iac.gatech.edu/hottech/

Institute for Technology Assessment and Systems Analysis (2001). http://www.itas.fzk.de/eng/tadbe/wasist_e.htm

Jantsch, Erich (1967). Technological Forecasting in Perspective Organization for Economic Co-operation and Development, Paris

Kaji, Hiroyuki; Morimoto, Yasutsugu; Aizono, Toshiko; Yamasaki, Noriyuki (1999). Navigation in an Association Thesaurus Automatically Generated from a Corpus. Proceedings of the Sixteenth Joint Conference on Artificial Intelligence-IJCAI99 Workshop IRF-3: Text Mining: Foundations, Techniques, and Applications

Katz, J S; Hicks, D (1997, August) Bibliometric indicators for national systems of innovation. Prepared for IDEA Project funded by the EC TSER Programme. Brighton: SPRU

Keim, Daniel A. (2001, August) Visual Exploration of Large Data Sets. Communications of the ACM, Vol.44, No.8, pp 39-44

Kilmann, Ralph H., Thomas, Kenneth W., Slevin, Dennis P., Nath, Raghu, Jerrell, S. Lee (1994). (eds) Producing Useful Knowledge for Organizations. Jossey-Bass Publishers, San Fransisco

Kinzey, Bruce and Johnson, Anne. (1997). Using Databases to Gather Competitive Intelligence. Keeping Abreast of Science and Technology: Technical Intelligence for Business/ edited by W. Bradford Ashton, Richard A. Klavans. Battelle Press, Columbus, Ohio.

Kosara, Robert; Miksch, Silvia; Hauser, Helwig (Jan/Feb 2002). Focus + Content Taken Literally, IEEE Computer Graphics and Applications, pp 22-39.

Kostoff, Ronald. (1998). Science and Technology Metrics. http://www.dtic.mil/dtic/kostoff/Metweb5_I.htm

Kostoff, R.N. (1997). Database Tomography for Technical Intelligence: Analysis of the Research Impact Assessment Literature. Competitive Intelligence Review. 8(2) 63-79.

Kostoff, R.N., DeMarco, R.A. (2001). Science and Technology Text Mining. Analytical Chemistry. 73: (13). 370-378

Larsen, Bjorner and Aone, Chinatsu (1999), Fast and Effective Text Mining Using Linear-time Document Clustering, KDD-99, San Diego, California

Lee, J., M. Kim, and Y. Lee, (1993) Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. Journal of Documentation,. 49(2): p.188-207

Leouski, Anton V.; Croft, W. Bruce. (1996). An Evaluation of Techniques for Clustering Search Results. University of Massachusetts at Amherst, Computer Science Department, Amherst

Linstone, Harold A and Sahal, Devendra (1976)Technological Substitution: Forecasting Techniques and Applications. Elsevier, New York

Lowden, Barry G. T.**;** Robinson, Jerome (2002) An Analysis of File Space Properties using Clustering. Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics: International Conference on Information Systems, Analysis and Synthesis: Computer Science I. v(5)

Martino, Joseph P. (1972). Technological Forecasting and Decision-Making. American Elsevier Publishing Inc. New York

McCabe, Catherine (2000, July 7). Advancing Information Retrieval through Databases, Fusion, and Information Extraction. Thesis Defense, George Mason University

McDonald, William D. and Richardson, John L. (1997). Designing and Implementing Technology Intelligence Systems. Keeping Abreast of Science and Technology Ashton, W. Bradford, Klavans, Richard A. (eds) Battelle Press Columbus Richland

Merkyl, Dieter & Rauber, Andreas (1999). Uncovering Associations between Documents. Institut fur Softwaretechnik, ARGE Information Retrieval, Technische Universitat Wiem, Wiem

Mirel, Barbara (1998). Visualizations for Data Exploration and Analysis: A Critical Review of Usability Research. Technical Communication. Fourth Quarter 1998, pp 491-509.

Mockler, R J (1992). Strategic Intelligence Systems: Competitive Intelligence systems to support strategic management decision making. SAM Advanced Management Journal, 57(1), Winter, 4-9.

Muller, Adrian; Hamp, Birgit (1999). The Information Traveler - A Graphical Browser for Multi-Dimensional Information Spaces. Proceedings of the Sixteenth Joint Conference on Artificial Intelligence-IJCAI99 Workshop IRF-3: Text Mining: Foundations, Techniques, and Applications

Nasukawa, T.; Nagano, T. (2001). Text Analysis and Knowledge Mining System. IBM Systems Journal, Vol. 40 Issue 4, p967, 18p

Newman, Nils (11/12/1999). Intelligent Information Services Corp.: Interview

Parthasarathy, Sathyan (n.d.) TPAC internal report. Retrieved from www.tpac.gatech.edu

Parker, D. (2000). Can Government CI Bolster Regional Competitiveness? Competitive Intelligence Review, Vol. 11(4). 57-64

Porter, Alan; Newman, Nils; Zhu, Donghua; Courseault, Cherie; Myers, Webb; Yglesias, Elmer (2000) Why Don't Technology Managers Want Our Knowledge. Conference Proceeding of the International Association of Management of Technology (IAMOT), Miami, FL, February 20-25

Porter, Alan L. (2003). Text Mining for Technology Foresight. AC/UNU Millennium Project. Futures Research Methodology. Version 2.0. Editors Jerome C. Glenn and Theodore J. Gordon

Porter, Alan L.; Yglesias, Elmer; Kongthon, Alisa; Courseault, Cherie R.; Newman, Nils C.(n.d.) "TIPing the Scales: Technology Information Products for Competitive Advantage" (accepted) Research Technology Management Retrieved March 11 from Georgia Tech Technology Policy and Assessment Center website. www.tpac.gatech.edu/public_papers/Cal_MOTI-abs.shtml

Porter, Alan L. (2003) Iraqi Engineering: Where has all the Reseach Gone?. Retrieved from Georgia Tech Technology Policy and Assessment Center website. www.searchtech.com/articles/IraqiEngineeringIntro.htm

Prescott, J. E. & D.C. Smith. (1989). A Survey of Leading-Edge Competitor Intelligence Managers. *Planning Review*, (May/June): 6-13

Rajman and Besanon, (1998). *Text Mining - Knowledge Extraction from Unstructured Textual Data*, 6th Conference of International Federation of Classification Societies (IFCS-98), Rome

Rauber, Andreas; Paralic, Jan; Pampalk, Elias. (2000). Empirical Evaluation of Clustering Algorithms. Vienna University of Technology, Department of Software Technology. Technical University of Kosice, Department of Cybernetics and Artificial Intelligence

Raymond, Louis. (2003, September). Globalization, the Knowledge Economy, and Competitiveness: A Business Intelligence Framework for the Development SMEs. Journal of American Academy of Business, Cambridge Hollywood: Vol. 3, Iss. ½; pg 260

R&D Magazine/ Battelle Memorial Institute (2002, January). Smaller Increase Forecast for US Research in 2002, R&D Magazine

Robertson, George G.; Card, Stuart K. ; Mackinlay, Jock D. (1993, April). Information Visualization Using 3D Interactive Animation. Communications of the ACM, 36(4), pp. 57-71

Rossini, Fred; Porter, Alan L.; Newman, Nils (1998, September 10-11). Competitive Technological Intelligence Workshop, Georgia Tech: Distance Learning, Continuing Education, and Outreach, Atlanta, GA

Sebrechts, Marc M.; Cugini, John; Laskowski, Sharon J.; Vasilakis, Joanna; Miller, Michael S. (1999): Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. SIGIR 1999: 3-10

Shah, Chirag. (2002). Automatic Organization of Text Documents in Categories Using Self-Organizing Map (SOM). IEEE's Regional Student Paper Contest

Smith, Daniel C. and Prescott, John E. (1987, September-October). Demystifying Competitive Analysis. Planning Review.

Text mining (n.d.) Retrieved from the SRA International website http://www.sra.com/fs_search.html

Stasko, John; Catrambone, Richard; Guzdial, Mark; McDonald, Kevin (February 2000). An Evaluation of Space-Filling Information Vusualizations for Depicting Herarchical Structures, Technical Report GIT-GVU-00-03, pp. 1-24.

Tellis, Winston (1997). Introduction to the Case Study, The Qualitative Report, 3(2)

Teo, Thompson S. H. (2000). Using the Internet for Competitive Intelligence in Singapore. Competitive Intelligence Review, Vol. 11(2), 61-70.

Tibbetts, Jean. (1997). Technology Scouting. Keeping Abreast of Science and Technology: Technical Intelligence for Business/ edited by W. Bradford Ashton, Richard A. Klavans. Battelle Press, Columbus, Ohio.

Tjaden, Gary S. (1998). Measuring Information Age Business. The Information Revolution: Current and Future Consequences, Ablex Publishing Corporation, Greenwhich, Connecticut, pgs 3-22

Tsuda, Hiroshi, (1999). WIND: Hyper Keyword Index as a Web Document Directory. Proceedings of the Sixteenth Joint Conference on Artificial Intelligence-IJCAI99 Workshop IRF-3: Text Mining: Foundations, Techniques, and Applications

Van Raan, A.F.J. (1992). Advanced Bibliometric Methods to Assess Research Performance and Scientific Development: Basic Principles and Recent Practical Applications. Research Evaluation. 3(3): 151-166

Vedder R.G., Vanecek M.T., Guynes C.S. & Cappel, J.J. (1999). CEO and CIO Perspectives on Competitive Intelligence. Communications of the ACM.

Vinkourov, Alexei; Girolami, Mark. (2000). A Probabillistic Hierarchical Clustering Method of Organizing Collections of Text Documents. University of Paisley, Department of Computing and Information Systems, Computational Intelligence Research Unit, Paisley

Watts, Robert J. and Porter, Alan L.(1997) Innovative Forecasting. Technological Forecasting and Social Change v56 p25-47

Watts, R.J., Porter, A.L., and Newman, N.C. (1998). Innovation Forecasting Using Bibliometrics. Competitive Intelligence Review, Vol. 9, No. 4, p. 11-19

Watts, Robert J., Porter, Alan L., Courseault, Cherie C. (1999). Functional Analysis: Deriving Systems Knowledge from Bibliographic Information Resources. Information. Knowledge. Systems Management

Watts, R., Courseault, C., Kapplin, S. (2000). Identifying Unique Information Using Principal Component Decomposition. Management of Technology: The Key to Prosperity in the Third Millennium Edited by Tarek. Khalil, Louis A. Lefebvre, R.M. Mason, Elsevier Science

Watts, Robert J., Porter, Alan L., Zhu, Donghua (2002). Factor Analysis Optimization: Applied in Natural Language Knowledge Discovery. Proceedings of Codata. 18th International Conference, Montreal, Canada

Wenk, E. and Kuehn, T.J. (1977): Interinstitutional Networks in Technological Delivery System. In: Haberer, J. (ed.): Science and Technology Policy. Lexington Books, Lexington, Massachusetts

What is TOA? (n.d.) Retrieved from July 20, 2004 Georgia Tech Technology Policy and Assessment Center website. www.tpac.gatech.edu/toa.shtml

White, Justin (1999) Word List. Retrieved on March 12, 2004 from http://calendarhome.com/wordlist.html

Wills, G. and Taylor, A. (1969). Technology Forecasting and Corporate Strategy. New York: Elsevier Publisher Record.

Wolfram, Dietmar (2000) Applications of Informetrics to infomration retrieval research. Informing Science Special Issue on Information Science Research Vol 3 No2

Yin, Robert K. (1994). Case Study Research Design and Methods Applied Research Methods Series 2nd Edition (5) Thousand Oaks: Sage Publications

Zamir, Oren & Etzioni, Oren. (1998). Web document clustering: a feasibility demonstration. Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York, pp. 46-54

Zhu, Donghua, Porter, Alan L. (n.d.) Knowledge Discovery in Databases for Data Mining. Technology Policy and Assessment Center Internal Paper. Retrieved from www.tpac.gatech.edu