A Frequent Itemsets Data Mining Algorithm Based on Differential Privacy

Qingpeng Li, Longjun Zhang, Haoyu Li and Wenjun Sun
Department of Information Engineering, Engineering College of the Chinese Armed Police Force, Xi'an, Shaanxi, China

***Abstract*—Differential privacy is a new privacy protection technology, which defines a strict and strong privacy protection model, by adding noise data distortion to achieve the purpose of privacy protection. Frequent pattern mining is an important field in data mining, and its purpose is to find frequent patterns in data set, but the content of the model itself, rules, and counting information is likely to lead to leaking sensitive information. This paper presents a frequent item sets mining method based on differential privacy, named DPFM, which adopts the mining strategy combined with Laplace system and index system, realizing the difference privacy under the premise of guaranteeing performance calculation of privacy protection. Experiments demonstrate that the proposed algorithm, DPFM has an advantage in decreasing error rate, and the convergence rate under two indexes is better than TF method.**

Keywords-differential privacy; data mining; frequent itemsets; privacy protection

I. INTRODUCTION

With the rapid development of information science and technology, how to find useful knowledge from massive information has become an urgent problem. Data mining is capable of extracting valuable models or rules from a large amount of data, which can eventually guide management decisions, in scientific research and production practice.

Frequent Pattern Mining is an important subject in the study of data Mining, whose purpose is to find frequent patterns (e.g., itemsets, a sequence or substructure), association rules, classification, clustering, correlation analysis and the basis of other data mining data set [1]. Also, it is one of the main techniques in data analysis. As the simplest type of FPM, frequent itemsets mining was originally applied in the discovery of association rules in a database and it is the basis of pattern mining. Apriori and FP - growth algorithm [2] are the two classical algorithms of frequent itemsets.

Apriori algorithm is one of the most influential Boolean association rules mining algorithms. Most importantly, domestic and foreign scholars have done a lot of fruitful researches. Among them ,literature [3] proposed a packet statistical strategies, which can effectively reduce the production of key/value pairs, Literature [4] puts forward a kind of parallel association rules based on matrix Apriori_MMR, which combines the ideas of the data partitioning in parallel, simplifying the connection of generating candidate item steps. At the same time, it can also compress the transaction in the calculation process so as to further improve the performance.

Despite that frequent pattern mining technology has become a kind of effective method in knowledge discovery, but the content of the model, rules, and counting information is likely to lead to leaking sensitive information. This paper presents a frequent itemsets mining method based on differential privacy, named DPFM, in which the mining strategy combined with Laplace system and index system, realizing the difference privacy under the premise of guaranteeing performance calculation of privacy protection. By finding all support which is greater than the given threshold theta in maximum frequent itemsets, the transaction data set is effectively compressed based on mining the top - k frequent itemsets to find the corresponding frequency mode.

II. PRELIMINARIES

Differential privacy protection technology is recognized as one of the strictest and strongest privacy protection models. In essence, it is a kind of disturbance method, adding noise to protect sensitive data from leaking information security technology. Based on original data conversion, a certain amount of noise is added to the query or analysis result, ensuring the data on the basis of the original statistical properties , the record association, structure, corresponding relationship between extra sensitive information unchanged so as to realize the goal of protecting privacy.

Definition 1. ε-differential privacy[5]. For the given two adjacent data sets D and D', there is up to one record of difference between the data sets. Given a privacy algorithm A with the output domain R , for any subset $S \subseteq R$, if A meets

$$\Pr\left[AD \in S\right] \le \exp\exp\varepsilon \times \Pr\Pr\left[AD' \in S\right] \qquad (1)$$

then we say A provides ε-differential privacy protection, $\Pr\left[X\right]$ is the probability that event X, that is, the risk of privacy leakage, which is determined by the random attribute of algorithm A. Parameter ε is the privacy protection budget, and the smaller ε is, the closer the probability of the same output of the two adjacent data sets, and the higher degree of privacy protection we obtain is.

Definition 2. Global sensitivity. Given function $f : D \rightarrow R^d$, input data set D and output d-dimensional real number vector, then for any adjacent data sets D and D',the global sensitivity of function is

$$\mathrm{GS_f} = \max_{\mathrm{D},\ \mathrm{D'}} \mathrm{Pf}\left(\mathrm{D}\right) - \mathrm{f}\left(\mathrm{D'}\right)\mathrm{P_1} \qquad (2)$$

$\| f(D) - f(D') \|_1$ is the *1*-dimensional bound norm distance. The lower the sensitivity is, the smaller the impact of differential privacy protection to the practicability of data and the higher the accuracy of the output.

Noise perturbation is linked to the sensitivity and privacy protection budget of ε. This paper uses Laplace mechanism that can keep the original data statistical property, and its principles are as follows:

Theorem 1. For F: D → Rd, mechanism K adds independently generated noise with distribution Lap (Δf/ε) to each of the d output terms which enjoys ε-differential privacy. [6]

In addition, differential privacy preserving has sequence composition and parallel combination [7]. Built on the properties, this paper proves that our improved algorithm can satisfy differential privacy and guide the allocation process of the privacy budget.

III. DPFM ALGORITHM

A. DPFM Algorithm Design

The biggest challenge of transactional database processing and mining process is the high dimension of transaction data, namely the attribute of long transaction. The core idea of DPFM algorithm is to use an important attribute of frequent itemsets, namely Aprioriproperty, to compress the dimension of transaction data set. That is, the subsets in one frequent itemset are also frequent. Further, to find the maximal frequent itemsets from transaction data sets that meets the threshold θ is to build θ-base set to mine the top - k frequent itemsets.

Therefore, on the premise of ensuring the differential privacy, how to map the original data for quick solution and how to ensure the validity and accuracy of data during the process of mining frequent itemsets are two major problems to be solved in DPFM algorithm.

DPFM algorithm proposed in this paper mainly includes the following steps:

Step1. Obtain the value of λ , which is the number of the items whose support meets threshold θ

Step2. Build a node set F , which contains the most frequent itemsets I . That is, all frequent items whose support meets the threshold θ. Set F will contain all the frequent items appearing in the top - k itemsets.

Step3. Contribute edge set P based on set F . Set P consists of all the subsets that have the length of 2and meet the threshold *θ*. That is, set P contains all the frequent pairs(frequent bi-itemsets) appearing in top-k itemsets.

Step4. Find the maximum mass based on the spanning graph $G(F,P)$ of sets F and P to contribute θ-based set B . Each maximum mass has a corresponding θ-base and finally finds a θ-base set B whose length and width are as small as possible.

Step5. Contribute candidate set $C(B)$ based on set B and calculate the support of itemset $C(B)$ and dispose of support with differential privacy to finally get the relevant information of the top-k frequent itemsets that satisfies privacy constraints. The flowchart of DPFM is shown as following.

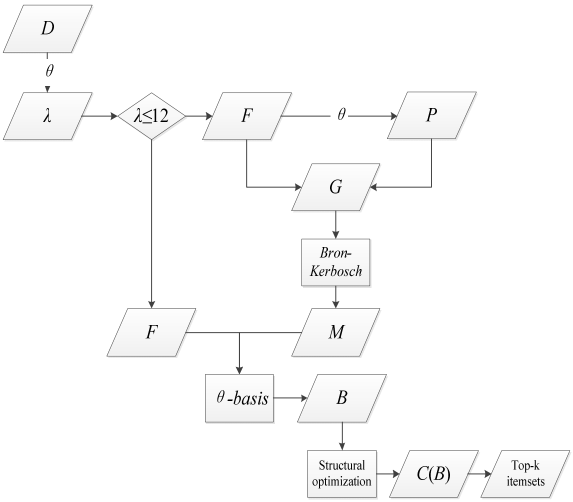

FIGURE I. FOLLOW CHART OF DPFM

B. Privacy Analysis

Step 1 is implemented through the GETLAMBDA function in DFPM. Using index mechanism $\{1.2\ldots, k\}$ form selection, through the following rate (utility) function

$$q(D, i) = \left(1 - \left|f_k - fitem_j\right|\right)N \qquad (3)$$

Among them, $fitem_j$ is the *j*th most frequent support items. Obviously, the scoring function sensitivity is 1 to f_k and at most $\frac{1}{N}$ for $fitem_j$ by adding or subtracting an item .

In Step2, DFPM will sample from all the items, so the size of the candidate set is $|I|$ and we can get set F . In step 3, we contribute candidate set U from set F . And with the same method, we contribute set from set U . Therefore, as there are only $\binom{\lambda}{2}$ elements in candidate set U , the size of candidate set should be controlled under proper size to ensure the practicability.

Above all, DPFM realizes privacy protection based on the exponential mechanism and the design of scoring function in this paper. Under the premise of ensuring safety and privacy, it also optimizes each step of the DFPM to ensure the practicability of calculation and accuracy. All in all, all steps are based on there alization of ensuring the implementation of differential privacy. Therefore, DPFM satisfies the differential privacy.

IV. EXPERIMENT AND ANALYSIS

In this paper, the experiment configurations is as following: AMD Athlon II X4 645 Processor 3.1 GHz Processor, 4 GB of memory. Software of Matlab with a series of Matlab toolbox is implemented on Windows 7 OS to execute related algorithms..

We should note that DPFM proposed in this paper has different treatment strategies under the condition of different λ values. In order to enhance persuasion, DPFM and traditional classic method of TF which proposed in the literature [8] are compared in three kinds of representative data, Through the mining results from FNR, and each group of frequent itemsets RE. It can be verified that the method in this paper is better than TF method which is shown in the table 1.

TABLE I. EXPERIMENTS USING REAL DATA SETS IN DETERMINING K VALUES OF RELATED FEATURES

The data set	N	$avg\lvert t\rvert$	λ	k
retail	8124	8.1	11	100
kosarak	90002	24	39	200
AOL	647377	46	171	300

According to the design of DPFM algorithm, a single θ-baseis constructed directly. The experiment compares the performance of DPFM with TF method. The results are shown as follows in Figure 2 and Figure 3.

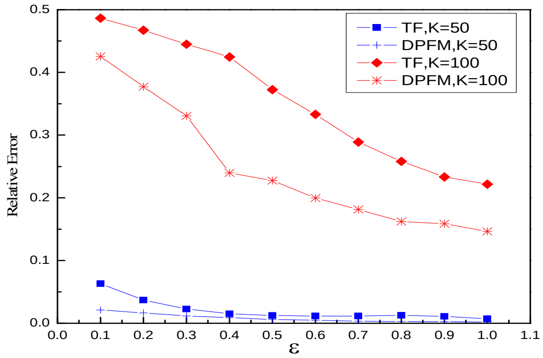

FIGURE II. RETAIL DATA OF DIGGING PERFORMANCE CONTRAST IN RE

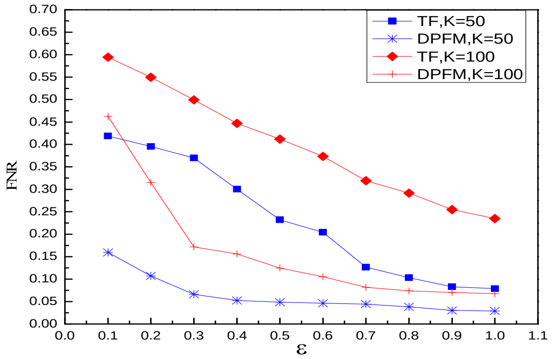

FIGURE III. RETAIL DATA OF DIGGING PERFORMANCE CONTRAST IN FNR

It can be seen that with the increasing of budget privacy ε, results of FNR and PE show a trend of declining, and budget privacy ε taken more than 0.6 will be gradually stabilized. Due to the narrow scope of the frequent itemsets mining, two algorithms on the error performance are relatively good. Taken together, the algorithm providing the results of the accuracy is higher, but the DPFM proposed in this paper, is better than TF method in RE, FNR with better convergence rate.

V. CONCLUSIONS

In view of the long transaction data mining efficiency and the relatively low accuracy, this paper proposes a differential privacy DPFM frequent itemsets mining algorithms, combined with great mass of theoretical base and mapping technology. According to the threshold, plenty of transaction data will be set. Compression superfluous and effective information are kept in mining business set of frequent closed itemsets to build the candidate set. And combined with Laplace mechanism, frequent items support privacy information noise disturbance by implementing the ε-differential privacy, ultimately satisfying candidate set refractory privacy security policy top - k support frequent itemsets. Effective to control the scale of candidate set, and reduce the amount of noise added and consume the privacy of our budget, DFPM can improve the mining frequent itemsets top - k performance and accuracy on the premise of guarantee the data privacy.

REFERENCES

[1] L. Ding，G. Lu,Survey of differential privacy in frequent pattern mining [J],Journal on Communication,2014,35(10):200-209

[2] Inokuchi A, Washio T, Motoda H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data[C]. European Conference on Principles of Data Mining&Knowledge Discovery,2000:13--23.

[3] L. Huang,Y. Liu, MapReduce based parallel Apriori algorithm improvement research [J]. Journal of Fuzhou University (NATURAL SCIENCE EDITION), 2011, 39 (5): 34-39.

[4] Z. Xie,P. Wang, a parallel matrix Apriori algorithm based on Reduce Map architecture [J], computer application research, 34 (1): 17-21.

[5] C. Dwork, C. Dwork. The Differential Privacy Frontier[J], 2009:496--502.

[6] Z. Xie, P. Wang, a parallel matrix Apriori algorithm based on Reduce Map architecture [J], computer application research, 34 (1): 17-21.

[7] B. Wang, D. Wang, S. Zhang. Short term distributed power load forecasting algorithm based on Spark and [J]. IPPSO_LSSVM electric power automation equipment, 2016,36 (1): 117-122.

[8] R. Bhaskar,S. Laxman,A Smith,et al,Discovering frequent patterns in sensitive data[C],The 16th ACM SIGKDD international conference on knowledge discovery and data mining ,2010:503-512.