A DESCRIPTIVE FRAMEWORK FOR THE FIELD OF DATA MINING AND KNOWLEDGE DISCOVERY

YI PENG†,§ and GANG KOU†,‡
School of Management and Economics
University of Electronic Science and Technology of China
Chengdu 610054, China
pengyi@uestc.edu.cn
kougang@yahoo.com

YONG SHI
CAS Research Center on Fictitious Economy
and Data Sciences, Beijing 100080, China
College of Information Science and Technology
University of Nebraska at Omaha
NE 68182, USA
yshi@gscas.ac.cn

ZHENGXIN CHEN
College of Information Science & Technology
University of Nebraska at Omaha
NE 68182, USA
zchen@mail.unomaha.edu

Despite the rapid development, the field of data mining and knowledge discovery (DMKD) is still vaguely defined and lack of integrated descriptions. This situation causes difficulties in teaching, learning, research, and application. This paper surveys a large collection of DMKD literature to provide a comprehensive picture of current DMKD research and classify these research activities into high-level categories using grounded theory approach; it also evaluates the longitudinal changes of DMKD research activities during the last decade.

Keywords: Grounded theory; knowledge discovery; descriptive framework; document clustering.

* A short 5-page version of this paper appeared previously at the IEEE ICDM workshops, 18–22 December 2006.
† These authors contributed equally to this work.
‡ Corresponding author.
1. Introduction

In the last two decades, the advances in digital data collection devices and data storage technology allow companies and organizations to store up huge amounts of data. Underneath these data are great opportunities for knowledge discovery. Data mining and knowledge discovery (DMKD), which utilizes methods, algorithms, and techniques from a variety of disciplines to extract useful information, emerged during 1990s and grew rapidly. In 1989, the first one-day KDD workshop was held at International Joint Conference on Artificial Intelligence (IJCAI) in Detroit with nine papers presented. In 2007, there are 17 data mining conferences. DMKD has attracted interest from various fields, including databases, machines learning, pattern recognition, artificial intelligence, statistics, data visualization, mathematics and optimization, and economics.

Despite the rapid development, the field is still vaguely defined. More often than not, it is perceived by outsiders as a collection of somewhat related algorithms and tools. The lack of integrated definitions and cohesive pictures of the DMKD field causes difficulties in learning and impedes the long-term advancement of the field. To move toward a scientific field, it is necessary to develop theoretic frameworks that describe the general subjects and provide integrated views of the research activities of DMKD.

A number of publications have focused on studying the nature of DMKD field and building conceptual frameworks. Fayyad et al. presented the first and most widely used KDD framework. It provides an overview of major knowledge discovery activities and connects them using a process flow, which has raw data as input and usable knowledge as the output. Within the KDD process are a set of interactive and iterative steps: data selection, preprocessing, transformation, data mining (DM), and interpretation. Kleinberg et al. proposed a microeconomic view framework to evaluate the value of extracted patterns. Geist modeled the KDD process by a three-view framework, consisting of a model view, a data view, and a process view. Pechenizkiy et al. viewed DM process in terms of the Information Systems Development and Knowledge Management paradigms. Yao et al. introduced a three-layered conceptual framework, consisting of the philosophy layer, the technical layer, and the application layer, to understand the characteristics of the field.

Each reference has its own focus, such as DM process, specific DM tasks, or DM techniques. Up to this point, no study has been conducted to survey large collection of DM publications to identify major categories and research trends of the field. This paper surveys DMKD publications to develop a descriptive framework that captures the main categories of this field using grounded theory and document clustering approaches. In particular, this study is intended to address the following issues: (1) What are the major subjects/categories of the field? (2) What is the central theme (i.e. core category)? (3) What are the connections among these subjects/categories? and (4) What are the longitudinal changes of DMKD research?

This study has two contributions: (1) provide a comprehensive picture of current DMKD research and classify these research activities into high-level categories for
students and business users who pursue basic understanding about DMKD; (2) summarize and evaluate the longitudinal changes of DMKD research and assist identification of areas in need of more research efforts.

The organization of this paper is as follows. Section 2 reviews the basics of grounded theory and document clustering, the two research approaches used in this study. Section 3 describes the process of identifying major concepts, categories, and developing a descriptive framework for the field of DMKD using grounded theory. Section 4 validates the framework by fitting new articles. Section 5 concludes this paper.

2. Research Approaches

This paper uses a combination of qualitative and quantitative methods. Grounded theory is a qualitative method and is used to discover categories, subcategories, interrelationship among these categories, and build a descriptive framework using the collected DMKD literature. Document clustering, a quantitative method, is utilized to validate the results of grounded theory. The input to both approaches is a collection of DMKD publications. The output of grounded theory is a descriptive framework that identifies major concepts, categories, and the relationships among the categories. The output of document clustering is a set of clusters; each cluster includes some documents with common features.

While grounded theory involves subjective experience and opinions, document clustering is determined by the types of algorithms used in the procedure. The methodology of combining qualitative and quantitative methods is called “triangulation” and it can enhance the belief that the results are valid when two different methods yield comparable conclusion.\(^{12}\)

2.1. Grounded theory

Grounded theory was created by sociologists Barney Glaser and Anselm Strauss. It is defined as “one that is inductively derived from the study of the phenomenon it represents. That is, it is discovered, developed, and provisionally verified through systematic data collection and analysis of data pertaining to that phenomenon” (Ref. 11, p. 23). Grounded theory has been effectively used in many disciplines, such as social science, organizational research, and information systems (IS), to generate theories and frameworks.\(^{13–18}\)

Grounded theory is adopted for here two main reasons. First, the objective of the proposed research is to identify the major categories and develop a descriptive framework for the field of DMKD using a collection of existing DMKD literatures. As an inductive theory discovery methodology, grounded theory was designed for such a purpose. Second, earlier studies indicate that it is feasible to apply grounded theory to generate theoretic frameworks in related fields. For example, Bacon and Fitzgerald\(^{19}\) adopted grounded theory in developing an integrated framework for the field of IS.
Three coding processes are fundamental to grounded theory: open coding, axial coding, and selective coding. As grounded theory was originated from sociology, the terminologies used in grounded theory are different from DMKD-related fields. The word “coding” is used in grounded theory to indicate “the operations by which data are broken down, conceptualized, and put back together in new ways” (Ref. 11, p. 57), while in computer science coding normally refers to computer programming. Because many targeted audiences of this research have background in computer science, the word “coding” is replaced by “analysis” to avoid confusion. Therefore, three coding processes of grounded theory are referred as open analysis, axial analysis, and selective analysis in the following discussion. Nevertheless, they are still the same procedures as described by Strauss and Corbin.

Open analysis creates building blocks for grounded theory. In this study, the input to open analysis is a collection of abstracts of DMKD publications and the output is a list of concepts and categories. Each abstract was first taken apart to sentences to find out the major ideas and each idea was saved as a concept. Concepts pertain to the same phenomena are then grouped into categories.

Axial analysis further develops categories by combining categories, developing subcategories, and linking subcategories to their higher-level categories.

Selective analysis integrates categories to form a framework. Strauss and Corbin offer a three-step guideline to conduct selective analysis. The first step is to explicate the story line, which is a descriptive story about the central phenomenon of the study. The central phenomenon is the core category of the study. The second step is to relate other categories around the core category. The third step validates those relationships against data. In practice, these steps are often used in a sequential and iterative manner.

In addition to the three processes, two analytic procedures, making comparisons and asking questions, are essential techniques of grounded theory and are used throughout the three processes. They represent an attitude of skepticism.

2.2. Document clustering

Document clustering partitions textual documents into clusters by grouping similar documents together. The major steps of document clustering include tokenization, stemming, elimination of stopwords, index term selection, the term-document matrix setup and clustering. Tokenization divides documents into words/terms. Stemming removes affixes of words. Stopwords are common English words such as articles, prepositions, and conjunctions and appear in almost every article. Stemming and eliminating stopwords can reduce the number of index terms and increase efficiency of document clustering algorithms. After these preprocessing steps, words/terms that are capable of representing documents are selected and the term-document matrix is established. This study chose k-means algorithm and CLUTO software to perform clustering. CLUTO is a free software package.
A Descriptive Framework for the Field DMKD

3. A Descriptive Framework for the Field of DMKD

This section describes the data source and how grounded theory was adopted to develop a descriptive framework for the field of DMKD. While this study follows most of procedures suggested by Strauss and Corbin, some of them are not applicable and therefore are modified.

3.1. Data sources

Data sources include abstracts of DMKD journals, proceedings, book, syllabi, and doctoral dissertations. Journals and proceedings reflect the overall developments of the field, including theoretical, technical, and practical aspects. A frequently cited book, Advances in Knowledge Discovery and Data Mining, was also included. Syllabi of DM courses were included to show the structures of current DM curriculum. PhD dissertations reflect the research trend and future directions. The total number of abstracts, including syllabi, is 1669.

3.2. Open analysis

Strauss and Corbin divide open analysis into four steps: labeling phenomena, discovering categories, naming categories, and developing categories in terms of their properties and dimensions. Three of these steps were followed, except the last one. The last step was left out for three reasons. First, it is difficult to distinguish between properties and subcategories. Secondly, with more than 1600 abstracts, thousands of concepts are expected from labeling phenomena. It would be impractical to develop and validate relationships on all combinations of properties. Thirdly, as the objective of this study is to identify the major areas of DMKD research, the properties and dimensions of categories are not the focus.

In what follows a detailed description of the three steps in open analysis, giving examples from each step describing how concepts and categories are identified and developed, is presented.

Step 1: Labeling phenomena

Labeling phenomena identifies concepts, the main ideas of documents, from literature. The major concepts of documents can be determined by sentence-by-sentence analysis and comparisons. Abstracts are dissected sentence by sentence to answer questions like What is the major purpose of this article? What is/are the problem(s) at which this paper is aimed? What is/are the proposed solution(s)? What is/are the result(s)? In DMKD publications, authors may use different terms for the same phenomena, technique, or idea. Therefore, concepts need to be compared with each other to determine their similarities and differences.
The following abstract was used to illustrate how the labeling phenomena step was performed.

“Title: Hierarchical Decision Tree Induction in Distributed Genomic Databases Classification based on decision trees is one of the important problems in DM and has applications in many fields. In recent years, database systems have become highly distributed, and distributed system paradigms, such as federated and peer-to-peer databases, are being adopted. In this paper, we consider the problem of inducing decision trees in a large distributed network of genomic databases. Our work is motivated by the existence of distributed databases in healthcare and in bioinformatics, and by the emergence of systems which automatically analyze these databases, and by the expectancy that these databases will soon contain large amounts of highly dimensional genomic data. Current decision tree algorithms require high communication bandwidth when executed on such data, which are large-scale distributed systems. We present an algorithm that sharply reduces the communication overhead by sending just a fraction of the statistical data. A fraction which is nevertheless sufficient to derive the exact same decision tree learned by a sequential learner on all the data in the network. Extensive experiments using standard synthetic SNP data show that the algorithm utilizes the high dependency among attributes, typical to genomic data, to reduce communication overhead by up to 99 percent. Scalability tests show that the algorithm scales well with both the size of the data set, the dimensionality of the data, and the size of the distributed system.”

The goal of Bar-Or et al.’s paper is to present a new decision tree algorithm for large distributed genomic databases. The key ideas are classification, decision tree algorithm, and large distributed genomic database.

Next, questions regarding the exiting problems, potential solutions, and results are examined. Their work is motivated by the fact that the existence of large distributed databases in healthcare and in bioinformatics needs efficient algorithms while current decision tree algorithms require high communication bandwidth. In order to solve this problem, they present an algorithm that reduces the communication overhead. Experimental results indicate that this algorithm reduces communication overhead by up to 99 percent and scales well with the size of the data set, the dimensionality of the data, and the size of the distributed system. Therefore, two additional concepts: communication overhead and scalable algorithm can be added.

The first two stages identify classification, decision tree algorithm, large distributed genomic database, communication overhead, and scalable algorithm as
A Descriptive Framework for the Field DMKD

the key concepts of this abstract. These concepts are not final. As open analysis goes on, documents are compared with each other and revisions may be made to the previously identified concepts. This stage requires theoretical sensitivity, which indicates “an awareness of the subtleties of meaning of data” (Ref. 11, p. 41). As an example, Bar-Or et al.’s paper was compared with Schuster and Wolff’s paper (quoted below):

“Title: Communication-Efficient Distributed Mining of Association Rules Mining for associations between items in large transactional databases is a central problem in the field of knowledge discovery. When the database is partitioned among several share-nothing machines, the problem can be addressed using distributed DM algorithms. One such algorithm, called CD, was proposed by Agrawal and Shafer and was later enhanced by the FDM algorithm of Cheung, Han et al. The main problem with these algorithms is that they do not scale well with the number of partitions. They are thus impractical for use in modern distributed environments such as peer-to-peer systems, in which hundreds or thousands of computers may interact. In this paper we present a set of new algorithms that solve the Distributed Association Rule Mining problem using far less communication. In addition to being very efficient, the new algorithms are also extremely robust. Unlike existing algorithms, they continue to be efficient even when the data is skewed or the partition sizes are imbalanced. We present both experimental and theoretical results concerning the behavior of these algorithms and explain how they can be implemented in different settings.”

This paper presents a set of scalable algorithms that solve the Distributed Association Rule Mining problem. The major differences between the two papers are as follows: (1) they focus on different DM tasks: Bar-Or et al.’s paper concentrates on classification problem, while Schuster and Wolff’s paper focuses on mining association rules; and (2) Bar-Or et al.’s paper works on genomic databases and Schuster and Wolff’s algorithm is designed for transactional databases. The similarities between the two papers are as follows: (1) focus on designing algorithms for distributed databases; and (2) try to increase efficiency of algorithms by reducing communication overhead. Schuster and Wolff call their algorithms “distributed algorithm.” Because both papers proposed approaches for distributed databases, does the concept “distributed algorithm” apply to Bar-Or et al.’s paper as well? A check of the full paper confirms that the decision tree algorithm proposed in Bar-Or et al.’s paper is indeed distributed. Therefore, distributed algorithm is added to the concept list of Bar-Or et al.’s paper.

Abstracts from conference proceedings, journals, and dissertations were analyzed using the same approach. Syllabi, however, have dissimilar layout and were
processed differently. The objectives, required text, and department information are useful, but course topics are the most important part for the labeling phenomena step because they reflect the main content of a course. These topics were collected as concepts to represent syllabi. Concepts are recorded and stored in Microsoft Office Access 2003. The query facility of Access allows flexible retrieval and summary of concepts. A total of 4230 concepts were identified after the labeling phenomena step, with 1892 distinct ones. The top three most frequent concepts are association rule mining, clustering, and classification.

Step 2: Discovering categories
It is difficult to comprehend the relationships and structures of the thousands of concepts discovered in labeling phenomena step. Discovering categories step organizes concepts into categories, which are higher-level concepts. The following paragraphs using Association rule mining as an exemplar of how to develop categories.

Agrawal et al.26 proposed the task of identifying significant association rules from transaction databases as association rule mining. Since association rule mining is one of the most acknowledged and widely studied DMKD topics, there is little disagreement or confusion about the terms used in this area. Thus, concepts with the key word “association (rule)” can be safely grouped under one category. For instance, association mining, association rule, representative association rule, and optimized association rules can be assigned to association rule mining category. Besides concepts with similar key words, concepts without those key words but derived from association rule mining articles are also potential concepts of association rule mining category. Fifty categories were developed after discovering categories step.

Step 3: Naming categories
Naming categories step assigns a name to each category. A category name should be abstract enough to represent the concepts within it. The process of naming a category involves reviewing concepts under each category, comparing categories, and restructuring categories when necessary. Reconstruction of categories is necessary in this research since one of our objectives is to provide students a high-level and concise description of DMKD research activities and 50 categories are too many to be easily understood. To make the resultant framework pellucid, categories should be further modified and combined. Through the use of asking questions and constant comparisons, eight categories were identified (see Fig. 1). The eight categories with illustrative subcategories are listed below:

1. DM tasks are basic functions of DMKD (e.g. association rule mining, classification and prediction, clustering, sequential pattern mining, time series analysis, concept description, exception detection, link analysis, and novel DM).
2. Learning methods and techniques include disciplines that involve and contribute to the DMKD research (e.g. databases, machine learning, artificial
intelligence, statistics, and optimization). Each discipline forms a subcategory and consists of a list of methods, techniques, and algorithms.

(3) Mining complex data include issues related to mining a diversity of data types (e.g. mining text, graph, temporal, spatial, stream, Web, biologic, and multimedia data).

(4) Foundations of DM involve theoretical and fundamental issues of DMKD (e.g. DM theories, frameworks, taxonomies, measures, privacy, security, and social impact issues).

(5) DM software and systems include various aspects of DM software and systems (e.g. software maintenance, software development environment, and DM systems).

(6) High-performance and distributed DM focuses on designing algorithms and techniques that can increase efficiency in handling massive data sets (e.g. parallel algorithms, parallel processing, and distributed DM algorithms).

(7) DM applications describe current DM application domains and related issues (e.g. fraud detection, e-commerce, basket analysis, and marketing).

(8) DM process and project includes various steps in the knowledge discovery process and DM project related issues (e.g. process models, preprocessing techniques, model and algorithm selection, post-processing techniques, and visual DM).

Categories are not mutually exclusive; rather, they have overlaps. For instance, the concept “clustering” appears in both DM tasks and learning methods and techniques. In learning methods and techniques category, clustering refers to an unsupervised method that groups similar objects together and includes specific clustering algorithms, such as $k$-means algorithm, hierarchical clustering algorithm, EM clustering algorithm, and medoid-based clustering. While in DM tasks
Table 1. Categories and number of concepts.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM tasks</td>
<td>1114</td>
</tr>
<tr>
<td>Learning methods and techniques</td>
<td>959</td>
</tr>
<tr>
<td>Mining complex data</td>
<td>874</td>
</tr>
<tr>
<td>DM process and project</td>
<td>506</td>
</tr>
<tr>
<td>Foundation of DM</td>
<td>180</td>
</tr>
<tr>
<td>DM application</td>
<td>172</td>
</tr>
<tr>
<td>High-performance and distributed DM</td>
<td>69</td>
</tr>
<tr>
<td>DM software and system</td>
<td>47</td>
</tr>
</tbody>
</table>

category, clustering refers to a DM task that utilizes clustering methods in various applications.

Table 1 summarized the eight categories and the number of concepts within each category. DM tasks and learning methods and techniques have the most number of concepts. Since the number of concepts is correlated with the number of articles, it implies that DM tasks and learning methods and techniques are the popular topics in the field during the last decade. One thing to note is that since most data sources in this study are research-oriented, the small number of concepts for DM applications may not be a true indicator of the status of DM applications. The output of open analysis answers the first research question: what are the major subjects/categories of the field DMKD?

3.3. Axial analysis

Axial analysis further develops categories by grouping similar concepts under each category into subcategories. Strauss and Corbin\(^\text{11}\) suggest a paradigm model for axial analysis, which can be simplified as

\[
\text{Causal conditions} \rightarrow \text{Phenomenon} \rightarrow \text{Context} \rightarrow \text{Intervening conditions} \rightarrow \text{Action/Interaction strategies} \rightarrow \text{Consequences}.
\]

While the example of studying pain discussed by Strauss and Corbin fits this model nicely, not all phenomena can be organized using the paradigm. In this study, there are no apparent causal conditions, context, actions and consequences relationships between categories and their responding subcategories. Therefore, this study adopts a revised two-step process\(^\text{23}\): the first step hierarchically relates the categories to subcategories and the second step creates an integrative memo for each category.

Use the category of learning methods and techniques as an example. It is linked to the subcategories such as algorithms and methods, statistics, AI, database technology, and optimization.\(^\text{27,28}\) Each subcategory may have its own sub-subcategories. For instance, the AI subcategory is linked to machine learning and soft computing. Machine learning has neural networks, SVM and kernel methods, decision
trees, and inductive logic programming (ILP), and soft computing has fuzzy logic, rough set, and evolutionary computation as their lower-level subcategories, respectively (Fig. 2). The logic and reasoning behind the organization of category and their subcategories are recorded in memos, which evolve from early stage of open analysis to selective analysis.

Presented below is the integrative memo for learning methods and techniques category:

Methods and techniques that originated from various fields is an essential component of DMKD. DMKD is concerned with the extraction of useful knowledge from data and this objective is achieved through learning methods and techniques. Although some concepts of “learning methods and techniques” can also be used in preprocessing and post-processing, the majority of concepts in this category concentrates on extracting patterns or builds predictive models from data. Preprocessing and post-processing techniques are placed in data mining process and project category. Methods and techniques are used to model data and discover patterns from data; they are used to support data mining tasks, mining complex data, and high-performance and distributed data mining; they are also used to support fundamental issues of data mining, such as privacy and security, attributes learning, and model building. From the analysis of the collected data, methods and techniques can be categorized into five areas: algorithms and models, database technology, artificial intelligence (AI), statistics, and optimization.
Algorithms, a subfield of computer science, are one of the core components of DMKD. Since the most influential group of researchers in DMKD come from Computer Science and Machine Learning, algorithms or mathematical models are used most often to represent solutions. More than half of the collected DMKD publications focus on algorithms or models formulation, implementation, and optimization for various data mining tasks. Concepts under this category concentrate on data mining tasks, such as classification, association, clustering, and complex data.

As a fundamental field to DMKD, database provides a multitude of methods and techniques that have been integrated with the data mining process. The major database technologies that have been utilized or designed for DMKD include data warehousing and OLAP, data mining query languages, multi-relational data mining (MRDM), inductive database, deductive database, indexing, database integration, and active data mining. Data warehousing and OLAP offer high-quality data and facilitate complex analytical queries for data miners. It has been the most widely studied database topic in DMKD. Data mining query languages are designed to support the data mining process. Data miners or end users can use data mining query languages to realize some knowledge discovery tasks. MRDM searches for knowledge and patterns from multiple tables in a relational database and can be expanded to many DM tasks, such as MR association analysis, MR classification, MR clustering, MR regression, and MR linkage analysis. Inductive database contains original data and the generalizations of the data. Deductive database is an extension of relational databases and contains the facts derived from original data and rules. Index structures are created to expedite database access and query processing. Data mining tasks often need to deal with multiple data sources and database integration techniques can help at this aspect. Active DM combines DM with active database systems and continuously mines data at a desired frequency.

AI has machine learning and soft computing subcategories. Machine learning methods and techniques applied in DM include unsupervised learning methods, decision trees, support vector machine (SVM) and kernel methods, neural networks, inductive logic programming (ILP), and reinforcement learning. Soft computing has concepts like fuzzy logic, rough sets, and genetic algorithms.

Statistics has offered numerous methods and techniques for every step of the KDD process. Major concepts include regression, Bayesian analysis, principle component analysis (PCA), probabilistic models, Markov chain Monte Carlo (MCMC), hidden Markov model (HMM), discriminant analysis, and Minimum Message Length (MML) principle.

Optimization is an emerging subcategory in learning methods and techniques. Though there are only 18 concepts in this subcategory, it may be a new direction of DM methods and techniques.
Integrative memos for other seven categories are summarized in Appendix A. In Appendix B, taxonomy of the eight categories, including their subcategories, was presented.

3.4. Selective analysis

Open and axial analysis identified concepts, categories, and subcategories from data. The objective of selective analysis is to develop a store about the phenomena of the study by selecting a core category and connecting other categories to the core category. This study follows a three-step selective analysis\textsuperscript{11}: (1) explicate the store line and identify a core category; (2) relate categories around the core category; and (3) validate the relationships against the data. The validation step is discussed in the following section: framework validation and document clustering.

Step 1: Explicating the story line

The story line is a descriptive story that represents the gist of the central phenomenon of the study with a few sentences. To write a general descriptive story is to record the thoughts developed during open and axial analysis. The following memo describes the story line of DMKD:

The main story of DMKD is the extracting of patterns and knowledge from massive data sets for decision support or scientific investigation through a set of interactive and iterative steps. A prominent feature of DMKD is that it needs handle massive data sets (e.g. millions of rows with hundreds of variables), which cannot be manually analyzed in a timely manner. Data to be processed and analyzed can be tabular (normally from relational databases) or other complex types (e.g. natural language, graph, DNA sequences, Web, and multimedia). Data mining is an interactive and iterative process that involves multiple steps. Theories, methods, and algorithms from related fields, such as databases, machine learning, statistics, data visualization, and optimization, are used to accomplish each step of the knowledge discovery process. DM has been applied successfully to business and scientific applications and software, query languages, and systems have been developed to assist these applications.

The primary issue of DMKD is the extraction of patterns and knowledge from large amounts of data. The motivation for extracting such knowledge is to support business decision making or explore and understand the phenomena underneath the data.\textsuperscript{32}

A core category should satisfy two criteria: emerges with high frequency in the analysis and is able to integrate other categories.\textsuperscript{33} Such a category may already exist in the results of axial analysis. If not, a new category needs to be created. In this study, two categories: learning methods and techniques and DM tasks are candidates for the core category since they both appear with high frequency in
the analysis and have broader scopes than other categories. Although a large proportion of DMKD research effort is devoted to developing learning methods and techniques, it concentrates on the knowledge extraction and fails to integrate other categories. Therefore, learning methods and techniques is not appropriate to be the core category.

DM tasks category reflects the primary issue of DMKD: extract knowledge from data. The top three most frequently appeared concepts: *association rule mining*, *clustering*, and *classification*, all belong to DM tasks and DM tasks have the most concepts (refer to Table 1). It certainly satisfies the high-frequency criterion of a core category. In order to become a core category, it must be able to integrate other categories into a whole story. The following memo describes how DM tasks combine other seven categories:

Having large amounts of data, which may contain complex types, changes the nature of traditional data exploration and analysis, raising new problems that require new or revised methods and algorithms. The major objective of *mining complex data* and *data mining applications* is to accomplish DM tasks with complex data types and/or in different application domains. *Learning methods and techniques* provide technical support to DM tasks. *DM process* defines a series of steps, which include data preparation, modeling, and results evaluation and utilization, to perform DM tasks. *High performance and distributed DM* develops scalable algorithms and systems for DM tasks that involve huge and/or distributed data sources. *DM software and system* allows users to implement DM tasks in their applications. *Foundations of DM* provides theoretic bases for DM tasks and deals with privacy and security concerns of DM tasks.

This memo indicates that DM tasks can meaningfully connect other categories and build a coherent unit. Thus the second research question of this study; what is the central theme of DMKD, is answered. *DM tasks* is the core category (Fig. 3) and the central theme of DMKD.

**Step 2: Relating other categories to the core category**
The main relationships between DM tasks and other seven categories are described below. In addition, Table 2 summarizes the main relationships between each pair of categories as a matrix. The results of this step answer the third research question: what are the connections among these subjects/categories?

- **DM tasks and learning methods and techniques**: Learning methods and techniques provide methods, techniques, and algorithms for the modeling/DM step of DM tasks.
- **DM tasks and foundation of DM**: Foundations of DM, such as privacy and security, focus on theoretical and fundamental issues that affect and support all DM tasks.
There are 28 relationships, one relationship between each pair of categories, in Table 2. To increase the understandability, this study adopts a general model of system and represents main categories and their relationships in a framework. A general model of system is defined by Nolan and Wetherbe as

A system consists of one or more inputs to a system, and one or more outputs from a system. Inputs are transformed into outputs as a result of their interaction with two or more subsystems. An additional and important dimension is feedback for system control. Feedback is a form of system input necessary
### Table 2. Matrix of main relationships between categories.

<table>
<thead>
<tr>
<th></th>
<th>Learning methods and techniques</th>
<th>DM tasks</th>
<th>Foundations of DM</th>
<th>Mining complex data</th>
<th>High-performance and distributed DM</th>
<th>DM process and project</th>
<th>DM applications</th>
<th>DM software and systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning methods and techniques</strong></td>
<td>Implemented by</td>
<td>Implement</td>
<td>Contribute to</td>
<td>Utilized by</td>
<td>Utilized by</td>
<td>Applied through</td>
<td>Utilized by</td>
<td>Utilized by</td>
</tr>
<tr>
<td><strong>DM tasks</strong></td>
<td>Implemented by</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Foundations of DM</strong></td>
<td>Utilize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mining complex data</strong></td>
<td>Utilize</td>
<td>Support</td>
<td>Extension of</td>
<td>Supported by</td>
<td>Support</td>
<td>Facilitated by</td>
<td>Support</td>
<td>Support</td>
</tr>
<tr>
<td><strong>High-performance and distributed DM</strong></td>
<td>Utilize</td>
<td>Facilitate</td>
<td>Supported by</td>
<td>Facilitate</td>
<td>Support</td>
<td>Follow</td>
<td>Support</td>
<td>Support</td>
</tr>
<tr>
<td><strong>DM process and project</strong></td>
<td>Utilize</td>
<td>Define the process for</td>
<td>Supported by</td>
<td>Define the process for</td>
<td>Supported by</td>
<td>Define the process for</td>
<td>Define the process for</td>
<td>Implemented by</td>
</tr>
<tr>
<td><strong>DM applications</strong></td>
<td>Utilize</td>
<td>Operationalize</td>
<td>Supported by</td>
<td>Operationalize</td>
<td>Supported by</td>
<td>Follow</td>
<td>Implement</td>
<td></td>
</tr>
<tr>
<td><strong>DM software and systems</strong></td>
<td>Utilize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
to evaluate the internal performance of this system, and to allow systems to adapt to environmental changes. (p. 3)

In Fig. 4, DMKD is viewed as an open system that transforms user requirements, raw data, and domain knowledge (inputs) into useful patterns or actionable knowledge (outputs) in the context of data-rich companies and organizations. The eight categories form the Data Mining and Knowledge Discovery System. DM task, the central theme, is located at the core of the system. Parts of “DM process and project” are in the inputs and outputs, such as domain knowledge, useful patterns, and actionable knowledge. There are two kinds of feedback in this system: one is from output to DMKD system to maintain and update existing descriptive and predictive models; the other is from output to input to add new knowledge to existing knowledge bases or domain knowledge.

3.5. Longitudinal changes of DMKD research

Previous sections answer the first three research questions. This section addresses the last question: what are the longitudinal changes of DMKD research? The longitudinal changes of DMKD research were analyzed from three perspectives: longitudinal changes of publications, longitudinal changes of categories, and the distribution of subcategories.

3.5.1. Longitudinal changes of DMKD publications and categories

3.5.1.1. Longitudinal changes of DMKD publications

Figure 5 delineates the number of DMKD publications from 1994 to 2005. There is an increasing trend over the years. The number of publications of 2005 is only
Fig. 5. DMKD publication trend.

69 because the data collection process was completed at the end of July 2005. If
the conference and the journal papers published between July 2005 and December
2005 were added, the number should be at least 295. Started from 2001, the publi-
cations become stable, with a standard deviation of 11.82 (using the adjusted 2005
publications).

Although the data collected in this study is not exhaustive, it represents a gen-
eral trend of DMKD publications. The trend can be summarized in two points.
First, DMKD has attracted steady contributions from related fields. Several inter-
national conferences that focus on DM and knowledge discovery have been held
annually since the first knowledge discovery in databases workshop, which was held
on 20 August 1989 in conjunction with IJCAI. Two academic periodicals, Data
Mining and Knowledge Discovery (DMKD) and SIGKDD Explorations newsletter,
were published since 1997 and 1999, respectively. Second, the DMKD field has
experienced a growing period from 1998 to 2001; after 2001, the field maintains
the enthusiasm and keeps drawing interests from researchers and practitioners.
Though the number of publications from 2001 to 2005 seems stable or even lower
(e.g. from 290 in 2003 to 280 in 2004), it actually received increasing attention
during this period. One reason for the stability in publication is that major DMKD
conferences maintain the number of accepted papers. Use ICDM as an example.
Table 3 summarizes the submission and acceptance data of ICDM from 2001 to
2005. The “submitted” column is the total submissions and the “accepted” column
is the accepted regular full sized papers. From 2001 to 2005, there is a submission
increase of 72.6%. This number, on the one hand, indicates that the interest in the
field has increased considerably from 2001 to 2005. This increase in submission was
not reflected in the number of publications because the number of accepted papers
is almost unchanged and even reduced (e.g. there were only 39 papers accepted in
2004).
Table 3. ICDM papers acceptance statistics.

<table>
<thead>
<tr>
<th>Year</th>
<th>Submitted</th>
<th>Accepted</th>
<th>Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>365</td>
<td>71</td>
<td>19.45</td>
</tr>
<tr>
<td>2002</td>
<td>369</td>
<td>72</td>
<td>19.51</td>
</tr>
<tr>
<td>2003</td>
<td>501</td>
<td>58</td>
<td>11.58</td>
</tr>
<tr>
<td>2004</td>
<td>445</td>
<td>39</td>
<td>8.76</td>
</tr>
<tr>
<td>2005</td>
<td>630</td>
<td>69</td>
<td>10.95</td>
</tr>
</tbody>
</table>

Table 4. Distribution of DMKD papers by category vs year.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DM tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>859</td>
</tr>
<tr>
<td>Learning methods and techniques</td>
<td>11</td>
<td>16</td>
<td>46</td>
<td>94</td>
<td>96</td>
<td>138</td>
<td>131</td>
<td>139</td>
<td>148</td>
<td>39</td>
<td></td>
<td></td>
<td>885</td>
</tr>
<tr>
<td>Mining complex data</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>17</td>
<td>27</td>
<td>80</td>
<td>74</td>
<td>96</td>
<td>72</td>
<td>20</td>
<td></td>
<td></td>
<td>396</td>
</tr>
<tr>
<td>Foundations of DM</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>15</td>
<td>19</td>
<td>26</td>
<td>35</td>
<td>30</td>
<td>41</td>
<td>7</td>
<td></td>
<td></td>
<td>189</td>
</tr>
<tr>
<td>DM software and systems</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>High-performance and distributed DM</td>
<td>3</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>12</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>DM applications</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>18</td>
<td>12</td>
<td>23</td>
<td>18</td>
<td>27</td>
<td>25</td>
<td>5</td>
<td></td>
<td></td>
<td>141</td>
</tr>
<tr>
<td>DM project and process</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>31</td>
<td>26</td>
<td>42</td>
<td>34</td>
<td>43</td>
<td>38</td>
<td>9</td>
<td></td>
<td></td>
<td>241</td>
</tr>
<tr>
<td>Total by year</td>
<td>2</td>
<td>44</td>
<td>47</td>
<td>131</td>
<td>286</td>
<td>290</td>
<td>461</td>
<td>447</td>
<td>504</td>
<td>503</td>
<td></td>
<td></td>
<td>125</td>
</tr>
</tbody>
</table>

3.5.1.2. Longitudinal changes of DMKD categories

To find out the longitudinal changes of the eight main categories, the collected DMKD articles were re-examined. According to its major contribution, each article was assigned to one or more of the categories.

The distribution of categories over the years is summarized in Table 4. Each cell in the table is painted with different levels of darkness: 25% darkness indicates that the number of articles presented in the dedicated year is less than 10; 50% darkness indicates that the number of articles presented in the designated year is between 10 and 50; 80% darkness indicates that the number of articles presented in the designated year is between 51 and 100; 100% darkness indicates that the number of articles presented in the designated year is above 100.

Over the last several years (1994–2005), the most widely studied DM categories are as follows: learning methods and techniques, DM tasks, and mining complex data. The “growing” areas of DMKD include learning methods and techniques, mining complex data, and foundations of DM. The number of publications of DM project and process category has been stable from 1999 to 2004. More than half of DM project and process publications are concentrated on data preprocessing, which is considered as the most time-consuming step in KDD process. DM community has
recognized the importance of this topic and laid a strong emphasis on it in the last several years. The growing number of publications in theoretical foundations, data preprocessing, and privacy-preserving DM indicate a healthy research development of the field.

Based on this distribution, it is reasonable to predict that DM tasks, learning methods and techniques, and mining complex data will continue to be the hot topics in DMKD. Foundations of DM and DM project and process will attract more attention in the next couple of years. The trend of high-performance and distributed DM category is not very clear from the data. Publications in this category increased from 1996 to 1999 (from 0 to 12) and decreased from 2000 to 2003 (from 7 to 2). In 2004, the number of publications increased again (from 2 to 12). The only category that exhibits a decreasing trend is DM software and systems, which focuses on DM software and systems development or related issues. As a matter of fact, many commercial and free DM software and packages were developed during the last decade. The decreasing number of publications of DM software and systems may due to three reasons. The first reason is the need for confidentiality. Details of commercial DM packages, such as architectures and specific algorithms, may provide competitive advantage. Second, the development of DM software or systems normally requires longer time than the development of DM methods or algorithms and this restricts the number of publications. Third, the DMKD field does not pay enough attention to this area. As an application-driven field, the ultimate goal of DMKD is to extract patterns from massive data sets for decision support or scientific investigation. DM systems and software packages are important tools that end users can use to realize their DM objectives. Since no single method or algorithm works perfectly for all data sets, the development of industry or application-specific DM software and systems is a promising direction.

3.5.2. Distributions of subcategories

Previous discussion gives a high-level description of the longitudinal changes of the eight categories. Except DM applications, all other categories have subcategories and the distribution of subcategories can provide more insights. The following paragraphs outline the distributions of subcategories.

**DM tasks**

Association rule mining, classification, and clustering are the top three studied subcategories of DM tasks. These three subcategories have wide applications in scientific and commercial fields.

Compare with them, semi-supervised learning and link analysis were far less studied. Semi-supervised learning utilizes both labeled and unlabeled data to make predictions. It is useful when there are a few labeled records and a lot of unlabeled data. It has great potentials since unlabeled raw data exist in many situations. Link analysis or link mining is a newly emerging area that intersects web and hypertext mining, social networks, security and law enforcement data, bibliographic citations...
and epidemiological records. It is a promising new area because it has wide application domains, such as World-Wide Web and bioinformatics. There has been an increasing interest in this area as evidenced by the series of link analysis/KDD workshops (2003–2006).

**Learning methods and techniques**

Learning methods and techniques category is the largest category among the eight. It indicates that designing, developing, and implementing learning methods and techniques are the most important task in current DMKD research. Many learning methods and techniques are originated from artificial intelligence, statistics, and database technology. Clustering algorithms, decision tree, SVM, and kernel methods are the most popular learning methods from machine learning; Bayesian analysis, ensemble, probabilistic analysis are the most popular statistic methods; data warehousing and OLAP, database query language, and multi-relational DM (MRDM) are the most popular learning methods from database technology.

**Mining complex data**

The top three popular subcategories are Web mining, text mining, and temporal DM. For Web mining, Web usage mining accounts for 65% of concepts and time series analysis accounts for 76% of temporal DM concepts. Text or document classification, information retrieval, and semi-structured DM are the most popular concepts in text mining.

Multimedia, stream, and graph mining, on the other hand, were less studied. This may be due to two reasons. First, extracting knowledge from stream, graph, video, and audio requires close collaboration of experts from different disciplines. Second, there are few openly accessible multimedia data sets. These two requirements make multimedia, stream, and graph mining difficult tasks. Researchers have recognized this issue and initiated Multimedia Data Mining (MDM)/KDD workshop series since 2000. Multimedia, stream, and graph mining represent opportunities for theoretic research and real-life applications. A promising direction in these areas is to build a set of multimedia, stream, or graph benchmarks that can be accessed publicly.

**Foundations of DM**

Measure, privacy and security, and theory are the most studied subcategories of foundations of DM. Measures include quality assessment and interestingness metrics. Scalability, efficiency, computational stability, and computational complexity of DM algorithms are tested and compared frequently to verify the quality of proposed algorithms. Similarity measures and interestingness measures are also considered as important measurements in DM.

There has been a growing interest in privacy and security issues during the past five years. Many concepts in this subcategory focus on developing privacy-protection DM techniques, which are also called privacy-preserving DM. Some specific techniques used in privacy-preserving DM include randomization, hashing, cryptographic techniques, distributed computing, and algebraic techniques.
DM software and systems

The number of concepts for DM software (24) and DM systems (23) subcategories are almost equal. DM software includes concepts such as DM software tools, software development and support environment, software maintenance, intelligent agents, and code synthesis. DM systems include DM system architecture, different-methods-based systems (e.g. rough-sets-based-system and hybrid DM system), and specific DM systems (e.g. behaviometric security system, fault tolerant teaching (FTT) systems, and sailing system).

High-performance and distributed DM

Distributed DM has almost twice as many concepts as high-performance DM. Main concepts of distributed DM are distributed databases, distributed mining, distributed association mining, and distributed algorithms. High-performance DM includes parallelizing DM, parallel algorithms, parallel computing, and scalable DM algorithms.

DM process and project

Though DM process and project category has two subcategories, DM project subcategory has only three concepts: key success factors for DM projects, DM project goal development, and interdisciplinary projects. There are two reasons to have a separate subcategory for DM project: first, these three concepts do not fit DM process subcategory; second, they represent a set of challenging problems and potential new direction. How are DM projects conducted? Do they follow DM process models, such as CRISP-DM or KDD process? What are the similarities and differences between DM projects and other IT projects? What are the lessons we have learned from past DM projects? How many DM projects, have been finished on time? How many of them have been successfully completed within budget? How can we improve the quality of DM projects? These problems are beneficial to DM research and applications and are mostly unanswered.

DM process, which includes data preprocessing, model/algorithm selection, post-processing, and visualization techniques, has long been recognized as the most time-consuming yet significant part of the whole knowledge discovery process. The research efforts in this area are stable. Data preprocessing and visualization subcategories account for 78% of the DM process concepts. Within data preprocessing, feature selection and data reduction are the most studied topics.

4. Framework Validation

This section validates the framework developed in Sec. 3 using two methods: (1) fitting new DMKD articles to the framework and (2) document clustering.

The reason to use the first validation method is straightforward: if new DMKD articles can be assigned to the eight categories, it verifies the comprehensiveness of the framework. The reason behind the second validation method is the concept of triangulation, which is an approach that combines dissimilar methodologies (e.g.
qualitative and quantitative methods) in the study of the same phenomenon. Triangulation can enhance the belief that the results are valid when two or more different methods yield comparable conclusion. Grounded theory is a qualitative method whereas document clustering is a quantitative method. If the clusters generated by document clustering can be successfully assigned to the eight categories, it increases our confidence in the validity of the framework developed using grounded theory.

4.1. Relating DMKD literature to the framework

Abstracts of 49 regular papers from PAKDD 2005, which were not used in the framework building process, were collected and analyzed to see if they fit in the categories identified by grounded theory. The abstract of each article was analyzed to identify the primary objective(s) the author(s) is addressing.

Take the article “Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis” by Si and Jin as an example. The abstract of this article is

Mixture models, such as Gaussian Mixture Model, have been widely used in many applications for modeling data. Gaussian mixture model (GMM) assumes that data points are generated from a set of Gaussian models with the same set of mixture weights. A natural extension of GMM is the probabilistic latent semantic analysis (PLSA) model, which assigns different mixture weights for each data point. Thus, PLSA is more flexible than the GMM method. However, as a tradeoff, PLSA usually suffers from the overfitting problem. In this paper, we propose a regularized probabilistic latent semantic analysis model (RPLSA), which can properly adjust the amount of model flexibility so that not only the training data can be fit well but also the model is robust to avoid the overfitting problem. We conduct empirical study for the application of speaker identification to show the effectiveness of the new model. The experiment results on the NIST speaker recognition data set indicate that the RPLSA model outperforms both the GMM and PLSA models substantially. The principle of RPLSA of appropriately adjusting model flexibility can be naturally extended to other applications and other types of mixture models.

This article proposes a variation of mixture models: regularized probabilistic latent semantic analysis (RPLSA) model, to overcome the overfitting problem that the probabilistic latent semantic analysis (PLSA) model suffers. The main contribution of this article is the development of a new analysis model. Therefore, it can be grouped under “algorithms and models,” a subcategory of learning methods and techniques. Since mixture models belong to mathematical statistics, this article can also be grouped under “statistics,” a subcategory of learning methods and techniques (see entry number 40 in Appendix C).
Following the similar process, all 49 articles were successfully classified using the framework. Twenty-four articles belong to learning methods and techniques; 14 belong to DM tasks; 10 belong to DM process and project; 10 belong to mining complex data; three belong to DM applications; and two belong to high-performance and distributed DM. Because articles may be classified into multiple categories, the total number of these articles exceeds 49. Appendix C summarizes the classification of these 49 articles.

The ability of the framework to categorize existing DMKD publications is a verification of its comprehensiveness.

4.2. Document clustering

Document clustering refers to methods and algorithms that can gather similar documents into clusters according to some criterion functions. By providing a form of abstraction for large amounts of electronic documents, document clustering helps optimize indexing and retrieval operations in databases and assists in organizing, navigating, and presenting large amounts of electronic documents. Various clustering algorithms and techniques have been developed over the years. For a comprehensive review of these algorithms and applications, refer to Refs. 42 and 43.

Many document clustering algorithms use single terms to cluster documents. However, a previous study of using single term to cluster the same data source turned out to be unsuccessful: the clusters have low intra-cluster similarity and high inter-cluster similarity and it is difficult to assign a single name to a cluster. This is because the literatures collected in this study share many key words. For example, articles that develop new classification algorithms and articles that apply classification algorithms to practical applications share many key words, but they belong to different research areas of DMKD. Using the eight categories developed by grounded theory, the first kind of articles should belong to DM tasks and the second kind of articles should belong to DM applications. In single-word clustering, however, these two kinds of articles are very likely grouped into the same cluster. One solution is to use phrases instead of single words since context information is important in clustering and phrases may catch more context information than individual words. Still use the previous example: if phrases are used, then the phrase “new algorithm” or “novel algorithm”, which appears frequently in algorithm development articles, can help distinguish algorithm development articles from application-oriented articles.

The following hybrid document clustering procedure summarizes the document clustering strategy used in this study. Each step of the procedure is explained in the subsequent paragraphs. Since the format of syllabi is not appropriate for document clustering, they were excluded in document clustering. Therefore, the total number of documents used in document clustering is 1626.

Hybrid document clustering procedure

Input: 1626 DMKD publications; eight predefined categories using grounded theory.
Output: A set of clusters and an organization of these clusters by manually assigning name(s) to each cluster.

Step 1: Extract abstracts and titles of these articles and tokenize them into a term (i.e., words or phrases) list.

Step 2: Drop off numbers, abnormal characters, and common stopwords from the term list.

Step 3: Use the Porter stemming algorithm to narrow down the term list.

Step 4: Calculate the TF/IDF weights and create vector-space model\(^{47}\) that represents documents as vectors of terms. Terms from titles are assigned twice as much weight as those from abstracts.

Step 5: Documents are clustered using CLUTO toolkit.

Step 6: High-frequent terms in each cluster are summarized.

Step 7: Manually inspect the top frequent terms/key words of each cluster to find out the main theme of that cluster and check whether the cluster can be assigned to one or more of the eight predefined categories.

END

A self-developed C++ program is used to extract titles and abstracts from the documents. Titles and abstracts are tokenized using Microsoft SQL Server.\(^{48}\) Tokenization converts documents into words/phrases and the length of phrases is determined by a threshold. According to the nature of the documents, the phrase threshold is set to three. That is, SQL Server generates a list of words/phrases with no more than three words. The threshold is set as such because identical phrases with more than four continuing words are infrequent and an increase in the threshold requires more computing resources. The term list was diminished by removing stopwords and stemming in order to increase efficiency and quality of clustering algorithms. Stopwords are common English words such as articles, prepositions and conjunctions that appear in every article and thus not useful in document clustering. In addition, the Porter stemming algorithm\(^{49}\) is used to further reduce the size of the list by removing suffixes of words.

Each document is represented by a widely used \(tf\)-\(idf\) vector, where \(tf\) is the term-frequency and \(idf\) is the inverse document frequency. Term-frequency counts the occurrences of terms in a document and inverse document frequency is calculated by \(\log (N/df_i)\), where \(N\) is the total number of documents in the data set and \(df_i\) is the number of documents containing the term \(i\).\(^{47}\) The inverse document frequency is used to de-emphasize terms appearing frequently in many documents because their discrimination power is limited.\(^{50}\) Highly frequent words and phrases from each cluster are extracted to represent that cluster.

CLUTO is a free software package developed by the Department of Computer Science and Engineering, University of Minnesota. There are two programs in CLUTO: vcluster and scluster. The vcluster program treats inputs as high-dimensional vectors, whereas scluster operates on the similarity space between the objects. Both require a predefined number of clusters as input. The vcluster program was used in this study because documents were represented as vector-space
model. CLUTO provides a set of parameters to control clustering algorithms and solutions. The parameter used in this study is “rbr” because it has been proven to produce high-quality clusters in Refs. 43 and 51. The “rbr” is a two-step method that utilizes a randomized incremental optimization algorithm to compute a $k$-way clustering solution. The first step generates $k$ clusters via repeated bisections and each bisection is used to optimize a particular criterion function. The second step repeats the incremental refinement procedure on the $k$ clusters obtained from the first step. CLUTO manual\textsuperscript{21} provides detailed description of the toolkit and the “rbr” method.

Table 5 summarizes 8-way and 100-way clustering results. It presents three internal quality measures and computing time. ISim, intra-cluster similarity, is the average similarity between objects of each cluster. ESim, inter-cluster similarity, is the average similarity of the objects of each cluster and the rest of the objects.\textsuperscript{51} ISim\textsubscript{avg} and ESim\textsubscript{avg} are the average ISim and ESim, respectively. A clustering solution that maximizes the intra-cluster similarity and minimizes the inter-cluster similarity is the desired result. Consequently, the higher the ISim\textsubscript{avg}/ESim\textsubscript{avg} ratio, the better the overall clustering solution is. Table 5 shows that 100-way clustering has higher ISim, lower ESim, and higher ISim\textsubscript{avg}/ESim\textsubscript{avg} ratio than 8-way clustering, which indicate that 100-way clustering solution is better than 8-way.

Following Step 7 of hybrid document clustering procedure, one or more names are assigned to the resultant clusters of 8-way and 100-way clustering to reflect the main theme of each cluster. The process for categorizing clusters is straightforward: the top frequent words/phrases are used to identify the key issue(s) of each cluster. The categorization results indicate that 100-way clusters are much more meaningful than 8-way clusters. Tables 6 and 7 summarize the clusters categorization for 8- and 100-way clustering, respectively. For 8-way clustering, five clusters were assigned to learning methods and techniques category; three clusters were assigned to DM tasks; and three clusters were assigned to mining complex data. This is because these clusters are so large that the top frequent words/phrases are very similar. Take the top 10 frequent words/phrases for instance. Among the 10 most frequent words/phrases from each cluster, “algorithm”, “data”, and “DM” appear in all clusters; “mining” appears in seven clusters; “approach”, “paper”, and “problem” appear in six clusters; “method” appears in five clusters; and “model” appears in four clusters. Highly similar top frequent words/phrases make it difficult to distinguish between clusters.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Computing time (s)</th>
<th>ISim\textsubscript{avg}</th>
<th>ESim\textsubscript{avg}</th>
<th>ISim\textsubscript{avg}/ESim\textsubscript{avg} ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-way</td>
<td>3.875</td>
<td>0.24932</td>
<td>0.01327</td>
<td>18.78824</td>
</tr>
<tr>
<td>8-way</td>
<td>0.797</td>
<td>0.139625</td>
<td>0.014375</td>
<td>9.713043</td>
</tr>
</tbody>
</table>
Table 6. Eight-way clusters categorization.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Learning methods and techniques</td>
</tr>
<tr>
<td>2</td>
<td>Learning methods and techniques</td>
</tr>
<tr>
<td>3</td>
<td>Mining complex data</td>
</tr>
<tr>
<td>3</td>
<td>Learning methods and techniques</td>
</tr>
<tr>
<td>4</td>
<td>Mining complex data</td>
</tr>
<tr>
<td>4</td>
<td>DM tasks: association rule mining</td>
</tr>
<tr>
<td>5</td>
<td>DM tasks: association rule mining</td>
</tr>
<tr>
<td>6</td>
<td>Learning methods and techniques</td>
</tr>
<tr>
<td>6</td>
<td>Mining complex data</td>
</tr>
<tr>
<td>7</td>
<td>DM tasks: classification</td>
</tr>
<tr>
<td>8</td>
<td>Learning methods and techniques</td>
</tr>
</tbody>
</table>

Table 7. 100-way clusters categorization.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of clusters</th>
<th>Number of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM tasks</td>
<td>45</td>
<td>860</td>
</tr>
<tr>
<td>Learning methods and techniques</td>
<td>44</td>
<td>678</td>
</tr>
<tr>
<td>Mining complex data</td>
<td>22</td>
<td>294</td>
</tr>
<tr>
<td>DM project and process</td>
<td>10</td>
<td>117</td>
</tr>
<tr>
<td>DM applications</td>
<td>6</td>
<td>63</td>
</tr>
<tr>
<td>Foundations of DM</td>
<td>6</td>
<td>119</td>
</tr>
</tbody>
</table>

Table 7 summarizes the categorization of 100 clusters. The column “Number of Clusters” indicates how many clusters are grouped within a particular category and the column “Number of Papers” indicates how many papers are grouped within a particular category. The distribution of number of clusters and documents in each category resembles the category distribution identified by grounded theory (refer to Table 1). In addition, the top frequent words/phrases of these clusters are quite indicative, which make the categorization process much easier than 8-way clustering. Take cluster 86 as an example. This cluster has 17 documents and is the sole cluster that belongs to high-performance and distributed DM category. Frequent words/phrases of this cluster include: DM, model, computation, computational time, parallel DM, parallel algorithm, parallel system, parallelism, and shared memory, and these words/phrases exhibit likeness to the concepts identified by grounded theory for this category.

The results of CLUTO and human assessment validate that the eight categories identified by grounded theory do reflect the major topics of DMKD research. This conclusion is supported by 8- and 100-way clustering: the resultant clusters can be categorized as one or more of the eight categories using the top frequent word/phrases.
5. Conclusion

This paper presented a descriptive framework for the field of DMKD using grounded theory. To develop the framework, a large collection of DMKD publications was analyzed following the analytical procedures of grounded theory. The framework identified eight main areas of DMKD: DM tasks, learning methods and techniques, mining complex data, foundations of DM, DM software and systems, high-performance and distributed DM, DM applications, and DM process and project. The connections among these areas were also described. Two approaches were used to validate the framework: first, 49 DM publications were categorized into the framework. The results indicate that the framework is able to classify all 49 articles successfully. Second, the same collection of articles was analyzed by document clustering and resultant clusters are comparable to the major categories identified by grounded theory.

The contribution of this study is two-fold. First, the framework that describes the major subjects of DMKD research and the relationships among these subjects contributes to the theoretic foundation of DMKD field. It provides a holistic view of the field and can be used in DM courses to give students a clear perspective of DMKD research. Second, the longitudinal changes of DMKD publications and areas summarized in this paper point out that some DMKD areas, such as foundations of DM, high-performance and distributed DM, and DM software and systems, were less developed and indicate some future research directions.

Acknowledgments

This research has been partially supported by grants from the Youth Fund of University of Electronic Science and Technology of China, National Natural Science Foundation of China (Nos. 70621001, 70531040, 70472074), 973 Project No. 2004CB720103, Ministry of Science and Technology, China, and BHP Billiton Co., Australia.

Appendix A. Axial Analysis Memos

A.1. DM task

DM tasks include characteristics and relevant concepts of major DM tasks. Methods, techniques, and algorithms that are developed to solve DM tasks are separated from DM tasks category and form a new category: learning methods and techniques. In many situations (such as topics in DM conferences or chapters of DM text books), learning algorithms and techniques are placed within DM tasks for which they are designed. This study creates separate categories for DM tasks and Learning methods and techniques for two reasons. First, a separate category of learning method and technique allows us to identify disciplines that have contributed to DMKD and how deeply they have been involved. Second, a DM task may be accomplished by different learning methods. A separate category of learning methods and techniques, with a list of algorithms and approaches, can help people understand which...
methods from this discipline have been applied to which DM tasks. For example, rough set approach, a subcategory of learning methods and techniques, has been mainly used in classification task. However, from the list of concepts under rough set approach, we notice that rough set can also be applied to clustering and feature selection. This information may be useful not only for beginners to understand learning methods, but also for researchers to find new research directions.

Subcategories within DM tasks include association rule mining, classification and prediction, clustering, semi-supervised classification, frequent pattern mining, sequential pattern mining, time series analysis, concept description, exception detection, link analysis, and novel DM.

According to this study, association rule mining is the most widely studied DM task, which includes 321 concepts. The majority of concepts in association rule mining task aim at finding association rules from transaction databases. Association rule mining algorithms, mining associations from various types of data (e.g. numerical association rules, temporal association rules), filtering techniques for association rule mining, and interestingness measures for association rules are the main concepts in association rule mining. Besides, there are several extended association rules that are different traditional association rules. These extended association rules include peculiarity rules, informative rules, optimized association rules, representative association rules, high-utility rules, positive and negative association rules, substitution rules, and linguistic association rules.

Classification and prediction construct classifiers/models using training and test data and predict future data using the constructed classifiers. The difference between classification and prediction is that classification predicts categorical classes while prediction models continuous target variables. Classification and prediction are the second widely studied DM task, which include 281 concepts (251 concepts for classification tasks and 29 concepts for prediction task). Classification tasks include concepts such as classification of gene expression data, classification of rare events, classification chemical compound data sets, classifier construction, cost-sensitive classification models, adversarial classification, fuzzy classification rules, and classification measurements. Prediction includes prediction models, prediction of DNA structure, predictive rules, and prediction system for bio-data.

Clustering task includes clustering transactional data, document clustering, conceptual clustering, high-dimensional clustering, spatial clustering, clustering image data, clustering sequential data, clustering workflows, multi-view clustering, clustering time series, and projected clustering.

Semi-supervised classification task uses both labeled and unlabeled data to improve classification performance. Though semi-supervised classification has only four concepts, it is a different kind of DM task and may attract more attention in the future because labeled data are often expensive and hard to get.

Frequent pattern mining task attempts to find patterns that appear frequently in data sets. A frequent pattern can be a set of items, subsequences, and substructures. Frequent pattern mining serves as a foundational task for other DM tasks. For instance, frequent itemset mining is a basic task for association rule mining.
Frequent pattern mining include frequent itemset mining, frequent structure patterns mining, frequent tree pattern mining, mining frequent patterns from graphs, mining top $k$-frequent closed patterns, and frequent set mining.

Sequential pattern mining task tries to find frequent subsequences or important sequences from sequence databases. One difference between frequent pattern mining and sequential pattern mining is that sequential pattern mining considers the order in the pattern. Sequential pattern mining includes frequent sequences mining, mining sequence of events, mining frequent episodes, mining asynchronous sequences, and evolutionary sequence mining.

Time series analysis involves analyzing sequences of data that are recorded at regular time intervals. Major concepts include similarity search in time series, periodic patterns in time series, segment time series data, and time series classification.

Concept description task includes characterization and comparison. Characterization summarizes the given collection of data and comparison provides descriptions comparing two or more data set. Concept description task includes concepts like generating characteristic rules, mining conceptual hierarchies, concept decompositions, data summarization, and concept construction.

Exception detection task attempts to identify exceptions or outliers from databases. Exception detection task can be either a single DM task or a preprocessing task. As a stand-alone task, exception detection is used to identify outliers that are useful in fraud detection. As a preprocessing task, it helps remove outliers from data.

There are only seven concepts in link mining subcategory. It is listed as a stand-alone because it is inappropriate to classify these concepts to other subcategories and link mining is a new promising area. To mine knowledge from complex data types is an important and challenge issue in DM. These complex data objects are normally linked in some manner and the linkage information can be used to improve classification accuracy. Mining social networks, network values of customers, and mining linkage between Web sites are main concepts of link mining.

A.2. Mining complex data

DM at its early stage is mostly concerned about mining knowledge from numerical data that are stored in relational databases or data warehouses. However, a great amount of data exists in complex forms, such as text, image, spatio, Web, multimedia, and combination of various data types. Complex data have certain characteristics and require new DM strategies at each step of DM process to manage these characteristics. As an emerging subfield of DM, mining complex data has attracted
a lot of research effort of late. The first international workshop on Mining Complex Data was held last year in conjunction with ICDM’05. Concepts that describe the characteristics and relevant ideas are included in this category. Algorithms and techniques that are used to mining complex data are placed under “learning method and technique” category.

The category of mining complex data has the following subcategories: text mining, spatial DM, graph mining, temporal DM, stream mining, Web mining, bioinformatics, and multimedia DM.

The major concepts in text mining are text classification, information extraction, information retrieval, text categorization, topic identification, and natural language processing.

Spatial DM refers to the extraction of knowledge from space-related data. Space-related data can be geometric, geographic, or spatial data. Spatial DBMS, spatial OLAP, frequent neighboring class sets in spatial DB, and spatial outlier detection are main concepts in spatial DM. In addition, spatio-temporal DM, which concerns the extraction of spatial and temporal relationships and studies both spatial and temporal relations simultaneously, is grouped under spatial DM. The study of spatio-temporal DM has been an interesting topic in geographic information science recently.

Temporal DM is “a single step in the process of Knowledge Discovery in Temporal Databases that enumerates structures over the temporal data, and any algorithm that enumerates temporal patterns from, or fits models to, temporal data”. Major concepts in temporal DM include temporal event sequence mining, temporal association mining, and temporal rule induction.

Graphs exist in various applications, including chemical compounds, proteins, DNA, workflow, Web, and social networks. Major concepts of graph mining include mining frequent subgraphs/graph patterns, graph partitioning, graphic modeling, mining connection subgraphs, and predictive graph mining.

Stream mining refers to the extraction of knowledge or structures from continuous, ordered, fast, and huge amount of data records. Major concepts in stream mining include finding correlations from streams, stream classification, and stream clustering.

Web mining task utilizes DM techniques to extract knowledge or useful information from Web data. Web data combine text documents, hypertext documents, links structures, and logs (server logs and browser logs). Concepts in Web mining task fit into the three categories proposed. Web content mining, Web structure mining, and Web usage mining. Web usage mining has the most concepts among these three subcategories.

Bioinformatics involves computational management and analysis of biological data. Major concepts in bioinformatics are microarray analysis, mining gene expression data, DNA sequence analysis, and mining protein sequences.

Multimedia DM deals with images, animation, sound, video and any other type of digital information. Major concepts in multimedia DM include image
mining, multimedia DM framework, video mining, and spoken and written data classification.

A.3. DM software and system

DM software and systems category has two subcategories: DM software and DM systems. DM software includes DM software and tools, DM software tools classification and recommendation, DM toolbox, software maintenance, SAS EM, IBM intelligent miner, code synthesis, DM software development environment, and intelligent agents. Data mining systems refer to various systems designed for DM tasks, such as FTT systems, intelligent information systems, rough-sets-based system for generating if-then rules, hybrid DM system, and sailing system.

A.4. High-performance and distributed DM

This category combines high-performance DM and distributed DM. One of the characteristics of DM is the huge size of data that need to be analyzed and these data are often physically distributed. Mining from large and distributed data sets requires high-performance computers and algorithms. Several important conferences also include high-performance and distributed mining program (e.g. SIAM conference on DM, International Parallel and Distributed Processing Symposium, European Conference on Parallel Processing). Therefore, these two subcategories are combined together.

High-performance DM aims at providing high-performance DM algorithms. It includes concepts like parallel algorithms, parallel processing, and scalable DM algorithms. Distributed DM deals with distributed databases and distributed DM algorithms.

A.5. DM application

As an application-driven field, most DM methods and techniques are designed for real-life applications. Among the numerous applications, e-commerce and market basket analysis are the top two DM applications. Other major applications include market analysis, network intrusion detection, fraud or anomaly detection, customer retention, change detection, churn analysis in telecommunications, stock trend prediction, drug design, spam filtering, and credit scoring. There are also some fresh DM applications, such as brain activation classification, chip testing, analyze expressive music performance, earthquakes prediction, DM in hydrogeology, environmental monitoring, and DM for reverse engineering. The broad spectrum of industries that utilize DM techniques indicates that more and more business and science areas have recognized the value of DM.

A.6. DM process and project

This category has two subcategories: DM process and DM project. The KDD process proposed by Fayyad et al.\textsuperscript{2,5,22} has been a \textit{de facto} standard in DM.
The process includes data selection, preprocessing, transformation, DM, and results interpretation/evaluation. Transformation is often considered as a part of preprocessing and results that interpretation/evaluation is often considered as post-processing. DM process has six sub-subcategories: preprocessing, model and algorithm selection, domain knowledge, post-processing, process models, and visualization techniques.

Preprocessing consists of data reduction, data integration and transformation, attributes learning, and data cleaning. Major concepts in data reduction are feature selection, data compression, sampling, and data/instance selection. Data integration and transformation includes data transformation strategy, attributes transformation, discretization methods, normalization, and integrate multiple data sources. Attributes learning concentrates on learning attribute relationships (e.g. correlated variables, attribute dependencies, and causal relationships among variables), feature interaction, feature ranking, and attribute generation. Data cleaning includes noisy data handling, missing data problem, incomplete data, data transformation, and data duplication.

Model and algorithm selection helps users to select the appropriate models or learning algorithms for specific DM tasks.

Understanding and utilizing domain knowledge is a critical success factor for DM activities. Domain knowledge include concepts like representation of background knowledge, exploit domain knowledge, extract domain knowledge, integrate domain knowledge into algorithms, background knowledge for managing uncertainty, represent experts' decision processes, and domain knowledge acquisition.

Post-processing includes model evaluation, maintenance of discovered knowledge (e.g. maintaining association rules, maintenance of sequence patterns, incremental maintenance algorithm, and knowledge base maintenance), update existing knowledge, results interpretation and exploration, evaluation techniques, and validation techniques.

Process models include KDD process, CRISP-DM model, DM process template, and platform for KDD process deployment.

Visualization techniques is listed as a separate sub-subcategory of DM process because they are designed for difference DM steps, from preprocessing to post-processing, and visual DM has become a new subfield of DM. Major concepts in visualization techniques include visualization of attributes, visualizing large data sets, visualization tools for presenting discovered knowledge (e.g. visualization of classifiers, visualize association rules, SOM results representation, and visualizing SVM results), and interactive visual DM.

DM project has only three concepts: DM projects goal development, interdiscipli-
enough attention to this issue. The only article concentrates on developing successful DM project is the dissertation by Sim (2003).

A.7. Foundation of DM

The term “foundation” is used here to refer to any fundamental principles, theories, or bases that support DM activities. Foundations of DM have the following subcategories: theory, measures, privacy and security issues, experiments, benchmarks, challenges and surveys and reviews.

Theory includes theories of DM, frameworks, taxonomies, theoretical proofs, theoretic position papers, and description of theories that can be used in DM. Examples in theory subcategory are as follows: theory for parameter-free DM, microeconomic view framework for DM, resource description framework, summarization of propensity score methods, statistical theory for quantitative association rules, rough set perspective, taxonomy for dirty data, and taxonomy for SOM.

Measures subcategory contains concepts that attempt to measure the quality and interestingness metrics of DM results. Concepts in measures include similarity measures, interestingness measures, significance measures, robustness measures, scalability and efficiency measures, computational complexity, computational stability, closeness measures, and reliability measures.

Privacy and security issues were brought up when the first KDD workshop was held in Detroit, Michigan. Major concepts in privacy and security include privacy protection, data privacy, privacy-preserving DM, cryptographic techniques, randomization technique, privacy-enhancing DM, and privacy protection techniques for specific DM tasks (e.g. privacy-preserving association rule mining and privacy-preserving outlier detection).

Most DM research papers use experiments to text their methods and algorithms. Experiments subcategory provides theoretic support for designing and evaluating these experiments. Concepts in experiments include experimental evaluation, experimental design, replicability of significant test, and significance tests for patterns.

Benchmarks subcategory is closed related to experiments. Benchmark data sets are important for two reasons: first, they enable researchers to test their algorithms; second, they provide standards that allow researchers to compare their experimental results and reduce potential empirical design problems. UCI machine learning and UCI KDD archive are two most popular online repositories of data sets. The Sloan Digital Sky Survey (SDSS), which has so far generated tens of millions of images of distant galaxies, is another data set.

Challenges concentrate on problems that DM community is facing and future development of DM. Concepts include DM grand challenges, future DM tools, practical development and DM, and advances and trends in DM.

Surveys and reviews provide introductory or thorough investigation of certain DM topics.
Appendix B. Taxonomy of Categories (Part A)

Data Mining and Knowledge Discovery

Data Mining Tasks
- Association rule mining
- Classification and prediction
- Clustering
- Semi-supervised classification
- Frequent pattern mining
- Sequential pattern mining
- Time series analysis
- Concept description
- Exception detection
- Link analysis
- Novel data mining

Learning methods and techniques
- Algorithms and models
- Database technology
  - Data warehouse and OLAP
  - Query language
  - Multi-relational DB
  - Indexing
  - Inductive and deductive DB
  - Database integration
  - Active data mining
- Artificial intelligence
  - Machine learning
    - Supervised
    - Unsupervised
    - Decisional
    - SVM and kernel
    - Neural
    - Inductive Logic
    - Reinforcement
    - Soft computing
      - Fuzzy logic
      - Rough sets
      - Genetic algorithms
- Statistics
  - Regression
  - Bayesian analysis
  - Principle Component Analysis
  - Probability models
    - Markov chain Monte Carlo
    - Hidden Markov Model
    - Discriminant analysis
- Optimization
Taxonomy of Categories Analysis (Part B)

Data Mining and Knowledge Discovery

- Mining complex data
  - Text mining
  - Spatial data mining
  - Graph mining
  - Temporal data mining
  - Stream mining
  - Web mining
  - Bioinformatics
  - Multimedia data mining

Data mining software & systems

- Data mining software
- Data mining systems

High performance and distributed data mining

- High performance data mining
- Distributed data mining

Data mining applications

Data mining process and project

- Data mining process
- Data mining project

Foundation of data mining

- Data mining theories
- Data mining measures
- Privacy and security issues
- Data mining experiments
- Benchmarks
- Data mining challenges
- Survey and review
Appendix C. Framework Validation: PAKDD 2005

<table>
<thead>
<tr>
<th>Learning methods and techniques</th>
<th>DM tasks</th>
<th>Foundations of DM</th>
<th>Mining complex data</th>
<th>High-performance and distributed DM</th>
<th>DM process and project</th>
<th>DM applications</th>
<th>DM software and system</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data mining of gene expression microarray via weighted prefix trees</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. A kernel function method in clustering</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Extraction of frequent few-overlapped monotone DNF formulas with depth-first pruning</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Rule extraction from trained support vector machines</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Pruning derivative partial rules during impact rule discovery</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. A new informative generic base of association rules</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Finding sporadic rules using a priori inverse</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Pushing tougher constraints in frequent pattern mining</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. An efficient compression technique for frequent itemset generation in association rule mining</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. A DNA index structure using frequency and position information of genetic</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Bayesian sequence learning for predicting protein cleavage points</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix C. (Continued)

<table>
<thead>
<tr>
<th>Learning methods and techniques</th>
<th>DM tasks</th>
<th>Foundations of DM</th>
<th>Mining complex data</th>
<th>High-performance and distributed DM</th>
<th>DM process and project</th>
<th>DM applications</th>
<th>DM software and system</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Threshold tuning for improved classification association rule mining</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Retrieval based on language model with relative entropy and feedback</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Text classification for DAG-structured categories</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Sentiment classification using word subsequences and dependency subtrees</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. Improving rough classifiers using concept ontology</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. QED: An efficient framework for temporal region query processing</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20. A MPAA-based iterative clustering algorithm augmented by nearest neighbors search for time-series data streams</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. Locating motifs in time-series data</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Description</td>
<td>Learning methods and techniques</td>
<td>DM tasks</td>
<td>Foundations of DM</td>
<td>Mining complex data</td>
<td>High-performance and distributed DM</td>
<td>DM process and project</td>
</tr>
<tr>
<td>-----</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td>----------</td>
<td>------------------</td>
<td>---------------------</td>
<td>-------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>22</td>
<td>A neighborhood-based clustering algorithm</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Speeding-up hierarchical agglomerative clustering in the presence of expensive metrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Dynamic cluster formation using level set methods</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>A vector field visualization technique for self-organizing maps</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Visualization of cluster changes by comparing self-organizing maps</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Approximated clustering of distributed high-dimensional data</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Improvements of IncSpan: Incremental mining of sequential patterns in large database</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Efficient sampling: Application to image data</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Cluster-based rough set construction</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Improving mining quality by exploiting data dependency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Progressive sampling for association rules based on sampling error estimation</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix C. (Continued)

<table>
<thead>
<tr>
<th></th>
<th>Learning methods and techniques</th>
<th>DM tasks</th>
<th>Foundations of DM</th>
<th>Mining complex data</th>
<th>High-performance and distributed DM</th>
<th>DM process and project</th>
<th>DM applications</th>
<th>DM software and system</th>
</tr>
</thead>
<tbody>
<tr>
<td>33.</td>
<td>ADenTS: An adaptive density-based tree structure for approximating aggregate queries over real attributes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td>Using consensus susceptibility and consistency measures for inconsistent knowledge management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35.</td>
<td>WLPMiner: Weighted frequent pattern mining with length-decreasing support constraints</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36.</td>
<td>A framework for incorporating class priors into discriminative classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37.</td>
<td>Increasing classification accuracy by combining adaptive sampling and convex pseudo-data</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38.</td>
<td>Kernels over relational algebra structures</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39.</td>
<td>SETRED: Self-training with editing</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40.</td>
<td>Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix C. (Continued)

<table>
<thead>
<tr>
<th>Learning methods and techniques</th>
<th>DM tasks</th>
<th>Foundations of DM</th>
<th>Mining complex data</th>
<th>High-performance and distributed DM</th>
<th>DM process and project</th>
<th>DM applications</th>
<th>DM software and systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved bayesian spam filtering based on co-weighted multi-area information</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An efficient framework for mining flexible constraints</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support-oriented discovery of generalized disjunction-free representation of frequent patterns with negation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USAID: Unifying signature-based and anomaly-based intrusion detection</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial data</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclic pattern kernels revisited</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subspace clustering of text documents with feature weighting K-means algorithm</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mining frequent trees with node-inclusion constraints</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Considering re-occurring features in associative classifiers</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References

41. J. Kogan, C. Nicholas and M. Teboulle, Clustering large and high dimensional data, *Tutorial of ACM Fourteenth Conf. Information and Knowledge Management (CIKM)* (Bremen, Germany, 2003).


