Preface

Since its definition a decade ago, the problem of mining patterns is becoming a very active research area, and efficient techniques have been widely applied to problems either in industry, government or science. From the initial definition and motivated by real applications, the problem of mining patterns not only addresses the finding of itemsets but also more and more complex patterns. For instance, new approaches need to be defined for mining graphs or trees in applications dealing with complex data such as XML documents, correlated alarms or biological networks. As the number of digital data are always growing, the problem of the efficiency of mining such patterns becomes more and more attractive.

One of the first areas dealing with a large collection of digital data is probably text mining. It aims at analyzing large collections of unstructured documents with the purpose of extracting interesting, relevant and nontrivial knowledge. However, patterns became more and more complex, and led to open problems. For instance, in the biological networks context, we have to deal with common patterns of cellular interactions, organization of functional modules, relationships and interaction between sequences, and patterns of genes regulation. In the same way, multidimensional pattern mining has also been defined, and a lot of open questions remain regarding the size of the search space or to effectiveness consideration. If we consider social network in the Internet, we would like to better understand and measure relationships and flows between people, groups and organizations. Many real-world applications data are no longer appropriately handled by traditional static databases since data arrive sequentially in rapid, continuous streams. Since data-streams are contiguous, high speed and unbounded, it is impossible to mine patterns by using traditional algorithms requiring multiple scans and new approaches have to be proposed.

In order to efficiently aid decision making, and for effectiveness consideration, constraints become more and more essential in many applications. Indeed, an unconstrained mining can produce such a large number of patterns that it may be intractable in some domains. Furthermore, the growing consensus that the end user is no more interested by a set patterns verifying selection criteria led to demand for novel strategies for extracting useful, even approximate knowledge.

The goal of this book is to provide an overall view of the existing solutions for mining new kinds of patterns. It aims at providing theoretical frameworks and presenting challenges and possible solutions concerning pattern extraction with an emphasis on both research techniques and real-world applications. It is composed of 11 chapters.

Often data mining problems require metric techniques defined on the set of partitions of finite sets (e.g., classification, clustering, data preparation). The chapter "Metric Methods in Data Mining" proposed by D. A. Simovici addresses this topic. Initially proposed by R. López de Màntaras, these techniques formulate a novel splitting criterion that yields better results than the classical entropy gain splitting techniques. In this chapter, Simovici investigates a family of metrics on the set of partitions of finite sets that is linked to the notion of generalized entropy. The efficiency of this approach is proved through experiments conducted for different data mining tasks: classification, clustering, feature extraction and discretization. For each approach the most suitable metrics are proposed.

Mining patterns from a dataset always rely on a crucial point: the interest criterion of the patterns. Literature mostly proposes the minimum support as a criterion; however, interestingness may occur in constraints applied to the patterns or the strength of the correlation between the items of a pattern, for instance. The next two chapters deal with these criteria.

In "Bidirectional Constraint Pushing in Frequent Pattern Mining" by O.R. Zaïane and M. El-Hajj, proposes consideration of the problem of mining constrained patterns. Their challenge is to obtain a sheer number of rules, rather than the very large set of rules usually resulting from a mining process. First, in a survey of constraints in data mining (which covers both definitions and methods) they show how the previous methods can generally be divided into two sets. Methods from the first set consider the monotone constraint during the mining, whereas methods from the second one consider the antimonotone constraint. The main idea, in this chapter, is to consider both constraints (monotone and antimonotone) early in the mining process. The proposed algorithm (BifoldLeap) is based on this principle and allows an efficient and effective extraction of constrained patterns. Finally, parallelization of BifolLeap is also proposed in this chapter. The authors thus provide the reader with a very instructive chapter on constraints in data mining, from the definitions of the problem to the proposal, implementation and evaluation of an efficient solution.

Another criterion for measuring the interestingness of a pattern may be the correlation between the items it contains. Highly correlated patterns are named "Hyperclique Patterns" in the chapter of H. Xiong, P. N. Tan, V. Kumar and W. Zhou entitled "Mining Hyperclique Patterns: A Summary of Results". The chapter provides the following observation: when the minimum support in a pattern mining process is too low, then the number of extracted itemsets is very high. A thorough analysis of the patterns will often show patterns that are poorly correlated (i.e., involving items having very different supports). Those patterns may then be considered as spurious patterns. In this chapter, the authors propose the definition of hyperclique patterns. Those patterns contain items that have similar threshold. They also give the definition of the h-confidence. Then, h-confidence is analyzed for properties that will be interesting in a data mining process: antimonotone, cross-support and a measure of association. All those properties will help in defining their algorithm: hyperclique miner. After having evaluated their proposal, the authors finally give an application of hyperclique patterns for identifying protein functional modules.

This book is devoted to provide new and useful material for pattern mining. Both methods aforementioned are presented in the first chapters in which they focus on their efficiency. In that way, this book reaches part of the goal. However, we also wanted to show strong links between the methods and their applications. Biology is one of the most promising domains. In fact, it has been widely addressed by researchers in data mining those past few years and still has many open problems to offer (and to be defined). The next two chapters deal with bioinformatics and pattern mining.

Biological data (and associated data mining methods) are at the core of the chapter entitled "Pattern Discovery in Biosequences: From Simple to Complex Patterns" by S. Rombo and L. Palopoli. More precisely, the authors focus on biological sequences (e.g., DNA or protein sequences) and pattern extraction from those sequences. They propose a survey on existing techniques for this purpose through a synthetic formalization of the problem. This effort will ease reading and understanding the presented material. Their chapter first gives an overview on biological datasets involving sequences such as DNA or protein sequences. The basic notions on biological data are actually given in the introduction of this chapter. Then, an emphasis on the importance of *patterns* in such data is provided. Most necessary notions for tackling the problem of mining patterns from biological sequential data are given: definitions of the problems, existing solutions (based on tries, suffix trees), successful applications as well as future trends in that domain.

An interesting usage of patterns relies in their visualization. In this chapter, G. Leban, M. Mramor, B. Zupan, J. Demsar and I. Bratko propose to focus on "Finding Patterns in Class-labeled Data Using

Data Visualization." The first contribution of their chapter is to provide a new visualization method for extracting knowledge from data. WizRank, the proposed method, can search for interesting multidimensional visualizations of class-labeled data. In this work, the interestingness is based on how well instances of different classes are separated. A large part of this chapter will be devoted to experiments conducted on gene expression datasets, obtained by the use of DNA microarray technology. Their experiments show simple visualizations that clearly visually differentiate among cancer types for cancer gene expression data sets.

Multidimensional databases are data repositories that are becoming more and more important and strategic in most of the main companies. However, mining these particular databases is a challenging issue that has not yet received relevant answers. This is due to the fact that multidimensional databases generally contain huge volumes of data stored according to particular structures called star schemas that are not taken into account in most popular data mining techniques. Thus, when facing these databases, users are not provided with useful tools to help them discovering relevant parts. Consequently, users still have to navigate manually in the data, that is—using the OLAP operators—users have to write sophisticated queries. One important task for discovering relevant parts of a multidimensional database is to identify homogeneous parts that can summarize the whole database. In the chapter "Summarizing Data Cubes Using Blocks," Y. W. Choong, A. Laurent and D. Laurent propose original and scalable methods to mine the main homogeneous patterns of a multidimensional database. These patterns, called blocks, are defined according to the corresponding star schema and thus, provide relevant summaries of a given multidimensional database. Moreover, fuzziness is introduced in order to mine for more accurate knowledge that fits users' expectations.

The first social networking website began in 1995 (i.e., classmates). Due to the development of the Internet, the number of social networks grew exponentially. In order to better understand and measuring relationships and flows between people, groups and organizations, new data mining techniques, called social network mining, appear. Usually social network considers that nodes are the individual actors within the networks, and ties are the relationships between the actors. Of course, there can be many kinds of ties between the nodes and mining techniques try to extract knowledge from these ties and nodes. In the chapter "Social Network Mining from the Web," Y. Matsuo, J. Mori and M. Ishizuka address this problem and show that Web search engine are very useful in order to extract social network. They first address basic algorithms initially defined to extract social network. Even if the social network can be extracted, one of the challenging problems is how to analyze this network. This presentation illustrates that even if the search engine is very helpful, a lot of problems remain, and they also discuss the literature advances. They focus on the centrality of each actor of the network and illustrate various applications using a social network.

Text-mining approaches first surfaced in the mid-1980s, but thanks to technological advances it has been received a great deal of attention during the past decade. It consists in analyzing large collections of unstructured documents for the purpose of extracting interesting, relevant and nontrivial knowledge. Typical text mining tasks include text categorization (i.e., in order to classify document collection into a given set of classes), text clustering, concept links extraction, document summarization and trends detection.

The following three chapters address the problem of extracting knowledge from large collections of documents. In the chapter "Discovering Spatio-Textual Association Rules in Document Images", M. Berardi, M. Ceci and D. Malerba consider that, very often, electronic documents are not always available and then extraction of useful knowledge should be performed on document images acquired by scanning the original paper documents (document image mining). While text mining focuses on patterns involving words, sentences and concepts, the purpose of document image mining is to extract high-level spatial objects and relationships. In this chapter they introduce a new approach, called WISDOM++, for processing documents and transform documents into XML format. Then they investigate the discovery of spatio-textual association rules that takes into account both the layout and the textual dimension on XML documents. In order to deal with the inherent spatial nature of the layout structure, they formulate the problem as multi-level relational association rule mining and extend a spatial rule miner SPADA (spatial pattern discovery algorithm) in order to cope with spatio-textual association rules. They show that discovered patterns could also be used both for classification tasks and to support layout correction tasks.

L. Candillier, L. Dunoyer, P. Gallinari, M.-C. Rousset, A. Termier and A. M. Vercoustre, in "Mining XML Documents," also consider an XML representation, but they mainly focus on the structure of the documents rather than the content. They consider that XML documents are usually modeled as ordered trees, which are regarded as complex structures. They address three mining tasks: frequent pattern extraction, classification and clustering. In order to efficiently perform these tasks they propose various tree-based representations. Extracting patterns in a large database is very challenging since we have to consider the two following problems: a fast execution and we would like to avoid a memory-consuming algorithm. When considering tree patterns the problem is much more challenging due to the size of the research space. In this chapter they propose an overview of the best algorithms. Various approaches to XML document classification and clustering are also proposed. As the efficiency of the algorithms depends on the representation, they propose different XML representations based on structure, or both structure and content. They show how decision-trees, probabilistic models, k-means and Bayesian networks can be used to extract knowledge from XML documents.

In the chapter "Topic and Cluster Evolution Over Noisy Document Streams," S. Schulz, M. Spiliopoulou and R. Schult also consider text mining but in a different context: a stream of documents. They mainly focus on the evolution of different topics when documents are available over streams. As previously stated, one of the important purpose in text mining is the identification of trends in texts. Discover emerging topics is one of the problems of trend detection. In this chapter, they discuss the literature advances on evolving topics and on evolving clusters and propose a generic framework for cluster change evolution. However discussed approaches do not consider non-noisy documents. The authors propose a new approach that puts emphasis on small and noisy documents and extend their generic framework. While cluster evolutions assume a static trajectory, they use a set-theoretic notion of overlap between old and new clusters. Furthermore the framework extension consider both a document model describing a text with a vector of words and a vector of n-gram, and a visualization tool used to show emerging topics.

In a certain way, C. J. Joutard, E. M. Airoldi, S. E. Fienberg and T. M. Love also address the analysis of documents in the chapter "Discovery of Latent Patterns with Hierarchical Bayesian Mixed-Membership Models and the Issue of Model Choice." But in this chapter, the collection of papers published in the Proceedings of the National Academy of Sciences is used in order to illustrate the issue of model choice (e.g., the choice of the number of groups or clusters). They show that even if statistical models involving a latent structure support data mining tasks, alternative models may lead to contrasting conclusions. In this chapter they deal with hierarchical Bayesian mixed-membership models (HBMMM), that is, a general formulation of mixed-membership models, which are a class of models very well adapted for unsupervised data mining methods and investigate the issue of model choice in that context. They discuss various existing strategies and propose new model specifications as well as different strategies of model choice in order to extract good models. In order to illustrate, they consider both analysis of documents and disability survey data.