Expert Systems With Applications 96 (2018) 271–283

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss

Ansel Y. Rodríguez-González[a,*], Fernando Lezama[a], Carlos A. Iglesias-Alvarez[b], José Fco. Martínez-Trinidad[a], Jesús A. Carrasco-Ochoa[a], Enrique Munoz de Cote[a]

[a] *Department of Computer Sciences, Institute of Astrophysics, Optics and Electronics (INAOE), Luis Enrique Erro 1, Tonantzintla, Puebla, 72840, Mexico*
[b] *Faculty of Mathematics and Computer Science, University of Havana, San Lazaro and L, Plaza de la Revolución, Havana, Cuba*

ARTICLE INFO

Article history:
Received 7 December 2016
Revised 7 December 2017
Accepted 8 December 2017
Available online 9 December 2017

Keywords:
Data mining
Frequent patterns
Mixed data
Similarity functions
Downward closure

ABSTRACT

Frequent pattern mining is considered a key task to discover useful information. Despite the quality of solutions given by frequent pattern mining algorithms, most of them face the challenge of how to reduce the number of frequent patterns without information loss. Frequent itemset mining addresses this problem by discovering a reduced set of frequent itemsets, named *closed frequent itemsets*, from which the entire frequent pattern set can be recovered. However, for *frequent similar pattern mining*, where the number of patterns is even larger than for Frequent itemset mining, this problem has not been addressed yet. In this paper, we introduce the concept of *closed frequent similar pattern mining* to discover a reduced set of frequent similar patterns without information loss. Additionally, a novel *closed frequent similar pattern mining* algorithm, named *CFSP-Miner*, is proposed. The algorithm discovers frequent patterns by traversing a tree that contains all the closed frequent similar patterns. To do this efficiently, several lemmas to prune the search space are introduced and proven. The results show that *CFSP-Miner* is more efficient than the state-of-the-art frequent similar pattern mining algorithms, except in cases where the number of frequent similar patterns and closed frequent similar patterns are almost equal. However, *CFSP-Miner* is able to find the closed similar patterns, yielding a reduced size of the discovered frequent similar pattern set without information loss. Also, *CFSP-Miner* shows good scalability while maintaining an acceptable runtime performance.

1. Introduction

Frequent pattern mining (Agrawal, Imieliński, & Swami, 1993; Agrawal, Srikant et al., 1994) is a technique that consists of finding patterns (i.e., feature sets with their corresponding values) that frequently occur (more than or equal to a minimum frequency threshold) in a dataset. It is considered a key task in data mining because of its application to discover useful information, such as risk factors (Li, Fu, & Fahey, 2009; Li et al., 2005; Nahar, Imam, Tickle, & Chen, 2013), user's profiles (Chiu, Yeh, & Lee, 2013), human behavior (Wen, Zhong, & Wang, 2015), malicious software (Fan, Ye, & Chen, 2016) among others. In addition, *Frequent pattern mining* can be used as a previous or internal step for other data mining tasks, like association rule mining (Alatas, Akin, & Karci, 2008; Kalpana & Nadarajan, 2008; Lopez, Blanco, Garcia, Cano, & Marin, 2008), classification(Hernández-León, Carrasco-Ochoa, Martínez-Trinidad, & Hernández-Palancar, 2012; Nguyen & Nguyen, 2015) and clustering (Beil, Ester, & Xu, 2002).

Since 1990, most of the *frequent pattern mining* algorithms were based on the exact matching of *boolean* features to compare and count patterns. This subclass of *frequent pattern mining* algorithms was called *frequent itemset mining* (considered as the traditional approach for frequent pattern mining). However, real life objects, such as objects in sociology (Ruiz-Shulcloper & Fuentes-Rodríguez, 1981), geology (Gómez-Herrera, Rodríguez-Morn, Valladares-Amaro et al., 1994), medicine (Ortiz-Posadas, Vega-Alvarado, & Toni, 2009) or information retrieval (Baeza-Yates, Ribeiro-Neto et al., 1999)), are rarely equal or they can be described by *non boolean* features. Thus, similarity functions different from the exact matching were proposed to compare object descriptions giving rise to a new approach named *frequent similar pattern mining* which can handle datasets containing *non boolean* features by using similarity functions (Danger, Ruiz-Shulcloper, & Llavori, 2004; Rodríguez-González, Martínez-Trinidad, Carrasco-Ochoa, & Ruiz-Shulcloper, 2008; 2011; 2013). This approach produces pat-

[*] Corresponding author.
E-mail addresses: ansel@ccc.inaoep.mx (A.Y. Rodríguez-González), f.lezama@inaoep.mx (F. Lezama), ciglsias@gmail.com (C.A. Iglesias-Alvarez), fmartine@inaoep.mx (J.Fco. Martínez-Trinidad), ariel@inaoep.mx (J.A. Carrasco-Ochoa), jemc@inaoep.mx (E.M. de Cote).

https://doi.org/10.1016/j.eswa.2017.12.018
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

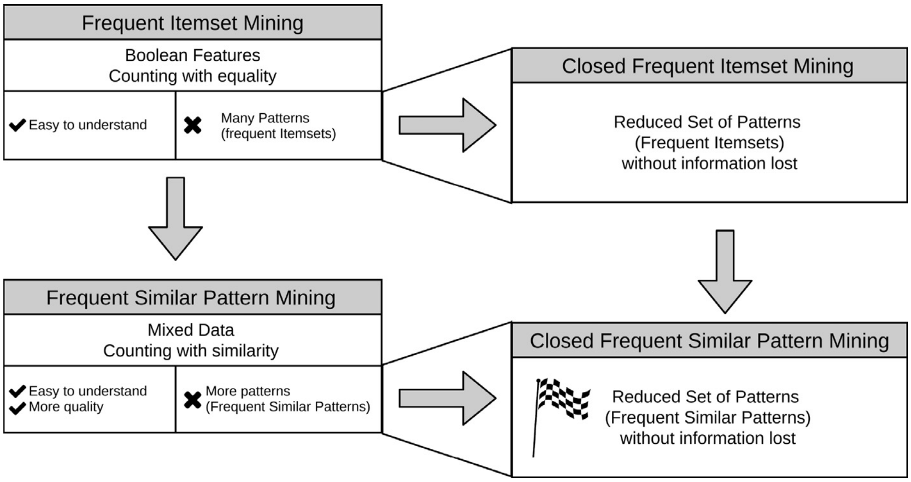
Frequent Itemset Mining
Boolean Features
Counting with equality
✓ Easy to understand
✗ Many Patterns (frequent Itemsets)
Closed Frequent Itemset Mining
Reduced Set of Patterns (Frequent Itemsets) without information lost
Frequent Similar Pattern Mining
Mixed Data
Counting with similarity
✓ Easy to understand
✓ More quality
✗ More patterns (Frequent Similar Patterns)
Closed Frequent Similar Pattern Mining
Reduced Set of Patterns (Frequent Similar Patterns) without information lost

Fig. 1. From frequent itemset mining to closed frequent similar pattern mining.

terns which can not be found by those algorithms based on exact matching. The frequent patterns found using a similarity function are named *frequent similar patterns* (Rodríguez-González, Martínez-Trinidad, Carrasco-Ochoa, & Ruiz-Shulcloper, 2013).

Despite the quality of solutions given by a *frequent itemset mining* algorithm or a *frequent similar pattern* algorithm, a critical drawback to both these approaches is that, although a complete set of frequent itemsets or frequent similar patterns can be found, the number of frequent patterns is often too big (Burdick, Calimlim, & Gehrke, 2001; Hu, Sung, Xiong, & Fu, 2008; Pei, Han, Mao et al., 2000; Rodríguez-González et al., 2013; Zaki & Hsiao, 2002).

It is helpful, therefore, to obtain a reduced set of all the frequent patterns without information loss (i.e., from which the entire frequent pattern set can be recovered). One way to do that, is through the use of *closed frequent itemsets mining* (Prabha, Shanmugapriya, & Duraiswamy, 2013). *Closed frequent itemsets mining* algorithms define that a frequent itemset is closed if it has no super-patterns with the same frequency, and use this definition to find the closed frequent itemsets. From such closed itemsets, the complete set of frequent itemsets can be generated without information loss. The so-called *closed frequent itemsets mining* algorithms also have more efficient runtimes than *frequent itemset mining* algorithms (Pei et al., 2000; Uno, Asai, Uchida, & Arimura, 2003; Zaki & Hsiao, 2002).

However, the concept of a closed patterns has not been exploited for the *frequent similar patterns* to the best of our knowledge. In this paper, we introduce the concept of a *closed frequent similar pattern* and a novel *closed frequent similar pattern mining* algorithm, named *CFSP-Miner*, that finds a reduced closed set of frequent similar patterns without information loss (see Fig. 1 to see the scope of our work). The results show that this proposed algorithm, *CFSP-Miner*, has more efficient runtimes than the state-of-the-art *frequent similar pattern mining* algorithms, except when the number of frequent similar patterns and the closed frequent similar patterns are almost equal. From the scalability point of view, the *CFSP-Miner* algorithm also finds the closed frequent similar patterns in an acceptable runtime, regardless of size.

The outline of this paper is as follows. In Section 2 related work is reviewed. Section 3 provides basic concepts. In Section 4 several concepts are introduced and redefined as a result of combining frequent similar pattern and closed pattern concepts. In Section 5 a novel algorithm for mining *closed frequent similar patterns* is proposed. Section 6 presents the experimental results and discussion, and finally, in Section 7 some conclusions and future work are discussed.

2. Related work

The frequent pattern mining problem has attracted the attention of the data mining research community because of its potential application in many different domains. One research line focuses on discovering frequent patterns in mixed data (datasets whose objects are described by numerical and non-numerical features) using similarity functions to compare and count objects (frequent similar patterns) (Danger et al., 2004; Rodríguez-González et al., 2008; 2013). Another research line focuses on obtaining a reduced set of all the frequent patterns in boolean datasets without information loss (closed frequent itemsets) (Prabha et al., 2013; Uno et al., 2003; Zaki & Hsiao, 2002).

The results of these research lines related to the current work are presented in the following subsections.

2.1. Frequent similar pattern mining

In the literature there are two algorithms for mining frequent similar patterns: *ObjectMiner* (Danger et al., 2004) and *STreeDC-Miner* (Rodríguez-González et al., 2013). Both algorithms find the whole set of frequent similar patterns by using boolean similarity functions. The discovered set of patterns usually contains patterns hidden to the traditional approach.

ObjectMiner (Danger et al., 2004) was the first algorithm for mining frequent similar pattern that used similarity functions different from equality, and it was inspired by the *Apriori* algorithm (Agrawal et al., 1994). *ObjectMiner* works by following a breadth first search strategy. Given a dataset $\mathcal{D}$, a similarity function, and a minimum frequency threshold, *ObjectMiner* finds the frequent similar patterns in $\mathcal{D}$ with only one feature. The frequency of patterns is computed by adding the occurrences of itself and the occurrences of its similar patterns. Each pattern with a frequency greater than the minimum frequency threshold is consider a frequent similar pattern. In the iteration k (starting with $k = 2$) *ObjectMiner* finds the frequent similar patterns in $\mathcal{D}$ with k features. This is done by merging the frequent similar patterns with $k - 1$ features. This process finishes after no frequent similar patterns are found.

The main weakness of *ObjectMiner*, is that the similarity between a pattern and its repetitions is computed in each iteration of the algorithm, causing an additional and unnecessary computation. *ObjectMiner* also stores the set of all similar subdescriptions (including its repetitions) of each frequent subdescription, which slows the performance as was shown in Rodríguez-González et al. (2013).