Engineering Applications of Artificial Intelligence 61 (2017) 161–172

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Extracting recent weighted-based patterns from uncertain temporal databases

Wensheng Gan[a], Jerry Chun-Wei Lin[a,*], Philippe Fournier-Viger[b], Han-Chieh Chao[a,c], Jimmy Ming-Tai Wu[d], Justin Zhan[d]

[a] *School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China*
[b] *School of Natural Sciences and Humanities, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China*
[c] *Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan*
[d] *Department of Computer Science, University of Nevada, Las Vegas, USA*

ARTICLE INFO

Keywords:
Weighted frequent itemset
Recency constraint
Uncertain data
Upper bound
SUBDC strategy

ABSTRACT

Weighted Frequent Itemset Mining (WFIM) has been proposed as an extension of frequent itemset mining that considers not only the frequency of items but also their relative importance. However, using WFIM algorithms in real applications raises some problems. First, they do not consider how recent the patterns are. Second, traditional WFIM algorithms cannot handle uncertain data, although this type of data is common in real-life. To address these limitations, this paper introduces the concept of Recent High Expected Weighted Itemset (RHEWI), which considers the recency, weight and uncertainty of patterns. By considering these three factors, more up-to-date and relevant results are found. A projection-based algorithm named RHEWI-P is presented to mine RHEWIs using a novel *upper-bound downward closure* (*UBDC*) property. An improved version of this algorithm called RHEWI-PS is further proposed based on a novel *sorted upper-bound downward closure* (*SUBDC*) property for pruning unpromising candidate itemsets early. An experimental evaluation against the state-of-the-art HEWI-Uapriori algorithm was carried out on both real-world and synthetic datasets. Results show that the proposed algorithms are highly efficient and are acceptable for mining the desired patterns.

1. Introduction

Pattern mining is an important research topic, which consists of discovering interesting patterns in databases such as association rules (Agrawal et al., 1993; Chen et al., 1996; Han et al., 2004) and sequential patterns (Agrawal and Srikant, 1995; Srikant and Agrawal, 1996) among others (Lin et al., 2015; Yun and Leggett, 2005). Agrawal and Srikant (1994) proposed the Apriori algorithm to mine association rules by discovering the set of frequent itemsets. A well-known *downward closure* property was introduced in that work (Agrawal and Srikant, 1994) to reduce the search space for mining association rules. Different from the generate-candidate-and-test approach of Apriori, a pattern-growth approach (Han et al., 2004) was then proposed, which outperforms the Apriori algorithm. Frequent Itemset Mining (FIM) is now recognized as a key data mining task having a wide range of real-world applications. However, traditional FIM (Agrawal et al., 1993; Chen et al., 1996; Han et al., 2004) relies on the frequency as sole measure to discover interesting patterns (item/sets) in transactional databases and ignore other important implicit criteria such as the weight, interest, risk or profit of patterns. Because of this limitation, traditional FIM algorithms cannot be used for several real-world applications.

Cai et al. (1998) proposed the first weighted-support model. Wang et al. (2000) assigned different weights to items and proposed an algorithm to mine Weighted Association Rules (WAR). Tao et al. (2003) developed the WARM (Weighted Association Rule Mining) algorithm and designed the *weighted downward closure* property to mine weighted frequent itemsets (WFIs). In recent years, WFIM and related issues have been extensively studied (Baralis et al., 2015; Cagliero and Garza, 2014; Lin et al., 2015; Lan et al., 2014; Nguyen et al., 2016; Yun and Leggett, 2005), and several extensions have been proposed in many other fields such as mining weighted association rules without preassigned weights (Sun and Bai, 2008), infrequent weighted itemset mining (Cagliero and Garza, 2014), multilingual summarization using WFIs (Baralis et al., 2015), weighted partial periodic pattern mining (Yang et al., 2014) and weighted sequential

[*] Corresponding author.
E-mail addresses: wsgan001@gmail.com (W. Gan), jerrylin@ieee.org (J.C.-W. Lin), philfv@hit.edu.cn (P. Fournier-Viger), hcc@ndhu.edu.tw (H.-C. Chao), wmt@wmt35.idv.tw (J.M.-T. Wu), justin.zhan@unlv.edu (J. Zhan).

http://dx.doi.org/10.1016/j.engappai.2017.03.004
Received 12 September 2016; Received in revised form 9 March 2017; Accepted 12 March 2017
Available online 21 March 2017

pattern mining (Lan et al., 2014). Although WFIM can reveal more useful information than FIM, WFIM still suffers from several limitations.

In general, knowledge found in a temporal database changes over time. Extracting up-to-date knowledge, especially from temporal databases, can provide timely and valuable information for decision making (Chen et al., 2013, 2016; Hong et al., 2009; Lin et al., 2015). However, WFIM does not consider how recent the discovered patterns are. As a consequence, WFIM algorithms may discover a large amount of WFIs that only appeared in the far past but are irrelevant for up-to-date decision-making. It seems unreasonable to measure the interestingness of patterns in time-sensitive databases without considering their recency. A database is said to be time-sensitive if it changes over time such that observations may be only relevant or valid for short periods of time. Many real-time control and analysis systems are time-sensitive. For example, for incident management and forecast systems, outdated information may lead to misleading risk analytics and incident forecast. In market basket analysis, obtaining information about recent or current sale trends is crucial, and much more important than obtaining information about previous sale trends. Managers and retailers may use up-to-date patterns to take strategic business decisions, while out-of-date patterns may be useless or even misleading for this purpose. In real-life applications of wireless sensor networks and location based-services, a large part of the collected data may be inaccurate, imprecise, or incomplete. This type of data is said to be uncertain (Aggarwal and Yu, 2009). Aggarwal et al. (2009) presented many real-life examples of uncertain databases. For example, uncertainty also exists for temperature readings and other measurements performed by sensors. For instance, consider a record $\{B, 75\%; C, 60\%\}$ representing the temperature and wind speed captured by sensors, where each item in the record is imprecise and is annotated with a probability distribution. In recent years, mining frequent itemsets by considering their existential probabilities in uncertain databases has been extensively studied. The two main proposed models are the expected support-based model (Aggarwal et al., 2009; Chui et al., 2007; Leung et al., 2008; Lin and Hong, 2012) and the probabilistic frequent model (Bernecker et al., 2009; Sun and Bai, 2008). However, most of the studies on uncertain itemset mining do not consider the relative importance of items in terms of weights. Although uncertain data is common in real applications, previous studies on WFIM aimed at extracting useful information from precise data, and only few studies have mined weighted frequent patterns in uncertain databases. Hence, it is a crucial challenge to design effective algorithms to solve the above limitations of previous work.

To the best of our knowledge, no algorithm has been proposed to address the problem of mining up-to-date weighted frequent itemsets in uncertain databases. In addition, previous techniques cannot be applied on uncertain data to reveal the valuable up-to-date patterns representing recent trends. Only one algorithm (Lin et al., 2016) was proposed to discover weighted itemsets in uncertain databases using a time-consuming level-wise approach for pattern generation. However, this algorithm may return a huge amount of useless out-of-date weighted itemsets since it does not consider the recency constraint. This study addresses this issue by proposing a novel type of patterns called Recent High Expected Weighted Itemsets (RHEWIs) to provide a smaller and more meaningful set of High Expected Weighted Itemsets (HEWIs) to the user, by considering the weight, uncertainty, and recency constraints. The major contributions of this study are summarized as follows:

- To the best of our knowledge, this is the first work to study the problem of mining recent high expected weighted itemsets in uncertain databases. The proposed Recent High Expected Weighted Itemset (RHEWI) patterns are extracted by considering the weight, uncertainty, and recency of patterns into account, to discover recent trends.
- Based on the concept of RHEWI, a time-decay strategy is defined to automatically assign recency values to transactions, which can be adjusted according to the user's preferences. Compared with previous approaches, the proposed model is more suitable for real applications such as real-time control and analysis systems. The concept of time-decay provides flexibility for specifying the user's requirements in terms of recency for interactive pattern mining in time-sensitive databases.
- The projection-based RHEWI-P algorithm is proposed to efficiently mine RHEWIs by utilizing the *upper-bound downward closure* (*UBDC*) property. It recursively processes projected sub-databases without rescanning the whole database. An improved version named RHEWI-PS is further proposed. It relies on a new *downward closure* property with sorted strategy to early prune unpromising candidates (itemsets that are not RHEWIs).
- Substantial experiments have been conducted on both real-life and synthetic datasets to evaluate the effectiveness and efficiency of RHEWI-P and RHEWI-PS in terms of number of patterns found, runtime and memory consumption for various parameter values. Results show that the proposed RHEWI patterns are acceptable and that the presented algorithms have excellent performance for discovering the complete set of RHEWIs in uncertain databases.

2. Related work

This section briefly reviews studies on frequent itemset mining in uncertain databases and weighted frequent pattern mining.

2.1. Frequent itemset mining in uncertain databases

Association rule mining (ARM) is one of the most widely used techniques to reveal useful and meaningful knowledge from various types of data. Frequent Itemset Mining (FIM) is a fundamental and challenging task in ARM, and many FIM approaches have been designed and used in numerous applications (Agrawal et al., 1993; Chen et al., 1996). In real applications, a huge amount of collected data may be inaccurate, imprecise, or incomplete. For example, data obtained from wireless sensor networks or location-based services is often uncertain (Agrawal and Srikant, 1994). To mine frequent itemsets with existential probabilities in uncertain databases, two main types of models have been proposed, which are the expected support model (Aggarwal et al., 2009; Chui et al., 2007; Leung et al., 2008; Lin and Hong, 2012) and the probabilistic frequent model (Bernecker et al., 2009; Sun and Bai, 2008).

The expected support model utilizes the expectation of support as its frequentness metric, while the probabilistic frequent model instead utilizes the frequentness probability as measure to mine probabilistic frequent itemsets. In the expected support model, a measure called expected support is used to measure the occurrence frequency of an itemset by considering the probability constraint. Chui et al. (2007) first introduced the problem of FIM in uncertain databases, and developed the UApriori algorithm by extending the Apriori algorithm. A tree-based algorithm named UFP-growth algorithm was then introduced (Leung et al., 2008), which derives frequent itemsets from a UF-tree structure. Aggarwal et al. (2009) then proposed the UH-mine algorithm to efficiently mine frequent itemsets using a hyper-structure called UH-Struct. As an extended generate-candidate-and-test approach, UH-mine provides a good trade-off between runtime and memory usage. Lin and Hong (2012) then designed an uncertain frequent pattern tree and the CUFP-growth algorithm to mine uncertain frequent itemsets. All the aforementioned algorithms including