A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management

A Dissertation
Presented to
The Academic Faculty

By

Alisa Kongthon

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

Georgia Institute of Technology
April 2004

A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management

Approved by:

Professor Alan L. Porter, Advisor	Professor Jye-Chyi Lu
Professor Xiaoming Huo	Professor Susan E. Cozzens
Professor Donghua Zhu	

March 16, 2004

To my parents, Ab and Suneejit Kongthon

ACKNOWLEDGEMENTS

I have had the most remarkable learning and experience during my Ph.D. program at Georgia Tech. First, I would like to thank my advisor Professor Alan L. Porter for his guidance, support, and encouragement on research and life. Thank you for taking me in as a student and giving me the opportunity to pursue the research topic I like. I would also like to thank Professors Jye-Chyi (JC) Lu, Xiaoming Huo, Susan E. Cozzens, and Donghua Zhu for their guidance and feedback while this thesis was being conducted.

This research would not have been possible without the financial support from the Royal Thai Government and the Technology Policy and Assessment Center (TPAC) at Georgia Tech.

I would also like to thank all my friends and colleagues at Georgia Tech and TPAC for sharing many memorable moments. Thanks to all the members of the Thai Student Organization at Georgia Tech for many nice potluck and Karaoke parties. Also, thanks to Devang Dave for his words of advice during our afternoon coffee break. I would like to give a very special thanks to Ralph Mueller for his support and companionship. Thanks for always being here for me. Also thanks to Ralph’s family for their support and warm hospitality during my two visits to Germany.

My family back in Thailand has always provided strong support for me. Finally, I would like to thank my parents, brother, and sister for their support and encouragement without which this work would have been impossible. Thanks for always believing that I could accomplish anything if I put my mind to it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Science and Technology (S&T) information presents a rich resource, essential for managing research and development (R&D) programs. Management of R&D has long been a labor-intensive process, relying extensively on the accumulated knowledge of experts within the organization. Furthermore, the rapid pace of S&T growth has increased the complexity of R&D management significantly. Fortunately, the parallel growth of information and of analytical tools offers the promise of advanced decision aids to support R&D management more effectively. Information retrieval, data mining and other information-based technologies are receiving increased attention.

In this thesis, a framework based on text mining techniques is proposed to discover useful intelligence implicit in large bodies of electronic text sources. This intelligence is a prime requirement for successful R&D management. This research extends the approach called "Technology Opportunities Analysis" (developed by the Technology Policy and Assessment Center, Georgia Institute of Technology, in conjunction with Search Technology, Inc.) to create the proposed framework. The commercialized software, called VantagePoint, is mainly used to perform basic analyses. In addition to utilizing functions in VantagePoint, this thesis also implements a novel text association rule mining algorithm for gathering related concepts among text data. Two algorithms based on text association rule mining are also implemented. The first algorithm called "tree-structured networks" is used to capture important aspects of both parent-child (hierarchical structure) and sibling relations (non-hierarchical structure) among related terms. The second algorithm called "concept-grouping" is used to construct term thesauri for data preprocessing. Finally, the framework is applied to Thai

S&T publication abstracts toward the objective of improving R&D management. The results of the study can help support strategic decision-making on the direction of S&T programs in Thailand.

CHAPTER 1

INTRODUCTION

1.1 Background

Science and Technology (S&T) are the major driving force of modern economies. Global S&T expenditures are approximately $500-800 billion annually, depending on one's definition of S&T (Kostoff, 2000). Science and Technology information presents a rich resource, vital for managing research and development (R&D) programs. Hence, cooperative S&T development efforts are required, if an organization or nation is to remain competitive. Management of R&D has long been a labor-intensive process, relying extensively on the accumulated knowledge of experts within the organization. Furthermore, the rapid pace of S&T growth and globalization has increased the complexity of R&D management significantly. Fortunately, the parallel growth of information and the cutting edge in computer and information technology offer the promise of advanced decision aids to support R&D management much more rapidly and with increasing robustness than in the past. Information retrieval, data mining and other information-based technologies are receiving increased attention. These effective tools should become an integral part of the organization's practices, but at the same time, the knowledge of subject matter experts is still needed to combine with the results of these information-based technologies (Porter, 2000). In other words, for an organization to manage technology well, derived knowledge (i.e., empirically based information such as results of data mining analyses) should be complemented with tacit knowledge (i.e., accumulated knowledge through experiences).

Various empirical methods are used to inform R&D management over the years. Notable examples are:

- Bibliometrics
- Technology forecasting
- Technology Assessment
- Competitive Technical Intelligence

The fields of bibliometrics and content analysis have offered ways to analyze large amounts of S&T information resources over the years, but acceptance has been limited. Recently, bibliometrics activity transitions into "text mining" – exploring the content of abstracts or full-text document sets. Such analytical approaches can aid management of technology and competitive technological intelligence in many ways. For instance, they can help researchers in mapping and profiling their whole research domain (Porter et al., 2002a, Borner et al., 2003) and technology managers in providing foresight analysis for their target technologies (Porter and Detampel, 1995, Watts and Porter, 1997).

In this thesis, a framework based on text mining techniques is presented in order to gather technological intelligence to support R&D management. This involves monitoring and analyzing (inter)relations between research domains based on scientific publications. However, the volume of scientific papers is growing exponentially (Cunningham, 1998). Fortunately, with the advance in speed and storage of computer technology today, this massive information can be gathered easily. For example, DIALOG, the leader in online-based information retrieval service, provides more than 12 terabytes of content (more than 800 million unique records within 900 different databases) from the most authoritative publishers (from http://www.dialog.com). Clearly, it is beyond the ability of any person or group to comprehend all of this information by digesting individual pieces. As a result, policy makers are in need of additional automatic tools to support the decision-making process.

1.2 Problem Description

The underlying motivation driving the research is to create a text mining framework that can extract technological intelligence from electronic text sources. This knowledge is a prime requirement for successful technology management. This text mining framework can help:

- identify technology infrastructure (i.e., experts and centers of excellence that are the main players in R&D concerning a target technology)
- discover overlapping or similar research activities among those centers
- identify and categorize the main research areas and sub-areas in a large body of technical literature
- construct innovation indicators (Watts and Porter, 1997)
- identify emerging technologies from related or disparate technical literature.

This research will refine and extend the approach called “Technology Opportunities Analysis” (TOA) developed by the Technology Policy and Assessment Center (TPAC), Georgia Institute of Technology, in conjunction with Search Technology, Inc., to create the proposed text mining framework. The commercialized

Windows-based software called VantagePoint, which combines bibliometrics and content analysis (Watts et al., 1997, also http://thevantagepoint.com) is mainly used to perform the analyses.

In addition to employing the basic profiling functions of VantagePoint, this research will focus on the text clustering process. One of the key contributions of this dissertation is the implementation of a new text clustering method. Currently VantagePoint performs multidimensional statistical analysis to identify clusters and relationships among concepts. However, its approach does not produce sets of hierarchically related features. As a result, two new algorithms, based on association rule mining are implemented. The first algorithm called “tree-structured networks” is used to capture important aspects of both parent-child (hierarchical structure) and sibling relations (non-hierarchical structure) among related terms. The second algorithm called “concept-grouping” is used to construct term thesauri for data preprocessing.

The framework is then applied to Thai S&T publication abstracts toward the objective of improving research management. The results provide the overview of research domains in Thailand. Analysis at the micro (discipline) level to identify the strongest research areas is presented. To illustrate the use of research profiling to support S&T management, a comparison between research publication profiles and export-oriented industrial activity for Thailand is made to identify the linkage between the research community and industry. The approach and results are expected to improve strategic decision-making processes for Thai technology managers and policy planners.

1.3 Organization of the Thesis

The organization of the thesis is as follows. In Chapter 2, the literature in the area of bibliometrics and text mining is reviewed. This overviews the current state of the art. Chapter 3 describes details of the proposed text mining framework. In Chapter 4, clustering techniques used in VantagePoint are presented. Chapter 5 presents a novel text association rule mining algorithm for generating related term clusters. Chapter 6 presents a tree-structured network algorithm that captures important aspects of both parent-child hierarchies (trees) and sibling relations (networks) among term clusters. Experimental results illustrate potential application of the algorithm. Comparisons to existing approaches are also presented. Chapter 7 presents the use of the association rule mining technique to construct abstract or title phrases thesaurus for data preprocessing. Phrase-clustering with and without the proposed technique is compared. In Chapter 8, the proposed framework is applied to Thai S&T publication abstracts. The target user, technologies to be monitored and the data are presented. The comparison between Thai research publication profiles and export-oriented industrial activity is discussed. Observation for the Thai R&D managers or policy planners to improve strategic decision making processes are presented. Finally, Chapter 9 presents the thesis conclusions and future research opportunities.

CHAPTER 2

LITERATURE REVIEW

The fields of bibliometrics and content analysis have been used to inform S&T management over the years. Recently bibliometric activity transitions into "text mining"—exploring content of abstracts or full-text document sets. Losiewicz et al. (2000) suggest that this approach treats retrieved texts as data, from which it seeks to discover patterns. The first part of this literature review discusses bibliometric analysis and related approaches for S&T study. The second part covers text mining and its techniques.

2.1 Bibliometrics

Bibliometrics is the study that uses statistical and mathematical methods to analyze the literature of a discipline as it is patterned in its bibliographies.[1] It uses counts of publications, patents, and citations to develop S&T performance indicators. These indicators are used to measure research outputs (Narin and Olivastro, 1994). In other words, the number of publications or citations generated by a research unit may be used to judge the productivity, or output, of that research group. Thus, bibliometrics fundamentally serves as a proxy or approximation of the outputs of R&D (Melkers, 1993).

Formal bibliometric analysis originated in the 1960's. Derek de Solla Price and Eugene Garfield were leaders in the movement to develop bibliometric indicators (Price,

[1] http://alexia.lis.uiuc.edu/~standrfr/beginner.html

1963, Garfield, et. al., 1964). Henry Small helped to refine the method with the development of co-citation analysis (Small, 1974). The invention of the Science Citation Index (SCI) in 1961 allowed bibliometric analysis to become more systematized (Garfield, 1979). The SCI was multidisciplinary, covering virtually all disciplines and fields of science. Prior to the development of SCI, publication counts and citation analysis were conducted manually. With computerized databases, researchers could quickly perform the analyses on larger and more complex data sets.

The first and most basic principle behind bibliometric analysis is activity measurement. This involves the counting of scientific publications or patents published by a researcher or a research group. It is the most basic of bibliometric analysis. Publication or patent counts are most useful for providing a measure of total research output. However, the drawback is that they do not recognize the quality of these outputs.

The second principle is impact measurement. This entails the number of times those patents or articles are cited in subsequent patents or articles. The citation counts provide valid indicators of the impact or importance of the cited patents and articles. The underlined assumption is that the more important or influential a work, the more often it will be cited. While the publication counts measure the quantity of research output, citation counts address questions of influence, and knowledge transfer.

The third principle is linkage measurement. This involves the co-citations or co-words used from articles to articles or from patents to patents. This measurement provides indicators of linkage among the organizations that are producing the patents and articles, and knowledge linkage among their subject areas. Small (1973) introduced the concept of co-citation analysis and defined it as “the frequency with which two items of earlier literature are cited together by the later literature." Co-citation analysis has been successfully applied to identify cognitive structure of many disciplines. This cognitive structure provides information on the direction and flow of scientific thought. Co-citation analysis helps monitor how different sub-domains of science change and evolve over time. Small and Greenlee (1989) applied co-citation analysis to ascertain the growth and diversification of research topics in AIDS research from 1981 to 1987. While co-citation analysis provides information on the direction and flow of research, it does not provide an immediate picture of actual research content within the literature. Co-word analysis, on the other hand, has the potential to address this issue. Co-word analysis uses the co-occurrence of keywords in the publications on a given subject to discern relationships among documents (Callon et al., 1983). If certain terms tend to appear together in documents, this is taken as evidence of possible relationship. Kostoff et al. (2001) presented a system called database tomography. It includes algorithm for extracting phrases frequencies and phrases proximities from any type of large text database in conjunction with expert opinion's interpretations to convert large volumes of disorganized data to meaningful information. Narin and Olivastro (1988) applied patent citations analysis to indicate linkages between companies, between technological areas, and between technology and science.

Strongly tied to co-citation and co-word analyses, scientific mapping involves developing a visual model of the domain of science under study. Mapping science is an attempt to arrive at a physical representation of fields and disciplines in which the relative locations of entities are depicted. Noyons and van Raan (1998) have contributed greatly to the notion of scientific mapping. Mapping of S&T aims to help researchers manage the vast number of scientific articles published every year in their fields, in order to keep track of the developments as a whole.

Although co-word and co-citation analyses appear to be very useful methods, there are limitations in publication and citation counts (Kostoff, 1997). Problems with publication counts include:

- publications indicate quantity of output, not quality
- publication patterns vary significantly among research fields and journals
- undesirable publishing practices may increase (e.g., synthetically increased numbers of co-authors)
- very few of the active researchers are producing the heavily cited papers

Citation counts can address more than the quantity of output, but they do not equate precisely to quality. The problems with citation counts include:

- intellectual link between citing source and reference article may not always exist
- incorrect work may be highly cited
- methodological papers tend to accrue more citations than substantive articles
- self-citation may synthetically increase citation rates

2.2 Text Mining

“Data Mining” involves the integration of concepts from computer science, mathematics, and statistics. It seeks to extract useful information and detect interesting correlation and patterns from any form of data, especially numeric data. Data Mining is most associated with the broader process of Knowledge Discovery in Databases (KDD), “the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data” (Fayyad et al., 1996). By analogy, this paper considers "text mining" as the process that exploits large text collections to obtain valid, potentially useful and ultimately understandable knowledge.

There are approximately five major technique categories in the text mining process: document retrieval, data extraction, data preprocessing (cleansing), data analysis, and data visualization (Losiewicz et al., 2000).

2.2.1 Document Retrieval

Information (or document) retrieval is a discipline concerned with the organizing, storage, searching, and retrieval of bibliographic information. Salton and McGill (1983) introduce the idea of the Vector Space Model (VSM) -- a vector, comprised of the keywords contained within the document, can represent a document. It is a powerful framework for analyzing and structuring documents. VSM model procedures can be divided into three stages: document indexing, term weighting, and computation of similarity coefficients.

In document indexing, each document is represented as a vector in a high dimensional space. It is obvious that many of the words in a document do not describe the content, words like “the” or “is.” By using a stopword list, consisting of common words, those non-significant words are removed from the document vector, so the document will only be represented by content-bearing words. In general, 40-50 percent of the total number of words in a document are removed with the help of a stop list (Salton, 1983).

The second stage is the weighting of the indexed terms to enhance retrieval of documents relevant to the user. Terms are weighted to indicate their importance for document representation. Most of the weighting designs, such as inverse document frequency, assume that the importance of a term is proportional to the number of documents the term appears in. Moreover, document length normalization is required since long documents usually have much larger term sets than do short documents. This makes long documents more likely to be retrieved than short documents.

The last stage ranks the document with respect to the query according to a similarity measure. The similarity between any two documents can be determined by the distance between document vectors in a high-dimensional space. Overlapping terms between the two vectors illustrate similarity. The most popular similarity measure is the cosine coefficient, which measures the angle between two document vectors. It is equal to the inner product of the two vectors, normalized (divided) by the products of the vector lengths (square root of the sums of squares.) Other measures are, for example, Jaccard and Dice coefficients (Salton, 1988).

The vector space model provides an easy way to assess document similarities based on word matches. However, the problem with polysemy (that is, many words have more than one distinct meaning) cannot be detected. This requires methods that examine the "semantic structure" of term-document association data such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990).

2.2.2 Data Extraction

Data extraction is the activity of automatically pulling out pertinent information from large volumes of texts. Extraction can take two forms; one is to identify the specific field of entity extracted such as name, date, or address, and the other one is to identify the parts of speech from text corpus using natural language processing (NLP) technology.

VantagePoint applies NLP to parse text into the part(s) of speech. It employs a combination of semantic and syntactic analyses. It processes text inputs as follows (TPAC, 1998):

- Distinguishes and separates each sentence
- Applies lexicon analysis to categorize nouns, verbs, etc., based on the underlying dictionary
- Refines word attribution based on syntactic inferences.

This then tags each word with the part(s) of speech it is likely to be.

2.2.3 Data Preprocessing

Data preprocessing, or data cleansing, is the algorithm that detects and removes errors or inconsistencies from data and consolidates similar data in order to improve the quality of subsequent analyses. This cleaned data will then be fed to the analysis process. Several methods could be used to clean the data. Three methods that are used in VantagePoint's list cleanup are stemming algorithm, elemental fuzzy logic to consolidate like terms, and thesauri. Word stemming or truncation can be used to achieve a quick approximation to the word root. A word for which one wants to find an exact or near match may by written as a stem or root word, and the retrieval system asked to find words that match the root. One approach used to determine the root of a word is to determine the semantic root such that “box” and “boxes” are equal. Fuzzy matching techniques can be used to identify, associate, and reduce data appropriately. For example, this will handle misspellings, alternative hyphenation and capitalization. A thesaurus is defined as a grouping of terms, into a certain concepts. This can be used for specialized data reduction.

2.2.4 Data Analysis

As stated before, each document can be represented as a vector in a high dimensional space. Hence, dimensionality reduction techniques are required to represent *n*-dimensional document data by a small number of significant dimensions. There are several techniques that have been used for dimensionality reduction, including Factor Analysis (FA)/Principal Component Analysis (PCA) and Cluster Analysis.

2.2.4.1 Factor Analysis/Principal Components Analysis

Factor Analysis is a statistical approach that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors) (Hair, 1992). A basic method in factor analysis is Principal Components Analysis (PCA), which projects the vectors into a lower-dimension linear space, and treats the projection, called a principal component, as a method of data reduction. The first principal component accounts for as much of the variability in the data as possible. Each succeeding component accounts for as much of the remaining variability as possible. An advantage of using factor analysis over traditional clustering techniques in text analysis is that it does not force each term/word into only one cluster. Words can be classified in multiple factors; unlike Hierarchical Agglomerative Clustering and K-means clustering where each term is only permitted to appear in one cluster. This property of factor analysis is very important for research domain study since research works now are multidisciplinary. The VantagePoint software uses factor analysis/principal component analysis to construct technology maps. The details of factor analysis will be presented in Chapter 4.

2.2.4.2 Cluster Analysis

Another technique that has been used in mining text data is cluster analysis. Cluster analysis, also called data segmentation, is an approach that groups or segments a collection of data items (e.g., documents or terms) into subsets or "clusters," such that those within each cluster are more closely related to one another than objects assigned to different clusters. Callon et al. (1983) use cluster analysis as a means of analyzing science. Schvaneveldt (1990) applies the pathfinder network approach to analyze the organization of knowledge within texts. Pathfinder network scaling is a structural and procedural modeling technique, which extracts underlying patterns in proximity data and represents them spatially in a class of networks. Pathfinder algorithms take estimates of proximities between pairs of items as input and define a network representation of the items that preserves only the most important links. The techniques of cluster analysis could be classified into two categories: hierarchical and partitional. The following figure shows a taxonomy representation of clustering methodologies.

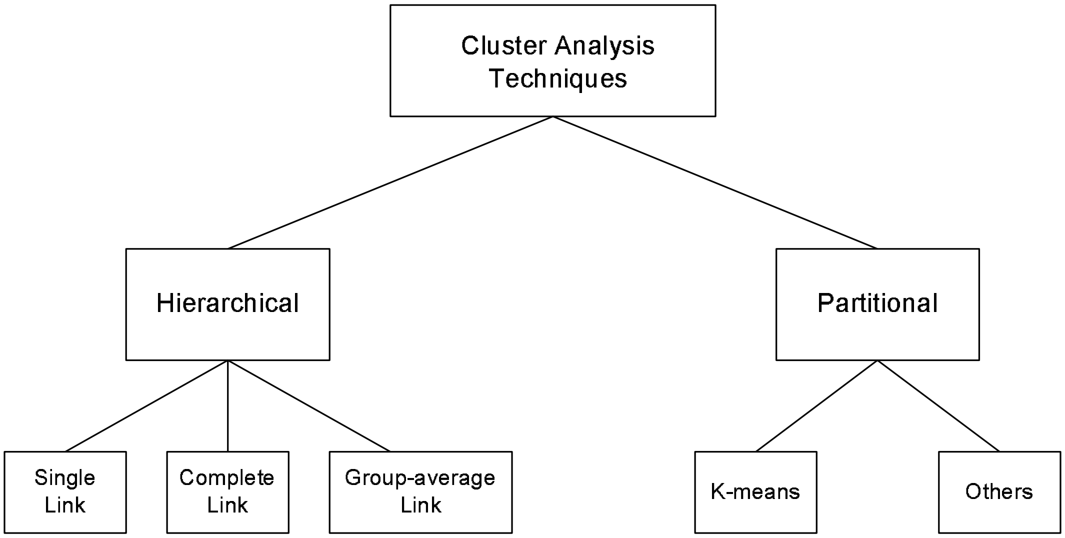


Figure 2.1: Taxonomy of Clustering Approaches

Hierarchical Clustering Algorithms

Hierarchical algorithms produce a nested sequence of data items, with a single, all-inclusive cluster at the top level and single clusters of individual data item at the bottom level. Each intermediate level can be viewed as combining two clusters from the next lower level. There are two basic approaches to generate a hierarchical clustering:

- Agglomerative: Begin with each data item in a single cluster, successively merge clusters together until a stopping criterion is satisfied.
- Divisive: Begin with all data items in a single cluster, at each step, perform splitting until a stopping criterion is met.

This thesis will compare Hierarchical Agglomerative Clustering (HAC) and the proposed text clustering algorithm. Details of the comparison are given in Chapter 6.

HAC starts with a proximity matrix (correlation/cosine coefficient matrix or a co-occurrence matrix), which represents the similarity measurement between data items (e.g., keywords) in the matrix. Most hierarchical clustering algorithms are variants of the single-link, complete link, and group-average link algorithms. These three algorithms differ in the ways they characterize the similarity between a pair of clusters.

In single-link clustering (Sneath and Sokal, 1973), the similarity between a pair of clusters is taken to be the similarity between the *most similar pair* of items (e.g., two clusters that have the highest correlation between them). Hence, a new candidate for cluster membership can be joined to an existing group based on the highest level of similarity of any member of the existing group. In the complete-link algorithm (Sokal and Michener, 1958), the similarity between the *least similar pair* of items from the two clusters is used as the cluster similarity. This method is the logical opposite of single-link clustering in that the complete linkage rule states that any candidate for inclusion into an existing cluster must be within a certain level of similarity to *all* members of that cluster. Group-average link clustering (Sokal and Michener, 1958) is a compromise between the extremes of the single-link and complete-link algorithms. This method computes an *average* of the similarity of a case under consideration with all data items in the existing cluster and, subsequently, joins to that cluster if a given level of similarity is achieved using this average value.

Partitional Clustering Algorithms

A partitional clustering algorithm produces a one-level (un-nested) partition of the data. This algorithm intends to partition data items into a good classification of *k* clusters. There are a number of partitional techniques, but this thesis will provide descriptions for the K-means algorithm, which is widely used in document clustering. The K-means algorithm is based on the idea that a centroid, which is the mean or median point of the data, can represent a cluster. The basic K-means algorithm for finding *k* clusters is presented below.

1. Select k points as the initial centroids.
2. Assign all data points to the closest centroid.
3. Recompute the centroid for each cluster.
4. Repeat steps 2 and 3 until the centroids do not change.

2.2.5 Data Visualization

2.2.5.1 Multidimensional Scaling (MDS)

A general goal of analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) among the investigated objects. This is accomplished by solving a minimization problem such that the distances among points in the conceptual low-dimensional space match the given (dis)similarities as closely as possible. In factor analysis, the similarities among objects (e.g., terms) are expressed in the correlation matrix. With MDS, one may analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices. However, a major weakness of MDS is that there are no quick and fast rules to interpret the nature of the resulting dimensions. In addition, often time, the steepest descent algorithm, a popular approach in MDS, can be trapped in a local minimum and never reach the global minimum (Zhu, 1998). To alleviate the problem, Zhu and Porter (2002) have devised a "step by step" search algorithm. This algorithm is effective at finding the global stress minimum, although it usually consumes more CPU times than the steepest descent algorithm. They also added an additional representational element, connecting links, based on a "path-erasing" algorithm to represent high dimensional spatial relations by displaying the elements in 2-D or 3-D spaces.

MDS has been one of the most widely used mapping techniques in information science, especially for technology mapping (Zhu and Porter, 2002; Noyons and Van Raan, 1998), science mapping (Small, 1999), and co-citation analysis (White and McCain, 1998).

2.2.5.2 Self-Organizing Map (SOM)

The Self-organizing map (SOM) approach, developed by Kohonen is one of the most insightful contributions made by artificial neural networks to information visualization (Kohonen, 1995, Kohonen et al., 2000, Kaski et al., 1998). SOM is often used as a statistical tool for multivariate analysis. It is both a projection method, which maps high-dimensional data space into low-dimensional space, and a clustering method so that similar data samples tend to be clustered together. During the learning phase, an SOM algorithm iteratively modifies weight vectors to produce a typically 2-dimension map in the output layer that will exhibit as best as possible the relationship of the input layer. The SOM is widely used as a data mining and visualization method for complex data sets. Lin (1997) was the first to adopt the Kohonen SOM for information visualization to document spaces. Application areas include, for instance, image processing and speech recognition, process control, economic analysis, and diagnostics in industry and in medicine. A summary of the engineering applications of SOM appears in Kohonen et al., (1996).

2.3 Conclusion

This chapter has summarized various approaches in bibliometric analysis and text mining. The proposed text mining framework will follow the steps that were previously discussed. The process starts with retrieving the relevant documents from the appropriate databases. Then the data will be extracted and cleaned to remove noises and errors. This cleaned data will then be fed to the analysis process. This thesis will add to this body of knowledge by implementing a new text clustering process. The last step is the representation/visualization of the results. The framework will be discussed in details in the following chapters.

CHAPTER 3

TEXT MINING FRAMEWORK FOR TECHNOLOGICAL INTELLIGENCE DISCOVERY

In this chapter, a text mining framing for knowledge discovery is presented. The framework uses several techniques in information retrieval and data mining to extract knowledge from technical text. Figure 3.1 illustrates the overall components and processes of the proposed framework. The overall processes divide into four main stages: data extraction and cleansing, text summarization, text clustering, and visualization. VantagePoint software has the capabilities to perform these four processes.

3.1 Text Extraction and Cleansing

The data set required for this framework can be obtained from electronic abstracts databases such as *INSPEC*[2], *Engineering Index*[3], and *Science Citation Index*[4]. Data extraction is the activity of automatically pulling out pertinent information from large volumes of text. Hence, for the underlying data set, the specific fields representing entities such as name, date, or address are extracted. VantagePoint also applies natural language processing to parse titles or abstracts into phrases and parts of speech for further analyses. This function is especially useful for analyzing certain databases that do not provide keywords (subject index terms) fields.

[2] Citations and abstracts of articles in physics, electronics, engineering, computer and information technology journals from 1970 – present.

[3] Citations and abstracts of articles from engineering journals from 1970 – present.

[4] Citations and abstracts of articles in multidisciplinary areas. SCI is published by the Institute for Scientific Information (ISI).

Moreover, title and/or abstract phrases with technical words sometime capture the technical themes, which the author intends to present, better than keywords indexed by a database provider.

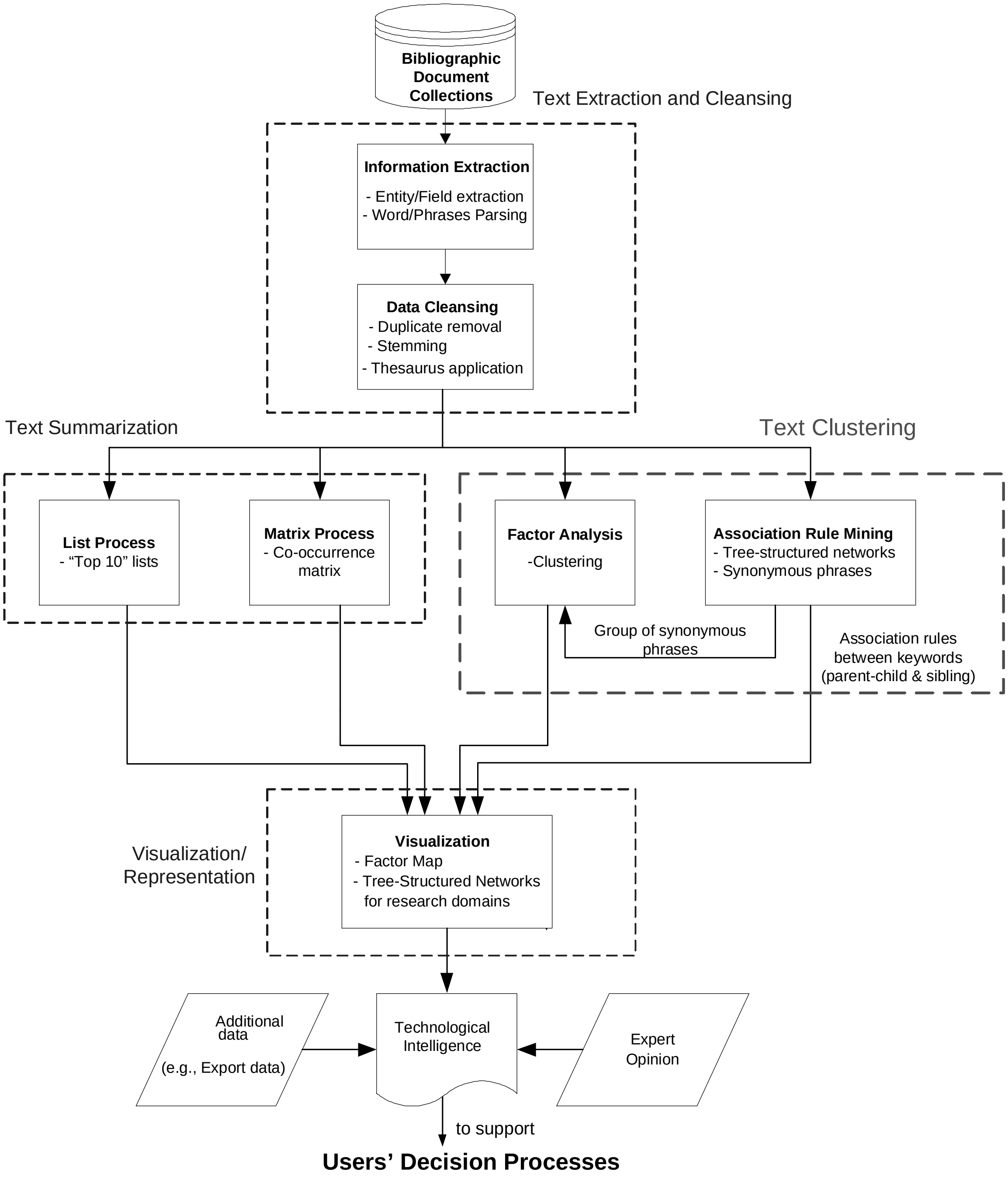


Figure 3.1: Text Mining Framework for Technological intelligence discovery

For the data cleansing step, VantagePoint applies fuzzy logic to remove duplicate records and consolidate similar terms. This function is important when combining data from multiple databases. The underlying data can be duplicates. In addition, VantagePoint uses fuzzy matching techniques to identify and combine similar entities among authors, keywords, or affiliations. For instance, an author name can be written in different way such as:

John S. White
J. S. White
White, JS
White, John

The articles with these author name variations should be combined before further analyses, if, indeed, they are the same person. Hence, fuzzy matching techniques help provide users with higher quality results.

VantagePoint also applies thesaurus and grouping capabilities based on regular expressions (RegEx) to allow examination of subsets of the set of records. For instance, using a thesaurus, a user can combine data elements into broad categories (e.g., “USA,” “Canada,” and “Mexico” into “North America”).

3.2 Text Summarization

VantagePoint creates a summary of available fields such as keywords (or authors or affiliations). The top keywords (or authors or affiliations) can be observed and the articles for each one can be browsed. Such lists can provide the first order of useful information –e.g., which affiliations or authors are most active in the field? Two such lists can be combined to create a co-occurrence matrix to show the concentration of activity. For instance, the co-occurrence matrix between keywords and affiliations can help identify which topics particular organizations mention frequently.

3.3 Text Clustering

VantagePoint applies statistical analyses, particularly Principal Components Analysis (PCA), to help find relationships among topics and concepts. However, these relationships do not discern hierarchical structure among the concepts. As a result, this research explores a new algorithm based on association rule mining that can automatically discern concept hierarchy, such that a general concept appears as a parent of a more specific one. Constructing concept hierarchies for research topics is one important goal of S&T study. This hierarchical structure can help identify emerging areas in an existing field of research.

Another algorithm used to construct abstract or title phrases thesaurus is also proposed. This algorithm can group similar abstract or title phrases before they are used for factor analysis. This process is important since it can improve the quality of clustering results.

3.4 Visualization/Representation

After relationships among concepts have been identified, VantagePoint applies Multidimensional Scaling (MDS) to create a low dimension representation that presents the original N dimensions as accurately as possible. Then a Path Erasing algorithm, a linkage technique, is applied to link concept clusters previously located via MDS (Zhu and Porter, 2002).

This thesis focuses on the text clustering process. Hence, the details of natural language processing, fuzzy matching techniques and thesaurus based on regular expressions are not discussed here. Instead, other statistical analyses, PCA particularly, will be discussed in more details in the next Chapter along with its advantages and disadvantages. In Chapters 5, 6 and 7, details on the proposed algorithm that can augment the existing text mining technique used in VantagePoint, are then presented.

CHAPTER 4

TEXT CLUSTERING

One can derive two kinds of relationships among a given set of documents. The first one measures similarity among documents in a database. Most research from the information retrieval (IR) community (Salton and McGill, 1983, Baeza-Yates and Ribeiro-Neto, 1999) concentrates on this area. Their purpose is to identify groups of documents in a database that can be retrieved faster and more accurately for a given type of user input query. The second relationship addresses similarity among terms in a set of documents. It receives much attention from bibliometrics and scientometrics communities. However, some research from the IR community also focuses on this area, essentially for constructing thesauri automatically. This thesis will focus on terms' relationships to understand the cognitive structure of research domains. The ultimate aim is to provide useful intelligence in support of S&T management.

4.1 Basic Statistical Analyses

The text mining process starts by considering a matrix of documents by terms (**X**). This follows the idea of Vector Space Models introduced by Salton and McGill (1983). A vector, comprised of the terms contained within that document, can represent a document. To illustrate, consider a sample data set consisting of six documents that share five common keywords. Table 4.1 presents an occurrence matrix of keywords across documents.

Table 4.1: An Occurrence Matrix **X** of Keywords across Documents

	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
D1	1	0	1	1	0
D2	1	0	1	0	1
D3	1	1	0	0	0
D4	0	0	1	0	1
D5	1	0	1	1	0
D6	1	1	0	0	0

In this matrix, a "1" indicates the occurrence of a keyword in a given document (D_i), and a "0" indicates its absence. If terms are occurring together more often in documents, these could reflect a strong relationship. In focusing on R&D-oriented abstracts, co-occurrence analyses can help reveal innovation relationships and prospects.

The simplest relationship analysis can be conducted on the terms co-occurrence matrix (**CO**). A terms co-occurrence matrix is simply a dot product of the input documents by terms matrix. Hence,

$$\mathbf{CO} = \mathbf{X}^{\mathbf{T}} \bullet \mathbf{X} \tag{4.1}$$

co_{xy}, the co-occurrences of the term *x* and term *y*, can be calculated by:

$$co_{xy} = \sum_{i=1}^{m} b_{ix} \cdot b_{iy} \tag{4.2}$$

In this formula "*b*" is a vector, a row of terms, which occur or do not occur in a given document. "co_{xy}" reflects the co-occurrence between two terms, *x* and *y*. If $x = y$, then co_{xx} is the occurrence of term *x*. "*m*" is the number of documents. The resulting co-occurrence matrix of terms is noted below.

Table 4.2: Terms' Co-occurrence Matrix

	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
Keyword 1	5	2	3	2	1
Keyword 2	2	2	0	0	0
Keyword 3	3	0	4	2	2
Keyword 4	2	0	2	2	0
Keyword 5	1	0	2	0	2

From the above example, each of the two times that Keyword 4 appears in a document, so does Keyword 3. This could be taken as evidence that Keyword 4 and Keyword 3 may be related.

VantagePoint applies singular valued decomposition (SVD) as a 'data reduction' technique to perform co-word analyses. SVD is based on a theorem of linear algebra: "any *m* x *n* matrix **X** whose number of rows *m* is greater than or equal to its number of columns *n*, can be written as the product of an *m* x *n* column-orthogonal matrix **U**, an *n* x *n* diagonal matrix **W**, and the transpose of an *n* x *n* orthogonal matrix **V"** (Press et al., 1986). In other words, **X** can be decomposed into the following form:

$$X = U \cdot W \cdot V^T \tag{4.3}$$

The **U** and **V** matrices are orthogonal and contain eigenvalues; $\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$**,** where **I** is an identity matrix; the **W** matrix is diagonal, containing eigenvalues. If the original matrix **X** is a matrix of documents by terms,

- **U** would be a matrix of documents by factors
- **W** would be a matrix of factors by factors
- **V** would be a matrix of terms by factors

Note that the term "factors" will be used for the linear combinations of the original data or "eigenvectors" or "principal components."

Deerwester et al. (1990) and Cunningham (1996) used Latent Semantic Indexing (LSI) –SVD of documents by terms matrices to help uncover underlying structures in the data for the purpose of indexing documents.

4.2 Principal Components Analysis

Principal Components Analysis (PCA) is a classical statistical technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables (called "principal components"). These principal components represent most of the information in the original set of variables. The first principal component accounts for as much of the variability in the data as possible. Each succeeding component accounts for as much of the remaining variability as possible. The goal is to reduce the dimensionality of the original large set of variables. Hence, it is potentially useful for resolving the correlation structure of a set of scientific terms.

VantagePoint uses PCA for dimension reduction and grouping to discover relationships among correlated scientific terms. Keep in mind that although PCA is traditionally performed on the correlation matrix, it can be performed on some other similarity matrix such as cosine transformed data. The details of classical PCA are presented below.

The documents by terms matrix **X** is first transformed to the correlations of terms by terms matrix **C**. The transformation is as follows:

$$c_{xy} = \frac{\sum_{i=1}^{m} (b_{ix} - \mu_x)(b_{iy} - \mu_y)}{(\sigma_x \cdot \sigma_y)} \tag{4.4}$$

$$\mu_x = \sum_{i=1}^{m} (b_{ix}) / m \qquad \mu_y = \sum_{i=1}^{m} (b_{iy}) / m \tag{4.5}$$

$$\sigma_x = \sqrt{\sum_{i=1}^{m} (b_{ix} - \mu_x)^2} \qquad \sigma_y = \sqrt{\sum_{i=1}^{m} (b_{iy} - \mu_y)^2} \tag{4.6}$$

In the above formula, c_{xy} represents the Pearson's correlation measure between term x and term y, μ is the mean of the term vector, and σ is the standard deviation. Table 4.3 presents the Pearson's correlations matrix of the original data from Table 4.1.

Table 4.3: Terms' Correlation Matrix C

	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
Keyword 1	1	0.32	-0.32	0.32	-0.63
Keyword 2	0.32	1	-1	-0.5	-0.5
Keyword 3	-0.32	-1	1	0.5	0.5
Keyword 4	0.32	-0.5	0.5	1	-0.5
Keyword 5	-0.63	-0.5	0.5	-0.5	1

Note that the correlation varies from one (maximally similar) to negative one (maximally dissimilar).

SVD is then performed on this correlation matrix. Since the correlation matrix is dimensioned terms by terms, **U** would be a matrix of terms by factors, **W** would be factors by factors, and **V** would be terms by factors. The following Tables (4.4-4.6) show the decomposition into **U**, **V** and **W** matrices of the Terms' correlation matrix in Table 4.3.

Table 4.4: The U Matrix from a Singular Value Decomposition

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Keyword 1	0.35	-0.44	0.83	0.01	0.00
Keyword 2	0.58	0.22	-0.12	-0.32	-0.71
Keyword 3	-0.58	-0.22	0.12	0.32	-0.71
Keyword 4	-0.15	-0.70	-0.30	-0.64	0.00
Keyword 5	-0.43	0.47	0.44	-0.63	0.00

Table 4.5: The W Matrix from a Singular Value Decomposition

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1	2.69	0.00	0.00	0.00	0.00
Factor 2	0.00	1.86	0.00	0.00	0.00
Factor 3	0.00	0.00	0.45	0.00	0.00
Factor 4	0.00	0.00	0.00	0.00	0.00
Factor 5	0.00	0.00	0.00	0.00	0.00

Table 4.6: The V Matrix from a Singular Value Decomposition

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Keyword 1	0.35	-0.44	0.83	-0.01	0.00
Keyword 2	0.58	0.22	-0.12	0.32	-0.71
Keyword 3	-0.58	-0.22	0.12	-0.32	-0.71
Keyword 4	-0.15	-0.70	-0.30	0.64	0.00
Keyword 5	-0.43	0.47	0.44	0.63	0.00

Since **C** (Pearson's correlation matrix) is a symmetric matrix, **U** = **V**. From the above tables, the matrices **U** and **V** contain information about term similarities. The **W** matrix is simply a weighting matrix that shows that the first two factors are more important in representing the data than the last three factors. Hence, by omitting the last three columns of the decomposed data, the resulting matrix would approximate the original data. For instance, using only the first two variables, the approximation will explain fully 91 percent of the original data.[5] Hence, the two factors can be used to approximate the original matrix.

[5] This calculation is performed using the eigenvalues from matrix **W**. The total sum of all five eigenvalues is 5. The sum of the first two eigenvalues, associated with the vectors used in the matrix approximation, is 4.55. Hence the variance explained in these first two eigenvalues is (4.55/5) = 91%

By looking at the **U** matrix, one can determine which term(s) is/are highly associated with each factor. For instance, consider only the first two factors, the highest loading term in factor one is “Keyword 2.” “Keyword 3” is as highly associated with factor one, but in a negative rather than a positive way. The highest loading term in factor two is “Keyword 4."

Sometime it is difficult to decide which term is a high loading term since the factor loading is somewhat in the middle range. VantagePoint applies Varimax Rotation (Harman, 1976) in order to facilitate the interpretation of the results without fundamentally changing the factors. The Varimax Rotation makes the interpretation easier by maximizing the variance of the squared factor loadings. For a given factor, high loadings become higher (close to 1), low loadings become lower (close to 0), and intermediate loadings become either lower or higher. To identify relevant terms for each cluster, VantagePoint applies a proprietary algorithm to automatically cut off high-loading terms from the resulting clusters (TPAC, 1999).

An advantage of using PCA to group similar terms together over traditional clustering techniques in text analysis is that PCA does not force each term into only one factor. Terms are allowed to occur in multiple factors, unlike hierarchical clustering or K-means clustering where each term is only permitted to appear in one cluster.

However, one disadvantage of using PCA to group similar terms is that PCA is performed on the transformed data (i.e., Pearson’s correlation), not on the raw occurrence data. Moreover, since the correlation coefficient is calculated based on deviations about the mean, it seems to be very sensitive to the total number of records in the data set.

Hence, estimates are likely to be unstable and are likely to remain sensitive to the introduction of new terms.

Principal components analysis, because of its correlation transformation, may not be the most appropriate technique for the clustering of terms. In the next chapter, another technique used to group similar terms and discover parent-child relationships, which are more complicated, is proposed.

CHAPTER 5

TEXT ASSOCIATION RULE MINING

In the data mining research area, the relationship among the data points can be based on correlations among a set of items (e.g., Factor Analysis), a distance metric defined for pairs of items (e.g., Vector Space Model), or association rules. The previous chapter presents the PCA approach. This chapter will explore the applications of association rule mining (ARM) for textual databases. This chapter proceeds in four parts. The first part presents the overall association rule mining activities. The second part discusses the definition of the association rule mining problem. The third part presents the classic Apriori algorithm used to generate frequent itemsets in ARM approach. The last part presents a novel algorithm based on object-oriented data structure for generating all frequent itemsets in text databases.

5.1 Association Rule Mining Research Profiling

Association rule algorithms were originally used for analysis of transaction data in the retail industry (market basket analysis). Agrawal et al. (1993) introduced what he called the “Apriori” algorithm to analyze these data. Mining association rules in transaction databases has been demonstrated to be useful and technically feasible in several application areas such as retail sales (Brin et al., 1997, Chen et al., 1996, and Cooley, et al., 1999). For instance, the association rules mined from point-of-sale (POS) transaction databases can be used to predict the purchase behavior of customers. Association rule mining can help answer the following questions:

- How can the sale of a specific item be increased?
- What is the impact if a certain item is discontinued?
- Is the sale of items on shelf A related to the sale of items on shelf B?

This thesis investigates association rule mining for bibliographic data. To profile the overall association rule mining research activity and to discover trend and domain applications of this area, VantagePoint is used to profile literature abstracts related to association rule mining. These records were first retrieved from the *INSPEC* and *EI Compendex (Engineering Index)* databases on July 22, 2003. The search[6] yielded over 700 records from *INSPEC* and over 400 *EI Compendex* records. Combining both search results, with duplicates removed, yielded 971 abstracts. Publications that are more recent[7] have been added to the existing data set, yielding now 1,107 abstracts. These consist of conference papers (62 percent) and journal articles (38 percent). Research domains that publish mainly through conferences tend to be faster moving than those that rely mainly on journals. Hence, the greater number of conference papers suggests that association rule mining research is fast moving. The data set is then imported and analyzed with the help of VantagePoint.

[6] Search terminology: ((association or associations) with mining) not (coal or coals or rock or rocks or explosion or geophysical or safety or power or energy), where “with” is in the same text unit (e.g., a sentence).

[7] The update is performed on February 28, 2004.

To monitor overall research activity, keywords indexed are examined first. Keywords provide a good summarization of the papers. Browsing through the list of keywords and clusters of keywords (Figure 5.1) suggests several ARM applications such as:

- Transaction processing
- Marketing data processing
- Business data processing
- Retail data processing
- Database systems
- Information retrieval
- Decision support systems
- Electronic commerce
- Text analysis

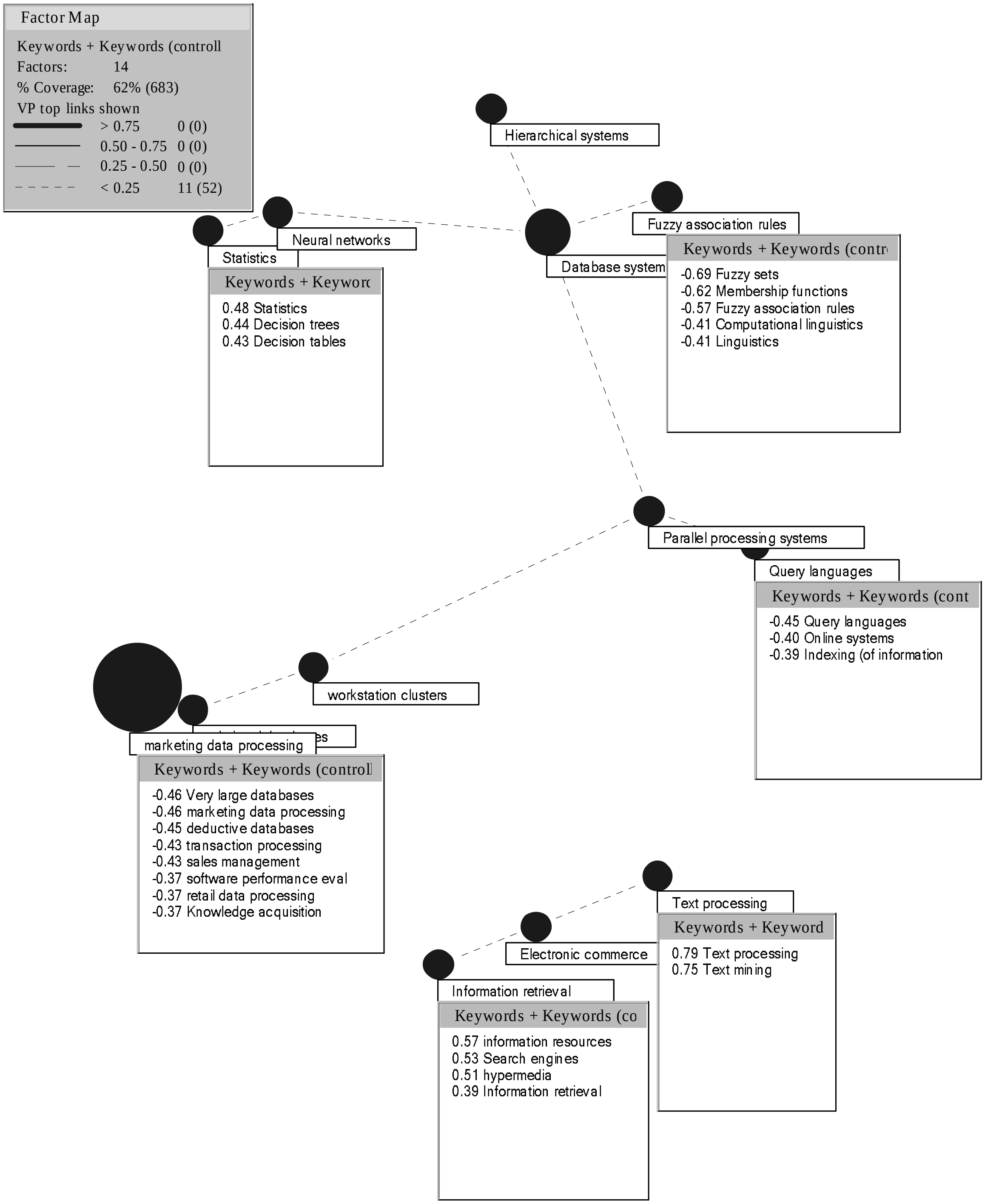


Figure 5.1: Clusters of Keywords related to Association Rule Mining

Figure 5.1 displays the clustering results of keywords relating to ARM research. In some cases, the terms that define the cluster are shown in the “pull-down” boxes. This clustering is based on principal components analysis (previously described in Chapter 4) to group keywords that appear often together in the records. Different nodes identify different factors or clusters of these highly correlated keywords. Node size reflects the frequency of documents represented by those terms. Placement of nodes is based on VantagePoint’s Multidimensional Scaling (MDS) routine. Topics that co-occur together will be placed near each other. Connecting lines represent the strength of the association between the two clusters, based on a Path Erasing algorithm. Note that the absence of a link suggests less association, not no association (Zhu and Porter 2002).

Figure 5.2 presents the comparison between selected topics in ARM research over the last decade. The majority of ARM research focuses on market data processing. The number of publications relating to ARM and market data processing increases sharply from 1993 to 1995. However, since 1995 the number of publications are mainly decreasing. The number of publications for fuzzy association rules and for database systems, on the other hand, appear to be increasing over time. The trend suggests these two areas are “hot” among the ARM research community. The number of ARM publications that link to text analysis is relatively small. The first paper appears much later than the other areas. This suggests that text association rule mining is quite an emerging research area.

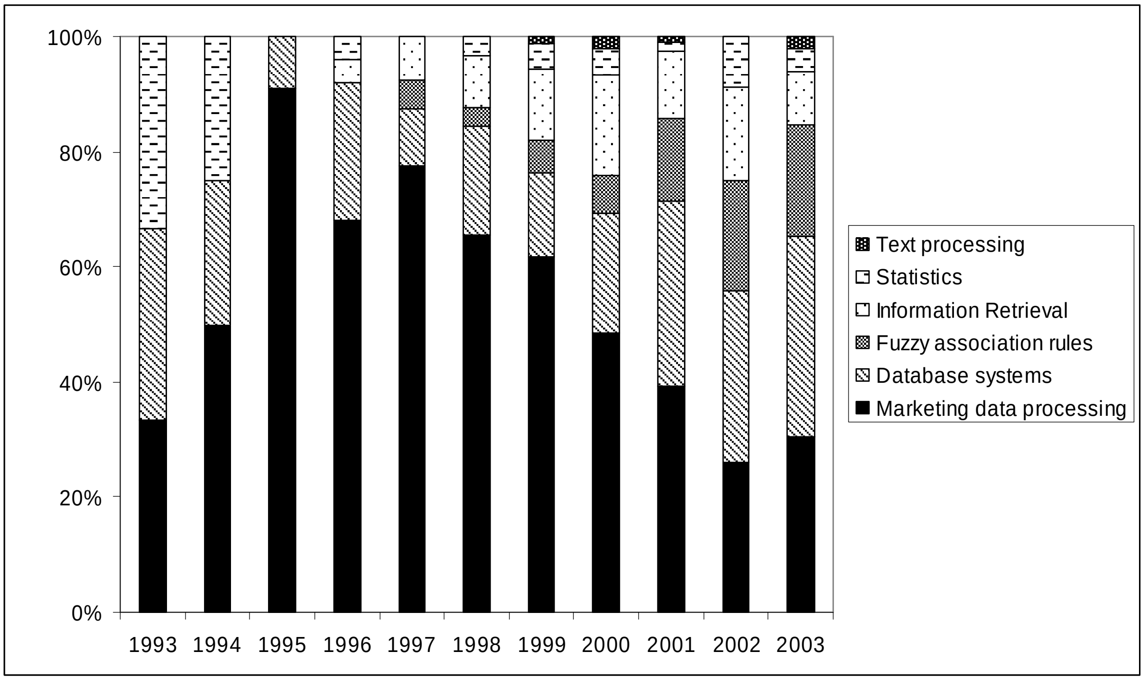


Figure 5.2: Comparison of topics in ARM research during 1993-2003

5.2 Research Domain Profile: Text Association Rule Mining

As discussed in Section 5.1, ARM is being applied most heavily to transaction databases. The application of ARM to text databases is still very limited. This section profiles text association rule mining to discover topic relationships and research trends. Of 1,107 publication abstracts retrieved, there are only 63 abstracts[8] (approximately 6 percent) that relate association rule mining with text databases. This suggests that text association rule mining is quite a new research area. A quick patent search of the United

[8] These records are determined by extracting records that contain the word ‘text’ in their abstracts.

States Patent and Trademark Office (USPTO)[9] also agrees that text association rule mining is a new area since this search yields only 14 patents relating to ARM and none relating to text ARM.

The sixty-three articles related to association rule mining in text databases are investigated further. Again browsing through a list of leading keywords and clusters of keywords (Figure 5.3) suggests text association mining has been mainly applied for information retrieval (e.g., bibliographic systems, query processing, and search engines).

[9] http://www.uspto.gov

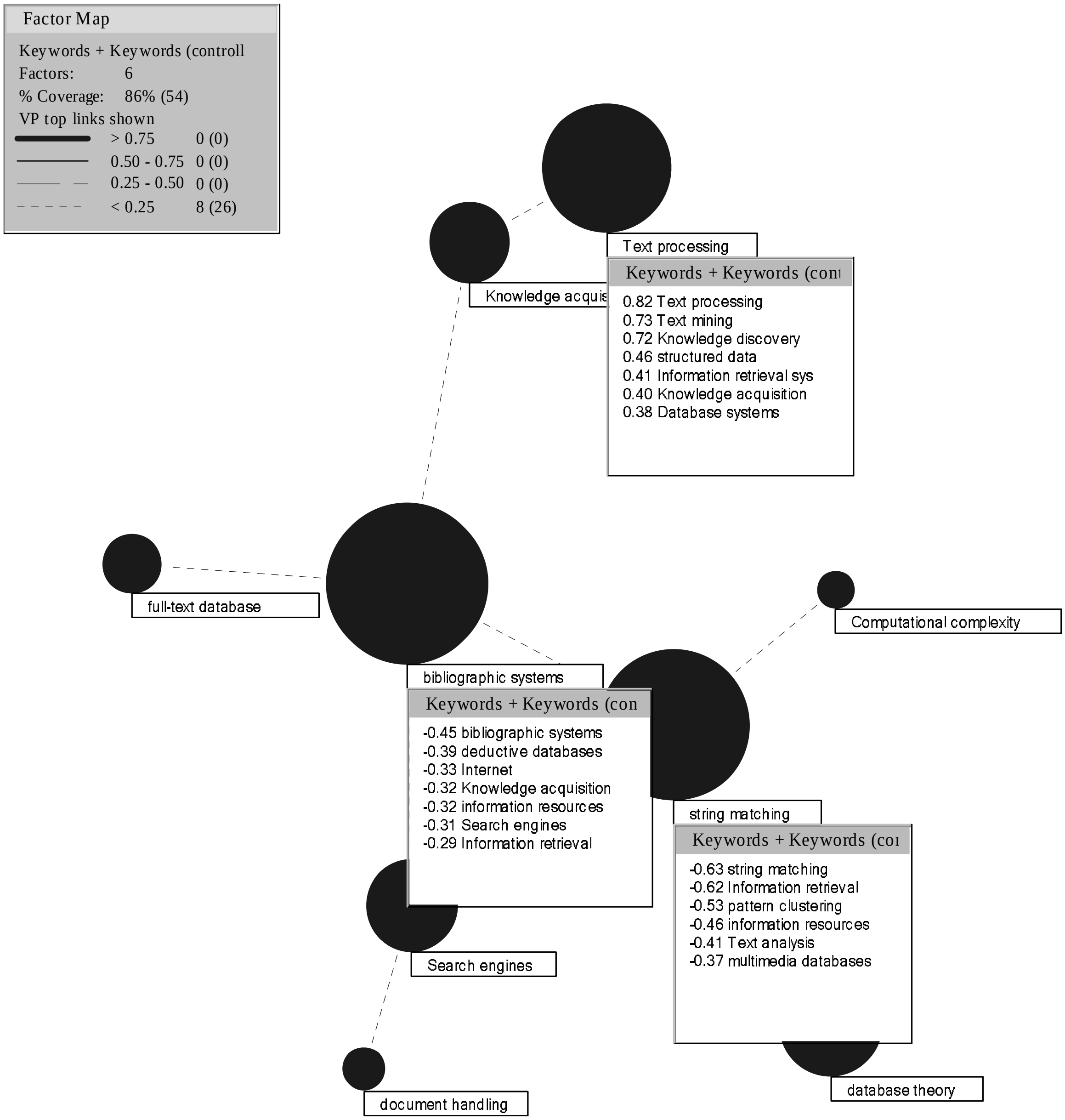


Figure 5.3: Clusters of Keywords related to Text Association Mining

To identify what's "hot or new," a new data subset for each topic is created, then a co-occurrence matrix between keywords and publication years is constructed for trend analysis to identify 'hot' topics. From this trend analysis, the "hot" sub-topics in text association rules mining are full-text databases, search engine, and data visualization. Moreover, two time slices are formed: one for 1996-2000, another for 2001-2003. Then the title phrases and keywords between these two groups are compared to identify 'new or unique' topics. These include:

- decision support systems
- hypermedia markup languages
- bilingual parallel corpora
- bibliographic data mining
- e-journal articles

To identify major research organizations and their emphases, a cross-correlation map between research organizations and keywords is created. Figure 5.4 presents the major research groups and their emphases. In this figure, each node represents a research organization. Proximity of organizations and links between them show higher correlation among keywords used by the organizations. For instance, the figure suggests, Stanford University, Kyoto University, and Bar-IIan University focus on similar research areas, which include search engines, deductive databases, and database management systems. Northwestern University and Pacific Northwest National Lab focus on text mining, structured data and information retrieval systems.

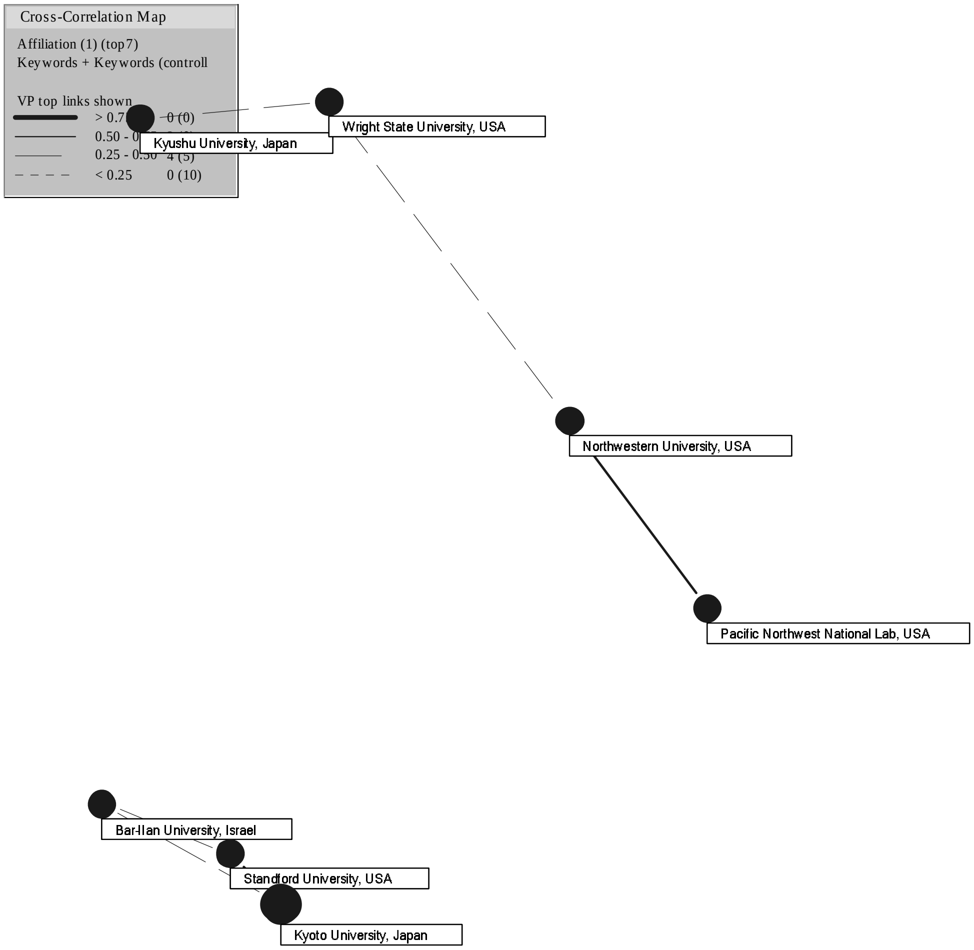

Cross-Correlation Map
Affiliation (1) (top7)
Keywords + Keywords (controll
VP top links shown
> 0.7 0 (0)
0.50 - 0
0.25 - 0.50 4 (5)
< 0.25 0 (10)
Kyushu University, Japan
Wright State University, USA
Northwestern University, USA
Pacific Northwest National Lab, USA
Bar-Ilan University, Israel
Standford University, USA
Kyoto University, Japan

Figure 5.4: Common Interests Among Research Organizations

The results of this “research profiling” (Porter et al., 2002a) suggest association rule mining is a very active area. Mining association rules in transaction databases has been demonstrated to be useful in several application areas. However, its application on text databases is still very challenging because characteristics of text and transaction databases are different. This leads to the motivation of this research, that is to apply association rule mining to bibliographic text databases to capture the relationships among words (terms).

The problem description of association rule mining is explained in the next section. Details of the proposed method to exploit association rule mining techniques to derive tree-structured networks and to construct term thesaurus in text sources will be discussed in Chapter 6 and 7 respectively.

5.3 Association Rule Mining

The general statement of the association rule mining problem is as follows:

Let $I = \{i_1, i_2, ..., i_m\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An ordered set of items is called an itemset. For instance, if a transaction contains the items $\{i_1, i_3, i_6\}$ then the itemsets present in this transaction are:

- $\{i_1\}$, $\{i_3\}$, $\{i_6\}$ (1-itemsets)
- $\{i_1,i_3\}$, $\{i_1,i_6\}$, $\{i_3,i_6\}$ (2-itemsets)
- $\{i_1,i_3,i_6\}$ (3-itemsets)

Definition 5.1. An **association rule** *r* is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y =$ Ø. *X* is called an **antecedent** of *r* and *Y* is a **consequent** of *r*.

A transaction *T* is said to contain *X*, a set of some items in *I*, if $X \subseteq T$. Two measures that are most often used to derive association rules are **support** and **confidence**.

Definition 5.2. Support *sup* of rule $X \Rightarrow Y$ in a transaction set *D* is the percentage of transactions in *D* containing both *X* and *Y*.

$$sup(X \Rightarrow Y) = \frac{count(XY)}{|D|} \tag{5.1}$$

Support can be viewed as an estimate of the probability of simultaneously observing both itemsets Pr(*X* and *Y*) in a randomly selected market basket.

Definition 5.3. Confidence *conf* of rule $X \Rightarrow Y$ in a transaction set *D* is the percentage of transactions in *D* containing *X* that also contain *Y*.

$$conf(X \Rightarrow Y) = \frac{sup(X \Rightarrow Y)}{sup(X)} = \frac{count(XY)}{count(X)} \tag{5.2}$$

Confidence can be viewed as an estimate of conditional probability: P(*Y*|*X*).

Definition 5.4. Association Rule Mining is the process of generating all association rules that have support and confidence greater than the user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*) respectively.

Rules with high confidence (i.e., close to 100 percent) are important because they denote a strong relationship of the items in the rule. Rules with high support are also important since they are based on a significant fraction of transactions in the database.

Association rule mining is a two-step process.

- **Find all frequent itemsets.** This is to find each set of items (called *itemsets*) that have a co-occurrence rate above the *minsup*. An itemset with at least the *minsup* is called a *frequent itemset*. The size of an itemset represents the number of items contained in the itemset, and an itemset containing *k* items will be called a *k*-itemset.
- **Generate association rules** from the frequent itemsets. For every frequent itemset *f*, find all non-empty subsets of *f*. For every such subset *a*, generate a rule of the form $a \Rightarrow (f - a)$ if the ratio of support(*f – a)* to support(*a*) is at least *minconf*.

5.4 Classic Apriori Algorithm

The literature has proposed various algorithms for discovering association rules. Most of the previous studies (Houtsma and Swami, 1993, Agrawal and Srikant, 1994a, and Brin et al., 1997) implement an *Apriori*-like approach, i.e., they require multiple passes over the database. The main idea of the *Apriori*-like approaches is to iteratively generate the set of *candidate itemsets* (potential *frequent itemsets*) of length (*k+1*) from the set of *frequent itemsets* of length *k* (for *k* ≥ 1), and count their supports (occurrence frequencies) in the database. However, because the size of the database can be very large, it is very costly to scan the database repeatedly to count support for candidate itemsets. To reduce the combinatorial search space, all algorithms exploit the following property: if any length *k* itemset is not frequent in the database, its length (*k+1*) extension itemsets can never be frequent. This property could greatly reduce the size of candidate itemsets. The classic Apriori algorithm (Agrawal and Srikant, 1994a) can be summarized as follows:

1. Create 1-itemsets
2. *k*=2
3. Generate candidates for *k*-itemsets
4. Count support of candidates in database
5. *k*:=*k*+1, repeat step 3 and 4 until no more candidates can be created.

In order to create the 1-itemsets, one has to scan each transaction in the database and count the occurrences of each item. Items that do not reach the minimum support are discarded. Larger candidate itemsets of size *k* are then formed based on the (*k*-1) frequent itemsets. This is done by joining two distinct (*k-1)*-itemsets, where the first *k-2* items are identical. As next step, a pruning algorithm is employed. The pruning step will check if all (*k*-1) subsets of the candidate *k*-itemsets are present in the list of (*k*-1) frequent itemsets. If this is not the case, the candidate cannot be frequent, since not all of its subsets are frequent. Hence, without checking the actual support of the *k*-itemset, one can conclude, if the itemset can be a frequent itemset. If it is a candidate, one must determine its support by scanning the data set. It is possible that the candidate will not have minimum support, although all its subsets have. See the next section for an example.

5.4.1 Candidate Generation

Let L_k be the set of frequent *k*-itemsets (those with minimum support). Let C_k be the set of candidate *k*-itemsets (potentially frequent itemsets). The 1-itemset generation is straightforward. One has to scan each transaction in the database and count the occurrences of each item. Items that do not reach the minimum support are discarded. L_1 then holds all frequent 1-Itemsets. For k>2 the algorithm to generate candidates is:

```
insert into $C_k$
select $p.item_1, p.item_2, .., p.item_{k-1}, q.item_{k-1}$
from $L_{k-1}\ p, L_{k-1}\ q$
where $p.item_1 = q.item_1, \ldots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1};$
```

In other words two distinct *k-1*-itemsets are joined together to form a *k*-itemset if their first *k-2* items are identical. As an example consider $L_3 = \{\{i_1, i_2, i_3\}, \{i_1, i_2, i_4\}, \{i_1, i_3, i_4\}, \{i_1, i_3, i_5\}, \{i_2, i_3, i_4\}\}$. These five 3-itemsets are used to create the candidate 4-itemsets. First $\{i_1, i_2, i_3\}$ is joined with $\{i_1, i_2, i_4\}$ to create the candidate $\{i_1, i_2, i_3, i_4\}$, because the first two items in theses itemsets are identical (i_1, i_2). For the same reason $\{i_1,$

i_3, i_4} is joined with {i_1, i_3, i_5} to create {i_1, i_3, i_4, i_5}. Therefore, the candidate set C_4 will be: {{i_1, i_2, i_3, i_4}, {i_1, i_3, i_4, i_5}}.

As next step, Algrawal explains a pruning algorithm, which will eliminate the itemsets in Ck that cannot be frequent, based on the Apriori-property. This algorithm checks if all the *k*-1 subsets of the *k*-itemset candidate are frequent. If they are not frequent, the candidate is discarded.

```
forall itemsets $c \in C_k$ do
    forall $(k\text{-}1)$-subsets s of c do
        if $(s \notin L_{k-1})$ then
            delete c from $C_k$
    end
end
```

Continuing the example from above, the prune step will delete {i_1, i_3, i_4, i_5} from the candidate set, because the subset {i_3, i_4, i_5} is not contained in L_3, therefore we know "apriori " that{i_1, i_3, i_4, i_5} cannot be frequent. Clearly this pruning step is only possible for generating k>2 itemsets. The above algorithm is named Apriori-gen by (Agrawal and Srikant, 1994a). With *D* as the set of transactions, the complete Apriori algorithm is:

```
$L_1$ = {frequent 1-itemsets}
for($k$=2; $L_{k-1}$; k++) do begin
    $C_k$ = apriori-gen($L_{k-1}$) // candidate generation, see above
    forall transactions $t \in D$ do begin
```

```
                $C_t$ = *subset*($C_k$,*t*); // $C_t$ is the set of candidates that are in *t*
                **forall** *candidates* $c \in C_t$ **do**
                    *c.count++;*
                **end**
                $L_k\{c \in C_k \mid c.count \geq minsup\}$
            **end**
        **end**
```

After the generation of the candidate itemsets, each transaction has to be scanned and the itemsets that are present in the transaction have to be counted. One problem of the Apriori algorithm is the high number of possible candidates. For n *(k-1)*-itemsets the number of possible candidates for *k*-itemsets is n(n-1)/2. For 10,000 frequent *1*-itemsets, this translates to almost 50 million candidates for each of which the support has to be determined. In this case, pruning will also not help, since it only is applicable for larger itemsets. For itemsets with k>2 this number will generally be smaller because of the pruning step. In situations with a large number of itemsets or very low minimum support thresholds, the Apriori algorithm may still suffer from the following costs:

- It is not reasonable to repeatedly scan the database and count supports for candidate itemsets.
- It is expensive to count the occurrences of a huge number of candidate itemsets, when only a small fraction is likely to be frequent.

To overcome these problems, this thesis proposes a new fast algorithm to generate all frequent itemsets for sparse text data sets in one pass.

5.5 Object-Oriented Data Structure for Text Association Rule Mining

The first papers on the topic of association rule mining were very concerned about memory management and the extent of I/O operations required. At the beginning of the 1990s when the first papers (Agrawal et al., 1993 and Agrawal and Srikant, 1994a, 1994b) on association rule mining appeared, main memory was relatively scarce in computers. The data sets were very large compared with the available memory. In addition, CPU speed was much slower. These algorithms typically use rather primitive data structures such as arrays and hash tables in order to use the memory as efficiently as possible.

The computing power available to the average researcher and student nowadays is increasing steadily. Furthermore the typical main memory of a standard PC now is 1GB or higher. This allows storing more data in main memory to avoid waiting for slower disk I/O operations. Due to the higher availability of memory that can be accessed at a much higher speed than data on disk, it is possible to create a more complex data structure that allows one to keep track of all relevant information without having to scan over the data set. Here a new algorithm called Object-Oriented Association Rule Mining (OOARM), which uses a special data structure that holds all relevant information, is presented. After only a single scan over the database, the algorithm can generate all frequent itemsets.

5.5.1 Justification for New Algorithm

The existing algorithms were developed especially for Basket Data Analysis. These transaction data typically have a limited number of different items, but a potentially very large number of transactions. The ratio of items/transactions therefore is typically << 1. A typical data set for a basket analysis might have 1,000 items and more than 100,000 transactions.

For the new domain of Association Rule Mining in bibliographic databases, these algorithms are not very efficient. The support of keywords often does not reach 1 percent and is usually much lower. The number of different keywords though can be relatively high. The typical ratio of items/transaction therefore is approximately 1 or larger.

Since the number of keywords can be relatively high, the number of 2-itemsets and larger itemsets can be also very high, depending on the minimum support that was set. The traditional Apriori-like algorithms will generate a very large number of potential candidates for which the support has to be determined. This means scanning over the database and counting the occurrences.

In order to overcome these difficulties, a special object-oriented data structure is implemented. It is specifically designed for itemsets with low support. The new data structure and algorithm has several advantages. It allows easy access and better organization of data, as well as fast generation of frequent itemsets with low support and their association rules. A comparison of the performance between the proposed algorithm and the Apriori-like algorithm is presented later. The data structure can also be used for more complex tools such as graphical representation, since it contains all the information needed.

5.5.2 Outline of Algorithm

During an initial scan, all items (i.e. keywords) are read from the database/file. Each bibliographic record (i.e. transaction in the traditional sense) is read and each keyword that exists in the record is held in memory. Each single keyword represents a 1-itemset. For each 1-itemset the references to the records where it occurs are stored as an attribute. After the initial scan, all 1-itemsets are held in memory. The next step is to remove all 1-itemsets that do not reach minimum support. After a record is read during the initial scan, the record's information will also be held in main memory. Each record holds the references to the itemset it contains. With this data structure, it is potentially easy to look up all itemsets that are occurring together in a record. By keeping track of all references of the transaction in which an itemset is occurring, it is possible to obtain all other possible itemsets with which this itemset can occur together.

All k+1-itemsets can be generated from k-itemsets, for example, each 1-itemset is combined with any other possible itemset based on the record references. If 1-itemset {A} occurs in record 1,2,3, and 1-itemset {B} occurs in record 4,5,6 then the 2-itemset {AB} cannot exist, since they do not share common records. The algorithm will automatically not consider this combination, because there is no reference to {B} in the records 1, 2, 3. Since the support in a text database is usually not very high (<1 percent) keeping all the references will not create a memory shortage. The following Figure 5.5 shows the basic concept:

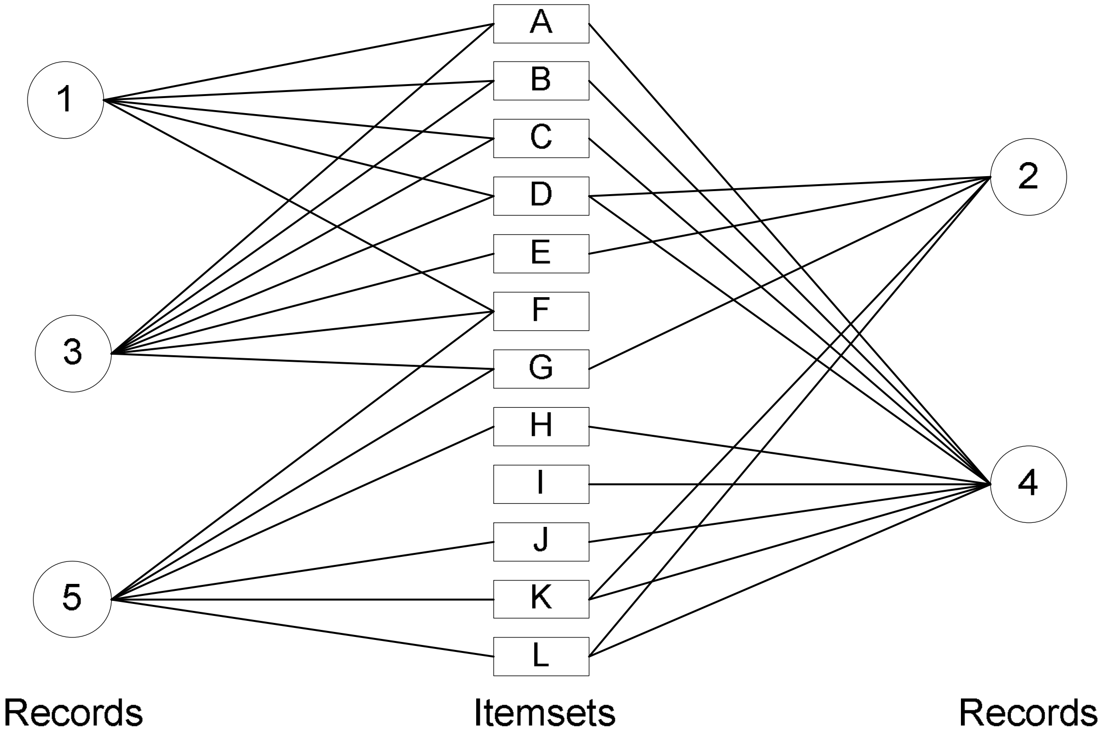


Figure 5.5: Example instance of data structure

The middle of Figure 5.5 shows all 1-itemsets of a sample instance. On the outside are the objects that represent the records (transactions). The 1-itemset containing A has the record references 1, 3 and 4. This means {A} is present in record 1, 3 and 4. Also note that the count of record references is equal to the support of the itemset (in this case support is 3). The record objects hold the references to the itemsets they contain. For example, record 1 contains A, B, C, D and F. This relation is represented as an arc between record 1 and A, etc. A detailed description of each data structure will follow in the next section. After creating the 1-itemsets, the 2-itemsets can be "assembled" from the 1-itemsets. The new algorithm will only consider itemsets as candidates that actually exist in the database and check their support. For example, the 2-itemset {G, I} does not exist in any record, so the algorithm will not create this itemset. This can easily be seen by looking at the records. The 1-itemset {I} has a reference to record 4 and the 1-itemset {G} has a reference to records 2, 3, 5. The classical Apriori algorithm would have chosen {G, I} as a candidate, since the Apriori property holds for all its subsets (assuming a *minsup* of 1 for this example).

The algorithm works as follows: The 1-itemset {A} is the first itemset to be examined. It holds references to the records 1, 3, and 4. By looking at record 1, the 1-itemset {B} is used to create the candidate itemset of {A, B}. For this 2-itemset candidate the intersection of the references is created, i.e., {1, 3, 4}. This new 2-itemset with support 3 is added now to the set of frequent itemsets. Also a reference of this new 2-itemset is added to the relevant records {1, 3, 4}. After creating all 2-itemsets we have the scenario in Figure 5.6.

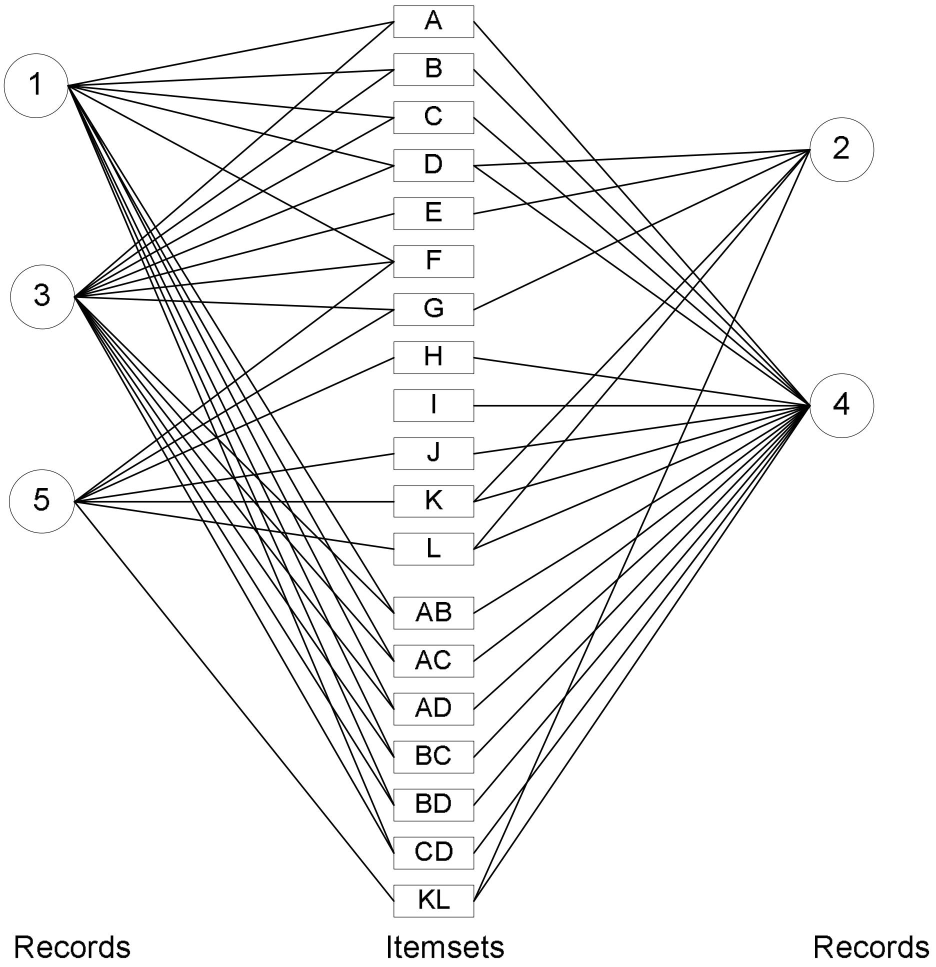


Figure 5.6: Data structure with new 2-itemset

5.5.3 Description of Data structure

The data are structured as an object-oriented structure in JAVA. Although traditional primitive data structures, such as arrays or hash tables, might have a slightly faster run time, they are more error prone and harder to debug. The proposed data structure is built with the JAVA Collections framework.

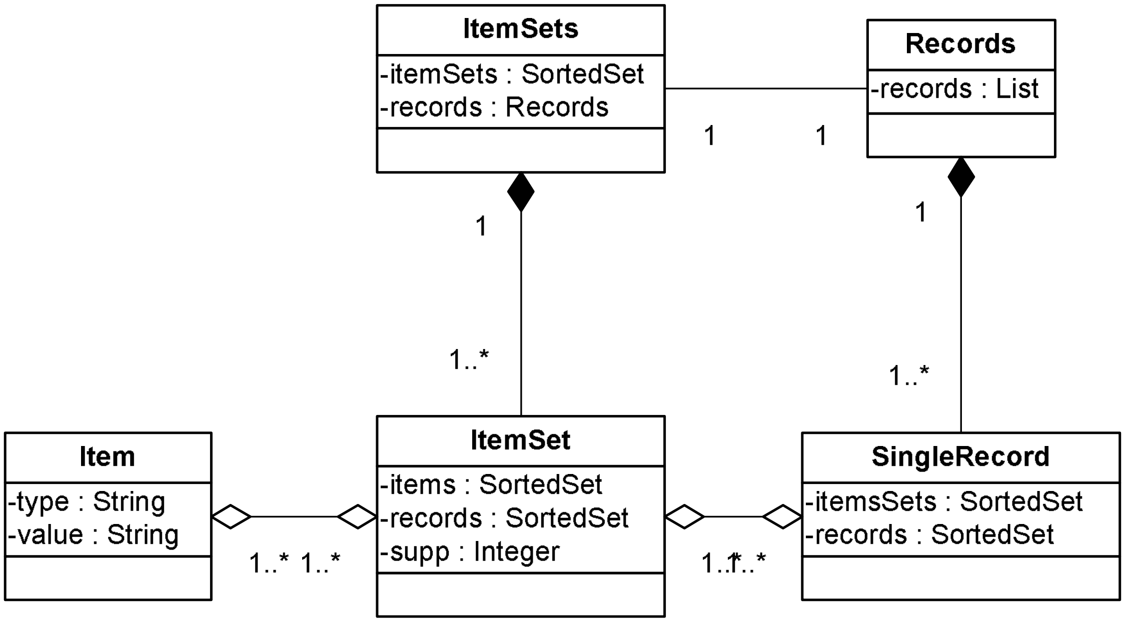


Figure 5.7: UML Class Diagram

Figure 5.7 shows a UML diagram of only the basic classes with the relevant attributes. UML stands for Unified Modeling Language and is a worldwide standard used to describe software. The smallest data unit is modeled with the `Item` class. The `Item` class represents an item such as a keyword, author, year, etc. It has two attributes: type and value. The type attribute holds a short description of the type of item, e.g., “kw” for keyword. The value attribute holds the actual information, i.e., “data mining” for type “kw”. Each item that exists in the data set will be represented by exactly one instance.

At the next level stands the `ItemSet` class. The `ItemSet` class consists of one or more items, where each item is an instance of the `Item` class. A *k*-Itemset consists of *k* different items. The `ItemSet` class also holds the references to the records in which the itemset is present.

The items are held in a data structure that implements the `SortedSet` interface (e.g. the `TreeSet` Class of the Java Collections Framework). This means there will be no duplicates and the items will be stored in ascending order. The references to the records are also stored in a data structure implementing the `SortedSet` interface.

The `SingleRecord` class allows storing the information of each single record (transaction). The `itemSets` attribute holds all the itemsets that occur in this record. All itemsets that have minimum support will be stored here. This also includes 2, 3,..*k*-itemsets. All itemsets are ordered -- first, according to their size and, second, lexicographically, i.e. first 1-itemsets are ordered lexicographically, then 2-itemsets are ordered lexicographically etc. This allows easy retrieval of specified itemsets.

At the top level, there is a single instance of the `ItemSets` Class, which serves as a container for all `ItemSet` objects. They are stored in a `SortedSet` array. At index 1, 2, …, *k*, all 1, 2, …, *k*-itemsets are stored respectively. It also contains the methods that are used to generate the itemsets. The `ItemSets` class also has an attribute that refers to the `Records` class. There is exactly one object of this class, which holds all the `SingleRecord` objects in an object of type `List.`

In sum, the data structure(s) have references for each itemset to the records where the itemset occurs and in reverse, each record holds the references to the itemsets that it contains.

At first sight, this seems like a wasteful use of memory, but as mentioned before, this is intended and is also acceptable due to the increasing availability of main memory. Also, note that only the references are stored. There will be only one copy of an itemset in memory. The `SingleRecord` class only contains the reference to this itemset, not the actual keyword. References only consist of a 32-bit address and therefore the size of a `SingleRecord` depends mainly on the number of itemsets that refer to it. The same is true for the record references of each `ItemSet` object.

5.5.4 Object-Oriented Association Rule Mining (OOARM) Algorithm

The following notation is used: As before let L_k be the set of frequent k-itemsets (those with minimum support). Let T_k be the temporary set of itemsets used to create the new k-itemsets. The itemsets in L_k and T_k will be ordered lexicographically. Furthermore *i.records* will be used for the set of record references where itemset *i* is present, *r.itemsets* will be used for the set of itemsets associated with record *r, minsup* represents the minimum support. The algorithm works as follows:

```
1. Scan in data from file and store 1-itemsets in $L_1$.
2. forall itemsets $i \in L_1$ do begin
       if(|i.records| < minsupp) // check for minimum support
           delete i from $L_1$
3. $k$=2
4. while($|L_{k-1}| > 0$)
     forall itemsets $i \in L_{k-1}$ do begin
       $T_k = \emptyset$
       forall records $r \in$ i.records do begin
           forall itemsets $c \in$ r.itemsets do begin
               if($c>i$) //check if c is lexicographically larger than i
                   add c to $T_k$
```

```
                end
            end   //now $T_k$ contains all itemsets that occur together with i
            forall itemsets $c \in T_k$ do begin
                if(|i.records $\cap$ c.records| > minsupp)
                    create new itemset j with items of i and c
                    j.records= i.records $\cap$ c.records
                if(j is k-itemset)
                    add j to $L_k$
            end
        end
    k++
    end while
```

Only at the beginning, the algorithm reads from the database or file. After deleting the itemsets from L_1 that do not have minimum support, there will be no further scanning of the database. The itemsets with k>1 are created based on the (k-1) itemsets. This is similar to most other algorithms. The generation of the candidates is however entirely different. The algorithm goes through the list L_{k-1} one by one to generate the (temporary) candidates T_k . Starting with an itemset *i* from L_{k-1}, potential candidates for T_k are found by following the references from *i.records* and then going back to the itemsets from *r.itemsets* for all $r \in$ *i.records*. The itemsets found this way are only added if they are lexicographically larger than *i*. The itemsets of T_k are then used to create the k-itemset candidates. This can be done because the itemset *i* and the itemsets in T_k both occur in at least one record together. Hence the *k*-itemset candidates will have at least support 1. If the intersection of *i.records* and *c.records*, where *c* is an itemset from T_k, has at least *minsup* records, a new itemset *j* is created with the items from *i* and *c*. It can happen that this new itemset has more than *k* items. If that is the case it will be discarded, else the new itemset is added to the set of frequent items L_k.

This algorithm relies heavily on set operations. The algorithm was implemented with the JAVA collections framework, which provides good support for set operations. It can easily be seen that the algorithm is not suitable for very large transactional databases, where itemsets have high support. For a database with 100,000 records and an itemset that has 50 percent support, this itemset would have to keep record references to 50,000 of the records. If there is a large number of such itemsets, memory will quickly become insufficient. Nevertheless for the purpose of mining bibliographic data, the algorithm works very well. The number of record references of an itemset is typically in the range of 5-50.

5.5.5 Rule Generation

Another advantage of this data structure is that all rules related to a specific itemset can easily be obtained. With this new algorithm and data structure it is possible to directly generate the desired rules. Since the number of potential association rules can be very high, it might not be reasonable to generate all rules at once but only rules that are of interest. The data structure can hold all relevant information and allows quickly extracting all association rules of interest. The researcher can pick itemsets, i.e., keywords that are of interest, and generate all rules related to it.

The usual procedure of generating rules is to take a k-itemset, where $k \geq 2$ and generate all subsets. The subsets then are used to calculate the rules. For instance, when looking at {ABC}, we have the subsets {AB}, {BC} and {AC}. If the support is known, we can easily calculate the confidence of the rules $AB \Rightarrow C$, $BC \Rightarrow A$, $AC \Rightarrow B$. However one might be interested in all the rules that involve {AB}. With traditional approaches this would involve looking up all itemsets that have {AB} as a subset.

Using the new data structure the rule generation is greatly simplified. Now it is easily possible to generate all rules with certain itemsets as antecedent or consequent. Let R be the set of Rules with minimum confidence *minconf* having the itemset x as antecedent or consequent. The algorithm to generate all rules with itemsets x as antecedent or consequent is as follows:

```
$R = \emptyset$
**forall** *records* $r \in x.records$ **do begin**
    **forall** *itemsets* $c \in r.itemsets$ **do begin**
        *generate the rules:* $c \Rightarrow x, x \Rightarrow c$
        *add the rules with confidence* $\geq \min conf$ *to R*
    **end**
**end**
```

The algorithm to generate the rules with itemset c as antecedent and x as consequent is:

```
create new itemset $j$ with items of $c$ and $x$
```

```
    $j.records = c.records \cap x.records$
        if($|c.records \cap x.records| < minsup$)
            stop, rule does not have minimum support
        else
            $confidence = \frac{|c.records \cap x.records|}{|c.records|}$
end
```

5.5.6 Extended Example

5.5.6.1 Scanning in the Data and Creating 1-Itemsets

To clarify the algorithm the following records are given

Table 5.1: Sample records

Record ID	Items
1:	A,B,C,D,F
2:	D,E,G,K,L
3:	A,B,C,D,E,F,G,
4:	A,B,C,D,H,I,J,K,L
5:	F,G,H,J,K,L

After initial scan, the list of 1-itemsets is as follows:

Table 5.2: List of 1-itemsets

1-Itemset	Record IDs	Support
A	1,3,4	3
B	1,3,4	3
C	1,3,4	3
D	1,2,3,4	4
E	2,3	2
F	1,3,5	3
G	2,3,5	3
H	4,5	2
I	4	1
J	4,5	2
K	2,4,5	3
L	2,4,5	3

After deleting the itemsets that do not achieve minimum support of 3, the list will look like this:

Table 5.3: List of 1-itemsets with minimum support

1-Itemset	Records	Support
A	1,3,4	3
B	1,3,4	3
C	1,3,4	3
D	1,2,3,4	4
F	1,3,5	3
G	2,3,5	3
K	2,4,5	3
L	2,4,5	3

The records are also updated. For this, all the references to the items that were deleted are also deleted from each record object.

Table 5.4: List of records

Record ID	Items
1:	A,B,C,D,F
2:	D,G,K,L
3:	A,B,C,D,F,G
4:	A,B,C,D,K,L
5:	F,G,K,L

This is the basis to create the 2 itemsets

5.5.6.2 Generating 2-Itemsets

First step is to iterate through the list of 1-itemsets: A, B, C, D, F, G, K, and L. Starting with A the algorithm is as follows:

- Look up all records associated with A: 1,3,4
- For each record, add all itemsets, associated with it, and lexicographically larger than itself to the candidate set:
    - add: B,C,D,F from record 1
    - add: G from record 3
    - add: K,L from record 4
- Create Candidate Sets: AB,AC,AD,AF,AG,AK,AL
- Iterate though candidate set and calculate intersection of their record references:

Table 5.5: List of 2-itemset candidates with itemset A

2-Itemset	Records
AB	1,3,4
AC	1,3,4
AD	1,3,4
AF	1,3
AG	3
AK	4
AL	4

Only the 2-itemsets AB, AC and AD have the required minimum support. The other itemsets are deleted. Note that it was not necessary to scan through the database. The intersection is easily retrievable from the references stored in the each `ItemSet` object. The references are stored in a `SortedSet` object, which provides a standard method to calculate the intersection of two sets.

Doing the same with the next itemset B yields:

Table 5.6: List of 2-itemset candidates with itemset B

2-Itemset	Records
BC	1,3,4
BD	1,3,4
BF	1,3
BG	3
BK	4
BL	4

The itemsets BF, BG, BK, and BL do not have minimum support and are deleted. Going on in this fashion, the list of 2-itemsets will be:

Table 5.7: List of frequent 2-itemsets

2-Itemset	Records
AB	1,3,4
AC	1,3,4
AD	1,3,4
BC	1,3,4
BD	1,3,4
CD	1,3,4
KL	2,4,5

After all frequent 2-itemsets are created; their references must be added to the `SingleRecord` objects:

Table 5.8: Updated list of records

Record ID	Items
1:	A,B,C,D,F,{AB},{AC},{AD},{BC},{BD},{CD}
2:	D,G,K,L,{KL}
3:	A,B,C,D,F,G,{AB},{AC},{AD},{BC},{BD},{CD}
4:	A,B,C,D,K,L,{AB},{AC},{AD},{BC},{BD},{CD},{KL}
5:	F,G,K,L,{KL}

5.5.6.3 Generating 3-Itemsets

In order to create all frequent 3-itemsets the procedure is repeated using the newly created list of 2 itemsets. The results are:

Table 5.9: List of frequent 3-itemsets

3-Itemset	Records
ABC	1,3,4
ABD	1,3,4
ACD	1,3,4
BCD	1,3,4

Now the `SingleRecord` objects have to be updated as before.

5.5.6.4 Generating 4-Itemsets

Generating the 4-itemsets from the 3-itemsets yields:

Table 5.10: List of 4-itmesets

4-Itemset	Records
ABCD	1,3,4

Since there is only one 4-itemset, there cannot be any larger itemset. Updating the `SingleRecord` objects:

Table 5.11: Updated Records

Record ID	Items
1:	A, B, C, D, F, {AB}, {AC}, {AD}, {BC}, {BD}, {CD}, {ABC}, {ABD}, {ACD}, {BCD}, {ABCD}
2:	D, G, K, L, {KL}
3:	A, B, C, D, F, G, {AB}, {AC}, {AD}, {BC}, {BD}, {CD}, {ABC}, {ABD}, {ACD}, {BCD}, {ABCD}
4:	A, B, C, D, K, L, {AB}, {AC}, {AD}, {BC}, {BD}, {CD}, {KL}, {ABC}, {ABD}, {ACD}, {BCD}, {ABCD}
5:	F, G, K, L, {KL}

Finally the complete list of itemsets is:

Table 5.12: List of all frequent itemsets

Itemset	Records	Support
A	1,3,4	3
B	1,3,4	3
C	1,3,4	3
D	1,2,3,4	4
F	1,3,5	3
G	2,3,5	3
K	2,4,5	3
L	2,4,5	3
AB	1,3,4	3
AC	1,3,4	3
AD	1,3,4	3
BC	1,3,4	3
BD	1,3,4	3
CD	1,3,4	3
KL	2,4,5	3
ABC	1,3,4	3
ACD	1,3,4	3
BCD	1,3,4	3
ABCD	1,3,4	3

In a graph representation this table looks like:

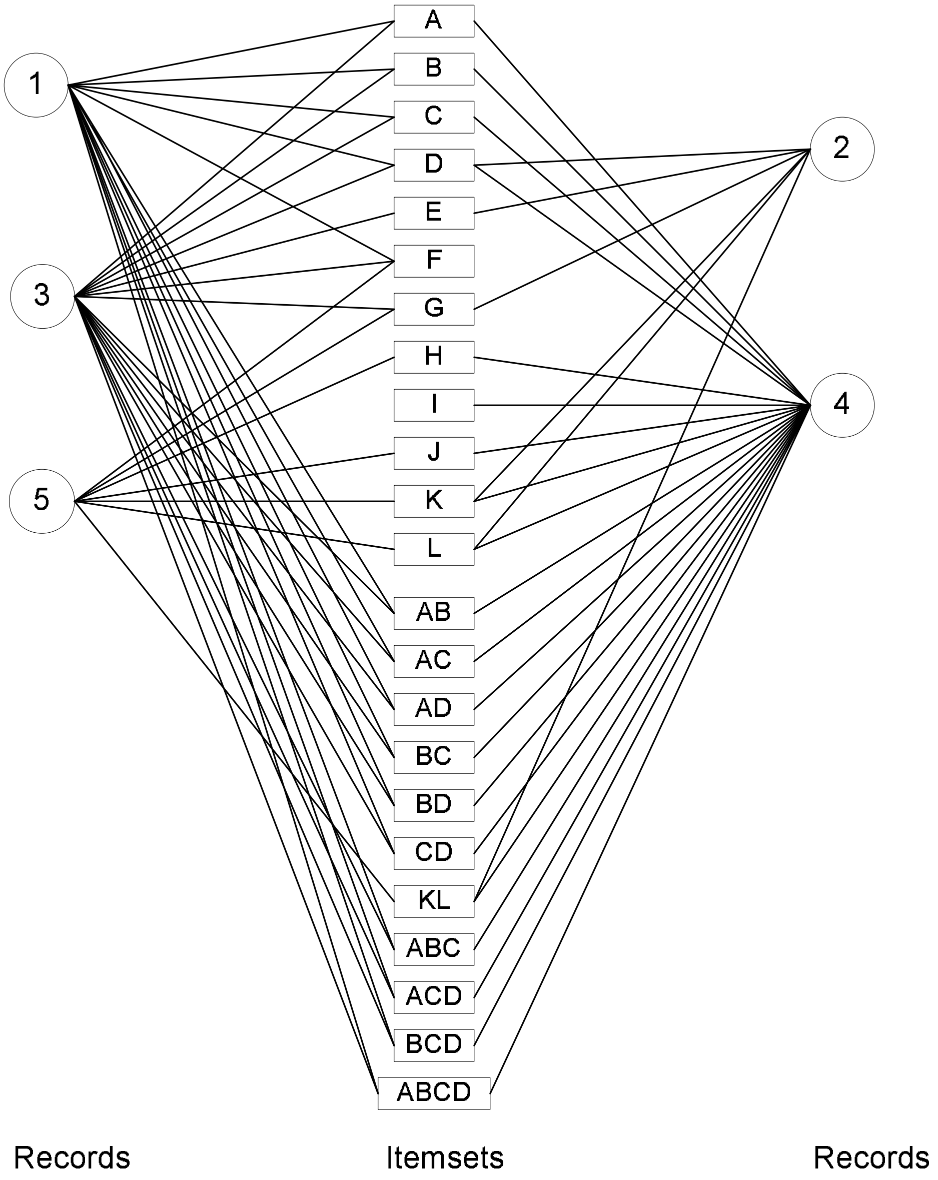


Figure 5.8: Frequent itemsets

5.5.6.5 Generating Association Rules

Rules for each itemset can now easily be generated. For instance, one might be interested in all the rules involving itemset D as antecedent with minimum confidence of

50 percent. The first step is to get all record references from itemset D: 1, 2, 3, 5. Then all items from these records are added to a list *L*: A, B, C, D, F, G, K, L, {AB}, {AC}, {AD}, {BC}, {BD}, {CD},{KL}, {ABC}, {ACD}, {BCD}, {ABCD}.

For each itemset *X* in *L* calculate the confidence of the rule: $D \Rightarrow X$ (See formula in previous section). The support for item D can be directly obtained from the number of record references (here 4). For the support of DX the intersection of the record references of both itemsets has to be calculated. For example to calculate the confidence of the rule $D \Rightarrow AB$, the 3-itemset ABD is created by joining D and AB. The support of ABD is 3. The confidence of the rule therefore is 75 percent.

5.5.7 Runtime Comparison with Apriori-like Algorithm

A comparison was made with an apriori-like algorithm based on pseudo code provided by Agrawal and Srikant (1994b). Three different data sets were tested. The summary of these data sets is shown in Table 5.13.

Table 5.13: Summary of data sets used to compare OOARM and Apriori algorithms

Data Set	Source	Number of Records	Number of Keywords Included
1	INSPEC and ENGI	971	266
2	INSPEC	2686	1927
3	INSPEC, ENGI, and SCI	16209	2330

The two algorithms were run for different minimum supports on standard Windows PC with P4 2.2 GHz and 640 MB main memory were as follows:

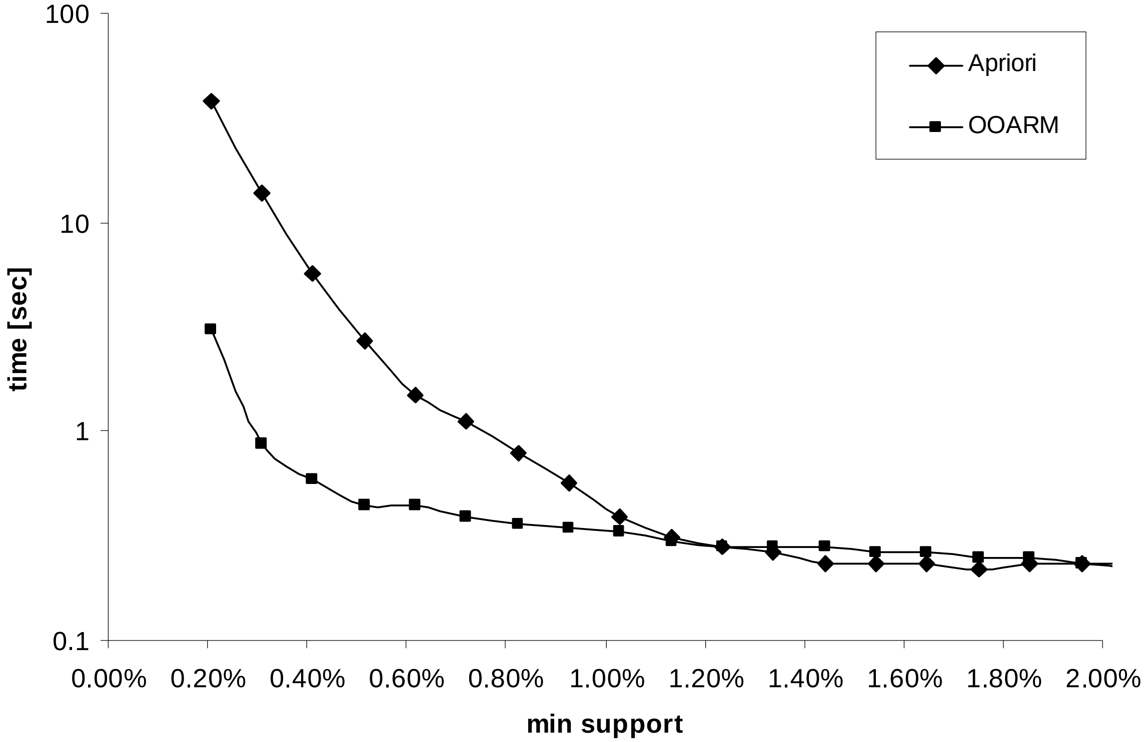


Figure 5.9: Runtime comparison for data set with 971 records

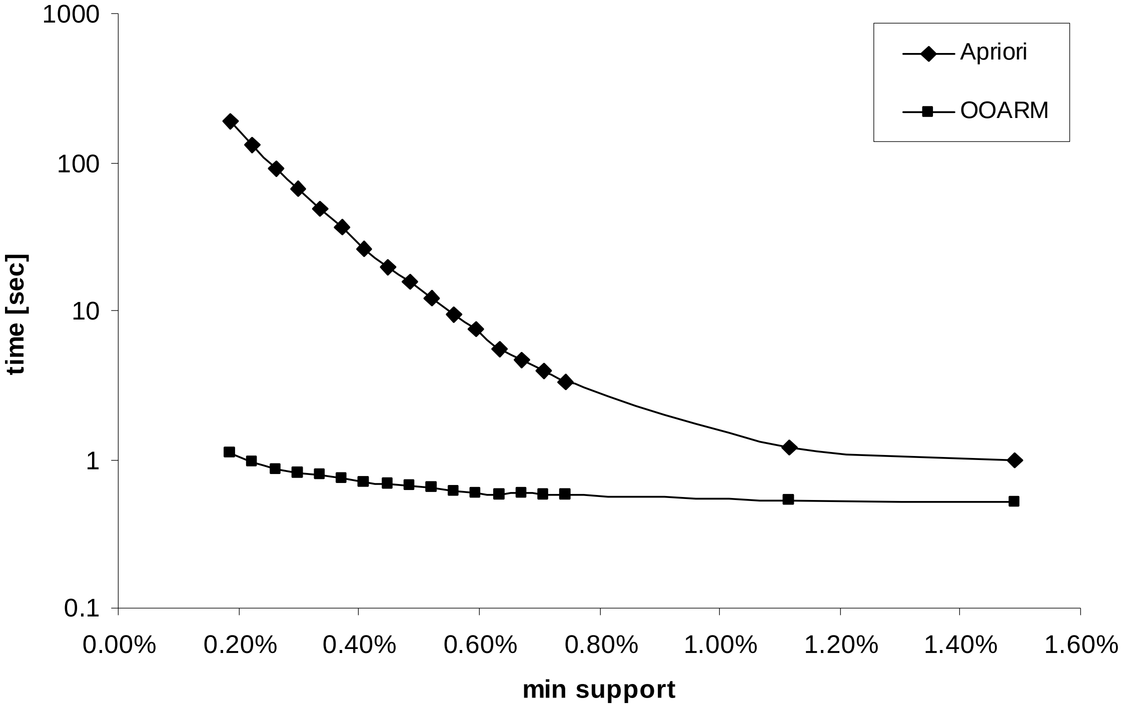


Figure 5.10: Runtime comparison for data set with 2686 records

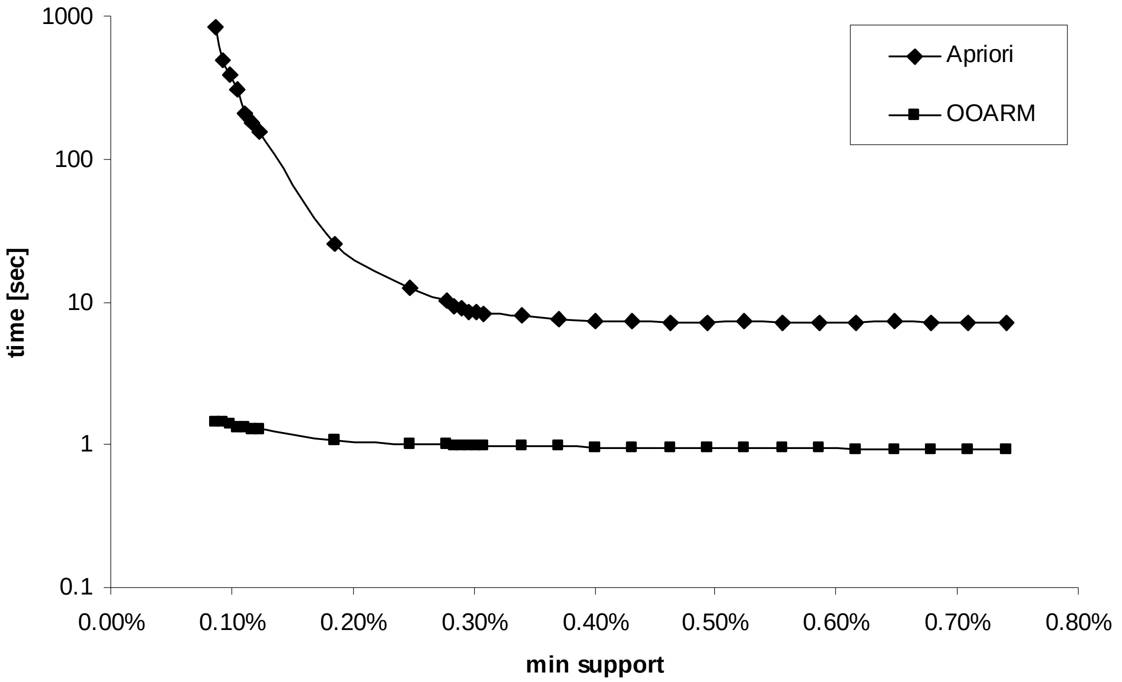


Figure 5.11: Runtime comparison for data set with 16209 records

The runtime of the OOARM algorithm is clearly smaller. For higher support values this advantage decreases. Note that the Y-axis is logarithmic scale. Also note the relative low support values that were used.

5.5.8 Differences between OOARM and Apriori Algorithm

The main difference is that for the proposed algorithm, after an initial scan, all data are held in main memory. The existing algorithms usually require several scans over the whole data set in order to determine the support. This advantage is very substantial for itemsets with low support. However, for itemsets with high support, e.g. 20 percent, which can be found in retail data, the proposed algorithm requires a lot more memory.

For the domain of Association Rule Mining in bibliographic databases, this is usually not a problem, since the data are preprocessed and keywords that are too general and would have a very high support are usually not used. Keywords or phrases with very high support can also be removed easily, since their value in discovering new relations among terms is very limited. (A keyword that appears in almost every record, for example, should not be considered to generate rules, since this keyword will be too frequently occurring in the rules.)

5.5.9 Conclusion

The OOARM algorithm allows generating all itemsets very quickly. It is specially designed to work on itemsets with low support. The experimental results show that improvements in run time of up to 500 times can be realized compared to a traditional Apriori-like algorithm. This is due to the fact that during the candidate itemset generation the number of potential candidates is much smaller than with the traditional Apriori-algorithm. However for higher support values, this difference is becoming smaller. Currently the disadvantage of the proposed algorithm is that all data have to be held in main memory. This limitation could be overcome by using an object-oriented database. This removes the limits on the memory requirements. Object-oriented databases allow storing the objects persistently. In the future, when computers become even more powerful, this data structure with larger data sets will be easily utilized. Another advantage of this data structure is that parallelization can be easily implemented. In order to create the (k+1)-itemsets, only the k-itemsets have to be read. Therefore, on a multiprocessor system the tasks can be easily distributed over several processors.

Another advantage is that the data structure can hold all relevant information for association rule mining. This allows for a new way to generate rules. After the initial creation of all large itemsets, it is possible to obtain rules that involve a chosen itemset in real time. This includes all rules where the itemset is an antecedent or a consequent. The data structure can also be seen as a new way to organize and store itemsets in memory and access rules for ARM.

CHAPTER 6

DERIVING TREE-STRUCTURED NETWORKS FROM TEXT

Clustering of data in a high-dimensional space is one of the most interesting topics among many other data mining applications. Various text clustering algorithms have been previously discussed in Chapter 2. This Chapter presents the use of the association rule mining technique to cluster related concepts and discern tree-structured networks from a set of technical documents. Tree-structured networks capture important aspects of both parent-child hierarchies (trees) and sibling relations (networks). It appears that most standard information retrieval and bibliometric analysis approaches using vector spaces or data reduction (e.g., Principal Components Analysis (PCA) or Latent Semantic Indexing) are able to identify relationships but not hierarchy. For instance, the "marketing data processing" in the factor map of keywords related to association rule mining topic (Figure 5.1) clusters together keywords that tend to occur together in the abstract records. Keywords: "marketing data processing," "transaction processing," "deductive databases," "very large databases," and others within a cluster are highly correlated. However, the cluster does not present the hierarchical structure of the terms such that higher hierarchy (parent) contains general terms and the lower (descendant of the parent) contains more specific terms.

Constructing tree-structured networks of research topics is one important goal of S&T study. Parent-child relationships can help identify emerging areas in an existing field of research. Sibling relationships are interesting as well since they may represent interdisciplinary structures among related topical areas.

6.1 Related Work

This section discusses work in discerning hierarchical structure from text corpora. Lexical techniques have been applied to derive various concept relationships from text (Miller, 1995, Hearst, 1992 and 1998, Caraballo, 1999). Hearst (1998) describes a method for the discovery of lexicosemantic relations by identifying a set of lexicosyntactic patterns in large text collections. She suggests that certain key phrases could be an indicator of a hyponym/hypernym relation. A concept represented by a lexical item A is said to be a hyponym of the concept represented by a lexical item B if native speakers of English accept that sentences constructed from A are a (kind of) B. B is called a hypernym of A if the reverse is true. Some of the key phrases she finds are:

- "such as", e.g., "…my favorite sports such as tennis and golf. "
- "or other", e.g., "…tennis, golf, or other sports."
- "including", e.g., "...any outdoor sports including tennis and golf."

These patterns are used to describe hyponym relationships, for instance tennis and golf are types of sports. Caraballo (1999) describes a technique to automatically construct a hypernym-labeled noun hierarchy from text. Lexical relationships such as those described by Hearst and Caraballo are applicable for large, full text corpora. However, for bibliographic text databases, these approaches might not perform well because of the small text body in the abstracts.

Sanderson and Croft (1999) describe a statistical technique based on subsumption relations. In their model, they focus on the relationship between two terms X and Y; X is said to subsume Y if the probability of X given Y is one[10], and the probability of Y given X is less than one. In contrast, this thesis will apply association rule mining to capture parent-child and sibling relations among more terms. This method is believed to be more efficient since the relationships among multiple (>2) terms cannot always be captured by the pairwise relationships of its subsets. For example, consider 3 terms: “zoo,” “tiger,” and “lion” in a given set of documents. The conditional probability between term pairs (e.g., between “zoo and tiger” and “zoo and lion”) presented in Sanderson and Croft can be quite low; hence this parent-child relationship cannot be detected. However, if one considers the relationship among the three terms all together, one can detect the hierarchy between “zoo” and “tiger and lion.”

This proposed method is important to S&T study since it can identify emerging areas that could be a fusion of two or more techniques.

6.2 Term Relationships

Relationships among terms can be categorized into four types: uncoupled, synonymy, parent-child, and sibling. The following Venn diagrams illustrate these relationships between two sets graphically.

[10] They actually used 0.8 instead to reduce the noise.

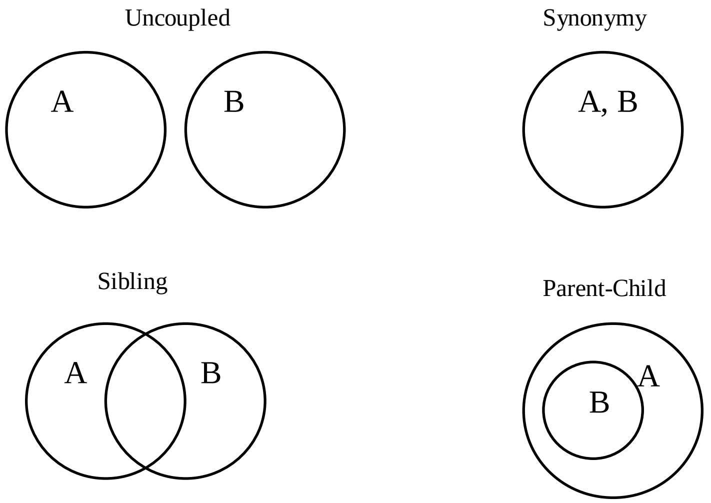


Figure 6.1: Venn Diagrams illustrating four different relationships of Two Sets, A and B

Note: Set A and B are determined by the number of documents containing the terms A and B respectively.

Uncoupled sets occur when both sets do not share any records. Synonymous sets occur when both items always occur together. Sibling sets share some records, and have other records not in common. In contrast, the last sets have a "parent-child" relationship; one set is entirely subsumed by the other.

Synonymous sets are interesting as part of data preprocessing (e.g., data cleansing or building a statistical thesaurus). Details on applying ARM to capture synonymous terms used for data preprocessing will be discussed in Chapter 7.

6.3 Tree-Structured Networks Algorithm

The algorithm for deriving parent-child and sibling relations using ARM consists of two steps. During the first step, clusters of related terms are generated. Then in the second step, parent-child and sibling relations among term clusters are identified.

Each *k*-frequent itemset that is generated by the algorithm described in Section 5.4, is used to represent each term cluster. These clusters are called *frequent term clusters*.

Parent-child relationships among clusters can be identified as follows:

1. Let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n distinct terms.
2. For $X, Y \subset T$ and $X \cap Y = \varnothing$, the set of terms Y is said to be 'parent of' (or 'more general than') the set of terms X if the rule X ⇒ Y has 100 percent confidence and the set of terms X is smaller than that of terms Y. These constraints can be expressed as,

$$X \Rightarrow Y = P(Y|X) = 1 \tag{6.1}$$

and

$$sup(X) < \varepsilon * sup(Y) \text{ , where } 0 < \varepsilon < 1 \tag{6.2}$$

In other words, Y contains X if the documents that X occurs in are a subset of the documents that Y occurs in. Because Y contains X and because it is more frequent, in the hierarchy, Y is the parent of X. On rare occasion when P(Y|X) = P(X|Y) = 1, Y and X are defined to be synonymous terms.

Although a good number of parent-child terms satisfy equation 6.1., many just fail to be included because a few occurrences of the 'child' term, X, do not co-occur with Y, hence P(Y|X) will rarely be equal to 1. As a result, the equality constraint is relaxed and redefined to be greater than or equal to a cutoff value called *minParent*. It is noted that the type of hierarchy obtained by the proposed algorithm depends critically on this cutoff value used to distinguish the various kinds of relations. The bigger *minParent* cutoff is, the stronger hierarchy relations are obtained. This number can be determined by empirical analysis between related term clusters. Note that using different data sets can result in different values of *minParent.*

Sibling relations between X and Y can be determined if the confidence of both ($X \Rightarrow Y$) and ($Y \Rightarrow X$) falls in between two cutoff values called *minSibling* and *maxSibling*. Again, the (*minSibling*, *maxSibling*) range is determined by empirical analysis between related term clusters. If the user needs to obtain strong sibling relations, the range should be narrower, namely the values for both *minSibling* and *maxSibling* should be high and close to each other. Note that *maxSibling* can be as high as 1. However, to distinguish between parent-child and sibling terms, *maxSibling* is set to equal *minParent.* As stated previously, terms X and Y are said to be synonyms if P(Y|X) = P(X|Y) =1. In the case when terms almost always occur together, this strict equality constraint is relaxed such that the two conditional probabilities are greater than or equal to a threshold called *minSynonym,* which is slightly less than 1. The basic tree-structured networks algorithm works as follows:

1. Find all frequent term clusters
2. For each cluster, generate association rules with any other clusters
3. To obtain Parent-Child relations, find association rules that satisfy:

   $X \Rightarrow Y = P(Y|X) \geq minParent$

   and

   $sup(X) < \varepsilon * sup(Y)$, where $0 < \varepsilon < 1$

   Y is then said to be 'parent' of X
4. To obtain sibling relations, find association rules between term clusters with $(minSibling \leq confidences < maxSibling)$.

6.4 Pruning the Discovered Relationships

After the tree-structured networks algorithm has been applied to the data set, many redundant and insignificant relationships are also captured. To effectively utilize the information, these redundant relationships should be removed. The pruning criteria include:

- If the term already appears as the parent or sibling, the combination that includes this specific term cannot appear as the child(ren).
- For a parent-child relationship, if there is a tie between two relationships where one term appears to be the parent in one relationship and the child in the other relationship, the relationship that has a parent with higher support is selected. This can be illustrated in the following figure.

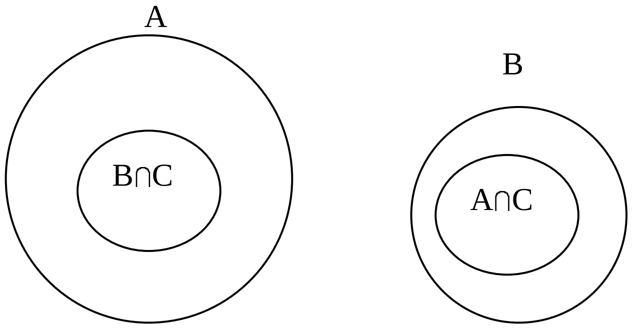


Figure 6.2: Example of Parent-Child relationships where a Term appear as Parent and Child

The first relationship shows A is the parent of (B and C). The second relationship shows B is the parent of (A and C). Since A has the higher frequency (i.e., A is more general than B), the first relationship will be selected and the second one will be discarded.

6.5 Experimental Results

The proposed tree-structured networks algorithm is tested on a set of publication abstracts. The results from the proposed method and those from Principal Component Analysis (PCA) and Hierarchical Agglomerative Clustering (HAC) are then compared. The factor map function in VantagePoint is used to generate the results of PCA. To obtain results for HAC using group average links, VantagePoint is linked with CLUTO-2.1 Clustering Toolkit library (Karypis 2002).

The data used in this experiment are the 971 association rule mining research abstracts retrieved from the *INSPEC* and *Engineering Index* databases. There are 158 selected keywords used for the experiment. Minimum support is set to be equal to 5 records, whereas maximum support is equal to 140 records. With this maximum support constraint, the top three most frequent keywords are removed. These terms consist of data mining (711), algorithms (166), and very large database (148). The number in parentheses represents the frequency of that particular keyword. The frequent keywords are excluded in the analysis because they do not contain other new and important information about the data set.

Using this data set, the cutoffs used in the tree-structured networks algorithm are varied to illustrate that different cutoff values yield different results. The user can determine which cutoff values are preferred based on these results. Table 6.1 and Table 6.2 show the results of parents-children and siblings term(s) using the tree-structure networks algorithm described in Section 6.3.

From Table 6.1, with the same parent term, the higher the *minParent* is set, the fewer children terms are generated from the proposed algorithm. This implies that with higher *minParent*, only the strongest hierarchy relations remain. However, some important relations might not be able to be captured if *minParent* is too high. For instance, the algorithm cannot detect the children terms for “association rules” if *minParent* is set to 1, while “apriori algorithms,” is identified as a child if *minParent* is reduced to 0.8. Hence, the user can decide which cutoff value to use based on the nature of the underlying data set and the results from the empirical analysis.

Table 6.1: Samples of concept hierarchy (Parent-Child) relations between term clusters using ARM for *minParent* = 1.0, 0.9 and 0.8

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
database systems	data acquisition, data structure	data acquisition, data structure	apriori algorithm
	knowledge discovery, learning systems	knowledge discovery, learning systems	data acquisition, data structure
			knowledge discovery, learning systems
knowledge acquisition			computer system programming
			association rule mining, data reduction, data structures
knowledge based systems			approximation theory fuzzy sets, learning systems
association rules			apriori algorithms
query processing			relational databases, software performance evaluation
fuzzy sets	membership functions	fuzzy association rules	fuzzy association rules
	database systems, membership functions	membership functions	membership functions
	database systems, fuzzy association rules	database systems, membership functions	database systems, membership functions
	fuzzy association rules, knowledge based systems	database systems, fuzzy association rules	database systems, fuzzy association rules
	knowledge based systems, membership functions	fuzzy association rules, knowledge based systems	fuzzy association rules, knowledge based systems
		knowledge based systems, membership functions	knowledge based systems, membership functions
			linguistic
			computational linguistic, fuzzy association rules
			approximation theory
relational databases			SQL

Table 6.2: Samples of sibling relations between term clusters using ARM for various cutoff levels

Term cluster	Term cluster's Siblings *minSibling* = 0.4 *maxSibling* = 0.8	Term cluster's Siblings *minSibling* = 0.3 *maxSibling* = 0.8
knowledge acquisition	deductive databases	deductive databases
asynchronous transfer mode	workstation cluster	workstation cluster
online system	data reduction, query languages	data reduction, query languages
	decision support system, query languages	decision support system, query languages
belief network	databases theory, learning (artificial intelligent)	databases theory, learning (artificial intelligent)
database systems		association rules
association rule mining, data reduction, data structure		data storage equipment, parallel processing systems
fuzzy association rules		computational linguistics
membership function		boolean function
		linguistic
database systems, membership function		boolean function
linguistic		membership function
		knowledge based systems, membership function
information resources		internet
parallel processing systems		data storage equipment
artificial intelligence		neural networks
interactive systems		user interface
bibliographic system		medical information system
computer software		distributed computer systems
sales management		deductive databases, software performance evaluation
		market data processing, transaction processing
		deductive databases, transaction processing, knowledge acquisition
		deductive databases, knowledge acquisition, software performance evaluation
statistic databases		distributed databases, transaction processing
hypermedia		information resources, search engine

Table 6.2 shows that more sibling terms can be discovered if the interval between *minSibling* and *maxSibling* is wider. Again the user can decide which value to use based on these empirical results.

Table 6.3 and Table 6.4 illustrate some of the term clusters generated from PCA and HAC respectively. The complete results can be found in Appendix A1 and A2. From the experimental results, ARM and PCA yield more similar term clusters than HAC. A cluster derived from HAC tends to consist of more terms than that from ARM and PCA. For instance, Table 6.1 and Table 6.2 show that, using ARM with *minParent* = 0.8, *minSibling* = 0.3, *maxSibling* < 0.8, “database systems” relates to “data acquisition,” “data structure,” “knowledge discovery,” “learning systems,” “association rules,” and “ apriori algorithms.” These terms are quite similar to terms generated from PCA (see Cluster 5 in Table 6.3). HAC also generates similar terms and many more as can be seen from Cluster 2 in Table 6.4. The reason for HAC to generate bigger clusters is that HAC begins with each term in a cluster and successively merges clusters together until a stopping criterion is satisfied; hence, every term will potentially be included in the cluster results. PCA and ARM, on the other hand, cluster only terms that are highly correlated/co-occurring.

The difference between PCA and ARM is that the former identifies related terms (non-hierarchical structure) using their correlations while the latter uses raw co-occurrence between terms to capture both parent-child (hierarchical structure) and sibling relationships (non-hierarchical structure). For instance, Cluster 5 in Table 6.3 suggests “database systems,” “association rules,” “knowledge discovery,” “theorem proving,” and “Apriori algorithms” are highly correlated based on PCA. ARM could discern a hierarchical relation suggesting, “database systems” is the parent of 1) “data acquisition and data structure,” 2) “knowledge discovery and learning systems,” and 3) “Apriori algorithms”. Moreover, ARM can capture the sibling relation between “database systems” and “association rules.” As a result, ARM promises to offer richer structural information on relationships in text sources.

Table 6.3: Examples of the Term clusters generated from PCA

Cluster 1	Query languages, Data reduction, Distributed database systems, Indexing (of information), Online systems, EiRev
Cluster 2	Fuzzy sets, Computer simulation, Computational linguistics, Fuzzy association rules, Membership functions, Linguistics, Approximation theory
Cluster 3	query processing, relational databases, Data warehouses, software performance evaluation, Decision support systems, SQL, Parallel programming, Asynchronous transfer mode, parallel databases, workstation clusters
Cluster 4	information resources, Information retrieval, Internet, Search engines, hypermedia
Cluster 5	Database systems, Association rules, Knowledge discovery, Theorem proving, Apriori algorithms
Cluster 6	Knowledge acquisition, deductive databases, transaction processing, marketing data processing, retail data processing, sales management
Cluster 7	Knowledge discovery, Classification (of information)

Table 6.4: Examples of the Term clusters generated from HAC with group average linking

Cluster 1	information resources, Information retrieval, Internet, Electronic commerce, World Wide Web, data visualization, Search engines, Text analysis, Visualization, document handling, Software agents, classification, constraint handling, distributed algorithms, Graphical user interfaces, hypermedia, multimedia databases
Cluster 2	Database systems, Knowledge based systems, Associative processing, Association rules, Data structures, Fuzzy sets, Set theory, Association rule mining, Computational complexity, Query languages, Data reduction, Learning systems, Knowledge discovery, Relational database systems, Computer simulation, Trees (mathematics), Parallel processing systems, Computational linguistics, tree data structures, Data acquisition, Data processing, Data storage equipment, Fuzzy association rules, Information retrieval systems, Classification (of information), Knowledge engineering, Learning algorithms, Problem solving, Statistical methods, Boolean functions, Decision making, Membership functions, Semantics, Theorem proving, Response time (computer systems), Security of data, Apriori algorithms, Computer networks, Distributed database systems, Indexing (of information), Online systems, Performance, Computational methods, Computer architecture, Computer software, Distributed computer systems, Inference engines, Linguistics, Sequences, Storage allocation (computer), Approximation theory, Computer systems programming, Expert systems, Hierarchical systems, Remote sensing
Cluster 3	pattern clustering, fuzzy set theory, pattern classification, fuzzy logic, Temporal databases, visual databases, uncertainty handling, learning by example
Cluster 4	Knowledge acquisition, database theory, query processing, deductive databases, relational databases, transaction processing, Data warehouses, software performance evaluation, data analysis, Decision support systems, very large, marketing data processing, retail data processing, SQL, business data processing, distributed databases, optimization, data models, Management information systems, parallel databases, time series, database indexing, sales management, statistical databases, Risk management, Text processing

The difference between HAC and ARM is that HAC decomposes terms into several levels of nested partitioning (tree of sub-clusters). Each leaf node in an HAC tree represents each single term. This nested tree shows how the clusters are merged hierarchically. The topmost cluster is considered a ‘parent’ cluster. However, given a sub-cluster (child), there is no easy way to discern which terms are meaningful parent names. ARM, on the other hand, derives parent-child relationships between terms and labels them according to the conditional probability between those terms. Some remarkable features of ARM that are worth noticing include:

- ARM does not require transformation of input data, instead it uses raw occurrence between input data.
- Tree-structured networks can be derived from a set of documents without use of data reduction or standard clustering techniques.
- ARM can control the quality of clusters according to the requirements of users and domains. With different levels of minimum support and confidence, the amount of relationship captured in ARM algorithm can be adjusted.

However, a drawback of ARM is that the appropriate support level and the threshold values to distinguish parent-child, synonymous, and sibling terms are subjective and largely depend on the application domain.

6.6 Threshold Testing

To illustrate how the nature of the data, such as original source of the data (database) and specific vs. broader topic, affects the threshold values, an experiment is performed by varying level of thresholds with different data sets. The results are shown in the following section:

6.6.1 Experimental Data sets

6.6.1.1 Data Set1

ARM (INSPEC only): 725 records, minsup = 5 records, maxsup = 100 records (the top two most frequent keywords -- data mining and very large databases -- are removed)

Table 6.5: All concept hierarchy (Parent-Child) relations between term clusters of ARM data set (from INSPEC only) for *minParent* = 1.0, 0.9 and 0.8

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
relational databases	software performance evaluation, SQL	software performance evaluation, SQL	software performance evaluation, SQL
		query processing, SQL	query processing, SQL
			SQL
parallel algorithms			resource allocation
query processing			software performance evaluation, SQL

For *minParent* = 1.0 and 0.9, the derived parent-child relationships are almost the same. However, when *minParent* is reduced to 0.8, more parent-child relationships are detected and more reasonable children terms are captured within the same parents. The *minParent* was also reduced to 0.75 and 0.70 (results not shown in the table). Despite the lower level, the results remained the same as when *minParent* = 0.8.

6.6.1.2 Data Set 2

ARM (ENGI only): 420 records, minsup = 5 records, maxsup = 170 records (the top two most frequent keywords -- data mining and algorithms -- are removed)

The results are shown in Table 6.6. More parent-child relationships are derived from ENGI data than from INSPEC. For *minParent* = 1.0 and 0.9, the derived parent-child relationships are the same. However, when *minParent* is reduced to 0.8, more parent-child relationships are detected and more reasonable children terms are captured within the same parents.

Table 6.6: All of concept hierarchy (Parent-Child) relations between term clusters of ARM data set (from ENGI only) for *minParent* = 1.0, 0.9 and 0.8

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
database systems	data acquisition, data structure	data acquisition, data structure	apriori algorithms
	knowledge discovery, learning systems	knowledge discovery, learning systems	associative processing, computational complexity
			computer systems programming, knowledge acquisition
			data acquisition, data structure
			knowledge discovery, learning systems
fuzzy sets	membership functions	membership functions	membership functions
	fuzzy association rules, database systems	fuzzy association rules, database systems	fuzzy association rules, database systems
	membership functions, database systems	membership functions, database systems	membership functions, database systems
	fuzzy association rules, knowledge based systems	fuzzy association rules, knowledge based systems	fuzzy association rules, knowledge based systems
	membership functions, knowledge based systems	membership functions, knowledge based systems	membership functions, knowledge based systems
		fuzzy association rules	fuzzy association rules
			computational linguistics, fuzzy association rules
			approximation theory
association rules			apriori algorithms
knowledge acquisition			computer systems programming
			data reduction, data structure, mining association rules
			data acquisition, data reduction
decision theory			decision trees
classification of information			database mining
knowledge based systems			approximation theory
			fuzzy sets, learning systems

The data set used to derive parent-child relationships in the above section is quite specific (i.e., related to ARM research). The results show that the preferred *minParent* value is 0.8. The next section will illustrate the results when a different data set is used. This second data set is a generally broader topic (i.e., related to Thai research).

6.6.1.3 Data set 3

Thai (INSPEC): 2686 records, minsup = 10 records, no maxsup (all keywords are used in the analysis)

Table 6.7: All concept hierarchy (Parent-Child) relations between term clusters of Thai data set (from INSPEC only) for *minParent* = 1.0, 0.9 and 0.8

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
superconducting transition temperature	barium compound, high-temperature superconductors	barium compound, high-temperature superconductors	barium compound, high-temperature superconductors
		BCS theory, superconducting energy gap	high-temperature superconductor
			BCS theory
			specific heat of solid
			strong-coupling superconductors
high-temperature superconductors	bismuth compound, strontium compound	bismuth compound, strontium compound	bismuth compound, strontium compound
	strontium, superconducting transition temperature	strontium, superconducting transition temperature	strontium, superconducting transition temperature
	bismuth compound, superconducting transition temperature	bismuth compound, superconducting transition temperature	bismuth compound, superconducting transition temperature
synthetic aperture radar	radar imaging, remote sensing by radar	radar imaging, remote sensing by radar	radar imaging, remote sensing by radar
III-V semiconductors	gallium arsenide	gallium arsenide	gallium arsenide
Solar absorber- convertors		solar heating	solar heating
Antenna radiation patterns		slot antenna arrays	slot antenna arrays
			electromagnetic wave polarisation
DC-AC power convertors		power conversion harmonic	power conversion harmonic
light emitting diodes		amorphous semiconductors	amorphous semiconductors
			silicon compounds
silicon compounds			amorphous semiconductors
photovoltaic power systems		solar cell arrays	solar cell arrays
hydrogen		silicon compound	silicon compound
		amorphous semiconductors	amorphous semiconductors
machine control			velocity control
harmonic distortion		dc-ac power converters, power conversion harmonic	power conversion harmonic
			dc-ac power converters, power conversion harmonic
geophysical techniques			geophysical signal processing
integral equations			method of moments

The derived parent-child relationships are very similar when *minParent* is set to 0.9 or 0.8. The user can decide whichever level to use.

6.6.1.4 Data set 4

Thai (ENGI): 2224 records, minsup = 10, max sup = 170 records (the top two most frequent keywords -- mathematical models and computer simulation-- are removed)

Table 6.8: All concept hierarchy (Parent-Child) relations between term clusters of Thai data set (from ENGI only) for *minParent* = 1.0, 0.9 and 0.8 and *minsup* = 10

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
High temperature superconductors	oxide superconductors	oxide superconductors	oxide superconductors
rubber		latexes	latexes
			vulcanization
			crosslinking, vulcanization
synthetic aperture radar		radar imaging	radar imaging
moisture			grain (agriculture product)
drying			dryers (equipments)
plastic fillers			calcite

Since there are far fewer relations derived, *minsup* is then lowered to 7 records in order to observe the different behaviors. As expected, more parent-child relationships are captured when *minsup* is reduced. Also more children terms are derived when *minParent* is 0.8.

Table 6.9: All concept hierarchy (Parent-Child) relations between term clusters of Thai data set (from ENGI only) for *minParent* = 1.0, 0.9 and 0.8 and *minsup* = 7

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
high temperature superconductors	oxide superconductors	oxide superconductors	oxide superconductors
rubber		latexes	vulcanization
			latexes
			crosslinking, vulcanization
			carbon black
			rheology, viscosity
synthetic aperture radar		radar imaging	radar imaging
moisture			grain (agriculture product)
			fluidization
drying	fluidization	fluidization	fluidization
			dryers
plastic fillers			calcite
communication channels (information theory)	convolutional codes	convolutional codes	phase shift key
			convolutional codes
speech recognition	speech analysis	speech analysis	speech analysis
code division multiple access			bit error rate, code (symbolic)
water filtration			deep-bed filtration
robustness (control systems)			uncertain systems

6.6.1.5 Data set 5

Thai (SCI): 12126 records, minsup = 10, no maxsup (all keywords are used)

Table 6.10: All concept hierarchy (Parent-Child) relations between term clusters of Thai data set (from SCI only) for *minParent* = 1.0, 0.9, 0.8 and 0.7 and *minsup* = 10

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)	Children cluster (with *minParent* = 0.7)
plasmodium-falciparum		artesunate, chemotherapy	artesunate, chemotherapy	artesunate, chemotherapy
scrub typhus		orientia tsutsugamushi	orientia tsutsugamushi	orientia tsutsugamushi
phamacokinetics				phamacodynamics
heavy metals				sewage sludge

With *minsup* = 10, if *minParent* = 1 then there are no relations generated. If *minParent* = 0.9, there are 2 relations generated. *Minsup* is again lowered to 7 records in order to produce more relationships.

Table 6.11: All of concept hierarchy (Parent-Child) relations between term clusters of Thai data set (from SCI only) for *minParent* = 1.0, 0.9 and 0.8 and *minsup* = 7

Parent cluster	Children cluster (with *minParent* = 1.0)	Children cluster (with *minParent* = 0.9)	Children cluster (with *minParent* = 0.8)
plasmodium-falciparum	artesunate, mefloquine	artesunate, mefloquine	artesunate, mefloquine
	artesunate, artemether	artesunate, artemether	artesunate, artemether
	chemotherapy, mefloquine	chemotherapy, mefloquine	chemotherapy, mefloquine
		artesunate, chemotherapy	artesunate, chemotherapy
			drug resistance, mefloquine
penaeus-monodon	black tiger shrimp	black tiger shrimp	black tiger shrimp
scrub typhus		orientia tsutsugamushi	orientia tsutsugamushi
mefloquine			drug interactions, phamacokinetics
guttiferae			xanthones
human immunodeficiency virus			acquired immunodeficiency syndrome
apocynaceae			indole alkaloids

The above experiment shows that with different kinds of data sets, the preferable level of *minParent* level is often 0.8 since reasonable numbers of parent-child clusters are generated. However, it cannot be concluded that 0.8 should be used for every data set. As mention previously, the user should decide which cutoff value to use based on the nature of the underlying data set and the results from the empirical analysis. The data behavior can be weighed against the analytical aims to determine suitable cutoffs.

6.7 Temporal Factor

Time can be taken into account as a factor for parent-child relationships. Ideally, a parent should always occur before its children. A parent-child relationship derived from ARM with *minParent* level equal to 100 percent is generated such that both parent and its child(ren) always co-occur together. Hence, it guarantees that child(ren) term(s) will never occur before their parent. However, when the *minParent* threshold is relaxed to allow some children terms that might not always occur with their parent, there is the possibility that some children terms might occur before the parent. However, this possibility is very small and its effect on the derived parent-child relationship is usually not significant. To illustrate this claim, the distributions of selected parent-child terms with *minParent* set to 0.8 are plotted in the following figure.

The following three figures show the distribution of the terms "association rules, relational databases, and database systems" with their derived children, respectively. The first two plots show that the children terms (Apriori algorithm and SQL) first appear at the same time as their parents (association rules and relational databases). In the last plot, the parent term, database systems, occurs before all of its children. In addition, the numbers of records of parent terms are much higher than the children terms most of the time in these three figures. This illustrates the emergence of new areas (children) from the existing research (parent).

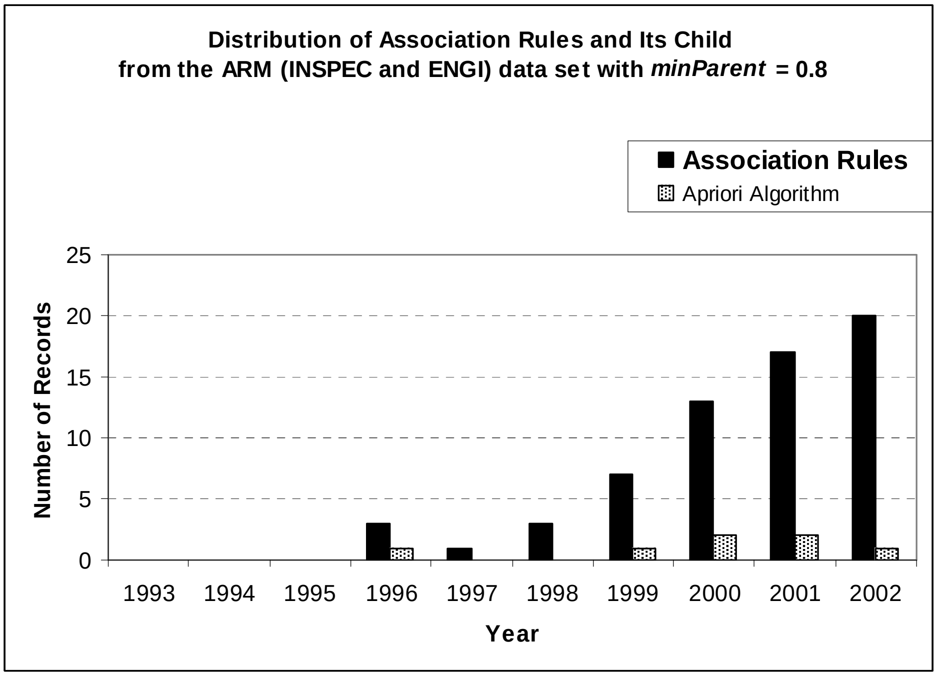


Figure 6.3: Distribution for “Association Rules” and Child

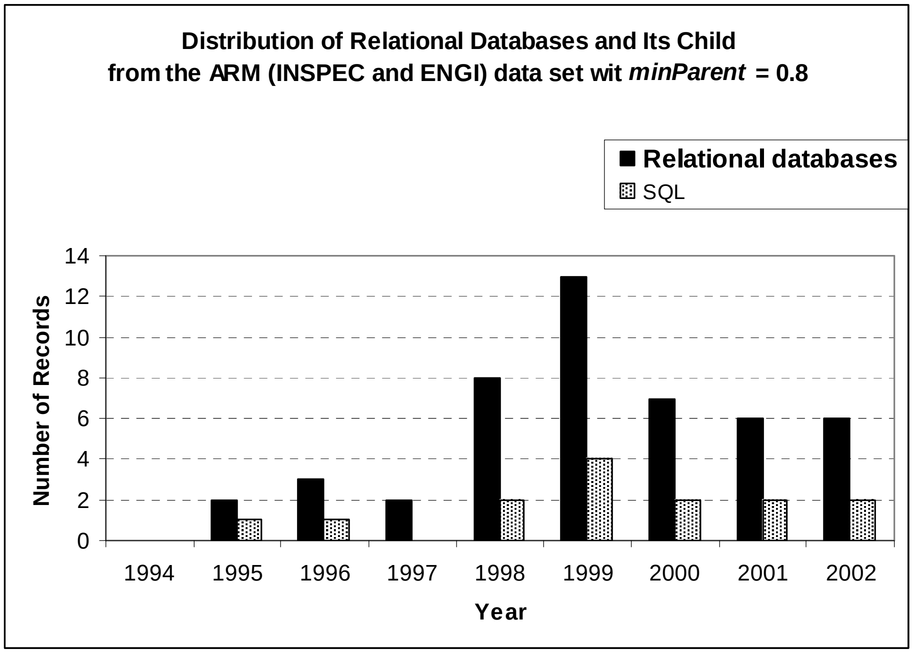


Figure 6.4: Distribution of “Relational Databases” and Child

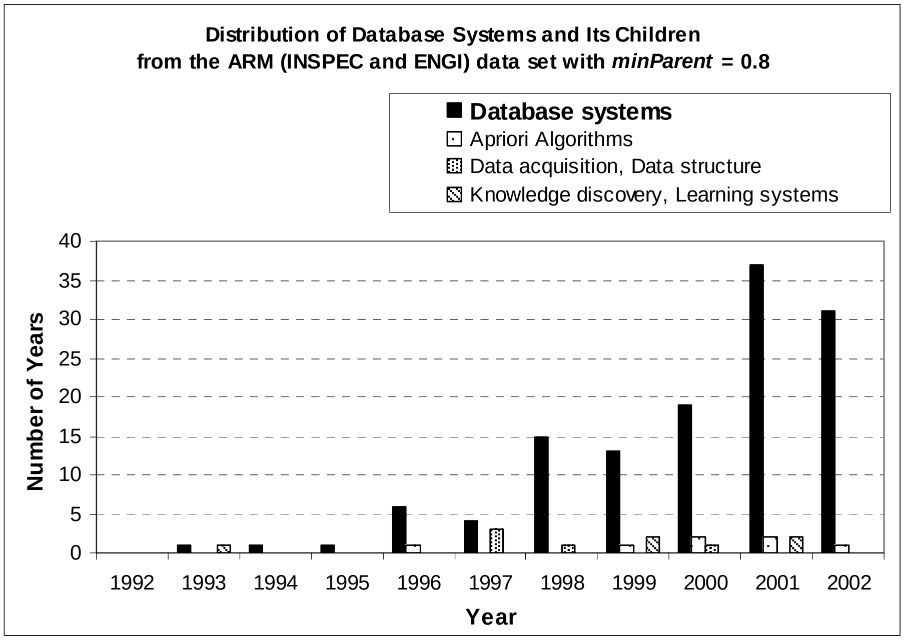
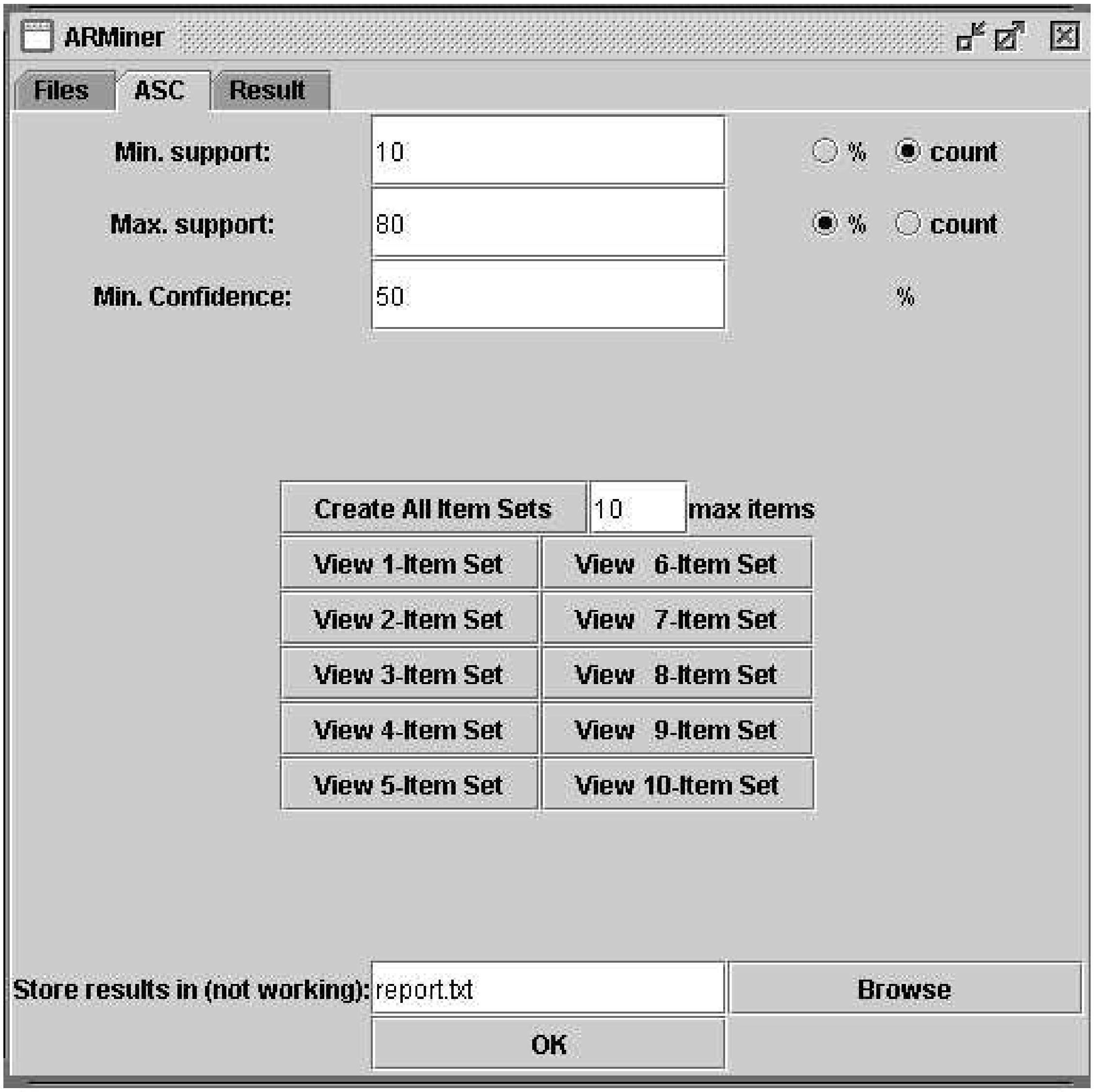


Figure 6.5: Distribution of “Database Systems” and Children

6.8 Presentation of Parent-Child and Sibling Relationships

It is simple to lay out a small tree-like graph structure on screen, however, the complete tree-structured networks being generated could be much larger. With current visualization techniques, it is possible to lay out the complete tree-structure on the screen. Although it might be visually pleasing to see the entire tree-structured networks structure laid out, it is not entirely necessary. An alternate way that only shows the tree of the topic the user is interested in might be sufficient.

Figure 6.6 shows the user interface developed from JAVA language. In this interface, the user can input his/her desire level of minimum support, maximum support, and minimum confidence.



Min. support: 10 ○ % ● count

Max. support: 80 ● % ○ count

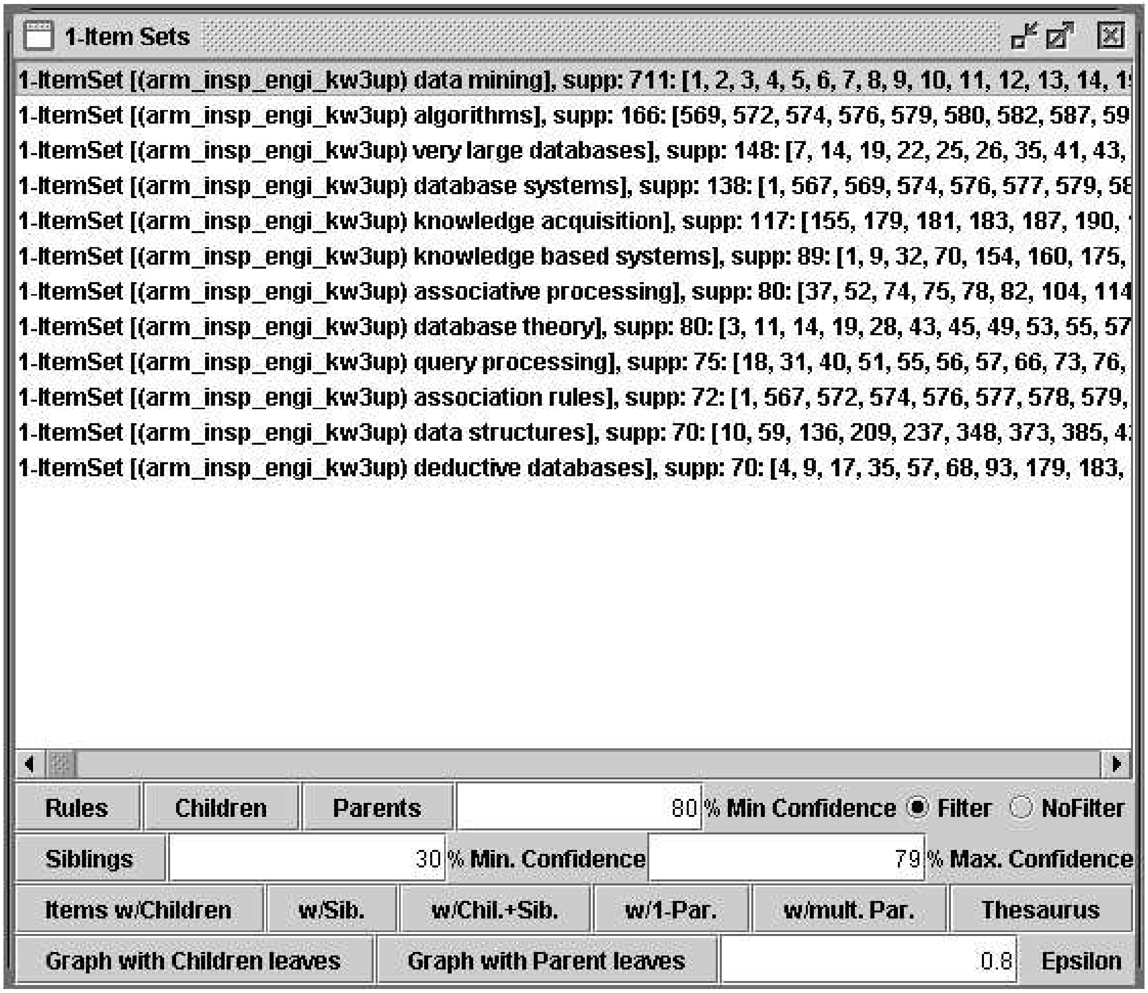
Min. Confidence: 50 %

Create All Item Sets 10 max items

Store results in (not working): report.txt

Figure 6.6: User Interface for the proposed ARM algorithm

Next, the user can choose to view the *k*-itemset generated from the algorithm. Figure 6.7 shows the sample of the viewer window for 1-itemset.



```
1-ItemSet [(arm_insp_engi_kw3up) data mining], supp: 711: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1
1-ItemSet [(arm_insp_engi_kw3up) algorithms], supp: 166: [569, 572, 574, 576, 579, 580, 582, 587, 59
1-ItemSet [(arm_insp_engi_kw3up) very large databases], supp: 148: [7, 14, 19, 22, 25, 26, 35, 41, 43,
1-ItemSet [(arm_insp_engi_kw3up) database systems], supp: 138: [1, 567, 569, 574, 576, 577, 579, 58
1-ItemSet [(arm_insp_engi_kw3up) knowledge acquisition], supp: 117: [155, 179, 181, 183, 187, 190,
1-ItemSet [(arm_insp_engi_kw3up) knowledge based systems], supp: 89: [1, 9, 32, 70, 154, 160, 175,
1-ItemSet [(arm_insp_engi_kw3up) associative processing], supp: 80: [37, 52, 74, 75, 78, 82, 104, 114
1-ItemSet [(arm_insp_engi_kw3up) database theory], supp: 80: [3, 11, 14, 19, 28, 43, 45, 49, 53, 55, 57
1-ItemSet [(arm_insp_engi_kw3up) query processing], supp: 75: [18, 31, 40, 51, 55, 56, 57, 66, 73, 76,
1-ItemSet [(arm_insp_engi_kw3up) association rules], supp: 72: [1, 567, 572, 574, 576, 577, 578, 579,
1-ItemSet [(arm_insp_engi_kw3up) data structures], supp: 70: [10, 59, 136, 209, 237, 348, 373, 385, 4
1-ItemSet [(arm_insp_engi_kw3up) deductive databases], supp: 70: [4, 9, 17, 35, 57, 68, 93, 179, 183,
```

80 % Min Confidence ● Filter ○ NoFilter

30 % Min. Confidence 79 % Max. Confidence

0.8 Epsilon

Figure 6.7: Viewer Window for 1-Itemset

From the above figure, the candidate itemsets that pass the minimum support and confidence are presented. The corresponding record numbers for each itemset are also available. The lower part of the screen contains the requirement for parent-child and sibling relationships. The user can input his/her desired level of *minParent*, *minSibling*, and *maxSibling*. The user can choose to display the list of itemsets that have children only, siblings only, both children and siblings, single parent only, and multiple parents.

Lastly, the user is able to visualize the hierarchical graph for each topic of his/her interest. There are two available types of hierarchical graphs. The first one displays a single parent and multiple children. This helps the user identify new or more specific areas in an existing broader research domain. The second graph shows a child and multiple parents. This graph helps the user detect the multidisciplinary research where some techniques, concepts, or results are used in different domains.

After the user selects these options, the new viewer window is created. Figure 6.8 and Figure 6.9 illustrate the example of a hierarchical graph for a single parent with multiple children and a single child with multiple parents, respectively. Figure 6.8 shows the new popup window, which presents some of the children of "data mining" when minimum support is set to 10. Note that "data mining" is the most frequent term in this data set. Thus, more children (not shown in the graph) are captured. The selected children terms are shown in Figure 6.8 to illustrate that a multi-level graph can be created.

Figure 6.9 shows the new popup window, which presents the two parents of "Apriori algorithms". From this graph, the user is able to see that Apriori algorithms can be used in both database systems and association rules research domains. These two presentations can help the user visualize the research domain of interest more effectively.

In this type of interface, the user is able to control the level of desired threshold. Each layer of the topic hierarchy can be examined clearly. However, if the user needs to visualize the entire tree-structured network, alternative means of displaying the structure have to be explored. This is beyond the scope of this dissertation. In the future, more effective algorithms that can display the entire tree-structure networks can be implemented.

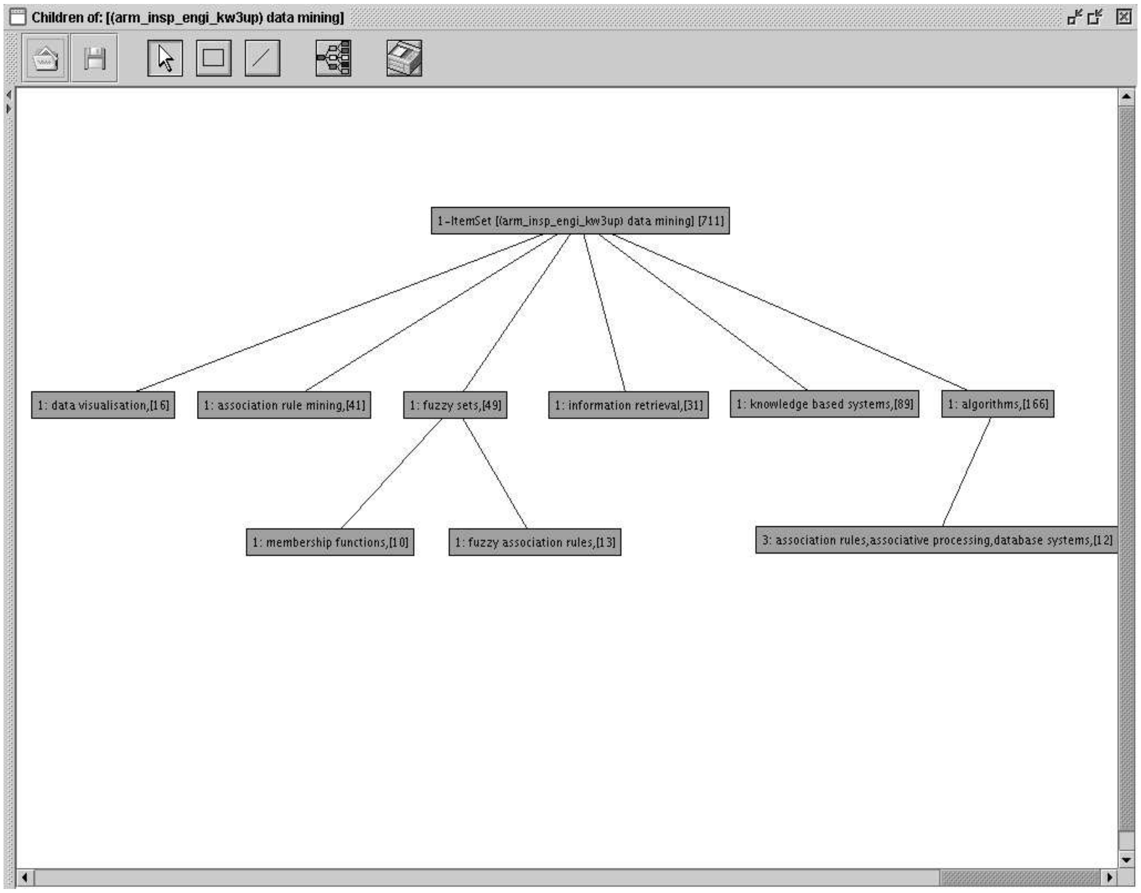

Children of: [(arm_insp_engi_kw3up) data mining]
1-ItemSet [(arm_insp_engi_kw3up) data mining] [711]
1: data visualisation,[16]
1: association rule mining,[41]
1: fuzzy sets,[49]
1: information retrieval,[31]
1: knowledge based systems,[89]
1: algorithms,[166]
1: membership functions,[10]
1: fuzzy association rules,[13]
3: association rules,associative processing,database systems,[12]

Figure 6.8: Example of Hierarchical Graph for a Single Parent with Multiple Children

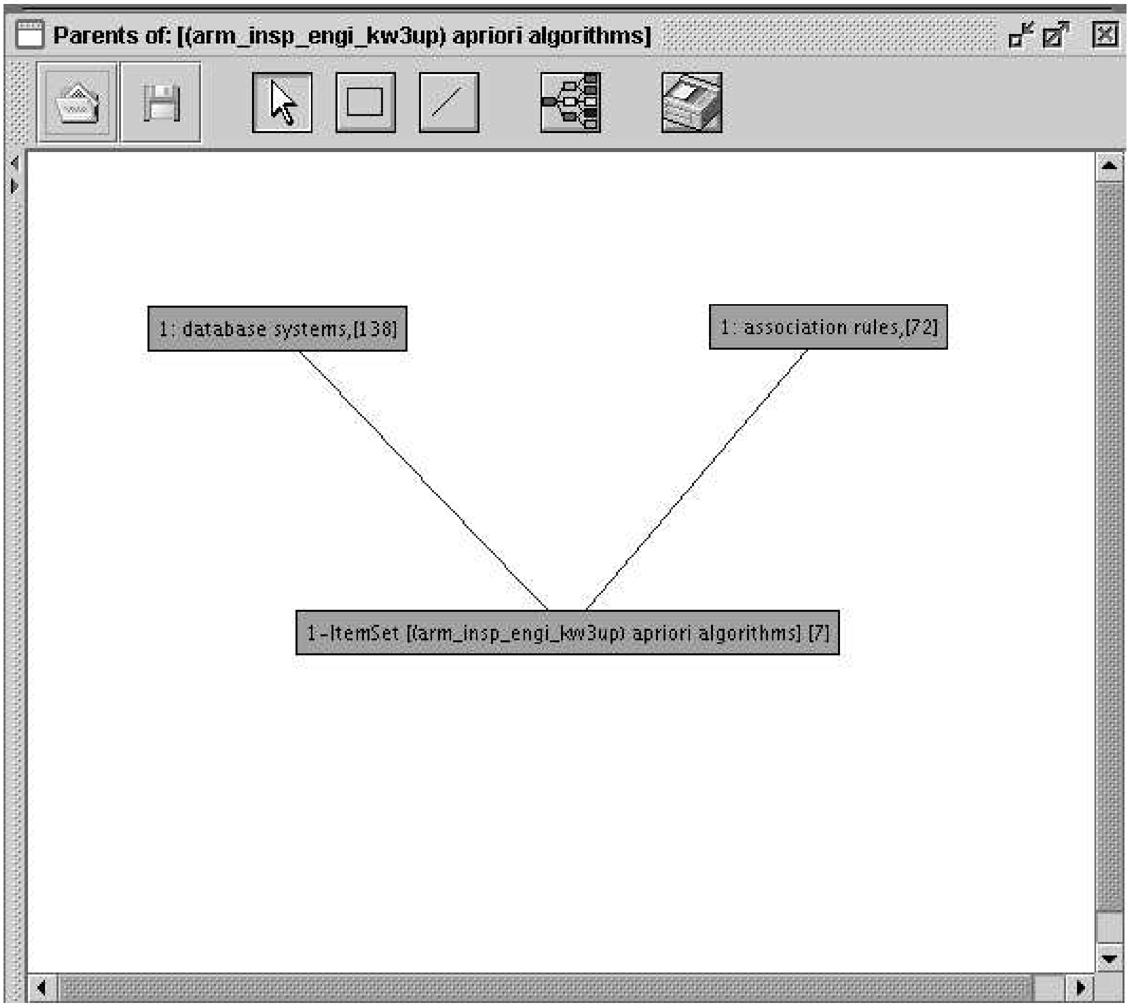

Parents of: [(arm_insp_engi_kw3up) apriori algorithms]
1: database systems,[138]
1: association rules,[72]
1-ItemSet [(arm_insp_engi_kw3up) apriori algorithms] [7]

Figure 6.9: Example of Hierarchical Graph for a Child with Multiple Parents

6.9 Conclusion

In this chapter, a tree-structured networks algorithm is proposed. The algorithm applies the association rule mining (ARM) technique to discern conceptual relationships (parent-child and sibling) from text data sets. Note that the derived parent-child relations will depend on the data set. For instance, if the data set covers general aspect of mathematics then relationship such as mathematics is the parent of mathematical statistics can be captured. However, if the data set covers only specific area in mathematical statistics, then one might not be able to capture mathematics as the parent of mathematical statistics. In stead, the child(ren) of mathematical statistics will be captured.

The results from the proposed algorithm are compared with those from Principal Component Analysis (PCA) and Hierarchical Agglomerative Clustering (HAC), two well-known clustering techniques. The advantage of ARM over PCA is that PCA can only identify related terms (non-hierarchical structure) using their correlations (or other defined similarity functions) while ARM uses raw co-occurrence between terms to discern both parent-child (hierarchical structure) and sibling relationships (non-hierarchical structure). As a result, ARM promises to offer richer structural information on relationships in text data sets. The advantage of ARM over HAC is that HAC decomposes terms into several levels of nested partitioning (tree of sub-clusters). This nested tree only shows how the clusters are merged hierarchically. Thus, given a child term, one cannot identify the meaningful parent names for that child. ARM, on the other hand, constructs parent-child relationships between terms and labels them according to the conditional probability between them. Regardless of the advantages of ARM, one drawback of this approach is that the appropriate support level and the threshold values to distinguish parent-child, synonymous, and sibling terms must be determined by empirical analyses among related terms. Hence they are subjective and largely depend on the data set.

CHAPTER 7

TERM THESAURUS CONSTRUCTION

Chapter 6 has already presented details on how to generate tree-structured networks from text sources by applying the association rule mining technique. This chapter will focus on developing a *concept-grouping algorithm* to construct a thesaurus from synonymous terms. For the purposes of this research, a "word" is a sequence of letters separated by spaces, a "phrase" consists of one or more words, and a "term" is a phrase that is extracted from title or abstract by VantagePoint. A thesaurus is defined as a grouping of terms, into certain concepts. It is my hypothesis that applying ARM to title or abstract phrases can enhance data preprocessing (e.g., data cleansing or building a statistical thesaurus). Similar abstract or title phrases, which are written differently, can be grouped together into a single concept based on their associations before they are used for further analysis. This data preprocessing is important since it has an impact on the quality of other data mining techniques such as clustering. VantagePoint already has a list cleanup function based on word stemming to group like terms into one. However, this method is insufficient since it cannot detect the similarity between "Internet" and "world wide web." The proposed method is believed to be more efficient since similar terms are grouped together into a concept based on the probability that they occur together. Combining both approaches is most promising.

The intentions for using a concept-grouping algorithm to construct thesaurus for abstract or title phrases are to combine similar terms into a pertinent concept and to increase the frequency (number of relevant documents) of abstract or title phrases before they are used in clustering. Abstract or title phrases with technical words sometime capture the technical themes, which the author intends to present, better than keywords indexed by a database provider. The bias and error introduced by the database providers can affect the performance of the techniques using keywords only. However, the problem with using abstract or title phrases in the analysis is that the frequencies of the phrases are quite low because of the variation in the terms that are used. The low frequency terms might be discarded in other analyses such as clustering.

7.1 Concept-Grouping Algorithm

To make the algorithm more efficient, only the informative phrases should be used. Hence, stopwords (i.e., terms that occur too frequently and are unimportant to the database content) should be removed from the list of phrases. These terms are determined by a published list of the most frequently used words from two to ten letters (see http://www.calendarhome.com/wordlist.html).

The algorithm for deriving synonymous terms using ARM consists of two steps. The first step is to generate all *frequent term clusters* as explained in Section 5.4. In the second step, synonymous terms can be captured as follows:

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n distinct terms. For $X, Y \subset T$ and $X \cap Y = \varnothing$, the set of terms Y is said to be a synonym of the set of terms X if both rules $X \Rightarrow Y$ and $Y \Rightarrow X$ have 100 percent confidence. These constraints can be expressed in terms of conditional probability as,

$$P(Y|X) = P(X|Y) = 1 \tag{7.1}$$

In other words, Y and X are synonyms if they always occur together. However, the constraint in Equation 7.1 should be relaxed to cover some cases when terms almost always occur together. The relaxed constraint is defined as,

$$P(Y|X) \text{ and } P(X|Y) \geq \mathit{minSynonym} \tag{7.2}$$

The quantity *minSynonym* is determined by empirical analysis among a set of related terms.

7.2 Experimental Results

The proposed algorithm is tested on the same set of publication abstracts as Section 6.4. Minimum support is set to 4 records and *minSynonym* is set to 0.9.

Table 7.1: Samples of synonymous terms for association rule mining research

	Synonymous Terms	Final Frequency
Group 1	conventional scientific calculation (4), extra memory (4), main memory (9), node executing applications (4), parallel data mining (5), pc clusters (10), real memory content (4), real memory size (4), storage device (4), swap area (4)	15
Group 2	data mining functions (4), interactive mining (6), wide sprectrum (5)	7
Group 3	handle text databases (4), large text databases (6), mining text databases (4), text databases (6)	8
Group 4	mining association rules focuses (4), association graphs (6), sufficient number (10)	12
Group 5	novel technique (4), exceptions (7), fuzzy set theory (10), user-supplied thresholds (4)	13
Group 6	representative ones (4), cover operator (5), important database mining problem (9)	10
Group 7	software system (4), subsystems (4)	4

Table 7.1 shows some promising results from the test data set. The number in parentheses shows the frequency of each single phrase. The final frequency count represents the frequency of records after the synonymous terms are merged. By considering Group 3 from the above table, “handle text databases, large text databases, mining text databases, and text databases” are considered synonyms based on the proposed algorithm. Despite the apparent differences, a document identified by these four terms can nevertheless be assigned to the same group. From this data set, there are 4, 6, 4, and 6 documents containing the phrase “handle text databases,” “large text databases,” “mining text databases,” and “text databases” respectively. If these four terms are grouped together, the total frequency is equal to 8 documents. Note that the total frequency need not be equal to the sum of each term’s frequency since two or more terms can occur in the same document. Suppose the minimum frequency to be used in a factor map is equal to 5, without synonymous term detection, only six documents will be included while eight documents will be used if these synonymous terms are considered. Hence, using ARM to detect synonymous terms could help increase the frequency of important phrases used in factor analysis. This will greatly affect the quality of the factor analysis performance if the data set is larger.

However, some groups of abstract phrases from Table 7.1 consist of common terms such as “novel technique” and “important database mining problem.” Since these terms do not provide any useful information to the derived synonymous group, they should be removed before applying the proposed algorithm. It appears that there might be two minor problems associated with the proposed algorithm. The first problem is that, for some data sets, the minimum support might have to be set at a very low level for the synonymous terms to be captured. In such cases where the minimum support is very low (e.g., terms occurring with one another in 2 records), one cannot be confident to conclude that these terms are synonyms. Moreover, derived synonymous terms tend to be very noisy.

To illustrate this problem, the proposed algorithm is tested on a new data set. This data set consists of 817 records that relate to information visualization research. Abstract phrases that occur in 4 or more records (405 phrases) are selected for the experiments. When minimum support is set to 5 records and *minSynonym* is set to 0.9, no synonymous terms can be identified. Table 7.2 presents the result when the minimum support is lowered to 4 records.

Table 7.2: Result of synonymous terms for information visualization research with *minsup* = 4

	Synonymous Terms
Group 1	SFA[11], similar documents

The result shows that only one group can be derived from this experiment. This implies that the minimum support level could be too high. In order to see how minimum support level can affect the outcome, this user-specified threshold is set to be lower than 4 records. Table 7.3 presents the results after the minimum support is set to 3 records.

[11] SFA stands for Stereoscopic Field Analyzer

Table 7.3: Result of synonymous terms for information visualization research with *minsup* = 3

	Synonymous Terms
Group 1	SFA, similar documents
Group 2	Human conceptual system, cone trees, information structured hierarchically, interactive animation, large information spaces,
Group 3	Immersive interactive system, information displays, SFA, similar documents, textual documents, traditional 2D, two-handed interaction, user's three dimensional perception, uses glyph-based volume rendering
Group 4	stereoscopic viewing, new system, SFA, similar documents

From the above table, the results seem to be noisy after the minimum support is reduced to 3 records. For instance, Group 1, Group3 and Group 4 share the same terms, SFA and similar documents, hence it appears that these groups might represent the same concept. Therefore it is ambiguous to tell which terms are the synonyms of "SFA and similar documents."

The second problem associated with the proposed algorithm is that it appears to be problematic to assign a name or concept to each group of synonymous terms. As a result, an improved method to group similar terms into a meaningful concept is proposed. The next section details this improved algorithm.

7.3 Improved Concept-Grouping Algorithm

The improved algorithm will probe for strong parent-child relationships between keywords and abstract phrases instead of using only the probability that phrases are occurring together, as described in Section 7.1. Note that this analysis therefore uses two distinct data fields together. It is my hypothesis that synonymous terms can be grouped into a meaningful concept by discovering strong parent-child relationships between keyword and abstract phrases where keyword is a “parent” and abstract phrases are its “children.” Since a parent is considered to be the more general topic compared to children terms, it is reasonable to represent that particular keyword as the parent concept to those synonymous terms (children). The improved concept-grouping algorithm is summarized as follows:

1. Find all parent-child relationships between keywords (parents) and abstract phrases (children).
2. Keep only 1-itemset parent and 1-itemset child relationships, discard the others.
3. If phrases (children) occur with multiple parents, choose the one with higher confidence. If both confidences happen to be equal, choose the one with less support parent.
4. If again both supports are equal, discard those phrases.

The reason to choose the parent-child with higher confidence in step 3 is that the child has more probability to occur with the parent. If this probability is equal between two parent-child relations, then the less support parent is chosen because, the child is closer to that parent. This can be illustrated in the following diagram.

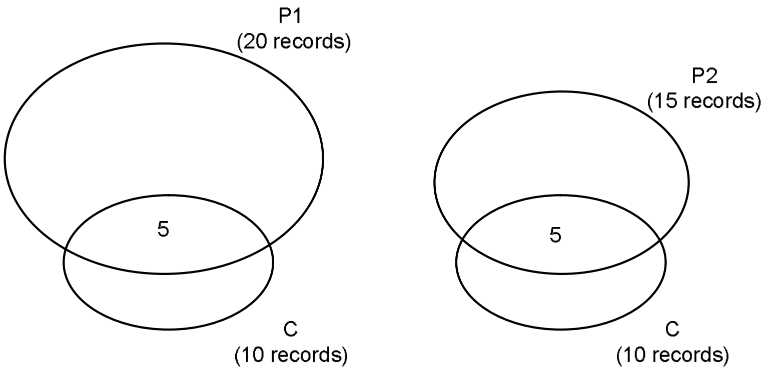


Figure 7.1: Comparison between higher and lower support parents, with equal confidence

From Figure 7.1, P1 and P2 denote the two parents, where support of P1 is higher than support of P2. C denotes the child term. Suppose there are 20, 15, and 10 records containing P1, P2, and C respectively. There are five overlapping records between C and both parents (see the intersection area). From this scenario, the confidence of P1-C relationship is equal to the confidence of P2-C, meaning the conditional probability Pr(P1|C) = Pr(P2|C) = 5/10 = 0.5. Since the support of P2 is less than that of P1, the proportion of C in P2 (5/15) is higher than that in P1 (5/20). Hence, C is closer to P2 than P1.

Table 7.4 and Table 7.5 illustrate experimental results for the same data sets: association rule mining data and information visualization data as in Section 7.2. The results are obtained from applying parent-child algorithm presented in Chapter 6. For both data sets, *minsup* and *minParent* are set to 5 and 0.8 respectively.

Table 7.4: Synonymous terms for association rule mining research using the improved algorithm

Group name	**Synonymous Terms**
knowledge acquisition	characterizations, computational costs
query processing	mining query
deductive databases	wide sprectrum
parallel algorithms	parallel algorithms, skewed data, processors
graph theory	association graphs
electronic commerce	e-commerce
asynchronous transfer mode	data-intensive applications, parallel data mining

Table 7.5: Synonymous terms for information visualization research using the improved algorithm

Group name	**Synonymous Terms**
user interfaces	spreadsheets, large information spaces, virtual worlds
Internet	Internet, world wide web
virtual reality	virtual reality, VR, virtual environments
graphical user interfaces	graphical user interfaces, references
information retrieval	search engines
multimedia computing	mobile environments
geographic information systems	GIS
splines(mathematics)	smoothness
spreadsheet programs	visualization spreadsheets
computer science education	educational courses

The above tables illustrate that most of the synonymous terms can be grouped together into meaningful concepts even when the minimum support is set to be higher (i.e., *minsup* = 5). For instance, the virtual reality group, from Table 7.5, consists of similar terms: “VR” and “virtual environments.” These terms would not be captured from the previous proposed algorithm (see Table 7.2 where *minsup* = 4). Hence, this improved concept-grouping algorithm promises to outperform the previously proposed method by reducing the noisy data and assigning appropriate concept to the derived synonymous terms.

From Table 7.4 and Table 7.5, some groups seem to capture obviously synonymous terms. For instance, the algorithm is able to detect “geographic information systems and GIS,” “electronic commerce and e-commerce,” and “virtual reality, VR, and virtual environment”. These synonymous terms would not be captured with the list cleanup function in VantagePoint. However, in some groups, it is not so obvious why the algorithm combined some terms together. Note that this algorithm is performed on the domain of interest. Hence, it does not imply that all the synonymous concepts captured from one data set can be generalized and used for a different data set. Consider the “parallel algorithms” group in Table 7.4 as an illustration. The algorithm suggests that “parallel algorithms,” “skewed data,” and “processors” should be combined together and this group should be called “parallel algorithms”. One might be surprised why “skewed data” should be included. It does not seem to relate to the “parallel algorithms” concept. However, within this data set, there are six records containing the term “skewed data” and within those,

- 5 records contain the term “parallel algorithms”
- 3 records contain the term “processors”

Therefore, within this specific domain, it is reasonable to group these terms together based on their co-occurrences.

7.4 Performance Evaluation

To assess the performance of this new concept-grouping algorithm, a factor map (generated from VantagePoint) of abstract phrases with and without the proposed algorithm will be compared. Factor analysis seeks to maximize similarity within clusters and minimize similarity between clusters. However, factor analysis allows overlapping terms, meaning terms can be included in multiple clusters. This method is suitable for analyzing multidisciplinary research domains.

There are two types of cluster validation studies. External criterion functions derive the clustering solution by focusing on optimizing a function that is based on how the various clusters are different from each other. Internal criterion functions try to determine if the structure is intrinsically appropriate for the data. This means it focuses on producing a cluster solution that optimizes a function defined only over the data within each cluster and does not take into account the data assigned to different clusters.

This research will not utilize the external quality measure since the cluster solution from factor analysis is allowed to have the same terms in multiple clusters. Hence, there will always be linkages between clusters, which are not suitable for external assessment. As a result, only internal assessment will be performed. A commonly used internal measurement, the cohesion (Steinbach et al. 2000 and Watts et al. 2002), will be employed to evaluate the accuracy of the produced clustering solutions.

7.4.1 Cohesion

Borrowing the concept of the vector-space model, the cosine similarity can be used to compute the cohesion between two terms (abstract phrases). In this evaluation, each term, *t*, is considered to be a vector, ***t***, in the document space. **t** = (df_1, df_2, …, df_n), where df_i is "1" if document i contains term t and "0" if it does not. To account for vectors of different lengths, the length of each term vector is normalized so that it is of unit length.

There are a number of possible measures for computing the similarity between term vectors, but the most common one is the cosine measure, which is defined as

$$\cos(\mathbf{t}_1, \mathbf{t}_2) = \frac{(\mathbf{t}_1 \bullet \mathbf{t}_2)}{\|\mathbf{t}_1\| \|\mathbf{t}_2\|} \tag{7.5}$$

where ● indicates the vector dot product and $\|\boldsymbol{t}\|$ is the length of vector $\boldsymbol{t}$. The cosine formula can be simplified to vectors' dot product when the term vectors are of unit length. This measure becomes one if the terms are identical, and zero if there is no document in common between them (i.e., the vectors are orthogonal to each other).

The average pairwise similarity between all points in a cluster will be used to determine each cluster's cohesion, which is defined as

$$cohesion_r = \frac{1}{m_r^{\,2}} \sum_{t_i, t_j \in S_r} cos(t_i, t_j) \tag{7.6}$$

where m_r is the number of terms in cluster r and S_r is the set of terms in cluster r. Note that this equation includes the similarity of each point with itself, which is equal to 1.

For an entire factor map, the total cohesion can be calculated as the weighted average of cohesions for each cluster:

$$Cohesion_{total} = \sum_{r=1}^{k} m_r \left(\frac{1}{m_r^{\ 2}} \sum_{t_i, t_j \in S_r} cos(t_i, t_j) \right) \qquad (7.7)$$

where *k* is the number of clusters in the factor map. This evaluation attempts to maximize the total cohesion.

7.4.2 Evaluation Results

Four data sets are used to evaluate the clustering results. The summary of these data sets is shown in Table 7.6. For all data sets, a stop-list is used to remove common words, and the words were stemmed using the list clean up function in VantagePoint.

Table 7.6: Summary of data sets used to evaluate the clustering results

Data Set	Source	Number of Records	Number of Terms used in Factor Map
ARM	INSPEC and ENGI	971	162
InfoViz	INSPEC	817	147
FuelCell	ENGI	1002	164
Thai-SW	INSPEC and ENGI	820	85

The *ARM* and *InfoViz* data sets are the same as the data sets used in Section 7.2. The *FuelCell* data set is retrieved from the ENGI database. These records relate to fuel cell research in the year 2003. The *Thai-SW* data set is retrieved from INSPEC and ENGI. These records are software research abstracts that have been published by Thai organizations. For each data set, I perform factor analysis on:

- Cleaned abstract phrases (resulted from the list cleanup function of VantagePoint).
- Concept grouped abstract phrases (results from the proposed algorithm in Section 7.3)

My hypothesis is that the clusters' quality for concept grouped abstract phrases will be better than the quality for cleaned abstract phrases. The relevant terms can be captured and combined by the proposed algorithm. This would increase the number of relevant documents in a cluster. As a result, the cohesion between terms within that cluster will increase.

Table 7.7 shows the cluster cohesion results for each data set. The results suggest that the quality of clusters generated from concept grouped abstract phrases is better than the cleaned abstract phrases. Even though, the cohesion of the concept grouped abstract phrases for the ARM data set is smaller, it is not significantly worse. The reason why the performance for the ARM data set is not improving much could be that the proposed algorithm only detects a few synonymous phrases in this data set (see Table 7.4). With a few synonymous phrases, the clusters of the concept grouped abstract phrases and the cleaned abstract phrases can be very similar to each other (see Table 7.8). As can be seen, the members in each cluster are very comparable. Only Cluster 12 contains all different members. As a result, this similarity leads to the indifferences between the total clusters' cohesion.

In contrast, for the FuelCell data set, the clusters' quality of concept grouped abstract phrases is better than the cleaned abstract phrases. The clusters generated from using two types of abstract phrases are shown in Table 7.9. From this table, the members in each cluster are significantly different. On average, the clusters generated from concept grouped abstract phrases are more compact and less noisy because the relevant phrases are grouped together into a meaningful concept before the analysis. This leads to the larger cohesion between terms in the clusters.

Table 7.7: Comparison of Clusters' Cohesion

Data set	Concept grouped abstract phrases	Cleaned abstract phrases
ARM (12 factors)	0.378	**0.383**
InfoViz (9 factors)	**0.418**	0.386
FuelCell (12 factors)	**0.438**	0.396
Thai-SW (8 factors)	**0.425**	0.407

Table 7.8: Comparison between clusters of concept grouped and cleaned abstract phrases for the ARM data set

	Concept grouped abstract phrases	Cleaned abstract phrases
Cluster 1	• advantageous • definitions • fuzzy associations rules • fuzzy sets theory • linguistic terms • mechanisms • mining fuzzy association rules • quantitative attributes • quantitative values • regularities • **fuzzy sets**	• advantageous • definitions • fuzzy associations rules • fuzzy sets theory • linguistic terms • mechanisms • mining fuzzy association rules • quantitative attributes • quantitative values • regularities
Cluster 2	• customer transactions • empirical evaluations • large databases • large itemsets • sales transactions • sufficient number • transaction	• customer transactions • empirical evaluations • large databases • large itemsets • sales transactions • sufficient number • transaction
Cluster 3	• concept hierarchy • decision trees • fuzzy rules • Genetic Algorithms • neural networks • predictive • principles • regularities • visualization	• concept hierarchy • decision trees • fuzzy rules • Genetic Algorithms • neural networks • predictive • principles • regularities • visualization
Cluster 4	• Apriori • frequent itemsets • itemsets • **candidate itemsets** • **magnitudes**	• Apriori • frequent itemsets • itemsets
Cluster 5	• candidates • data warehousing • experimental results • PC clusters • proposed methods • requirements	• candidates • data warehousing • experimental results • PC clusters • proposed methods • requirements
Cluster 6	• OLAP • query processing • scalable	• OLAP • query processing • scalable • **data warehousing**

Table 7.8 (continued).

Cluster 7	• algorithms • important tasks	• algorithms • important tasks
Cluster 8	• iteratively • parallel algorithms • **execution times**	• iteratively • parallel algorithms • **processors** • **candidates** • **reductions**
Cluster 9	• performance improvements • **association-rule-mining algorithms** • **new approach**	• performance improvements • **structures**
Cluster 10	• interesting knowledge • proposed algorithms • quantitative values • real-world applications • transaction • **interesting patterns**	• interesting knowledge • proposed algorithms • quantitative values • real-world applications • transaction
Cluster 11	• architectures • constraints • **SQL**	• architectures • constraints • **mining algorithms**
Cluster 12	• **domain knowledge** • **recent years**	• **classifications** • **data mining techniques** • **businesses**

Note: items in boldface represent the different terms

Table 7.9: Comparison between clusters of concept grouped and cleaned abstract phrases for the FuelCell data set

	Concept grouped abstract phrases	Cleaned abstract phrases
Cluster 1	• conductivity • methanol permeability • Nafion • protonic conductivity • **methanol**	• conductivity • methanol permeability • Nafion • protonic conductivity
Cluster 2	• gas turbines • high-efficiency • SOFCs	• gas turbines • high-efficiency • SOFCs
Cluster 3	• SOFCs • Solid-oxide fuel cells SOFCs	• SOFCs • Solid-oxide fuel cells SOFCs • **electrolytes**
Cluster 4	• MEAs • Membrane-electrode-assembly MEA • Proton exchange membrane fuel cells PEMFCS	• MEAs • Membrane-electrode-assembly MEA • Proton exchange membrane fuel cells PEMFCS • **Electrochemical Society**
Cluster 5	• Pt • Ru • **CO**	• Pt • Ru
Cluster 6	• electricity conductivity • Sr	• electricity conductivity • Sr • **conductivity** • **microstructures** • **ion conductivity** • **C. &copy** • **AC impedance spectroscopy** • **electron microscopy SEM** • **electronic conductivity** • **electrical properties**
Cluster 7	• oxygen reduction • **catalyst layers**	• oxygen reduction • **electrochemical reactions** • **mechanisms** • **electronic conductivity** • **relationships**
Cluster 8	• PEMFCs • Single-cells	• PEMFCs • Single-cells • **electrochemical reactions** • **proton exchange membrane fuel cells PEMFCs** • **permeability**

Table 7.9 (continued).

Cluster 9	- x-ray diffraction - **x-rays diffraction XRD** - **morphology** - **scanning electron microscopy**	- x-ray diffraction - **electron microscopy** - **characterised**
Cluster 10	- H<sub>2</sub> - CO<sub>2</sub> - H<sub>2</sub>O - O<sub>2</sub> - **Cell voltage** - **Concentrations** - **Molten carbonate fuel cells MCFCs** - **Single-cells**	- H<sub>2</sub> - CO<sub>2</sub> - H<sub>2</sub>O - O<sub>2</sub> - **CO**
Cluster 11	- **Conversions** - **Hydrogenation**	- **direct-methanol fuel cells DMFCs** - **DMFCs**
Cluster 12	- **cm<sup>2</sup>** - **impedance spectroscopy** - **resistivity** - **thicknesses**	- **experimental data** - **high-current densities** - **liquid-water**

Note: items in boldface represent the different terms

7.5 Conclusion

In this chapter, a concept-grouping algorithm used for constructing a thesaurus from synonymous terms is proposed. A “term” is defined as a phrase extracted from a title or abstract by VantagePoint. This algorithm adapts an association rule mining (ARM) technique to capture the synonymous terms. Similar terms, which are written differently, can be grouped together into the same concept based on their associations before they are used for subsequent analyses. A performance evaluation for the proposed algorithm is conducted. To assess the performance, a factor map (generated from VantagePoint) of cleaned abstract phrases is compared to one based on the concept-grouped abstract phrases. The cleaned abstract phrases result from applying the list cleanup function of VantagePoint to the extracted abstract phrases. The cluster quality is then validated by a measure called cohesion, which focuses on maximizing the similarity between terms within the clusters. The evaluation results show higher cohesions for the concept-grouped abstract phrase clusters compared to the cleaned abstract phrase clusters. Hence, the proposed concept-grouping algorithm can help to improve the accuracy of the representation of abstract phrases.

CHAPTER 8

MANAGING R&D WITH TEXT INFORMATION EXPLOITATION

In this chapter, a text mining framework for intelligence gathering is applied to the Thai data set for the needs of Thailand's R&D management programs. This chapter proceeds in several parts. The first part provides a brief context of current status and development of Thailand's S&T system, noting its strengths and weaknesses. The target user and technology to be studied are presented next, followed by the details of the analyzed data. Then the Thailand's overall and biotechnology R&D profiles are presented. The last section presents the comparison between the publication data, the grant data, and the export data, along with the observations for the Thai R&D managers or policy planners to improve strategic decision-making processes.

8.1 Current Status of Thailand's S&T

A variety of studies of technology-based competitiveness of nations have sought to measure advances in S&T, identify emerging technologies, predict S&T trends, and suggest ways to maintain competitiveness. For example, the Technology Policy and Assessment Center at Georgia Institute of Technology has been generating High-Tech Indicators (HTI) (Porter et al., 2001) – measures of national technology-based export competitiveness since 1987. This study indicates "high-tech standing," measuring current high-tech production and export competitiveness. Thailand's standing has fluctuated notably from 1993-2003, and appears in the lower rank among the 33 nations covered. Similar studies of national competitiveness have been generated. For example, the US Council on Competitiveness presents the Innovation Index, which is a method for assessing the strengths of national innovation systems (Porter and Stern, 1999). The International Institute for Management Development (IMD) has analyzed the competitiveness of nations and publishes the World Competitiveness Yearbook. According to the 2000 edition, the rank of Thailand's scientific and technological capacity together with its success at basic and applied research is 47 (out of 47 countries participating)[12]. In addition, a study conducted by the Thailand Development Research Institute (TDRI)[13] concluded that of the four technological capabilities in Thai industry, operative capability is generally the highest, followed closely by acquisitive and adaptive capabilities, while there tends to be a very low level of innovative capability.

In general, the status of S&T of a country can be assessed by several universal indicators of which there are two main types: input and output indicators. By consensus of the Organization for Economic Co-operation and Development (OECD) countries (a group of industrialized countries), input indicators of interest include the number of R&D researchers and the amount of R&D expenditure per GDP. There are significant differences between the levels for developed and developing countries. The differences between developed and developing countries are due to the fact that it takes a long time to build up S&T manpower, and that R&D and other S&T activities require both investment and infrastructure.

[12] http://www02.imd.ch/wcy

[13] http://ww.tdri.or.th

Output indicators include the number of a country's scientific papers in international and local journals, the number of patents, and other measurements. The numbers of publications retrieved from INSPEC databases and numbers of patents issued to residents for Thailand and other countries for the year 2000 are shown in Table 8.1.

Table 8.1: Output indicators in selected country in the year 2000

Country	Publication Abstracts Retrieved from INSPEC database, 2000	Number of patent issued to residents, 2000
Industrialized Countries		
USA	77,730	85,071
Japan	35,155	112,269
Germany	21,918	16,901
Newly Industrialized Countries and Southeast Asian Countries		
South Korea	8,659	22,943
Singapore	3,050	110
Thailand	419	153
Malaysia	496	24

Source: Patent data are from World Intellectual Property Organization

Table 8.1 shows that only a small amount of publications and patents are published from Thailand compared to other countries. The low number of publications and patents could be partly due to the unfamiliarity of Thai researchers with the systems of international scientific communications and intellectual property protection practiced in industrially developed countries. However, more seriously, the statistics reveal a lack of innovative capability.

The situation is not helped by the relatively low financial input into research and development (R&D) and lack of R&D manpower. Table 8.2 shows R&D input indicators, measured by the number of R&D researchers per million population and the amount of R&D expenditure per gross domestic product (GDP), for selected countries. These indicators reflect the low level of innovative capability of Thailand compared to other industrialized and newly industrialized countries. Expectantly, this picture may soon change, with increasing awareness of the essential role of science and technology in development, coupled with the increased resources made available by high economic growth.

Table 8.2: R&D researchers and R&D expenditure as percentage of gross domestic product (GDP) in selected country (in 2000)

Country	R&D researchers per million population	R&D expenditure as percentage of GDP
Industrialized Countries		
USA (1997)	4,099	2.58
Japan	5,095	2.93
Germany	3,161	2.49
Newly Industrialized Countries and Southeast Asian Countries		
South Korea	2,319	2.68
Singapore	4,140	1.89
Thailand (1997)	74	0.10
Malaysia (1998)	160	0.40

Source: UNESCO Statistical Yearbook (http://www.unesco.org)

8.1.1 Development of S&T Systems in Thailand

8.1.1.1 S&T Institutions

The universities are the major performers of scientific research. While medical science is the strongest area of research at an international level, Thailand also has considerable achievements in agriculture, engineering and other areas, mostly at the adaptive levels. The government has increasingly realized that crucial factors for the success of S&T development in Thailand, in addition to adequate manpower, are financial support for research and the linkage between technology producers and users. In particular, the so-called technology producers in public research institutes and universities, and the technology users in industry, fail to interact with each other (Yuthavong and Wojcik, 1997).

8.1.1.2 S&T in the Private Sector

In Thailand, the private sector plays the dominant role in industrialization and is responsible for most of the present economic growth. However, when it comes to R&D, private sector spending is almost negligible. Growth in the industrial sector has been achieved largely through the successful utilization of imported technologies such as new machinery, equipment, product designs and managerial know-how. Other factors such as cheap labor and abundant natural resources yielded lower production costs and made Thailand an attractive location for manufacturing. As developed countries move towards more modern and automated production and make significant advances in new materials, certain exports become less competitive, and hence some industries may be at risk. Without substantial efforts to develop indigenous capability in S&T to support industrial development, Thailand may even face a major economic and social crisis in the near future. Problems and Lessons of S&T development in Thailand can be summarized as:

- Tendency of researchers to carry out S&T activities in a manner that is isolated and disconnected from the real world. In other words, these studies have often been unable to have an impact on industry.
- Industrial firms tend to utilize proven technologies from foreign sources rather than the local S&T development.
- "Industrial firms tend to doubt the ability and effectiveness of universities and public technical institutes to solve practical industrial problems" (Yuthavong and Wojcik, 1997).

In conclusion, current weaknesses in technological capabilities of Thailand appear to include: inadequate supply of technical human resources (see Table 8.2), weak linkage between the S&T community in universities and public institutes to the private sector, and reliance on imported foreign technology. Therefore, a challenging problem for the Thai Ministry is how to develop R&D policy and priority setting to increase the country's technological and innovative capabilities.

The above section has pointed out the current situation in Thailand's S&T systems along with some weaknesses. The purpose of this case study is not to solve all these problems but rather to apply a text mining framework to help the Ministry monitor and profile R&D activities in Thailand's S&T community.

8.2 Target User(s)

The target user of this research is Thailand's National Science and Technology Development Agency (NSTDA). NSTDA consists of four centers: National Center for Genetic Engineering and Biotechnology (BIOTEC), National Metal and Materials

Technology Center (MTEC), National Electronics and Computer Technology Center (NECTEC) and Technical Information Access Center (TIAC). NSTDA's three major goals are "to achieve an effective system for support of research and development (R&D), to perform its own R&D and services, and to invest in or support the private sector in investment, which leads to technology development and innovation." (http://www.nstda.or.th) This research is expected to help the NSTDA achieve the first goal.

TIAC attempted to conduct bibliometric analysis for Thailand R&D in the year 2000[14]. They used the Science Citation Index (SCI) as their source. The analyses performed include counting and ranking what type of affiliations (e.g., public and private universities, government units, or private companies) publish the most research articles. This study, however, was done manually by a group of analysts. One analyst[15] stated that the study was very time consuming and might not be very accurate since she and her colleagues had to count all the records manually. There were 1,354 records retrieved from SCI for the year 2000 alone. The project took about 2 months to finish. After being told about the proposed method and tool used for this thesis, she believes that it will be very valuable for the Ministry since more analyses can be done more accurately in less time.

[14] http://www.tiac.or.th/, the study has reported in Thai language.

[15] The author would like to thank Mrs. Rungsima Pedmedyai for the interview.

8.3 Technologies to be Studied

This research seeks to help NSTDA perform macro- and micro-level analysis of research domains. For macro-level analysis, overall R&D activities will be investigated. After discussing with two senior officers[16] of NSTDA, we agreed to study biotechnology more specifically for micro-discipline level analyses. Biotechnology is chosen because Thailand is a country traditionally rich in natural resources, and agriculture. Biotechnology has the potential to bring enormous benefits to Thailand and to maintain the competitiveness of its existing agriculture exports. Dr. Yuthavong, an expert in biotechnology, stated, “Thailand has conducted research on biotechnology for more than 20 years. Hence the number of publications should be more than other areas but it will be interesting to see if we have indigenous development capability.”

8.4 The Data

8.4.1 The Publications Data

The study begins by examining the records of Thailand’s technical publications and patents abstracted in both international and local R&D databases. INSPEC, Engineering Index (ENGI) and Science Citation Index (SCI) are potential international (English- based) databases. However, finding the potential local databases that have a complete collection for Thai scientific publications is not an easy task. Most of the academic libraries create their own collections in their automated library systems but their coverage is not complete and may be duplicated by one another. Currently, in

[16] The author would like to thank Dr. Chatri Sripaipan - Vice-President, NSTDA, and Dr. Yongyuth Yuthavong - President, Thai Academy of Science and Technology, and Senior Researcher, BIOTEC, NSTDA for their guidance.

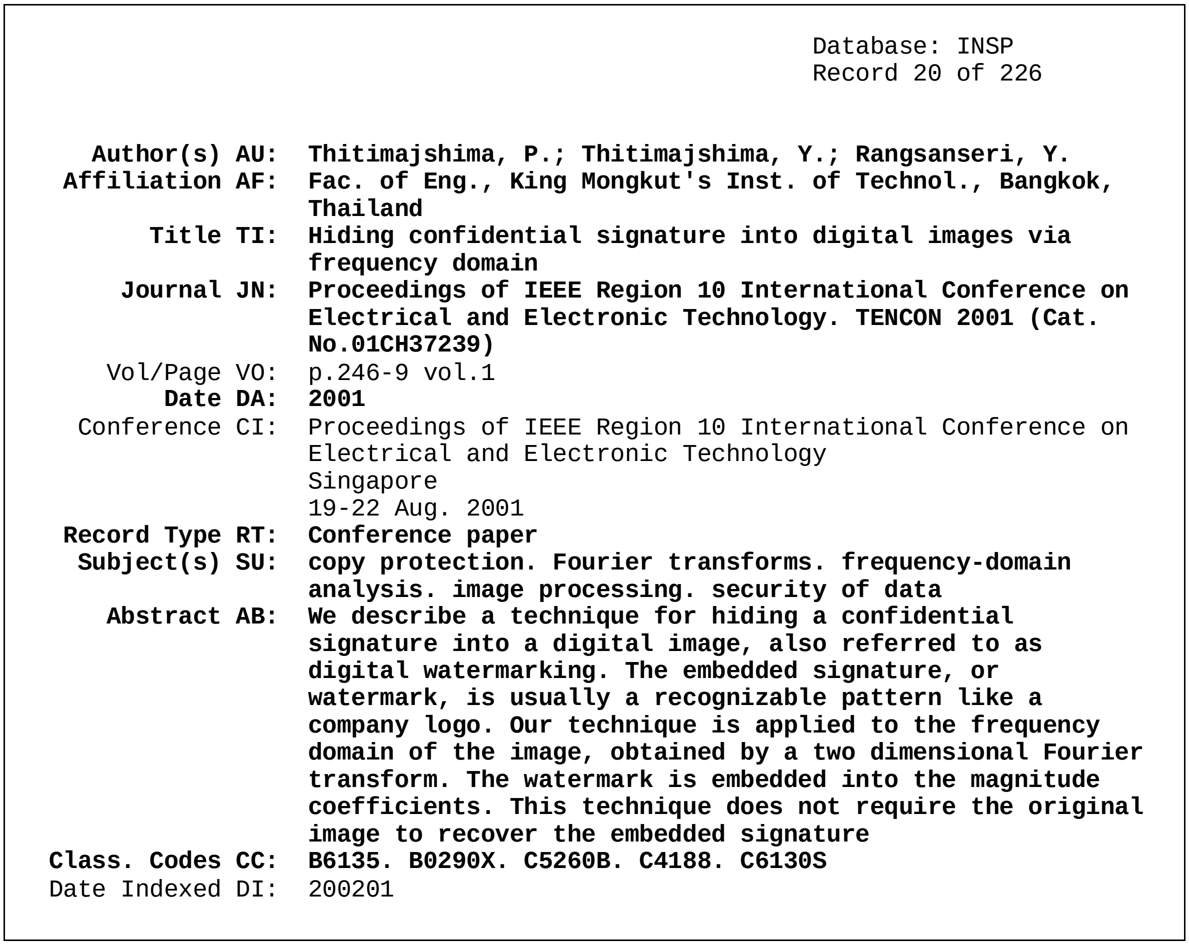
Thailand, there is no central organization responsible for this duty. One Thai database that is used for this research, is the project grants database (http://www.nstda.or.th/grants). This database combines all the research projects that are funded by the National Science and Technology Development Agency.

Another type of data that is usually included to perform bibliometric analysis is the Patent database. After searching for Thai assignees in The US Patents Bibliographic Database (PATS), it appears that there are only 91 patents filed from Thai assignees during the years 1982 to 2001. Furthermore, most of these patents are agriculture and handicrafts related. The Thai Patents Database (http://www.ipic.moc.go.th) has also been explored. This database contains the patent applications, published by the Thai Department of Intellectual Property. There are 5382 resident patent applications over the 1980-2002 period. However, the number of applications by non-residents is much higher than by residents (42,294 applications during that 1980-2002 period). Most patents are filed by foreign companies operating in Thailand to protect their rights. After browsing through the resident patent applications, most of these are also agriculture and handicrafts related. The low number of patents could be partly due to the unfamiliarity of Thai researchers with the systems of intellectual property protection in industrially developed countries. Since most of the patents filed by Thai organizations are less related to target technologies' R&D, patent analysis using VantagePoint will not be included in this research.

Both English and Thai language publication data will be imported to VantagePoint to profile trends in R&D and can help identify emerging or unfamiliar research that may intersect the functional interest of the Ministry.

8.4.1.1 English–Based Publications

The research is being performed on the INSPEC, Engineering Index (ENGI) and Science Citation Index (SCI) databases. Records in these databases are structured with fields such as title, author, affiliation, year of publication, keywords, and abstract. Figure 8.1 illustrates a record retrieved from INSPEC.



```
                                                    Database: INSP
                                                    Record 20 of 226


   Author(s) AU:  Thitimajshima, P.; Thitimajshima, Y.; Rangsanseri, Y.
 Affiliation AF:  Fac. of Eng., King Mongkut's Inst. of Technol., Bangkok,
                  Thailand
       Title TI:  Hiding confidential signature into digital images via
                  frequency domain
     Journal JN:  Proceedings of IEEE Region 10 International Conference on
                  Electrical and Electronic Technology. TENCON 2001 (Cat.
                  No.01CH37239)
    Vol/Page VO:  p.246-9 vol.1
        Date DA:  2001
  Conference CI:  Proceedings of IEEE Region 10 International Conference on
                  Electrical and Electronic Technology
                  Singapore
                  19-22 Aug. 2001
 Record Type RT:  Conference paper
  Subject(s) SU:  copy protection. Fourier transforms. frequency-domain
                  analysis. image processing. security of data
    Abstract AB:  We describe a technique for hiding a confidential
                  signature into a digital image, also referred to as
                  digital watermarking. The embedded signature, or
                  watermark, is usually a recognizable pattern like a
                  company logo. Our technique is applied to the frequency
                  domain of the image, obtained by a two dimensional Fourier
                  transform. The watermark is embedded into the magnitude
                  coefficients. This technique does not require the original
                  image to recover the embedded signature
Class. Codes CC:  B6135. B0290X. C5260B. C4188. C6130S
Date Indexed DI:  200201
```

Figure 8.1: A sample Record of the INSPEC database

ENGI and SCI contain similar fields to INSPEC. However, to conduct further analysis, only the relevant fields are selected. These fields (labeled in boldface) consist of Author(s), Affiliation, Title, Source (Journal), Year of publication, Keywords, Abstract, and Class Codes.

- **Author(s) AU:** The Author (AU) field contains the names of the authors or editors of a record.
- **Affiliation AF:** The Affiliation (AF) field contains the primary author's or editor's affiliation.
- **Title TI:**The Title (TI) field contains the bibliographic record, for example, titles of journal papers, book chapters, conference proceedings, conference papers, patents, dissertations, and report sections.
- **Journal JN:** The Journal Name (JN) field contains the full name of the journal in which the article was published.
- **Date DA:** The Date (DA) field contains the year in which an article or monograph was published.
- **Record Type RT:** The Record Type (RT) identifies whether a document is a conference paper or journal paper.
- **Subject SU:** The Subject (SU) field contains free-language words or phrases, which INSPEC indexers assign. These give a more exhaustive description of the content of the document than which is provided by the original title or by controlled index terms.

- **Abstract AB:** The Abstract (AB) field summarizes the content of the document and may provide a description of the background, methods, results, and/or conclusions.
- **Class Codes CC:** The Classification Codes (CC) field contains the appropriate code or codes from the INSPEC and ENGI Classification. For instance, INSPEC's codes begin with a letter identifying the section of the Classification they appear in. Any record may appear in more than one section. Sections include Physics (A), Electrical and Electronic Engineering (B), Computers and Control (C), and Information Technology (D). The general classification codes for INSPEC and ENGI are listed in Appendix A3 and A4.

The search string used to retrieve the data set is: (thailand.af. or thail.af. or thai.af.). Combining search results from these three databases, with duplicates removed, yielded 22,992 abstracts[17] over the 1970-2003 period. Certainly, this search does not capture all Thai research, but it does represent an important portion that has been published in recognized S&T journals abstracted by these international databases. Figure 8.2 shows the trend of Thai R&D publications over the 1970 – 2003 period.

[17] These records are completely retrieved on February 9, 2004.

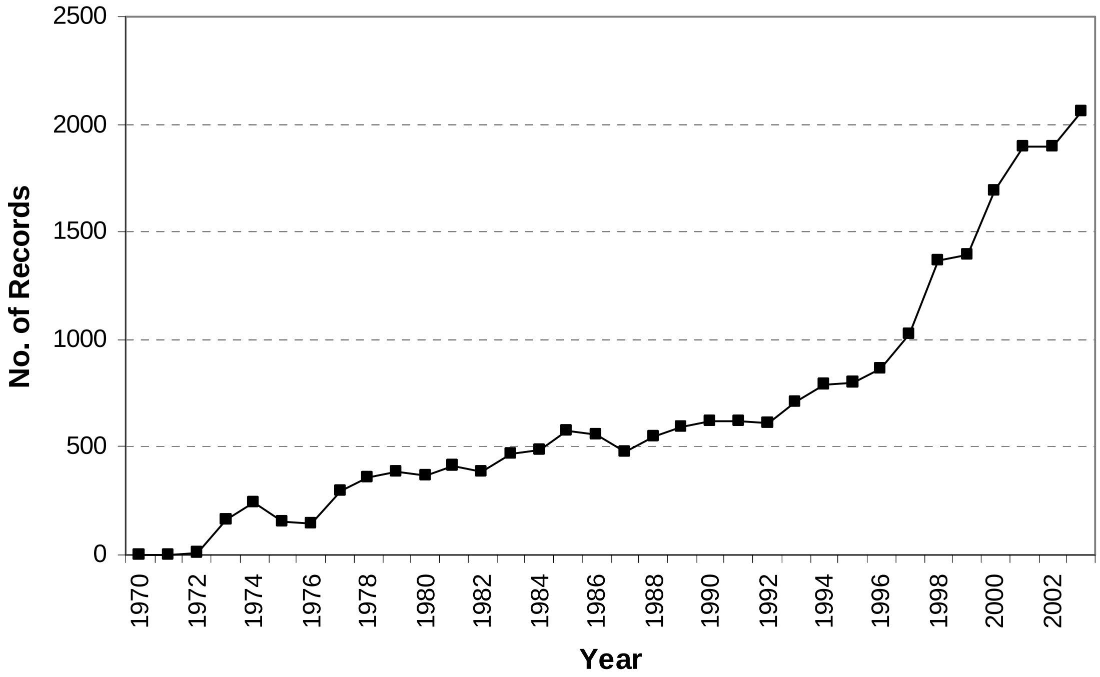


Figure 8.2: Trend of Thai R&D Publications during 1971 - 2003

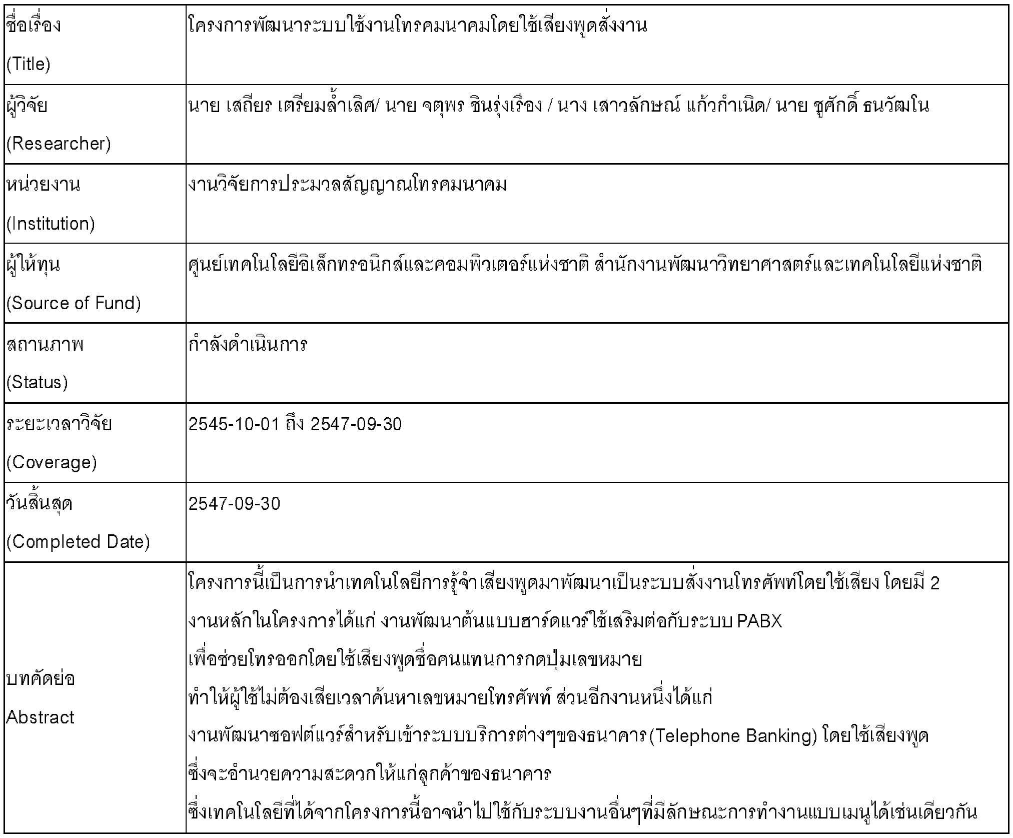
The above figure illustrates Thai publication in internationally indexed journals and conferences is generally rising. The overall trend (from 1973 to 2002) shows publication increasing at an average annual growth rate of 7.5 percent[18] .

8.4.1.2 Thai-Based Publications

The research will be performed on the grants database provided by NSTDA. Records in these databases are structured with fields such as title, author, institution, source of funds, project coverage, and abstract. A sample record retrieved from the Thai grants database is shown in Figure 8.3. Even though Thai language characters can be

[18] Linear regression using a log-linear model of growth. The dependent variable is the log of the number of records, and the independent variable is the year. The growth parameter of 0.075 was significant with $R^2 =$ 0.93.

read by VantagePoint, some functions such as performing natural language processing on title or abstract will not be valid because of the characteristic differences between the English and Thai languages. Thai is an alphabetic based language. The Thai alphabet uses forty-four consonants and fifteen basic vowel characters. These are horizontally placed, left to right, with no intervening space, to form syllables, words, and a sentence. Each sentence is separated by a space. Vowels are written above, below, before, or after the consonant, they modify.



ชื่อเรื่อง (Title)	โครงการพัฒนาระบบใช้งานโทรคมนาคมโดยใช้เสียงพูดสั่งงาน
ผู้วิจัย (Researcher)	นาย เสถียร เตรียมล้ำเลิศ/ นาย จตุพร ชินรุ่งเรือง / นาง เสาวลักษณ์ แก้วกำเนิด/ นาย ชูศักดิ์ ธนวัฒโน
หน่วยงาน (Institution)	งานวิจัยการประมวลสัญญาณโทรคมนาคม
ผู้ให้ทุน (Source of Fund)	ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ
สถานภาพ (Status)	กำลังดำเนินการ
ระยะเวลาวิจัย (Coverage)	2545-10-01 ถึง 2547-09-30
วันสิ้นสุด (Completed Date)	2547-09-30
บทคัดย่อ Abstract	โครงการนี้เป็นการนำเทคโนโลยีการรู้จำเสียงพูดมาพัฒนาเป็นระบบสั่งงานโทรศัพท์โดยใช้เสียง โดยมี 2 งานหลักในโครงการได้แก่ งานพัฒนาต้นแบบฮาร์ดแวร์ใช้เสริมต่อกับระบบ PABX เพื่อช่วยโทรออกโดยใช้เสียงพูดชื่อคนแทนการกดปุ่มเลขหมาย ทำให้ผู้ใช้ไม่ต้องเสียเวลาค้นหาเลขหมายโทรศัพท์ ส่วนอีกงานหนึ่งได้แก่ งานพัฒนาซอฟต์แวร์สำหรับเข้าระบบบริการต่างๆของธนาคาร(Telephone Banking) โดยใช้เสียงพูด ซึ่งจะอำนวยความสะดวกให้แก่ลูกค้าของธนาคาร ซึ่งเทคโนโลยีที่ได้จากโครงการนี้อาจนำไปใช้กับระบบงานอื่นๆที่มีลักษณะการทำงานแบบเมนูได้เช่นเดียวกัน

Figure 8.3: A sample record retrieved from the Thai grants database.

The following fields will be used for the analysis.

- **Researcher(s):** The Researcher field contains the name of the researchers for a project. In the case of multiple researchers, each one is separated by "/" sign with the name of the leading researcher in the first place.
- **Institution:** The Institution field contains the leading researcher's affiliation.
- **Source of Funds:** The Source of Funds filed contains the source of funding from:

1. The central office of the National Science and Technology Development Agency
2. National Center for Genetic Engineering and Biotechnology (BIOTEC)
3. National Metal and Materials Technology Center (MTEC)
4. National Electronics and Computer Technology Center (NECTEC)

- **Project Year:** Project year is equal to the starting year of the coverage field, which contains the duration of a project. This coverage field is displayed in "yyyy-mm-dd to yyyy-mm-dd" format. Note that the year shown is in Buddhist years. To convert to the western year, subtract 543 from the Buddhist year.

There are 1,970 research projects over the 1988 – 2003 period. As can be seen, the available database does not provide a keywords field. Therefore, some functions such as factor maps based on keywords cannot be performed. However, from the available data, the overview of R&D activities can be captured. For example, the distribution of research in different areas (such as electronics and computer technology, biotechnology, etc.) can be determined.

In the future, this research could be done in collaboration with the Text Processing Technology Group[19] at the National Electronics and Computer Technology Center. This group of people has developed the Thai word segmentation software. An attractive potential application can be to apply this algorithm to parse Thai sentences into

[19] http://www.links.nectec.or.th/itech/i4.html

words, then perform factor analysis in VantagePoint or using the proposed tree-structured networks algorithm on these words.

8.4.2 The Export Data

In order to profile Thai R&D activities and their linkage to the industrial sectors, the amount of export-oriented industrial activity will also be explored. These export data will be compared with the R&D publication profiles to identify the linkage between the research community and the respective industry. Results of the comparison will be presented in a later section.

8.5 Thailand's Research Profiling

As stated earlier, this research seeks to help NSTDA perform macro- and micro-level analysis of research domains in Thailand. For macro-level analysis, overall R&D activities will be investigated. Biotechnology will be explored for micro-discipline level analysis.

8.5.1 Overall Research Domain in Thailand

8.5.1.1 Types of Research Organization

The total number of bibliographic abstracts (22,992 records) retrieved from INSPEC, ENGI and SCI are first used to perform the macro analysis. However, SCI did not provide author abstracts and keywords until 1991. As a result, the bibliographic records from 1991 onward (16,209 records) will be used in the analysis. Using an organizational type thesaurus (i.e., a thesaurus that categorizes affiliations into industry, academia, and government) in VantagePoint, the proportion of research from various types of organization can be illustrated as in Figure 8.4. Note that this organizational type thesaurus uses general identifiers (e.g., "univ" is for university; "ltd" or "corp" for corporate, "national" or "ministry" for government, and "hospital" or "hosp" for hospital). Although this quick analysis cannot classify all Thai affiliations covered from the data set, it can provide the big picture of the R&D players in Thailand.

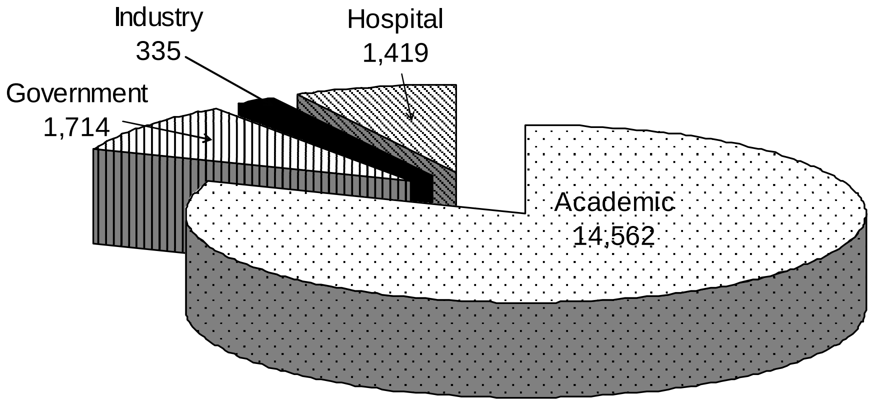


Figure 8.4: Organizational Types Publishing on Thai R&D

Figure 8.4 shows that the majority of research article author affiliations (80 percent --14,562 of 18,030 classified entities) are classified as universities, with 10 percent as governmental organizations, 8 percent as hospital and only 2 percent as industrial affiliations. These proportions indicate that universities are the major performers of scientific research while the private sector has the smaller role. Policy planners might be interested in setting priorities to promote more R&D in the industrial sector. Also, measures to enhance collaboration by industry with academia deserve exploration. Note that the publication data are retrieved from the publicly available databases. Hence, the number of publications published from the industrial organizations can be small because companies may not want to publish their proprietary research.

8.5.1.2 The Leading Research Organizations

Information within the retrieved abstract records includes titles of the articles, authors, author affiliations, keywords, date, class codes, and the abstracts themselves. VantagePoint can tally occurrences to generate lists of these fields. Such lists can provide the first order of useful information -- e.g., "who's doing what" in a research domain. The analyst can then concentrate on the aspects of most interest for further analyses. For instance, the list of leading affiliations (research organizations) or authors can be created to identify who are the most active players in the field. Table 8.3 illustrates the "Top-10" Thai R&D organizations.

Table 8.3: Top Thai R&D Organizations

Affiliation	Number of Records
Mahidol University	3776
Chulalongkorn University	2652
Asian Institute of Technology	1507
Chiangmai University	1225
King Mongkut's Institute of Technology, Bangkok	914
Prince of Songkla University	769
Khon Kaen University	700
Kasetsart University	673
Armed Forces Research Institute Medical Sciences	371
King Mongkut's University of Technology, Thonburi	327

As expected, the top R&D organizations are Thai universities. Mahidol University offers a strong program in medical science and biotechnology. Chulalongkorn University, Asian Institute of Technology, and King Mongkut's Institute of Technology are among the leading engineering schools. Thus, one might expect to discover the major research in medical science and engineering from the publications data.

8.5.1.3 The Focus of Leading Research Organizations

To identify activity concentration, two lists can be combined to create a 2-dimensional co-occurrence matrix. For instance, the co-occurrence matrix between keywords and affiliations can help identify which topics particular organizations mention frequently. Table 8.4 illustrates the research focus for the top Thai universities. Some interesting points that can be drawn from Table 8.4:

- Mahidol University is the main R&D player in medical science and biotechnology research, especially in malaria.
- Chulalongkorn University focuses on engineering, and also medical science and biotechnology research.
- The strong research area for the engineering schools, such as the Asian Institute of Technology and King Mongkut's Institute of Technology, are telecommunication systems and microelectronics, respectively.

Table 8.4: The Focus of Leading Thai Universities

Affiliation	keywords
top5	Top Terms

Mahidol University[3776]	MALARIA [122];
	PLASMODIUM-FALCIPARUM [93];
	PHARMACOKINETICS [36];
	CHEMOTHERAPY [31];
	MEFLOQUINE [28];
	PENAEUS-MONODON [27];
	Morphology [25]

Chulalongkorn University[2652]	Mathematical models [33];
	PENAEUS-MONODON [27];
	Algorithms [23];
	PLASMODIUM-FALCIPARUM [21];
	Computer simulation [21];
	Genetic algorithms [19];
	Synthesis (chemical) [18]

Asian Institute of Technology[1507]	Mathematical models [85];
	code-division multiple-access [60];
	Algorithms [40];
	Computer simulation [34];
	Production control [31];
	clays [28];
	OPTICAL COMMUNICATIONS [25]

Chiangmai University[1225]	Scanning electron microscopy [20];
	FLOW-INJECTION [16];
	PRENATAL DIAGNOSIS [15];
	Ion implantation [14];
	ULTRASOUND [11];
	Spectrophotometry [10];
	Heat transfer [10]

King Mongkut's Institute of Technology, Bangkok[914]	Computer simulation [37];
	Mathematical models [36];
	Algorithms [36];
	Circuit simulator [32];
	Induction motors [27];
	Integrated Circuits, CMOS [27];
	Fuzzy control [27]

8.5.1.4 An Overview Map of Research in Thailand

To look at the overview of research in Thailand, a map of keywords can be created. VantagePoint can construct technology maps to represent relationships graphically. The maps are generated by applying Principal Components Analysis (PCA) as described in Chapter 4 and clustering techniques.

While medical science is the strong area of research at an international level, Thailand also has considerable achievements in agriculture, engineering and other areas. A technology map (Figure 8.5) generated by clustering the leading 920 keywords (those appearing 10 or more times in the full 16,209 abstract records) reveals Thai research concentrations. Some of these include medical science (especially relating to malaria), materials science (especially relating to rubber and plastics), telecommunications and optical communications systems, integrated circuits, food science, power system control, electric power generation, and production control.

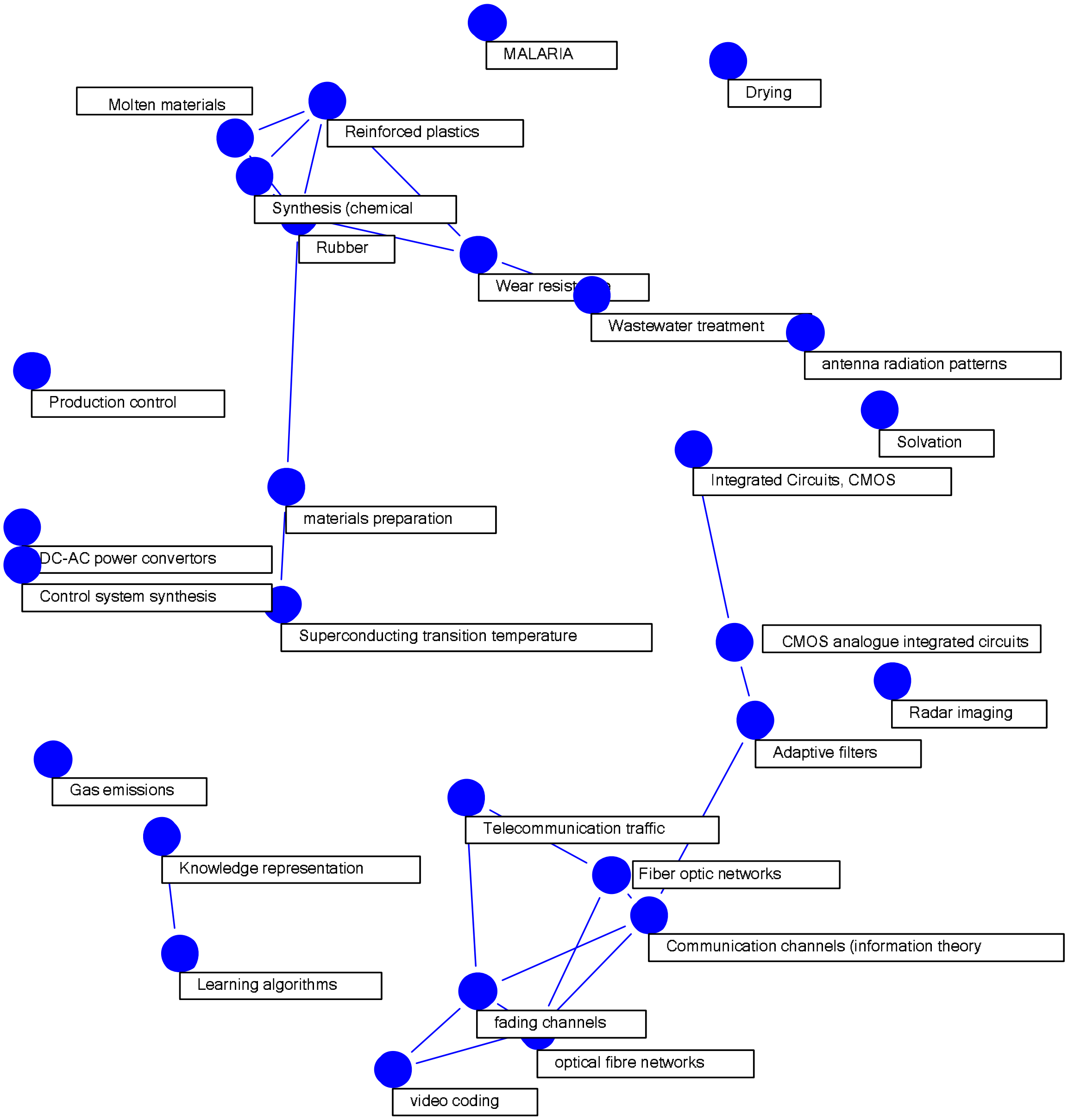
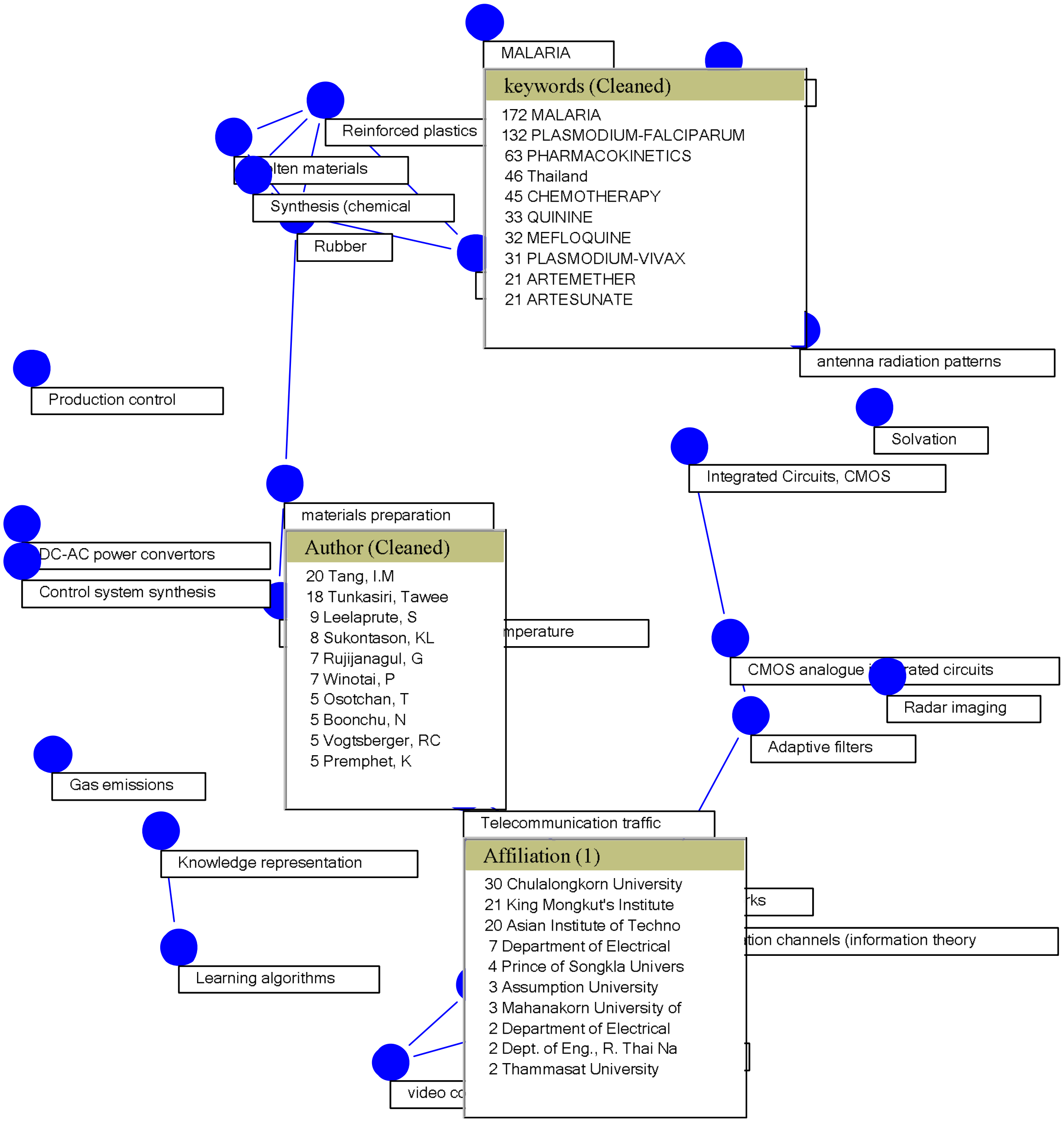


Figure 8.5: Overview Map of Research Domains in Thailand

To investigate cluster of interest, a “pull-down” box that contains the fields used to construct that particular clusters can be presented. For instance, the “pull-down” boxes are enabled for malaria, materials preparation, and telecommunication traffic clusters in Figure 8.6.

Within the same factor map, one can quickly identify the leading organizations for telecommunication traffics topic and the leading researchers for materials preparation topic.



keywords (Cleaned)	
172	MALARIA
132	PLASMODIUM-FALCIPARUM
63	PHARMACOKINETICS
46	Thailand
45	CHEMOTHERAPY
33	QUININE
32	MEFLOQUINE
31	PLASMODIUM-VIVAX
21	ARTEMETHER
21	ARTESUNATE

Author (Cleaned)	
20	Tang, I.M
18	Tunkasiri, Tawee
9	Leelaprute, S
8	Sukontason, KL
7	Rujijanagul, G
7	Winotai, P
5	Osotchan, T
5	Boonchu, N
5	Vogtsberger, RC
5	Premphet, K

Affiliation (1)	
30	Chulalongkorn University
21	King Mongkut's Institute
20	Asian Institute of Techno
7	Department of Electrical
4	Prince of Songkla Univers
3	Assumption University
3	Mahanakorn University of
2	Department of Electrical
2	Dept. of Eng., R. Thai Na
2	Thammasat University

Figure 8.6: Overview Map of Research Domains in Thailand with the “pull-down” boxes enabled

8.5.2 Biotechnology Analyses

There are 1,389 biotechnology records[20] in the original data set. The publication trend shows increasing interest as shown in Figure 8.7. Note that the 2003 data might not yet be complete.

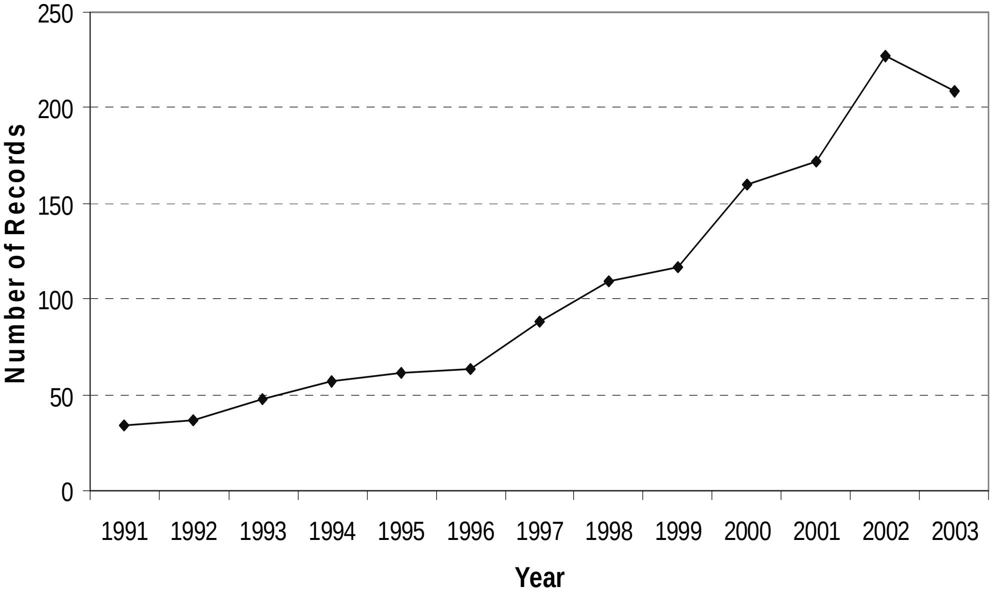


Figure 8.7: Trend of Thai Biotechnology Publications during 1991-2003

[20] These records are extracted from the 16,209 Thai records. Extraction is based on using keywords and classcodes that relate to biotechnology and bioengineering.

8.5.2.1 The Overall Map of Biotechnology Research in Thailand

Figure 8.8 and Figure 8.9 present the maps of topics in biotechnology research. The first map is constructed based on keywords indexed by the database provider. The second map is created based on concept-grouped phrases extracted from the abstracts. The keyword map tends to cover the broader research topics while the abstract phrases map tends to identify more specific research areas.

The keywords map suggests malaria, waste management, and shrimp (penaeus-monodon) are the dominating current research areas. It is interesting that the agricultural research is not captured in this map despite the fact that Thailand is a country traditionally rich in agriculture. Policy planners might try to promote more R&D in agricultural sector.

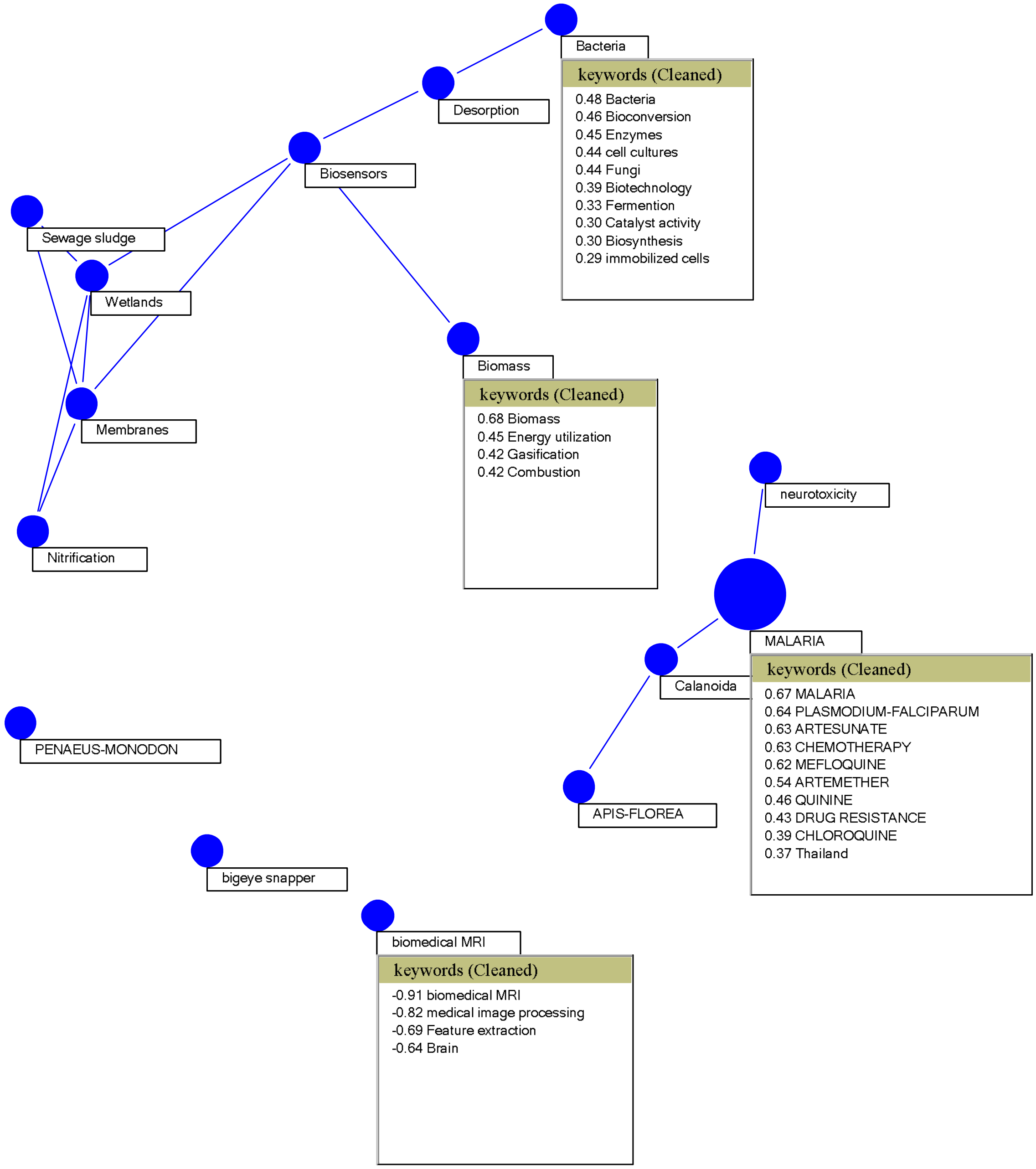


Figure 8.8: A Keywords Map of Biotechnology Research in Thailand

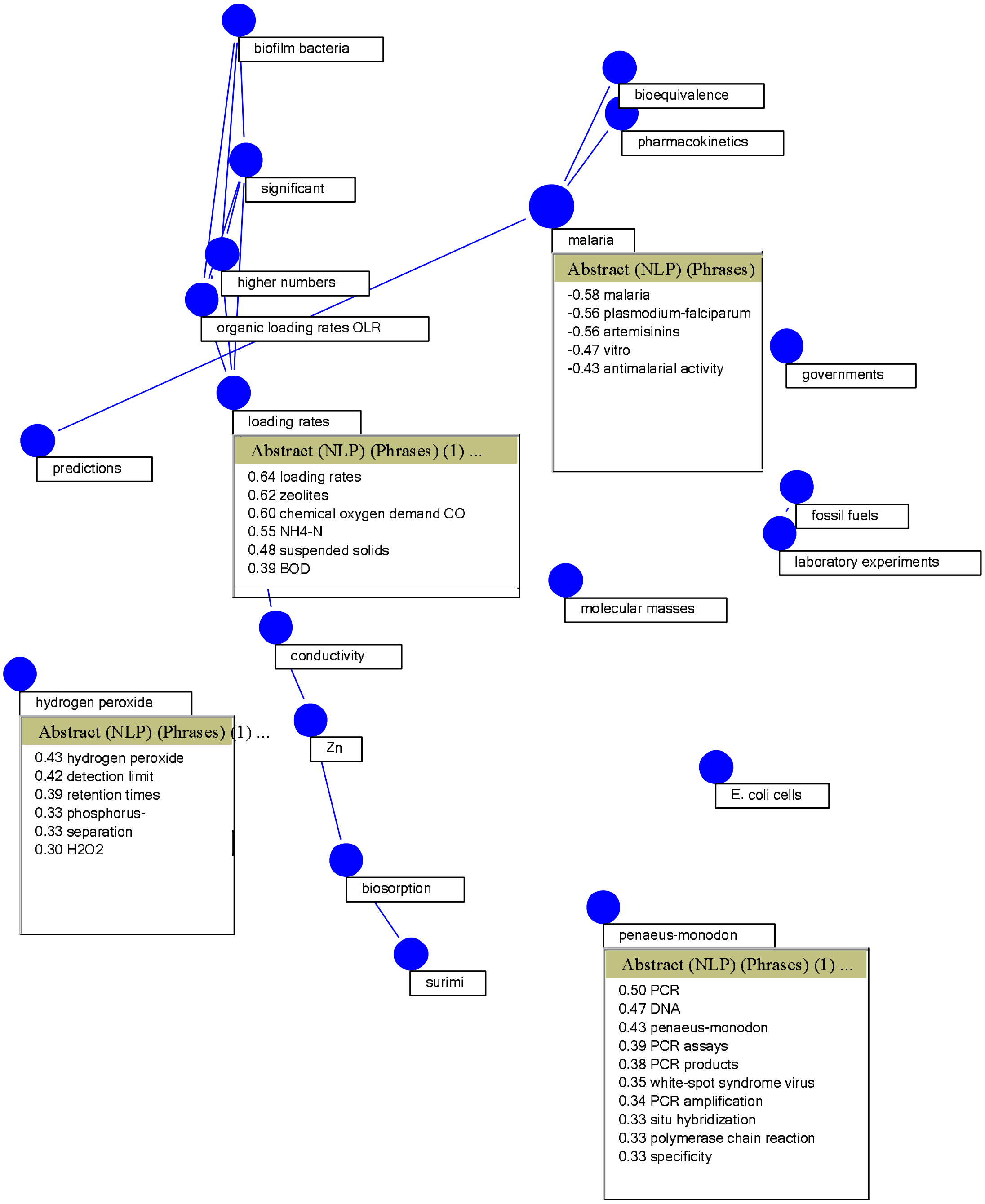


Figure 8.9: A Concept-Grouped Abstract Phrases Map of Biotechnology Research in Thailand

8.5.2.2 Concept Hierarchy of Biotechnology Research

To construct the concept hierarchy of Thai biotechnology research, the proposed tree-structured networks algorithm is performed on the combined keywords and abstract phrases contained in the Thai biotechnology literature. The following two figures present the parent-child relationships among topics in malaria and penaeus-monodon.

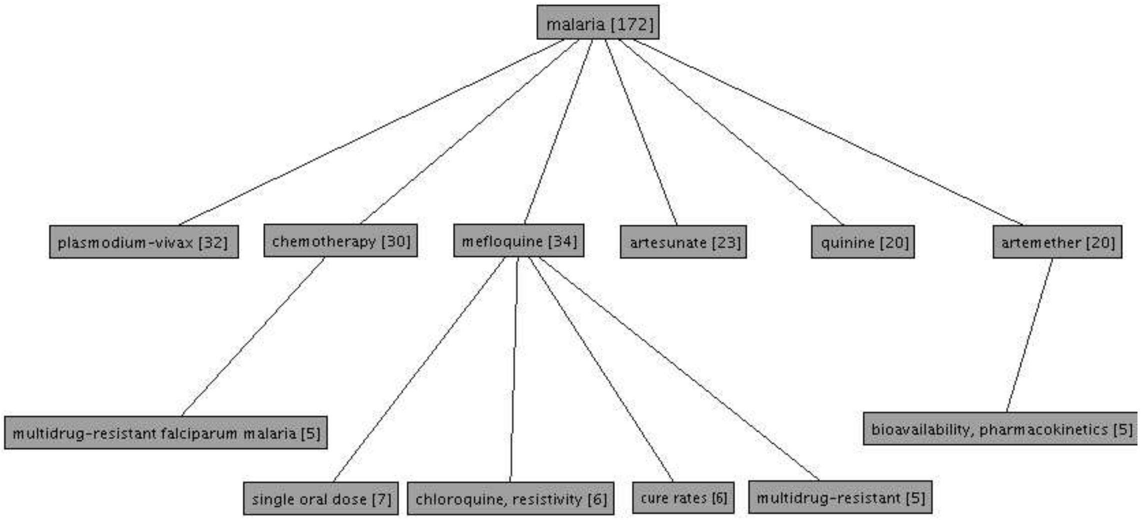


Figure 8.10: Parent-Child Relationships for Malaria Research

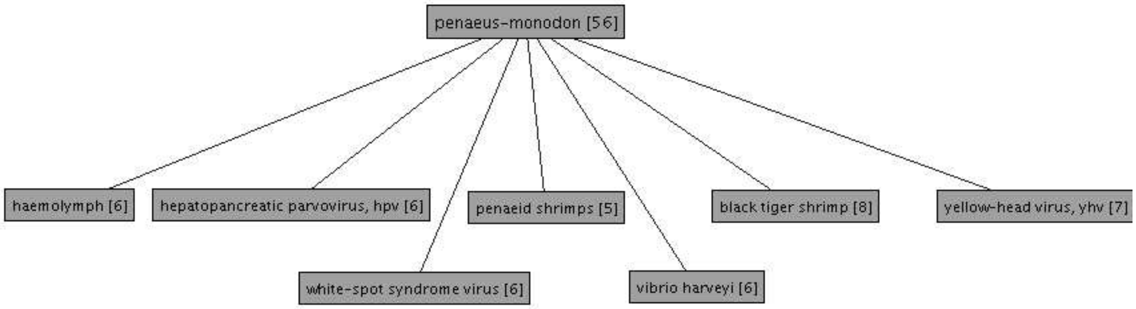


Figure 8.11: Parent-Child Relationships for Penaeus-Monodon Research

8.5.2.3 The Leading Research Organizations

The leading biotechnology research organizations in Thailand are listed below.

Table 8.5: Leading Research Organizations in Biotechnology Publications

Affiliation	Total Articles (record)	Longevity (year)	Publication rate (record/year)
Mahidol University	377	13	29
Chulalongkorn University	189	13	15
Asian Institute of Technology	138	13	11
Chiangmai University	109	12	10
Kasetsart University	102	12	9
National Center of Genetic Engineering and Biotechnology (BIOTEC)	99	8	13
Prince of Songkla University	79	12	7
Khon Kaen University	50	12	5
King Mongkut's Institute of Technology, Bangkok	46	13	4

Longevity indicates the total number of years that an organization published articles on biotechnology since 1991. The above table shows the main players in

Biotechnology research are again the universities. The leading governmental organization is the National Center of Genetic Engineering and Biotechnology (BIOTEC). BIOTEC is the newest institution among these. It was established in 1992. According to the publication rate, Mahidol University, Chulalongkorn University, Asian Institute of Technology, and BIOTEC dominate biotechnology research in Thailand.

8.5.2.4 The Focus of Leading Organizations

The following table presents the topics that each leading organization emphasizes.

Table 8.6: Focus Topics of Leading Organizations

Thai Affiliation	keywords	Concept-Grouped Abstract Phrases
	Top Terms	Top Terms

Mahidol University[377]	MALARIA [122];
	PLASMODIUM-FALCIPARUM [93];
	PENAEUS-MONODON [27];
	CHEMOTHERAPY [25];
	THALASSEMIA [21];
		malaria [98];
		plasmodium-falciparum [58];
		penaeus-monodon [26];
		conclusions [23];
		combinations [20];

Chulalongkorn University[189]	PENAEUS-MONODON [27];
	PLASMODIUM-FALCIPARUM [21];
	MALARIA [14];
	APIS-CERANA [11];
	Organic compounds [9];
		penaeus-monodon [22];
		plasmodium-falciparum [17];
		malaria [10];
		apis-cerana [8];
		concentrations [7];

Asian Institute of Technology[138]	Wastewater treatment [21];
	Biomass [15];
	Sewage sludge [11];
	Biofilms [11];
	Biodegradation [10];
		biomass [16];
		applicators [14];
		experimental results [13];
		wastewaters [13];
		productivity [10];

Chiangmai University[109]	Bacteria [6];
	POLYMERASE CHAIN REACTION [6];
	MONOCLONAL ANTIBODY [6];
	Ion implantation [5];
		conclusions [12];
		collections [9];
		Northern Thailand [9];
		new species [6];
		determined [6];

Kasetsart University[102]	acetic acid bacteria [11];
	CASSAVA [7];
	POLYMERASE CHAIN REACTION [5];
	Enzymes [4];
		acetic acid bacteria [9];
		efficiency [7];
		molecular masses [7];
		culture medium [6];
		cultivars [5];

BIOTEC[99]	Gelation [8];
	PENAEUS-MONODON [8];
	MALARIA [7];
	surimi [7];
	bigeye snapper [6];
		surimi [9];
		bigeye snapper [8];
		deformation [7];
		Chemical Industry [6];
		Escherichia coli [6];

Prince of Songkla University[79]	surimi [8];
	bigeye snapper [7];
	Gelation [7];
	Degradation [6];
	BIOLOGICAL CONTROL [5];
		surimi [9];
		bigeye snapper [8];
		Chemical Industry [7];
		deformation [7];
		concentrations [6];

8.5.2.5 Identifying Experts and Their Interests

The following table presents the leading research publishers in the field, along with their affiliation, and their area of expertise.

Table 8.7: Leading Experts and Their Area of Expertise

Thai Author	Author's Affiliation	keywords
topthai		Top Terms

LOOAREESUWAN, S[31]	Mahidol University	MALARIA [28];
		PLASMODIUM-FALCIPARUM [22];
		CHEMOTHERAPY [11];
		PLASMODIUM-VIVAX [7];
		ARTESUNATE [6];

PANYIM, S[30]	Mahidol University	PENAEUS-MONODON [11];
		Bacillus Thuringiensis [11];
		delta-endotoxins [5];
		WSSV [3];
		Mutagenesis [2];

Polprasert, Chongrak[28]	Asian Insitute of Technology	Wastewater treatment [8];
		Biofilms [6];
		Chemical oxygen demand [5];
		Tropics [4];
		Mathematical models [4];

Tanticharoen, Morakot[25]	BIOTEC	GAMMA-LINOLENIC ACID [7];
		SPIRULINA-PLATENSIS [5];
		Biosensors [4];
		Cyanobacteria [4];
		AMPEROMETRY [3];

KARBWANG, J[22]	Mahidol University	MALARIA [15];
		PLASMODIUM-FALCIPARUM [10];
		ARTEMETHER [9];
		MEFLOQUINE [8];
		PHARMACOKINETICS [8];

8.5.2.6 Collaboration among Research Organizations

The following maps present the collaboration among research organizations. The first map (Figure 8.12) shows the inter-organizational publishing activities. The connections present the organizations that have authors publishing together. From this map, Mahidol University and BIOTEC tend to interact with more organizations. The second map (Figure 8.13) shows organizations that write about the same things. This type of map can help discover overlapping or similar research activities among these centers. For instance, BIOTEC and Prince of Songkla University are both focusing their research on Penaeus-Monodon, gelation, and bigeye snapper.

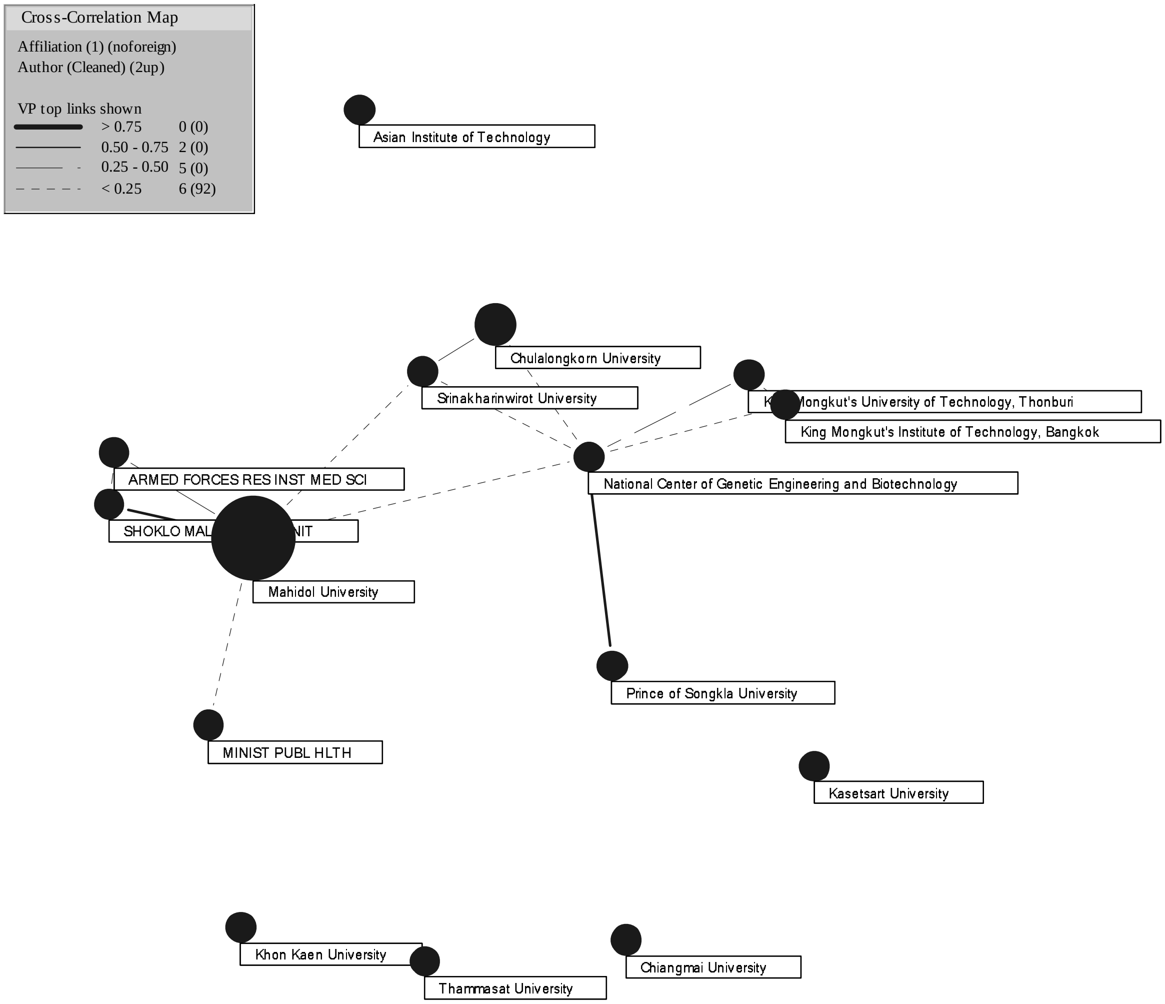

Cross-Correlation Map
Affiliation (1) (noforeign)
Author (Cleaned) (2up)
VP top links shown
> 0.75 0 (0)
0.50 - 0.75 2 (0)
0.25 - 0.50 5 (0)
< 0.25 6 (92)
Asian Institute of Technology
Chulalongkorn University
Srinakharinwirot University
K Mongkut's University of Technology, Thonburi
King Mongkut's Institute of Technology, Bangkok
National Center of Genetic Engineering and Biotechnology
ARMED FORCES RES INST MED SCI
SHOKLO MAL NIT
Mahidol University
Prince of Songkla University
MINIST PUBL HLTH
Kasetsart University
Khon Kaen University
Thammasat University
Chiangmai University

Figure 8.12: Collaboration among Research Organizations based on Authors

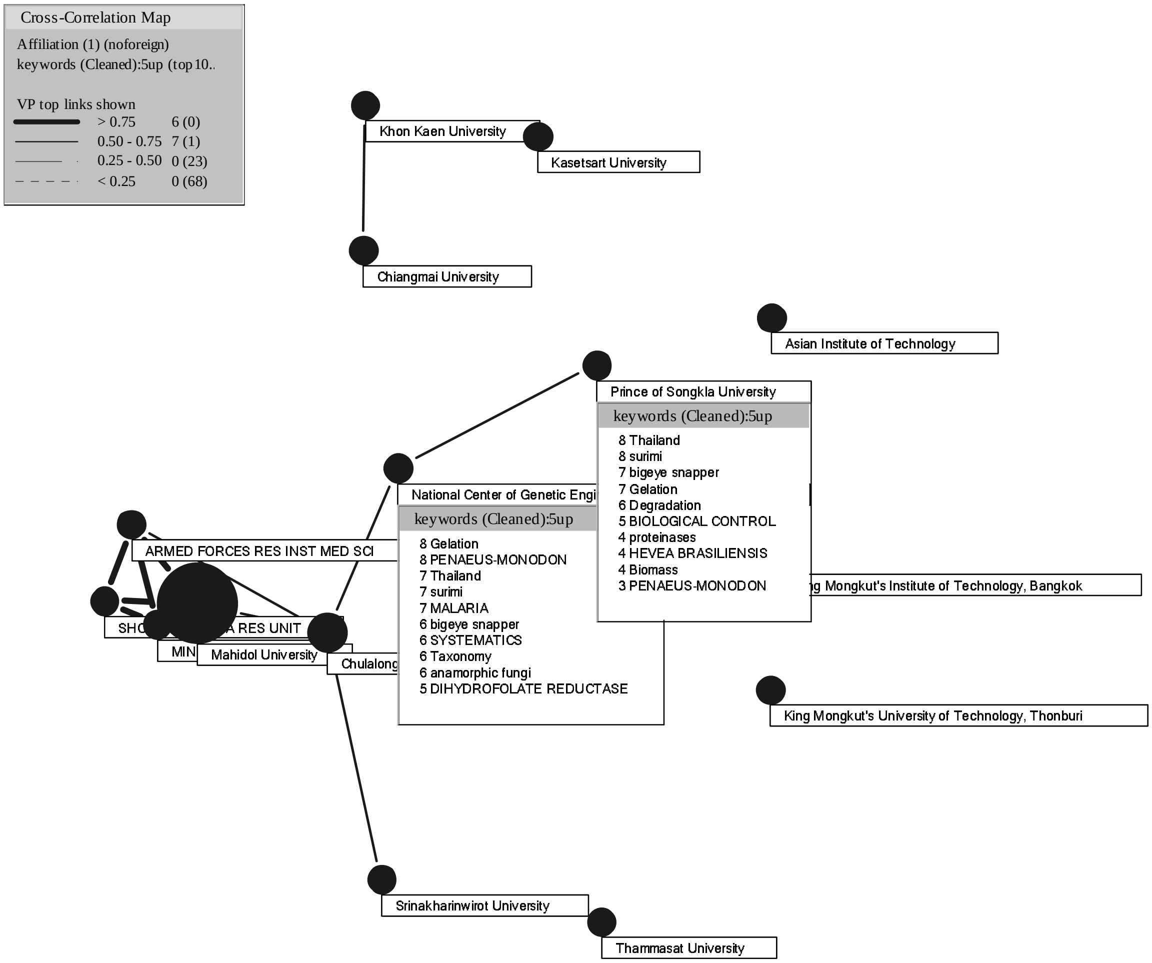


Figure 8.13: Common Interests Among Research Organizations

8.5.2.7 Expert Opinion

The results from literature mining for biotechnology in Thailand have been presented to Dr. Yongyuth Yuthavong, Senior Researcher, Thailand's National Center for Genetic Engineering and Biotechnology (BIOTEC). His feedback suggests that the results are very interesting and useful. He is not surprised that the results focus mostly on malaria and shrimp. He also stated, "The Thai researchers are not great publishers, especially in agricultural biotech. The main message from your work seems to be 'too few areas, too little connections'. However, something interesting could be discerned, for example, the two areas of waste/environment and malaria."

8.6 Comparison between Thai R&D Publications and Project Grants

Section 8.5.1 has already discussed the overall R&D situation in Thailand based on international publications. In this section, the Thai publication data are compared with the Thai-language grant data previously introduced in Section 8.4.1.2. Of the total 1,970 research projects, 860 (44 percent) are funded by the National Electronics and Computer Technology Center (NECTEC), 30 percent by the National Metal and Materials Technology Center (MTEC), 20 percent by the National Center for Genetic Engineering and Biotechnology (BIOTEC), and 4 percent by the central office of the National Science and Technology Development Agency. This distribution suggests high interest in electronics and computer technology. Table 8.8 tallies the leading institutions that are funded by the four funding organizations.

Table 8.8: Leading Funded Institutions from each Funding Agency

Source of Funds	Funded Institute

NECTEC [860]	NECTEC [301];
	Kasetsart University [23];
	Chulalongkorn University [17];
	Prince of Songkla University [16];
	King Mongkut's Institute of Technology, Bangkok [14]

MTEC [596]	MTEC [118];
	Chulalongkorn University [112];
	King Mongkut's Ins. of Tech. North Bangkok [45];
	Mahidol University [40];
	Prince of Songkla University [33]

BIOTEC [388]	BIOTEC [148];
	Mahidol University [63];
	Kasetsart University [39];
	King Mongkut's University of Technology, Thonburi [24];
	Chiangmai University [23]

The Central Office of NSTDA [127]	Mahidol University [15];
	King Mongkut's Institute of Technology, Bangkok [14];
	Chulalongkorn University [14];
	Khon Kaen University [13];
	Chiangmai University [12]

Some observations derive from these data. First, projects relating to electronics and computer technology and materials technology receive higher amounts of funding compared to the other areas. However, the number of international publications in electronics and computer technology and materials technology is not as high as in other areas, such as biotechnology. For instance, NECTEC and MTEC have 301 and 118 projects funded, but only 25 and 18 articles published in the international databases, respectively. BIOTEC, in contrast, has 148 projects funded and 170 articles published. The policy planner may want to encourage researchers in electronics and computer technology and materials technology to contribute more international publications. (Note that medical science is the most contributed area in the publications. However, the grants database does not cover projects in health and medical science.)

Second, it appears that most of the researchers receive the funding from their own institutes. The policy planner may want to examine the funding process more carefully.

The results from analyzing the publications from the international publication databases and the grants data suggest that there are misfits between them in some areas, such as electronics and computer technology and materials technology. This poses important issues for the policy planner to reconsider the existing funding process.

8.7 Comparison between Thai R&D Publications and Exports

This section examines how the Thai R&D community links with national industrial activity, especially that directed toward export. To achieve this goal, Thai R&D publications, classified into both scientific and industrial categories, and Thai export data are compared.

8.7.1 Thailand's Exports

While Thailand's economy traditionally has been agriculturally based, the industrial sector has begun to play a more significant role. Since the 1980s, the structure of Thailand's economy has shifted from an agricultural-based economy toward technology-based production and manufacturing. Table 8.9 presents Thailand's exports classified by Standard Industrial Trade Category (SITC) code from 1998-2002.

Table 8.9: Thailand's exports classified by SITC code from 1998-2000 (shown in percentage)

	1998	1999	2000	2001	2002
Food	17.8%	17.0%	14.4%	15.3%	14.4%
Crude materials	3.7%	3.3%	3.7%	3.4%	4.0%
Mineral fuel	1.4%	1.6%	2.7%	2.4%	2.4%
Animal and vegetable oils	0.1%	0.1%	0.1%	0.2%	0.1%
Chemicals	4.0%	4.7%	5.7%	5.5%	5.7%
Manufactures	15.4%	15.5%	15.4%	15.7%	16.2%
Machines	40.2%	42.1%	43.7%	42.3%	43.0%
Miscellaneous manufactures	13.1%	12.2%	11.3%	11.6%	10.8%
Miscellaneous transactions and commodities	4.1%	3.3%	2.8%	3.5%	3.2%
Re-exports	0.3%	0.2%	0.1%	0.1%	0.2%
Total export	100.0%	100.0%	100.0%	100.0%	100.0%

Source: Bank of Thailand (http://www.bot.or.th)

The above table suggests four sectors dominate Thailand's export profile in 2002:

- Machines (including electronic appliances and microelectronics, telecommunications equipment). This category alone accounts for 43 percent (up from 42.3 percent in 2001) of the total export.
- Basic manufactures (paper, textiles, iron and steel, etc.) accounted for 16.2 percent (up from 15.7 percent in 2001).
- Food (meat, dairy products, fish, cereals, vegetables, etc.) accounted for 14.4 percent (down from 15.3 percent in 2001)
- Miscellaneous manufactures (especially wearing apparel) accounted for 10.8 percent (down from 11.6 percent in 2001)

In 2002, the three categories of manufactures, miscellaneous manufactures, and machines together account for approximately 70 percent of the total exports. This implies that manufacturing dominates the economy. The natural resource sections, on the other hand, account for only a small percentage (approximately 6 percent in 2002).

8.7.2 Thailand's Industry Related Publications

The easy way to identify industry related publications is to link the main export trade categories with the classification codes provided from the INSPEC and ENGI databases. However, most of the Thai publications data are from the SCI database, which does not provide classification codes. Hence, instead of just matching INSPEC and ENGI classification codes, keywords and abstract phrases are employed as well. Table 8.10 categorizes the records according to the occurrence of classification codes, keywords, and abstract phrases that relate to the export trade categories. Note that this categorization is not mutually exclusive (i.e., one record can be categorized into both "crude materials" and "mineral fuels"). In addition, one more category called "information technology" is created for articles relating to software, information and communication systems, computing, etc., even though they are not identified in the export SITC categories. Information technology, such as software, which can be considered as a technology-intensive product, can help facilitate production and control processes.

Table 8.10: Thai Industry-Related Publication by Year

	1995	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	Total
Food	85	69	101	108	111	113	121	176	884
Crude Material	36	32	39	48	59	63	95	152	524
Mineral Fuels	11	16	23	27	19	22	66	98	282
Animal and Vegetable Oils	1	4	4	2	4	6	6	4	31
Chemicals	84	112	120	188	190	293	305	338	1630
Manufactures	11	8	25	9	24	25	14	32	148
Machines	62	91	154	208	219	263	325	377	1699
Miscellaneous Manufactures	0	0	0	0	1	1	0	1	3
Information Technology	35	49	72	147	153	178	170	211	1015
Non-Industry Research	475	484	486	627	615	728	797	961	5173
Total	800	865	1024	1364	1395	1692	1899	2350	11389
% of Industry-Related Research	40.6 %	44.0 %	52.5 %	54.0 %	55.9 %	57.0 %	58.0 %	59.1 %	

The non-industry related research abstracts (5,173 records) cover several categories for the following activities:

- Civil engineering
- Bioengineering
- Medical Science
- Environmental engineering
- Industrial engineering and management
- Superconductivity

The rest of the articles also cover fundamental research such as applied mathematics, applied physics, etc.

From Table 8.10, the average percentage of the industry-related publications captured from the total Thai publications is approximately 50 percent during 1995-2002. This is a relatively modest number. Figure 8.14 shows the trend of this percentage over this period.

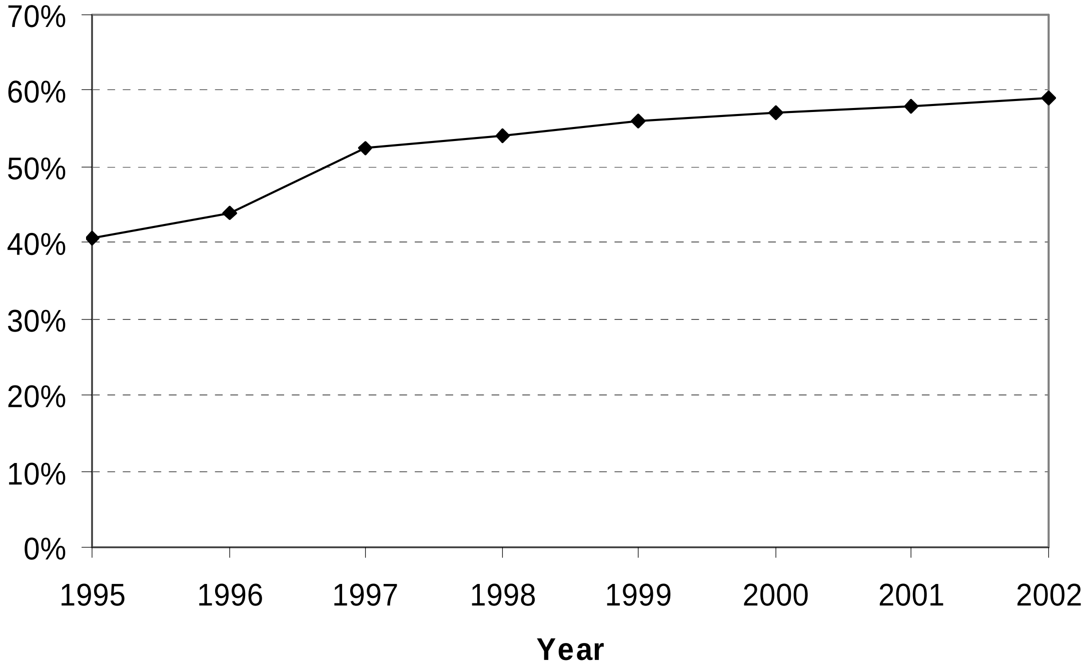


Figure 8.14: Percentage of Industry Related Publication Activity Per Year

The industry-related publications of Thai R&D appear to be increasing over time. Table 8.10 also suggests that the total number of articles in the machines, chemicals, and information technology categories dominate Thai industry-related publications during 1995-2002. From the previous section, the dominating export categories are machines, manufactures and food. Hence, there might be a linkage between research and industry activity in machines sector. Figure 8.15 profiles these industry-related publications and the respective export activities during the 1995-2002 period.

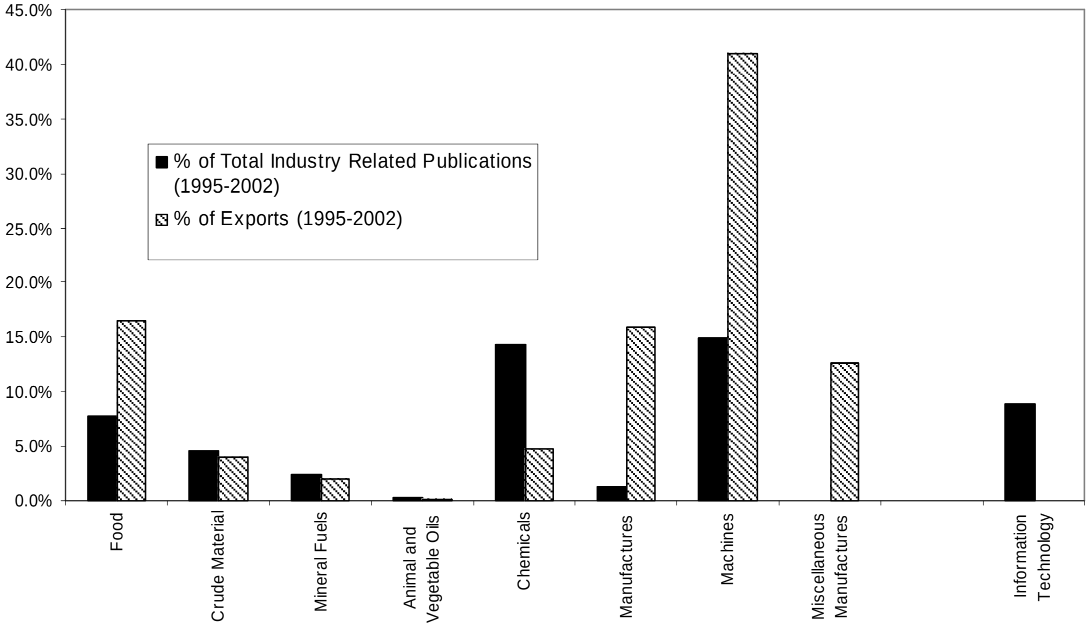


Figure 8.15: Comparison of Thai Export and R&D Publications Activity by SITC category

The figure presents several interesting observations. First, machine research, although accounting for the highest portion of R&D publications, does not match up to the relative level of the industry activity. Hence, policy planners might want to promote more R&D in this area. Second, chemicals research and information technology show active R&D programs with only fair exports. Conversely, food and manufactures dominate exports, but they show only moderate evidence of indigenous R&D, as measured by international publications. Third, R&D activity and export in crude materials, mineral fuels, and animal and vegetable oils appear to roughly correspond to the relative levels of export activity.

In addition to the Thai R&D publications and export data, expert opinion data obtained from Thai experts who participate in the Georgia Tech High Tech Competitiveness Indicators study[21] are considered. The study asks the experts to judge the nature of the competitiveness of products of selected industry sectors about 15 years in the future. The results obtained from the Thai experts suggest some indigenous production capability in chemicals and plastics, information technology (computers and software), and medicine and biologicals.

These results from analyzing the Thai R&D international publications, exports, and expert opinion data can help discern the following policy issues:

- There are mismatches between R&D emphases and industrial concentration for some sectors.
- If the long-term investments in machines, chemicals, and information technology are the goals, then Thailand could anticipate significant gains in these sectors in the future.
- R&D programs could be more actively promoted in the food and manufactures areas to improve the main current industrial domains.
- Decide who should be the national lead research institutions for certain areas. This can be achieved by analyzing the R&D publications. For instance, the publications suggest Asian Institute of Technology is the leading research

[21] Source: HTI reports (http://www.tpac.gatech.edu)

organization in information technology. Machine-related R&D is concentrated at the King Mongkut's Institute of Technology and Asian Institute of Technology.

In conclusion, Thailand is shifting the structure of its economy from an agricultural-based toward technology-based production and manufacturing. It shows considerable R&D capability based on international publications. Policy planning and priority setting should explore how best to link these capabilities with industrial targets.

8.8 Conclusions

This chapter presents the utilization of the proposed text mining framework with the Thai S&T publication abstracts toward the objective improving of R&D management. The overall R&D and biotechnology activities have been profiled. This text mining framework in conjunction with the grant data and the export data can help Thai R&D managers or policy planners discern the intelligence, which is a prime requirement for successful technology management. This information can serve R&D managers in several ways:

- identifying experts and leading research organizations who are the main players in R&D concerning a technology of interest
- identifying and classifying the main research areas and sub-areas in a large body of publication databases
- discovering overlapping or similar research activities among the R&D centers. This could point toward increasing research collaboration among them.
- discerning the mismatch between the R&D outputs and the project funding

- discerning the mismatch between the R&D emphases and industrial concentrations

The results of this sort of R&D profiling can support strategic decision-making on the direction and funding of S&T programs in Thailand.

CHAPTER 9

CONCLUSIONS AND FUTURE RESEARCH OPPORTUNITIES

In this concluding chapter, the text mining framework for discovering technological intelligence to support S&T management is summarized. Future research opportunities are discussed in the last section.

9.1 Conclusions

In this thesis, a framework based on text mining techniques is proposed to discern useful intelligence from the large body of electronic text sources. This intelligence is a prime requirement for successful S&T management. This research extends the approach called “Technology Opportunities Analysis” (TOA) developed by the Technology Policy and Assessment Center (TPAC), Georgia Institute of Technology, in conjunction with Search Technology Inc. (http://www.searchtech.com), to construct the proposed framework. The software called VantagePoint (http://thevantagepoint.com) is mainly used to perform the analyses. These include information extraction and cleansing, information summarization, data analysis, and data visualization. In addition to employing the basic functions of VantagePoint, this research advances the data clustering process. One of the key contributions of this dissertation is the implementation of a new text clustering method and derivative text cleansing method based on association rule mining approach.

In order to apply the association rule mining approach efficiently to the bibliographic data, a new algorithm called “object-oriented association rule mining (OOARM)” is developed. Mining association rules are being actively studied for transaction databases, but extension to text applications is relatively novel. Most of the previous studies implement an Apriori-like approach, which requires multiple passes over the database to find all the relevant data. However, for bibliographic databases where the word frequency distribution is different from the case of traditional sales transaction databases, the existing algorithm becomes too costly in time and resources. The proposed OOARM algorithm uses a special object-oriented data structure that holds all the requisite information. This enables making only a single scan through the database. The experimental results show that, by applying the OOARM algorithm, improvements in run time of up to 500 times can be realized compared to a traditional Apriori-like algorithm.

Two algorithms based on the OOARM are proposed in this research. The first algorithm called "tree-structured networks" can automatically construct parent-child (hierarchical structure) and sibling relationships (non-hierarchical structure) among concepts. Currently VantagePoint performs multidimensional statistical analysis to identify clusters and relationships among concepts. However, its approach does not produce sets of hierarchically related features. This hierarchical structure derived from the tree-structured networks can help identify emerging areas in an existing field of research, while the non-hierarchical structure can help understand interdisciplinary structures among related topical areas. The visualization of the hierarchical structure of a topic the user is interested in is also implemented. The experimental results on a publication data set show that tree-structured networks algorithm offers richer structural information on relationships in text sources than the existing approach (i.e., Principal Components Analysis) from VantagePoint.

The second algorithm based on the OOARM, called “concept-grouping,” is used to construct a thesaurus from synonymous phrases (terms). Similar abstract or title phrases, which are written differently, can be grouped together into the same concept based on their associations by this proposed algorithm. This data cleansing process is important since it has an impact on the quality of the subsequent analyses such as clustering. VantagePoint already has a list cleanup function based on word stemming and fuzzy matching to group like terms into one. However, this method is not sufficient since it cannot detect the similarity between terms such as “Internet” and “world wide web.” The concept-grouping algorithm, on the other hand, can be more efficient since similar terms are grouped together into a concept based on the probability that they occur together. Hence, combining stemming and concept-grouping approaches is most promising. To evaluate the effectiveness of using concept-grouped phrases in comparison to the cleaned abstract phrases (resulted from stemming algorithm), two factor maps generated from these two groups are created. The evaluation results show better cluster quality for the concept-grouped abstract phrases clusters compared to the cleaned abstract phrases clusters.

Using the above proposed algorithms, in conjunction with VantagePoint, a text mining framework for discovering technical intelligence to support S&T management is applied to the Thai S&T publication abstracts. The results provide the overview of research domains in Thailand. Moreover, the algorithms can aid in identifying the main technology infrastructure, overlapping research activities, prominent research topics and sub-topics area, etc. This knowledge derived from the publication data is compared with the grant and export data to discover the efficiency of the funding process and the linkage between the research community and industry. The results show strikingly different emphases. This poses challenging policy issues for the Thai R&D managers and/or policy planners.

This research presents the approach to help the Thai R&D manager/policy planner automatically gather the technological intelligence implicit in large bodies of electronic text sources to support S&T management. With the capability of computer technology today, these results can be produced in a timely fashion. This tool can help the NSTDA (the user of this study) achieve an effective system for support of R&D. Instead of taking months to conduct bibliometric analysis for Thailand R&D as they used to, such analysis can be done in a matter of days.

9.2 Future Research Opportunities

The possible extensions to the current proposed framework include:

- The concept-grouping algorithm could be combined with other data cleansing methods for more effective results. For instance, Cherie Courseault Trumbach is developing a data cleansing method based on shared words in each phrase for her dissertation. Terms such as “association rule” and “association rule mining” should be grouped together according to her algorithm.
- The text mining framework could be used in conjunction with the Thai word segmentation software. This is useful for the case that the Thai databases do not provide a keywords field. An attractive potential application is to apply this software to parse Thai sentences into words, and then perform the regular factor

analysis or tree-structured network analysis to discover relationships among these words.

APPENDIX

A1: Complete term clusters generated from Principal Components Analysis

Cluster 1	Query languages, Data reduction, Distributed database systems, Indexing (of information, Online systems, EiRev
Cluster 2	Fuzzy sets, Computer simulation, Computational linguistics, Fuzzy association rules, Membership functions, Linguistics, Approximation theory
Cluster 3	Genetic algorithms, Decision trees, Decision theory, Artificial intelligence, Neural networks, Computer aided software engineering
Cluster 4	query processing, relational databases, Data warehouses, software performance evaluation, Decision support systems, SQL, Parallel programming, Asynchronous transfer mode, parallel databases, workstation clusters
Cluster 5	Parallel algorithms, Parallel processing systems, Data storage equipment, Response time (computer systems), Asynchronous transfer mode, Computer networks, Computer software, Distributed computer systems, Storage allocation (computer), Computer systems programming
Cluster 6	information resources, Information retrieval, Internet, Search engines, hypermedia
Cluster 7	Database systems, Association rules, Knowledge discovery, Theorem proving, Apriori algorithms
Cluster 8	Knowledge acquisition, deductive databases, transaction processing, marketing data processing, retail data processing, sales management
Cluster 9	Decision making, Semantics, Computer architecture, Expert systems, Hierarchical systems
Cluster 10	learning (artificial intelligence), neural nets
Cluster 11	Knowledge discovery, Classification (of information)

A2: Complete term clusters generated from Hierarchical Agglomerative Clustering

Cluster 1	Graph theory, Optimization, Logic programming, Boolean algebra, interactive systems, User interfaces, Constraint theory, Online searching, Redundancy
Cluster 2	information resources, Information retrieval, Internet, Electronic commerce, World Wide Web, data visualization, Search engines, Text analysis, Visualization, document handling, Software agents, classification, constraint handling, distributed algorithms, Graphical user interfaces, hypermedia, multimedia databases
Cluster 3	Marketing, Financial data processing, Information analysis, Industrial management, Websites
Cluster 4	database management systems, Genetic algorithms, learning (artificial intelligence), Pattern recognition, Knowledge representation, Decision trees, Decision theory, statistical analysis, Probability, Artificial intelligence, Neural networks, Data handling, inference mechanisms, neural nets, Statistics, belief networks, Computer aided software engineering, unsupervised learning
Cluster 5	bibliographic systems, Medical computing, medical information systems
Cluster 6	pattern clustering, fuzzy set theory, pattern classification, fuzzy logic, Temporal databases, visual databases, uncertainty handling, learning by example
Cluster 7	Mathematical models, Rough set theory, search problems, Equivalence classes
Cluster 8	Computational geometry
Cluster 9	Database systems, Knowledge based systems, Associative processing, Association rules, Data structures, Fuzzy sets, Set theory, Association rule mining, Computational complexity, Query languages, Data reduction, Learning systems, Knowledge discovery, Relational database systems, Computer simulation, Trees (mathematics), Parallel processing systems, Computational linguistics, tree data structures, Data acquisition, Data processing, Data storage equipment, Fuzzy association rules, Information retrieval systems, Classification (of information), Knowledge engineering, Learning algorithms, Problem solving, Statistical methods, Boolean functions, Decision making, Membership functions, Semantics, Theorem proving, Response time (computer systems), Security of data, Apriori algorithms, Computer networks, Distributed database systems, Indexing (of information), Online systems, Performance, Computational methods, Computer architecture, Computer software, Distributed computer systems, EiRev, Inference engines, Linguistics, Sequences, Storage allocation (computer), Approximation theory, Computer systems programming, Expert systems, Hierarchical systems, Remote sensing
Cluster 10	Mining, Mineral industry

Cluster 11	Feature extraction, Correlation methods, Pattern matching
Cluster 12	Parallel algorithms, Parallel programming, Asynchronous transfer mode, storage management, workstation clusters, Resource allocation
Cluster 13	Knowledge acquisition, database theory, query processing, deductive databases, relational databases, transaction processing, Data warehouses, software performance evaluation, data analysis, Decision support systems, very large, marketing data processing, retail data processing, SQL, business data processing, distributed databases, optimization, data models, Management information systems, parallel databases, time series, database indexing, sales management, statistical databases, Risk management, Text processing

A3: Complete list of ENGI Classification Codes

400	CIVIL ENGINEERING, GENERAL
403	Urban and Regional Planning and Development
404	Civil Defense and Military Engineering
405	Construction Equipment and Methods; Surveying
406	Highway Engineering
407	Maritime and Port Structures; Rivers and Other Waterways
408	Structural Design
409	Civil Engineering, General
410	CONSTRUCTION MATERIALS
411	Bituminous Materials
412	Concrete
413	Insulating Materials
414	Masonry Materials
415	Metals, Plastics, Wood and Other Structural Materials
420	BUILDING MATERIALS PROPERTIES AND TESTING
421	Strength of Building Materials; Mechanical Properties
422	Strength of Building Materials; Test Equipment and Methods
423	Non Mechanical Properties and Tests of Building Materials
430	TRANSPORTATION
431	Air Transportation
432	Highway Transportation
433	Railroad Transportation
434	Waterway Transportation
440	WATER AND WATERWORKS ENGINEERING
441	Dams and Reservoirs; Hydro Development
442	Flood Control; Land Reclamation
443	Meteorology
444	Water Resources
445	Water Treatment
446	Waterworks
450	POLLUTION, SANITARY ENGINEERING, WASTES
451	Air Pollution
452	Sewage and Industrial Wastes Treatment
453	Water Pollution
454	Environmental Engineering
460	BIOENGINEERING
461	Bioengineering
462	Biomedical Equipment

470	OCEAN AND UNDERWATER TECHNOLOGY
471	Marine Science and Oceanography
472	Ocean Engineering
480	ENGINEERING GEOLOGY
481	Geology and Geophysics
482	Mineralogy
483	Soil Mechanics and Foundations
484	Seismology
500	MINING ENGINEERING, GENERAL
501	Exploration and Prospecting
502	Mines and Quarry Equipment and Operations
503	Mines and Mining, Coal
504	Mines and Mining, Metal
505	Mines and Mining, Nonmetallic
506	Mining Engineering, General
510	PETROLEUM ENGINEERING
511	Oil Field Equipment and Production Operations
512	Petroleum and Related Deposits
513	Petroleum Refining
520	FUEL TECHNOLOGY
521	Fuel Combustion and Flame Research
522	Gas Fuels
523	Liquid Fuels
524	Solid Fuels
525	Energy Management
530	METALLURGICAL ENGINEERING, GENERAL
531	Metallurgy and Metallography
532	Metallurgical Furnaces
533	Ore Treatment and Metal Refining
534	Foundries and Foundry Practice
535	Rolling, Forging and Forming
536	Powder Metallurgy
537	Heat Treatment
538	Welding and Bonding
539	Metals Corrosion and Protection; Metal Plating
540	METALLURGICAL ENGINEERING, METAL GROUPS
541	Aluminum and Alloys
542	Beryllium, Lithium, Magnesium, Titanium and Other Light Metals and Alloys

543	Chromium, Manganese, Molybdenum, Tantalum, Tungsten, Vanadium and Alloys
544	Copper and Alloys
545	Iron and Steel
546	Lead, Tin, Zinc, Antimony and Alloys
547	Minor, Precious and Rare Earth Metals and Alloys
548	Nickel and Alloys
549	Nonferrous Metals and Alloys
600	MECHANICAL ENGINEERING, GENERAL
601	Mechanical Design
602	Mechanical Drives and Transmissions
603	Machine Tools
604	Metal Cutting and Machining
605	Small Tools and Hardware
606	Abrasives
607	Lubricants and Lubrication
608	Mechanical Engineering, General
610	MECHANICAL ENGINEERING, PLANT AND POWER
611	Hydro and Tidal Power Plants
612	Internal Combustion Engines
613	Nuclear Power Plants
614	Steam Power Plants
615	Thermoelectric, Magnetohydrodynamic and Other Power Generators
616	Heat Exchangers
617	Turbines and Steam Turbines
618	Compressors and Pumps
619	Pipes, Tanks and Accessories; Plant Engineering Generally
620	NUCLEAR TECHNOLOGY
621	Nuclear Reactors
622	Radioactive Materials
630	FLUID FLOW; HYDRAULICS, PNEUMATICS AND VACUUM
631	Fluid Flow
632	Hydraulics, Pneumatics and Related Equipment
633	Vacuum Technology
640	HEAT AND THERMODYNAMICS
641	Heat and Mass Transfer; Thermodynamics
642	Industrial Furnaces and Process Heating
643	Space Heating and Air Conditioning
644	Refrigeration and Cryogenics
650	AEROSPACE ENGINEERING

651	Aerodynamics
652	Aircraft
653	Aircraft Engines
654	Rockets and Rocket Propulsion
655	Spacecraft
656	Space Flight
657	Space Physics
658	Aerospace Engineering, General
660	AUTOMOTIVE ENGINEERING
661	Automotive Engines and Related Equipment
662	Automobiles and Smaller Vehicles
663	Buses, Tractors and Trucks
664	Automotive Engineering, General
670	NAVAL ARCHITECTURE AND MARINE ENGINEERING
671	Naval Architecture
672	Naval Vessels
673	Shipbuilding and Shipyards
674	Small Craft and Other Marine Craft
675	Marine Engineering
680	RAILROAD ENGINEERING
681	Railroad Plant and Structures
682	Railroad Rolling Stock
690	MATERIALS HANDLING
691	Bulk Handling and Unit Loads
692	Conveyors and Elevators
693	Cranes and Derricks
694	Packaging
700	ELECTRICAL ENGINEERING, GENERAL
701	Electricity and Magnetism
702	Electric Batteries and Fuel Cells
703	Electric Circuits
704	Electric Components and Equipment
705	Electric Generators and Motors
706	Electric Transmission and Distribution
707	Illuminating Engineering
708	Electric & Magnetic Materials
709	Electrical Engineering, General
710	ELECTRONICS AND COMMUNICATION ENGINEERING
711	Electromagnetic Waves
712	Electronic and Thermionic Materials
713	Electronic Circuits
714	Electronic Components and Tubes
715	Electronic Equipment, General

	Purpose and Industrial
716	Electronic Equipment, Radar, Radio and Television
717	Electro-Optical Communication
718	Telephone and Other Line Communications
720	COMPUTERS AND DATA PROCESSING
721	Computer Circuits and Logic Elements
722	Computer Hardware
723	Computer Software, Data Handling and Applications
730	CONTROL ENGINEERING
731	Automatic Control Principles and Applications
732	Control Devices
740	LIGHT AND OPTICAL TECHNOLOGY
741	Light, Optics and Optical Devices
742	Cameras & Photography
743	Holography
744	Lasers
745	Printing & Reprography
750	SOUND AND ACOUSTICAL TECHNOLOGY
751	Acoustics, Noise. Sound
752	Sound Devices, Equipment and Systems
753	Ultrasonics and Applications
800	CHEMICAL ENGINEERING, GENERAL
801	Chemistry
802	Chemical Apparatus and Plants; Unit Operations; Unit Processes
803	Chemical Agents and Basic Industrial Chemicals
804	Chemical Products Generally
805	Chemical Engineering, General
810	CHEMICAL ENGINEERING, PROCESS INDUSTRIES
811	Cellulose, Paper & Wood Products
812	Ceramics, Refractories and Glass
813	Coatings and Finishes
814	Leather and Tanning
815	Polymers and Polymer Science
816	Plastics Processing and Machinery
817	Plastics, Products and Applications
818	Rubber and Elastomers
819	Synthetic and Natural Fibers; Textile Technology
820	AGRICULTURE ENGINEERING AND FOOD TECHNOLOGY
821	Agricultural Equipment and Methods
822	Food Technology
900	ENGINEERING, GENERAL
901	Engineering Profession
902	Engineering Graphics; Engineering Standards; Patents
903	Information Science
910	ENGINEERING MANAGEMENT
911	Cost and Value Engineering; Industrial Economics
912	Industrial Engineering and Management
913	Production Planning and Control; Manufacturing
914	Safety Engineering
920	ENGINEERING MATHEMATICS
921	Applied Mathematics
922	Statistical Methods
930	ENGINEERING PHYSICS
931	Applied Physics Generally
932	High Energy Physics; Nuclear Physics; Plasma Physics
933	Solid State Physics
940	INSTRUMENTS AND MEASUREMENT
941	Acoustical and Optical Measuring Instruments
942	Electric and Electronic Measuring Instruments
943	Mechanical and Miscellaneous Measuring Instruments
944	Moisture, Pressure and Temperature, and Radiation Measuring Instruments

A4: Complete list of INSPEC Classification Codes

A00	General physics
A01	Communication education history and philosophy
A02	Mathematical methods in physics
A03	Classical and quantum physics; mechanics and fields
A04	Relativity and gravitation
A05	Statistical physics and thermodynamics
A06	Measurement science general laboratory techniques and instrumentation systems
A07	Specific instrumentation and techniques of general use in physics
A10	The physics of elementary particles and fields
A11	General theory of fields and particles
A12	Specific theories and interaction models; particle systematics
A13	Specific reactions and phenomenology
A14	Properties of specific particles and resonances
A20	Nuclear physics
A21	Nuclear structure
A23	Radioactivity and electromagnetic transitions
A24	Nuclear reactions and scattering: general
A25	Nuclear reactions and scattering: specific reactions
A27	Properties of specific nuclei listed by mass ranges
A28	Nuclear engineering and nuclear power studies
A29	Experimental methods and instrumentation for elementary particle and nuclear physics
A30	Atomic and molecular physics
A31	Theory of atoms and molecules
A32	Atomic spectra and interactions with photons
A33	Molecular spectra and interactions with photons
A34	Atomic and molecular collision processes and interactions
A35	Properties of atoms and molecules: instruments and techniques

A36	Studies of special atoms and molecules
A40	Fundamental areas of Phynomenology
A41	Electricity and magnetism; fields and charged particles
A42	Optics
A43	Acoustics
A44	Heat flow thermal and thermodynamic processes
A46	Mechanics elasticity rheology
A47	Fluid dynamics
A50	Fluids plasmas and electric discharges
A51	Kinetic and transport theory of fluids; physical properties of gases
A52	The physics of plasmas and electric discharges
A60	Condensed Matter: structure thermal and mechanical properties
A61	Structure of liquids and solids; crystallography
A62	Mechanical and acoustic properties of condensed matter
A63	Lattice dynamics and crystal statistics
A64	Equations of state phase equilibria and phase transitions
A65	Thermal properties of condensed matter
A66	Transport properties of condensed matter (nonelectronic)
A67	Quantum fluids and solids; liquid and solid helium
A68	Surfaces and interfaces; thin films and whiskers
A70	Condensed matter: electronic structure electrical magnetic and optical properties
A71	Electron states in condensed matter
A72	Electronic transport in condensed matter
A73	Electronic structure and electrical properties of surfaces interfaces and thin films
A74	Superconductivity
A75	Magnetic properties and materials

A76	Magnetic resonances and relaxation in condensed matter; Mössbauer effect
A77	Dielectric properties and materials
A78	Optical properties and condensed matter spectroscopy and other interactions of matter with particles and radiation
A79	Electron and ion emission by liquids and solids; impact phenomena
A80	Cross-disciplinary physics and related areas of science and technology
A81	Materials science
A82	Physical chemistry
A86	Energy research and environmental science
A87	Biophysics medical physics and biomedical engineering
A90	Geophysics astronomy and astrophysics
A91	Solid Earth physics
A92	Hydrospheric and lower atmospheric physics
A93	Geophysical observations instrumentation and techniques
A94	Aeronomy space physics and cosmic rays
A95	Fundamental astronomy and astrophysics instrumentation and techniques and astronomical observations
A96	Solar system
A97	Stars
A98	Stellar systems; Galactic and extragalactic objects and systems; Universe
B00	General topics engineering mathematic and materials science
B01	General electrical engineering topics
B02	Engineering mathematics and mathematical techniques
B05	Materials science for electrical and electronic engineering
B10	Circuit theory and circuits
B11	Circuit theory
B12	Electronic circuits
B13	Microwave technology
B20	Components electron devices and materials
B21	Passive circuit components cables switches and connectors
B22	Printed circuits hybrid integrated circuits and molecular electronics
B23	Electron tubes
B25	Semiconductor materials and devices
B28	Dielectric materials and devices
B30	Magnetic and superconducting materials and devices
B31	Magnetic materials and devices
B32	Superconducting materials and devices
B40	Optical materials and applications electro-optics and optoelectronics
B41	Optical materials and devices
B42	Optoelectronic materials and devices
B43	Lasers and masers
B50	Electromagnetic fields
B51	Electric and magnetic fields
B52	Electromagnetic waves antennas and propagation
B60	Communications
B61	Information and communication theory
B62	Telecommunication
B63	Radar and radionavigation
B64	Radio television and audio
B70	Instrumentation and special applications
B71	Measurement science
B72	Measurement equipment and instrumentation systems
B73	Measurement of specific variables
B74	Elementary particle and nuclear instrumentation
B75	Medical physics and biomedical engineering
B76	Aerospace facilities and techniques
B77	Earth sciences
B78	Sonics and ultrasonics
B79	Military systems and equipment
B80	Power systems and applications
B81	Power networks and systems
B82	Generating stations and plants
B83	Power apparatus and electric machines
B84	Direct energy conversion and energy storage
B85	Power utilisation
B86	Industrial applications of power

C00	General and management topics
C01	General control topics
C02	General computer topics
C03	Management topics
C10	Systems and control theory
C11	Mathematical techniques
C12	Systems theory and cybernetics
C13	Control theory
C30	Control technology
C31	Control and measurement of specific variables
C32	Control equipment and instrumentation
C33	Control applications
C40	Numerical analysis and theoretical computer topics
C41	Numerical analysis
C42	Computer theory
C50	Computer hardware
C51	Circuits and devices
C52	Logic design and digital techniques
C53	Computer storage equipment and techniques
C54	Analogue and digital computers and systems
C55	Computer peripheral equipment
C56	Data communication equipment and techniques
C60	Computer software
C61	Software techniques and systems
C70	Computer applications
C71	Business and administration
C72	Information science and documentation
C73	Natural sciences computing
C74	Engineering computing
C78	Other computer applications
D10	Information Technology: General & management aspects
D20	Information Technology: Applications
D30	Information Technology: General systems and equipment
D40	Information Technology: Office automation - communications
D50	Information Technology: Office automation –computing

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In Proc. of ACM SIGMOD Conference on Management of Data, pp. 207-216, Washington, D.C.

Agrawal, R. and Srikant, R. (1994a). Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, Santiago, Chile, pp. 487-499.

Agrawal, R. and Srikant, R. (1994b). Fast algorithms for mining association rules in large databases. Research Report RJ 9839, IBM Almaden Research Center, San Jose, CA.

Baeza-Yates, R. and Ribeiro-Neto, B., eds. (1999). Modern Information Retrieval, ACM Press, Addison Wesley.

Borner, K., Chen, C. and Boyack, K. (2003). Visualizing Knowledge Domains. Annual Review of Information Science and Technology, Vol. 37.

Brin S., Motwani R., Ullman J., and Tsur S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. Proc of the ACM SIGMOD International Conference on Management of Data, pp. 255-264.

Callon, M., Courtial, J.P., Turner W.A. and Bauin S. (1983). From translations to problematic networks: an introduction to co-word analysis. Social Science Information, 22, pp. 191-235.

Caraballo, S.A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In Proceeding of the 37th Annual Meeting of the Association for Computational Linguistics.

Chen M.S., Han J., and Yu P.S. (1996). Data Mining: An Overview from a Database Perspective. IEEE Trans. On Knowledge and Data Engineering. Vol. 8, No. 6, pp. 866-883.

Cooley R., Mobasher B, and Srivastava J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems, Vol. 1, No. 1.

Cunningham, S.W. (1996). Ph.D. dissertation: The Content Evaluation of British Scientific Research, Science Policy Research Unit, University of Sussex, Brighton, U.K.

Cunningham, S.W. (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, pp. 149-160.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of American Society of Information Science, 41,pp. 391-407.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI magazine. 17(3), pp. 37-54.

Garfield, E. (1979). Is Citation Analysis a Legitimate Evaluation Tool?. Scientometrics, 1, pp. 359-375.

Garfield, E., Sher, I.H., and Torpie, R.J., (1964). The Use of Citation Data for Writing the History of Science, Philadelphia, Institute for Scientific Information.

Hair, J.F. et al. (1992). Multivariate data analysis (3rd ed.). New York: Macmillan.

Harman, H.H. (1976). Modern Factor Analysis (3rd ed.). Baltimore, MD: Johns Hopkins University Press.

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France.

Hearst, M.A. (1998). Automated discovery of WordNet relations. In Christiane Fellbaum, editor, Worldnet: An Electronic Lexical Database. MIT Press.

Houtsma, M., and Swami, A. (1993). Set-oriented mining for association rules in relational databases. Technical Report RJ 9567, IBM Almaden Research Center, San Jose, California.

Karypis, G. (2002). CLUTO 2.1: Software Package for Clustering High-Dimensional Data sets. http://www-users.cs.umn.edu/~karypis/cluto/index.html.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM- Self-organizing maps of document collections. Neurocompting. 21(1-3), pp. 101-117.

Kohonen, T. (1995), Self-Organizing Maps: Springer.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). Self organization of a massive document collection. IEEE Transactions of Neural Networks, 11(3), pp. 574-585.

Kostoff, R.N. (1997). The Handbook of Research Impact Assessment, Seventh Edition, DTIC Report Number ADA296021. available at (http://www.onr.navy.mil/sci_tech/special/technowatch/reseval.htm)

Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J. (1996). Engineering applications of the self-organizing map. Proceedings of the IEEE, 84(10), pp. 1358-84.

Kostoff, R. N. (2000). Science and Technology Text Mining. Keynote presentation at The Technical Cooperation Program (TTCP) meeting on the International Technology Watch Partnership (ITWP), 12 October 2000, Farnborough, UK.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., and Humenik, J. A. (2001). Text Mining Using Database Tomography and Bibliometrics: A Review. Technology Forecasting and Social Change, 68 (3), pp. 223-253.

Losiewicz, P., Oard, D. W., and Kostoff, R. N. (2000). Textual data mining to support science and technology management. Journal of Intelligent Information Systems, 15, pp. 99-119.

Lin, X. (1997). Map displays for information retrieval. Journal of the Americal Society for Information Science, 48(1), pp. 40-54.

Melkers, J. (1993). Bibliometrics as a tool for analysis of R&D impacts, in Evaluating R&D Impacts: Methods and Practice. B. Bozeman and J. Melkers, eds. Kluwer, Boston, pp. 43-61.

Miller, G.A. (1995). WorldNet: A lexical database for English. In the Communications of the ACM, 38(11), pp. 39-41.

Narin, F. and Olivastro, D. (1988). Technology Indicators Based on Patents and Patent Citations. Handbook of Quantitative Studies of Science and Technology, A.F.J. van Raan (ed.). North-Holland, Elsevier Science Publishers.

Narin, F. and Olivastro, D. (1994). Bibliometrics/Theory, Practice and Problems. Evaluation Review, Feb94, Vol. 18, Issue 1, pp. 65-77.

Noyons, E. C. M. and Van Raan, A. F. J. (1998). Advanced mapping of science and technology. Scientometrics, 41(1-2), pp. 61-67.

Porter, A. L. (2000). Text Mining for Technology Foresight. http://www.tpac.gatech.edu.

Porter, A. L., Roessner, J. D., Jin, X-Y, and Newman, N. C. (2001). Changes in National Technological Competitiveness:1990, 1993, 1996 and 1999. Technology Analysis & Strategic Management. 13 (4).

Porter, A. L. and Detampel, M. J. (1995). Technology Opportunities Analysis. Technological Forecasting & Social Change, Vol. 49, pp. 237-255.

Porter, A. L., Kongthon, A., and Lu, J. C. (2002a). Research profiling: Improving the literature review, Scientometrics, Vol. 53, No. 3, pp. 351-370.

Porter, A. L., Roessner, J. D., Jin, X-Y, Newman N. C. (2002b). Measuring national “emerging technology” capabilities. Science and Public Policy, Vol. 29, No. 3.

Porter, M. E. and Stern, S. (1999). The New Challenge to America’s Prosperity: Findings from the Innovation Index, Council on Competitiveness, Washington, DC, (also, see: http://www.compete.org/).

Price, D. de Solla (1963). Big Science, Little Science, New York: Columbia University Press.

Press, W.H., Flannery, B.P., Teukolsky, S. A., and Vetterling, W.T. (1986). Numerical Recipies in C: The Art of Scientific Computing, Cambridge University Press: Cambridge.

Salton, G. and McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Salton, G. (1988). Automatic Text Processing. Addison-Wesley Publishing Company.

Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In the 22nd ACM SIGIR Conference, pp. 206-213.

Schvaneveldt, R.W. (1990). Pathfinder Associative Networks: Studies in Knowledge Organization, Norwood, NJ: Ablex Publishing.

Singh, L., Scheuermann, P., and Chen, B. (1997). Generating Association Rules from Semi-Structured Documents: Using an Extended Concept Hierarchy. Proceedings of the International Conference on Information & Knowledge Management (CIKM).

Small, H. and Griffith, B. (1974). The structure of scientific literatures. Science Studies, 4, pp. 17-40.

Small, H. (1973). Cocitation in the scientific literature : a new measure of the relationship between two document. Journal of the American Society for Information Science, 24, pp. 265-269.

Small, H., and Greenlee, E. (1989). A Co-Citation Study of AIDS Research. Communication Research, Vol. 16, No. 5, October, 1989, pp. 642-666.

Small, H.(1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50(9), pp. 799-813.

Sneath, P. H. A., and Sokal, R. R. (1973). Numerical Taxonomy. Freeman, London, UK.

Sokal, R. R., and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 38, pp. 1409-1438.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. University of Minnesota, Technical Report #00-034. http://www.cs.umn.edu/tech_reports

TPAC (1998). Georgia Tech Research and Development on TOA. http://www.tpac.gatech.edu

TPAC (1999). Algorithms and Processes of Bigmap. Technical Report for TOAS Project. http://www.tpac.gatech.edu

Watts, R.J., and Porter, A.L. (1997). Innovation Forecasting. Technological Forecasting and Social Change, Vol. 56, pp. 25-47.

Watts, R.J., Porter, A.L., Cunningham, S.W., and Zhu, D. (1997). VantagePoint Intelligence Mining: Analysis of Natural Language Processing and Computational Linguistics. Principles of Data Mining and Knowledge Discovery (First European Symposium, PKDD'97, Trondheim, Norway), J. Komorowski and J Zytkow, eds., pp. 323-335: Springer.

Watts, R.J., Porter, A.L., and Zhu, D. (2002). Factor Analysis Optimization: Applied in Natural Language Knowledge Discovery. CODATA 2002 18th international conference: Frontiers of Scientific and Technical Data, Montreal, Canada.

White, H.D., and McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 50(13), pp. 1224-1233.

Yuthavong, Y. and Wojcik, A.M. (1997). Science and Technology in Thailand: Lessons from a Developing Economy. NSTDA/UNESCO Publishing.

Zhu, D. (1998). Technology Mapping: Extracting Patterns from S&T Databases. TPAC internal paper, http//www.tpac.gatech.edu.

Zhu, D., and Porter, A.L. (2002). Automated Extraction and Visualization of Information for Technology Intelligence and Forecasting. Technological Forecasting and Social Change, Vol. 69, pp. 495-506.

VITA

Alisa Kongthon was born on June 14, 1974 in Bangkok, Thailand. When she was 17 years old, she received a long-term scholarship from the Royal Thai Government to pursue her study in the U.S. During 1991-1992, she spent her senior high school year at Tabor Academy in Marion, Massachusetts. In September 1992, she entered the University of Rochester, New York (without knowing how cold it can be during the winter there). She received a Bachelor of Science in Electrical Engineering with a high distinctive honor in 1996. After four long winters in Rochester, she thought she has had enough with the cold so she decided to move across the country to California, the Golden State! She enrolled at the University of Southern California and received her Master of Science in Industrial Engineering in 1998. Driving across the country was very enjoyable for her, Alisa decided to make her last long journey from Los Angeles to Atlanta. In August 1998, she joined the Ph.D. program at the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia. She received a Master of Science in Operations Research in 1999 and her Ph.D. in Industrial Engineering in May 2004.