CHAPTER I

INTRODUCTION

Data mining, also known as Knowledge Discovery in Databases (KDD), is concerned with the extraction of previously unrecognized and interesting information contained within (usually large) data repositories. The interesting is of course a subjective concept, and a definition of what is interesting is required: it is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity (Fayyad et al., 1996). Almost always, what is being sought is some relationship which can be observed between categories of information in the data. A particular way to describe such a relationship is in the form of an association rule which relates attributes within the database.

The problem of mining association rules has been defined by Agrawal (Agrawal et al., 1993) as first proposed for market basket analysis in the form of association rule mining. It analyzes customers buying habits by finding associations between the different items that customers place in their "shopping baskets". For instance, if customers are buying milk, how likely are they going to also buy yogurt (and what kind of yogurt) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and arrange their shelf space. Association mining applications have been applied to many different domains including market basket and risk analysis in commercial environments, epidemiology, clinical medicine, fluid dynamics, astrophysics, crime prevention, and counter-terrorismall areas in which the relationship between objects can provide useful knowledge.

The process of mining association rules consists of two steps: (i) Find the frequent itemsets that have minimum support; (ii) Use the frequent itemsets to generate association rules that meet the confidence threshold. Among these two steps, step (i) is the most expensive since the number of itemsets grows exponentially with the number of items. Consequently, the task of frequent itemsets discovery (also called frequent patterns discovery) is widely studied in data mining as a mean of generating association rules (Agrawal et al., 1993), correlations (Brin et al., 1997a), sequential patterns (Agrawal and Srikant, 1995), emerging patterns (Dong and Li, 1999), dense regular patterns (Engler, 2008), frequent patterns with maximum length (Hu et al., 2008), frequent patterns with temporal dependencies (Tatavarty et al., 2007), negative rules (Wu et al., 2004), causality (Silverstein et al., 2000), weighted pattern mining (Tao et al., 2003)(Yun and Leggett, 2006) and classification rules (Li et al., 2001) (Liu et al., 1998).

Over the past decade a large number of research works have been published presenting new algorithms or improvements on existing algorithms to solve the frequent pattern mining problem more efficiently through the refinement of search strategies (depth first/breath first search (Agrawal and Srikant, 1994), top down/bottom up traversals (Grahne and Zhu, 2005)), pruning techniques, data structures (trees/other data structures (Han et al., 2004)), the use of alternative dataset organizations (vertical/horizontal formats (Zaki and Gouda, 2003)) and constraints ((Bonchi and Lucchese, 2005) (Pei et al., 2001a)). Recent surveys may be found in (Goethals, 2005) and (Han et al., 2007). However, two main bottlenecks exist: (i) A huge number of patterns are generated and (ii) Most of them are redundant or uninteresting. To tackle these problems, various approaches have been developed.

Frequent-closed pattern mining algorithms have been proposed to reduce redundant patterns (Pasquier et al., 1999) and to mine a compact set of frequent patterns which cover all frequent patterns (Pei et al., 2000). When a data set is dense, the number of frequent closed patterns extracted can be orders of magnitude fewer than the number of corresponding frequent patterns since they implicitly benefit from data correlations. Nevertheless, they concisely represent exactly the same knowledge. From closed patterns, it is in fact trivial to generate all the frequent patterns along with their supports. More importantly, association rules extracted from closed patterns have been proven to be more meaningful for analysts, because all redundancies are discarded (Zaki, 2004). A recent survey may be found in (Yahia et al., 2006).

While the previous approaches work at the algorithmic level, another strategy is to rank patterns in a post-algorithmic phase with objective measures of interest (Hilderman and Hamilton, 2000). A large number of interestingness measures have been proposed such as mining frequent patterns with tougher constraints (Bonchi and Lucchese, 2005), mining dense regular patterns (Engler, 2008), mining frequent patterns with maximum length (Hu et al., 2008), mining frequent patterns with convertible constraints (Pei et al., 2001a), mining frequent patterns with constraints using pattern growth approach (Pei and Han, 2002), mining temporal dependencies between frequent patterns (Tatavarty et al., 2007), integrating classification and association rule mining (Li et al., 2001)(Liu et al., 1998), etc. Interesting surveys and comparisons may be found in (Geng and Hamilton, 2006)(Lenca et al., 2008) and (Suzuki, 2008).

On constraint-based patterns mining, pushing the constraints using objective measures deeply into the patterns mining process is a very interesting approach (Bonchi and Lucchese, 2005) (Pei et al., 2001a). This approach uses efficient pruning strategies to discover interesting patterns such as optimal rule mining (Li, 2006)(Le Bras et al., 2009). It is important to notice that most of the previous mentioned works, except mainly (Li, 2006) and (Le Bras et al., 2009), are

always subject to the dictatorship of support for the frequent pattern mining step. Avoiding the use of support has been recognized as a major challenge, such as mining high confidence association without support pruning (Cohen et al., 2001), (Bhattacharyya and Bhattacharyya, 2007), (Le Bras et al., 2010), and mining rules without support threshold (Li et al., 1999)(Koh, 2008).

However, without specific knowledge, the setting of minimum support threshold is quite tricky and it leads to the following problem that may hinder its popular use. There are two challenges of minimum support based mining: (i) if the value of minimum support is set to be too small, the pattern mining algorithm may lead to the generation of thousands of patterns; (ii) if the value of minimum support constraint is set to be too big, the mining algorithm may often generate a few patterns or even no answers. In which case, the user may have to guess a smaller threshold and do the mining again, which may or may not give a better result. As it is difficult to predict how many patterns will be mined with a user-defined minimum support threshold, the top-k pattern mining has been proposed. As a consequence many works have focused on avoiding the use of a support threshold (e.g. (Li et al., 1999; Cheung and Fu, 2004; Koh, 2008)), or avoiding the use of the support itself (e.g. (Cohen et al., 2001), (Bhattacharyya and Bhattacharyya, 2007) and (Le Bras et al., 2010)). Another solution involves asking the number of desired outputs (Fu et al., 2000). Therefore, mining top-k patterns has become a very popular task. In particular, top-k frequent closed patterns (e.g. (Han et al., 2002), (Wang et al., 2005) and (Pietracaprina and Vandin, 2007)) and top-k patterns (e.g. (Fu et al., 2000) and (Yang et al., 2008)) have motivated a lot of works. Nowadays, mining from data streams also offers a new challenge because one cannot save all the patterns and their related information, due to the limitation of memory space. Thus mining top-k patterns from data streams becomes of great interest (e.g. (Metwally et al., 2005), (Li, 2009b), (Li, 2009a) and (Tsai, 2010)).

Recently, Tanbeer et al. (Tanbeer et al., 2009) proposed a pattern mining approach with a regular constraint on patterns appearance and a minimum support constraint. As pointed out by the authors, there are several applications to apply regular frequent patterns mining: in a retail market, among all frequently sold products, the sales manager may be interested only on the regularly sold products compared to the rest; for web site design or web administration, an administrator may be interested on the click sequences of heavily hit web pages; in genetic data analysis the set of all genes that not only appear frequently but also co-occur at regular interval in DNA sequence may carry more significant information to scientists; for stock market, the set of high stocks indices that rise regularally may be of special interest to traders, etc. Thus the occurrence regularity plays an important role in discovering some interesting frequent patterns in such applications (Engler, 2008) (Laxman and Sastry, 2006).

This dissertation here focuses on these two bottlenecks and extend the work of (Tanbeer et al., 2009). Thus, a new kind of pattern, namely the top-k regular-frequent itemsets, which is discovered from transactional databases is proposed. From this kind of pattern, the users can control the number of regular-frequent itemsets to be mined. At first, MTKPP algorithm (Mining Top$K$ Periodic(Regular)-frequent Patterns) is introduced. It based on the use of top-k list structure and a best-first search strategy to quickly discover the regular itemsets with high supports. To calculate support of each itemset, a set of transaction-ids (that each itemset occurs) is collected. MTKPP also applies intersection process to collect and calculate a set of transaction-ids, a regularity and a support of each larger itemset. Next, TKRIMPE algorithm (Top-K Regular Itemset Mining using database Partitioning and support Estimation) is presented. TKRIMPE based on the database partitioning and support estimation techniques. By using these techniques, TKRIMPE can achieve a good performance especially on sparse datasets. Further, a new concise representation named interval tidset representation is devised. Then, a new efficient algorithm, called TKRIMIT (Top-K Regular-frequent Itemsets Mining based on Interval Tidset representation), is also proposed. With interval tidset representation, TKRIMIT can reduce the processing time and memory usage on dense datasets. Finally, an efficient and scalable algorithm named H-TKRIMP (Hybrid representation on Top-K Regular-frequent Itemsets Mining based on database Partitioning), based on the combination between normal tidset and interval tidset representations and the database partitioning, is devised. By comparing with other algorithms, H-TKRIMP can achieve a good performance for the small and large values of desired results on both sparse and dense datasets.

1.1 Objectives of Study

The objectives of study are as follows:

- To develop algorithms to mine top-k regular-frequent patterns that are very efficient in the terms of computational time and memory consumption.
- To develop a new technique to collect tidset of each itemset which can be applied in various problems such as frequent pattern mining, frequent closed pattern mining, and weighted frequent pattern mining.
- To propose an analysis of the performance of various techniques to maintain and intersect tidsets by making a comparison to a see trade-off between time and space.

1.2 Scopes of Study

The scopes of this study are as follows:

- This work considers the problem of top-k regular-frequent pattern mining.
- The datasets from *UCI* repository are used as a benchmark to test the proposed algorithms.
- Performance measurement can be either an actual running time (an actual memory consumption) or a complexity analysis.

1.3 Research Methodology

- Survey literature and review related works about association rules mining, frequent itemsets mining, frequent closed-itemsets mining, top-k frequent itemsets mining and regular-frequent itemsets mining.
- Study the principle theories and various techniques to mine frequent and other kinds of itemsets.
- Study various proposed representations used to maintain the content of databases.
- Collect the sparse and dense datasets from standard and existing benchmark datasets.
- Design an appropriate algorithms and perform the experiments to validate the algorithms
- Conclude the experimental results by comparing the results with those from other methods

1.4 Organization

The remainder of this dissertation is structured as follows: In Chapter 2, general background on association rules mining and its variant are introduced. Further, the frequent pattern mining, top-k itemsets mining and regular-frequent itemsets mining problems and related works are described. In Chapter 3, the formal notations and definitions used to mine a set of top-k regular-frequent itemsets are mentioned. An efficient algorithm, named MTKPP (Mining Top-K Periodic(Regular)-frequent Patterns), used a normal-tidset representation and applied a best-first search strategy is introduced. Chapter 4 presented a new efficient algorithm, called TKRIMPE (Top-K Regular-frequent Itemsets Mining using database Partitioning and support Estimation), applied the database partitioning and support estimation techniques in order to reduce computational time of MTKPP algorithm. Besides, a new concise interval tidset representation named interval tidset representation and an efficient algorithm called TKRIMIT (Top-K Regular-frequent

Itemsets Mining using Interval Tidset representation algorithm) is also proposed in Chapter 5. As further extensions, in Chapter 6, the database partitioning technique and the interval tidset representation are merged in order to devise a new algorithm, named H-TKRIMP (Hybrid representation on Top-K Regular-frequent Itemsets Mining based on database partitioning), that have a good performance on both dense and sparse datasets. Finally, Chapter 7 concludes this dissertation and describes future extension of this work.