Frequent Itemset Mining Under Differential Privacy with Multiple Minimum Supports

1st Yum-Luh Wang
Graduate Institute of Electronics Engineering
National Taiwan University
Taipei, Taiwan
r07943161@ntu.edu.tw

2nd Yao-Tung Tsou
Department of Communications Engineering
Feng Chia University
Taichung, Taiwan
yttsou@fcu.edu.tw

3rd Sy-Yen Kuo
Graduate Institute of Electronics Engineering
National Taiwan University
Taipei, Taiwan
sykuo@ntu.edu.tw

***Abstract*—Frequent itemset mining is an extensively studied research domain of data mining. The aim is to find interesting correlations between items in a transactional database. However, malicious user might gain sensitive information in the mining process. Differential privacy is the de facto standard when it comes to protecting data. Combining differential privacy with frequent itemset mining can provide strong privacy guarantee while generating statistical information from sensitive data. In existing studies, most algorithms are focusing on finding frequent patterns using one predefined threshold with the protect of differential privacy. However, using single threshold to extract itemsets creates "rare item problem", and setting respective support thresholds for each item is more adequate to reflect the nature of widely varied items. In this paper, we propose a novel FP-growth-like solution DPCFP++ to solve the rare item problem of frequent itemset mining while guarantee differential privacy at the same time. The experiments illustrate that our algorithm acheieves high utility in privately solving rare item problem.**

***Index Terms*—Differential Privacy, data mining, multiple minimum supports, frequent itemset mining**

I. Introduction

In modern era, the development of information technology grows rapidly, more and more services aggregate data from individuals. With the evolution of machine learning and statistics, organizations utilize these enormous amount of data to improve the accuracy of their model, understand the past and predict the future. Frequent itemset mining is a popular research topic in data mining, and it was first introduced in 1993 by Agrawal et al. [1] The original purpose of frequent itemset mining was to analyze customer habits in a supermarket. They tried to find associations between different items that customers frequently bought together. According to the mined results, the management team of the supermarket can make business decisions to maximize profit including product rearrangement, bundle selling, etc. After decades of research, frequent itemset mining can also be extended to many other area besides market basket analysis. For example, text mining, disease diagnosis, and so on.

Majority of existing literatures endeavors on single threshold mining, that is, treat all the items equally and apply a single threshold to decide whether an itemset is frequent or not. However, setting one appropriate threshold for all items is a hard question. For example, consider two items: soda and television in a wholesale store. Soda is frequently bought while television is an uncommon purchase. If the threshold is set too high, we might consider TV as a less important item, but TV contributes thousand times more to the store revenue than the common purchased soda. On the other hand, if the threshold is too low, the computational cost becomes expensive, the mining result will cause combinational explosion and we will receive many meaningless itemsets. The above dilemma is the famous "rare item problem". In real life applications, some items tend to have more weights in natural than other items. Researchers [2] [3] [4] [5] have worked on this problem by allowing users to use "multiple minimum supports". Briefly speaking, rare itemset mining is a more advanced setting of frequent itemset mining, and it allows user to apply different threshold on each item .

The convenience which data mining brings comes with the price of privacy leakage. The preciser the model is, the more behavior. How to de-identification is a mainstream topic in related research. A straightforward approach is to Differential Privacy, first proposed by Dwork in [4], had come to rescue in 2006. Since then, many research based on differential privacy have been conducted. Tech giants such as FAANG(facebook, applem, amazon, netflix, google) have added features with different kinds of differential privacy model. It guarantees that a user In this paper, we study the problem of how to perform frequent itemset mining (FIM) on transaction databases while satisfying differential privacy.

II. Related Work

A. Frequent Itemset Mining

There are mainly two types of frequent itemset mining algorithms. One is Apriori, and the other is FP-growth.

B. Frequent Itemset Mining with Multiple Minimum Supports

C. Differentially Private Frequent Itemset Mining

Frequent itemset mining, also known as frequent pattern mining, is

III. Preliminaries and Problem Statement

In this section, we introduce and provide formal definitions of differential privacy and the rare item problem of frequent itemset mining. Finally, we give the problem statement in the last subsection.

A. Differential Privacy

Differential privacy is proposed by Dwork [6],
we have

B. Frequent Itemset Mining

Given a threshold λ and a transactional database $D = \{t_1, t_2, t_3, \ldots, t_n\}$ consisting of n transactions where each transaction consists of sets of distinct items in the item universe $I = \{i_1, i_2, i_3, \ldots, i_m\}$, i.e. $t_j \subseteq I$, Frequent itemset mining, also known as frequent pattern mining, refers to discovering all set of patterns while the support σ of which is greater than or equal to a user-specified threshold.

Definition 1. *(Support and Threshold)*
We denote support and threshold as σ and λ, respectively. The number of transactions in the database containing an itemset is known as the support of that itemset.

There are two ways to specify σ and λ. One is the percentage way, and the other is by absolute number. It is simple to convert between them, just multiple(divide) the number of total transactions n. For consistency, we choose the former one in this paper.

Definition 2. *(Frequent Itemset)*
An itemset X is frequent if and only if its support $\sigma(X)$ is greater than or equal to support threshold λ. This means that the presence of itemset X in the database is statistically significant.

Definition 3. *(k-Itemset)*
An itemset which contains k distinct items is called k-itemset.

C. Rare Item Problem

By the nature of items, some items appear more frequently than others. However, this sometimes causes meaningless itemsets. We have to find frequent itemsets without generating too many meaningless itemsets. To deal with rare item problem, Liu [2] first tackled the problem with minimum item supports (MIS). MIS allows user to specify respective threshold for every item.

Definition 4. *(Multiple Item Support)*
For an item j with support $\sigma(j)$, the Minimum Item Support(MIS) of j is

$$MIS(j) = max\{\rho * \sigma(j), \lambda\}, \tag{1}$$

where $\rho \in [0, 1]$ is a relevance parameter that controls how the MIS values for items should be related to their frequencies. Noted that when ρ=0, the MIS value equals λ, which is the same as the traditional setting of frequent itemset mining.

D. Problem Statement

Here we give the formal problem statement of

IV. Proposed Method

A. A Straight Forward Approach

B. Truncate Database

C. Assign MIS

V. Experiment

A. Metrics

We use F-score, a the widely used F-measure which is the harmonic mean of precision and recall.

B. Experimental Results

C.

D. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{2}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(2)", not "Eq. (2)" or "equation (2)", except at the beginning of a sentence: "Equation (2) is . . ."

E. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

F. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

G. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Fig. 1. Example of a figure caption.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami. "Mining association rules between sets of items in large databases," in Proceedings of ACM-SIGMOD, 1993.

[2] B. Liu, W. Hsu, and Y. Ma. "Mining association rules with multiple minimum supports," in Proceedings of ACM-SIGKDD, 1999.

[3] Y. Hu, Y. Chen. "Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism," in Decision Support System, 2006.

[4] R. Uday Kiran, P. Krishna Reddy. "Novel techniques to reduce search space in multiple minimum supports-based frequent pattern mining algorithms," in Proceedings of ACM-EDBT, 2011

[5] W. Gan, J. C. Lin, P. Fournier-Viger, H. Chao, J. Zhan. "Mining of frequent patterns with multiple minimum supports," in Engineering Applications of Artificial Intelligence, 2017

[6] C. Dwork. Differential privacy. In ICALP, 2006.