International Journal of Sociotechnology and Knowledge Development
Volume 12 • Issue 1 • January-March 2020

A Proposed Frequent Itemset Discovery Algorithm Based on Item Weights and Uncertainty

Hanaa Ibrahim Abu Zahra, Information Systems Department, Faculty of Computers and Informatics, Benha University, Benha, Egypt

Shaker El-Sappagh, Information Systems Department, Faculty of Computers and Informatics, Benha University, Benha, Egypt

Tarek Ahmef El Shishtawy, Information Systems Department, Faculty of Computers and Informatics, Benha University, Benha, Egypt

ABSTRACT

Most frequent itemset mining algorithms (FIMA) discover hidden relationships from unrelated items. They find the most frequent itemsets depending only on the frequency of the item's existence in the dataset. These algorithms give all items the same importance, and neglect the differences in importance of the items. They assume the full certainty of data, but in most cases, real word data may be uncertain. As a result, the data could be incomplete and/or imprecise. These two problems are the most common challenges that face FIMA algorithms. Some new algorithms proposed some solutions to face these two issues separately. In other words, some algorithms handle item importance only, and others handle uncertainty only. Few algorithms dealt with the two issues together. In this article, the single scan for weighted itemsets over the uncertain database (SSU-Wfim) is proposed. It depends on the single scan frequent itemsets algorithm (SS_FIM), and enhances it to deal with weighted items in an uncertain database. SSU_WFIM deals with the uncertainty of data by giving each item in a transaction an additional value to indicate occurrence likelihood. It gives the items different values to define the weight of them. It uses a table called Ptable to save the items and their probability values. This table is used to generate all possible candidates itemsets. The results indicate the high performance in aspects of runtime, memory consumption and scalability of SSU-Wfim comparing with the UApriori algorithm. The proposed algorithm saves time and memory with a percentage exceeds 70% for all tested datasets.

KEYWORDS

FIMA, Frequent Itemsets, Importance of Items, Uncertain Database, Weighted Itemsets

1. INTRODUCTION

Frequent itemsets mining (FIM) is a very important branch in data mining as it discovers the hidden related frequent itemsets in datasets. But it will be very necessary and precise that if the frequent itemsets are discovered based on their importance to users instead of the frequency of itemsets existence only. This is very useful for the recommendation system application based on market datasets as each item in market has different profit, and the application based on event logs to determine the related pages based on dwelling time of the users instead of the number of visiting these pages. We may use the real-world application data collected from different sensors. This data may face any data distortion

DOI: 10.4018/IJSKD.2020010106

problems such as noise problem. These problems may make the data incomplete or imprecise. So, it is necessary to find a good technique to treat with the uncertainty of data to get a precise mining result.

When the importance and the uncertainty of data are taken into account in FIMA, the result of finding the most frequent itemsets for all real-world application such as healthcare applications, recommendation system applications will be determined precisely. Some of the previous algorithms of FIM do not take into their considerations two critical issues, which are: (1) the importance of each item and, (2) the uncertainty of used dataset.

Regarding the first issue, FIM algorithms treat all items without taking into account the importance of different items. All items have the same importance or weight value. In some FIM algorithms, the number of existence frequency in the database is the only measure for determining the frequent itemsets, but this is not sufficient as some itemsets will be ignored because of infrequently presence although they have more importance than the itemsets which are more frequent. There is a solution to deal with this limitation by defining weights for items according to some criteria such as user preference, profits, interestingness, or dwelling time on websites. There are some proposed algorithms to mine the weighted frequent itemsets (Cai et al., 1998; Yun & Leggett, 2005; Sun& Bai, 2008; Lan et al., 2013; Lan et al., 2014; Lin et al., 2015; Zhao et al., 2018; Chee et al., 2018). However, these algorithms don’t consider the data uncertainty.

The second limitation of FIM algorithms is the neglecting of data uncertainty. The real-world data, such as data collected from practical sensor applications, may be inaccurate or imprecise. Some algorithms are proposed to find the frequent itemset from uncertain databases, which are classified into two categories of probabilistic frequent itemset mining (Bernecker et al., 2009) and expected support model (Chui et al., 2007; Sun & Bai, 2008; Aggarwal et al., 2009; Shah & Halim, 2018; Braun et al., 2018).

Recently, some algorithms tried to find out weighted frequent itemset from an uncertain database (Lin et al., 2016, Lin et al., 2016). For example, Lin et al. (2016) proposed high expected weighted itemsets-Uapriori (HEWI- UApriori) algorithm based on the UApriori algorithm. It is applied in two-phase to find the most weighted frequent itemsets. Lin et al. (2016) proposed a high expected weighted itemsets-Utree(HEWI-Utree) algorithm to decrease the multiple database scan without generating enormous candidates. The first algorithm is suffering from the multiple database scan which leads to increase the execution time and the second algorithm may consume more memory. So, we need to use algorithm which solves this problem of multiple scan of database high memory cost. Handling the two issues is very necessary for treating with the IOT applications and big data applications (Kinnunen et al., 2018; Galli, 2018; Krimpmann & Stithmeter, 2017; Mizuno & Odake, 2017)

Some of the mentioned algorithms (Chui et al., 2007; Sun& Bai, 2008; Aggarwal et al., 2009) read the database multiple time to find the most frequent itemsets. They have one of the most famous mining problem which is the multi scan database. This problem leads to execute much time to mine the database. One of the algorithms that proposed the smart solution to the multi scan problem is single scan frequent itemsets mining algorithm SS_FIM (Djenouri et al., 2017). This algorithm reads transactions contained in the database just once. This solution saves a lot of time. But this algorithm has the two limitations. It can’t work with uncertain database that contains inaccurate data. All items have an equal importance value. So the advantages of SS_FIM algorithm motivated us to extend it to face these two issues. As a result, it is necessary to represent a new technique to make this algorithm work with uncertain database to benefit from its abilities decreasing the run time of the mining process.

In this paper, the single scan weighted frequent itemsets over uncertain database (SSU-Wfim) algorithm is proposed to address the two issues of weighted itemsets and uncertainty data for enhancing SS-FIM algorithm. The proposed algorithm reads the transaction at once and constructs the temporary Ptable for this transaction to save the transaction items and its probabilities. It uses Ptable to generate all possible itemsets of this transaction, then calculates the probability of itemsets and their evaluated support. The itemsets and their evaluated support are saved in a hash table. After reading all transaction, it reads each itemset to calculate the itemset weight. It calculates the itemset