International Journal of Database Theory and Application
Vol.9, No.7 (2016), pp.195-206
http://dx.doi.org/10.14257/ijdta.2016.9.7.17

Candidate Pruning-Based Differentially Private Frequent Itemsets Mining

Yangyang Xu[1], Zhaobin Liu[2*], Zhonglian Hu[3] and Zhiyang Li[4]

School of Information Science and Technology, Dalian Maritime University
No. 1, Linghai Road, Dalian, P.R.China, 116026
[1,2*,3,4]*{yyx.dlmu, zhbliu, huzhongnian, lizy0205}@gmail.com*

Abstract

Frequent Itemsets Mining(FIM) is a typical data mining task and has gained much attention. Due to the consideration of individual privacy, various studies have been focusing on privacy-preserving FIM problems. Differential privacy has emerged as a promising scheme for protecting individual privacy in data mining against adversaries with arbitrary background knowledge. In this paper, we present an approach to exploring frequent itemsets under rigorous differential privacy model, a recently introduced definition which provides rigorous privacy guarantees in the presence of arbitrary external information. The main idea of differentially privacy FIM is perturbing the support of item which can hide changes caused by absence of any single item. The key observation is that pruning the number of unpromising candidate items can effectively reduce noise added in differential privacy mechanism, which can bring about a better tradeoff between utility and privacy of the result. In order to effectively remove the unpromising items from each candidate set, we use a progressive sampling method to get a super set of frequent items, which is usually much smaller than the original item database. Then the sampled set will be used to shrink candidate set. Extensive experiments on real data sets illustrate that our algorithm can greatly reduce the noise scale injected and output frequent itemsets with high accuracy while satisfying differential privacy.

***Keywords**: Differential Privacy, Frequent Itemsets Mining, Privacy Protection*

1. Introduction

Frequent Itemsets Mining (FIM) is a well-known problem in data mining. Since Agrawal Imielinski and Swami [1] introduced the concept of association rule mining in 1993, FIM tasks have gained a lot of attention [2]. Given an item data set, the goal of FIM is to discover itemsets that occur frequently in the data set. Those frequent itemsets can be used in many fields, such as predicting user behaviors [3] and discovering correlations [4-5]. For example, by the help of frequent itemsets mining, we may be amazed to find the fact that beer and diapers are often purchased together. However, due to the fact that some itemsets deal with sensitive data such as medical records [6, 7], revealing those frequent itemsets may pose a threat to individual privacy. No one can deny that discovering correlation in these data will bring much useful information. However, some sensitive data must be well preserved in this process.

Differential privacy [8-9] is a novel strategy used to protect individual privacy in data release. It gives an assurance that no matter whether any single item is in data set or not, the released data set is virtually indistinguishable. Compared with other privacy-preserving methods, differential privacy does not assume background knowledge of attackers. It preserves individual privacy by adding appropriate noise to query or analysis

[*] Corresponding author: zhbliu@gmail.com

ISSN: 2005-4270 IJDTA

results. And under the most extreme circumstances, the adversary cannot reason value of a specific record even though he knows all remaining records. Besides, differential privacy is based on rigorous statistical model, which greatly facilitates the quantitative analysis and proof of it.

In this paper, we investigate how to mine frequent itemsets while satisfying differential privacy. Previous works on this field mostly focus on injecting noise into the supports of itemsets. We found that the noise scale is proportional to the size of candidate frequent set. In order to reduce the noise scale, which will improve the utility of result, we propose an approach to squeezing the candidate set by removing unpromising itemsets from it. Our work is partly inspired by [10]. We used the method proposed in [10] to generate a super set of frequent itemsets and use it to decide whether an itemset should be removed from the candidate set.

Our contribution is that we propose a novel 2-stage mechanism to mine frequent itemsets under differential privacy model. We use a progressive sampling method to generate a super set of frequent itemsets to shrink the size of candidate set, after which the noise scale injected will be greatly lowered. Our algorithm would increase the utility of result while satisfying differential privacy.

The rest of this paper is organized as follows: Section 2 briefly discusses the related works. Section 3 presents preliminary information about differential privacy and frequent itemsets mining. In Section 4, we illustrate details of the proposed mechanism and give its privacy analysis. In Section 5, the performance of our algorithm is evaluated on a variety of data sets. Finally, Section 6 concludes our works.

2. Related Works

In this section we discuss related works on frequent itemsets mining under differential privacy briefly.

Bhaskar *et al*. [11] propose two differentially private FIM algorithms using Laplace mechanism [8] and exponential mechanism [12] respectively. Li *et al*. [13] introduce the PrivBasis algorithm to solve the problem of high dimensionality of transaction database by projecting the input high dimensional database onto lower dimensions. Algorithm proposed in [14] truncate transactions in the original database in order to improve the utility of result.

Different from previous works, we focus on improving the utility of result by diminishing the noise scale required in differential privacy. As mentioned in [15], the utility of result will be greatly improved by squeezing candidate set in frequent sequence mining. From this standpoint, we seek for an effective way to shrink scale of candidate set while making the result still satisfy differential privacy.

3. Preliminaries

In this section, we introduce the background knowledge and notations used in our paper. Table 1 summarizes the notations used in this paper.

3.1. Differential Privacy

Differential privacy [8-9] is a state-of-art model for protecting privacy. Compared with other privacy-preserving models, differential privacy can provide stronger privacy guarantee. The main idea of differential privacy is it can ensure that the output is insensitive to any particular tuple in the data set.

Table 1. Notations

Symbol	Description
$D=\{t_1,t_2,...,t_n\}$	Database of transactions
$\|D\|$	Number of transactions in database D
I	A set of items contained in D
X	The itemset, and it is a subset of I
$T_D(X)$	Subset of transactions in D that contain the itemset X
$S_D(X)$	Frequency of transaction containing X in D
ε	Privacy budget in differential privacy
$Lap(\lambda)$	Laplace distribution with mean 0 and scale factor λ

Definition 1 (NEIGHBORING DATA SETS): Let D_1, D_2 be two neighboring data sets, if and only if D_1 and D_2 differ in only one tuple (by adding or removing a tuple), *i.e.* $|D_1-D_2|\cup|D_2-D_1|=1$. And we denote this as $D_1=nbds(D_2)$ or $D_2=nbds(D_1)$.

Definition 2 (ε-DIFFERENTIAL PRIVACY): A randomized algorithm ξ satisfies ε-differential privacy if for all neighboring data sets $D_1,D_2\subseteq\chi$, and any set of possible subset of output $O\subseteq Range(\xi)$, the following holds:

$$\Pr[\xi(D_1)\in O]\le e^{\varepsilon}\Pr[\xi(D_2)\in O] \quad (1)$$

where χ is the input data set of ξ.

The parameter ε describes the privacy level and it is usually referred to as privacy budget. The value of ε should not be too large, for it is inversely proportional to the utility of result. Therefore, its value is commonly set to 0.1, ln2, ln3, but no larger than 1. We should consider both the privacy and utility of result when deciding the value of ε.

One effective method to achieve differential privacy is injecting appropriate noise into the output result. As we all know, large noise will do harm to the utility of result. So, we need a metric to decide the scale of noise needed to provide privacy protection. In differential privacy, sensitivity is an important metric of injected noise. It depicts the maximal change of a query result by removing any single record in the input data set.

Definition 3 (GLOBAL SENSITIVITY): Given a function $f:D\rightarrow R^d$ whose input is a data set D and output is a d-dimensional vector, for any neighboring data sets D_1 and D_2 in D, the global sensitivity of f, denoted as Δf is:

$$\Delta f=\max_{D_1,D_2}\left\|f(D_1)-f(D_2)\right\| \quad (2)$$

Among which $\left\|f(D_1)-f(D_2)\right\|$ is the L_1 distance between $f(D_1)$ and $f(D_2)$.

Laplace mechanism and Exponential mechanism are two fundamental mechanisms satisfying differential privacy. Laplace and Exponential mechanism are designed for numerical and non-numerical value respectively. In this paper, as we seek for a method to privately perturb the support of frequent itemsets, we will focus on the use to Laplace mechanism.

Laplace mechanism provides ε-differential privacy by injecting noise obeying Laplace distribution into the output result. We use $Lap(\lambda)$ to denote the Laplace probability distribution with mean 0 and scale λ. The probability density function of Laplace distribution is:

$$p(x) = \frac{1}{2\lambda}\exp(-\frac{|x|}{\lambda}) \tag{3}$$

Definition 4 (LAPLACE MECHANISM): Given a data set D, privacy budget ε, and a function f, the Laplace mechanism $L(D)$ is defined as:

$$L(D) = f(D) + Lap(\Delta / \varepsilon) \tag{4}$$

3.2. Frequent Itemsets Mining

FIM is a well-known data mining task. Given a database $D = \{t_1, t_2, ..., t_n\}$, where t_i representing transaction is a set of items, the goal of FIM is to discover all transactions in database which frequently appear.

For an itemset X, we denote the frequency of X as $S_D(X) =| T_D(X) | / | D |$.

Definition 5 (FREQUENT ITEMSET): For any itemset X, if $S_D(X) \geq \lambda$ where λ is a predefined threshold, we say X is a frequent itemset with respect to the threshold λ.

Definition 6 (SUPPORT): In data mining, $| T_D(X) |$ is usually called support of X. And $S_D(X)$ is called relative support of X.

4. Candidate Pruning-based FIM

4.1. A Straight Forward Approach

Although [15] deals with frequent sequences mining, which is different from our topic here, we find the idea proposed in it quite suitable for our research in this paper. A straightforward approach to finding frequent itemsets under differential privacy is adding noise directly to the support of all possible itemsets in candidate set. Then the noisy support will be used to determine whether itemsets in candidate set is frequent or not. According to [15, 16], we only need to output those itemsets whose noisy support exceeds the threshold λ.

Despite the fact that this method can achieve differential privacy, the output of it has poor utility which means that it is nearly useless. The root cause of this problem is the large number of itemsets in candidate set, which forces the differential privacy mechanism to inject too much noise to the support of itemsets. From this standpoint, we seek for an effective method to squeeze the size of candidate so as to improve the utility of result.

4.2. Progressive Sampling

Before introducing our algorithm, we need to introduce a sampling method proposed in [16]. Our method will use the generated super set of all frequent itemsets to remove unpromising frequent itemsets in candidate in order to diminish the noise scale injected.

Definition 5 ((γ, δ) -APPROXIMATION) [16]: For $(\gamma, \delta) \in (0,1)$, a (γ, δ)-approximation of $FI(D, I, \lambda)$ is a collection $\zeta = P\{(X, S_X) : X \in I, S_X \in (0,1)\}$ such that, with probability at least $1-\delta$:

(i) for any $(X, S_X) \in FI(D, I, \lambda)$, there is a pair $(X, S_X) \in \zeta$; and

(ii) for any $(X, S_X) \in \zeta$, it holds $S_D(X) \geq \lambda - \gamma$; and

(iii) for any $(X, S_X) \in \zeta$, it holds $|S_D(X)| - S_X \leq \gamma / 2$

Using rademacher averages, [16] proposed a progressive sampling method to get a (γ, δ)-approximation to $FI(D, I, \lambda)$ from random samples of D. In this paper, we are going to use this method to get a sampled super set of frequent itemsets. This sampled set will be used to shrink the candidate frequent set in our algorithm. The details of this method are beyond the scope of this manuscript. We recommend interested readers to read [16] for particulars of it.

4.3. Candidate Pruning-Based FIM

Algorithm 1 describes the steps in our algorithm.

Note that in the beginning of the algorithm, we firstly use progressive sampling method proposed in [16] to get a (γ, δ)-approximation of the database, and use it in the following steps. The sampled database definitely has smaller size than the original one. In the following steps, we use the sampled super set of frequent itemsets to remove unpromising itemsets from the candidate set in each iteration. The support of itemsets remained in the candidate set will be injected with a noise which is smaller than other methods do. In the end, the algorithm will determine which itemset is frequent according to its noisy support and output the frequent itemsets.

Compared with the straightforward method, the noise scale injected will be reduced as the noise is proportional to the size of candidate set. So we can say this method can improve the utility of result.

4.4. Privacy Analysis

Theorem 1: Our algorithm satisfies ε-differential privacy.

Proof: Differential privacy has an important composition property [17]. Specifically, given k algorithms $A_1, A_2, ..., A_k$, suppose the domain is divided into k arbitrary disjoint subsets and each of them corresponds to an algorithm A_i. The parallel execution of $A_1, A_2, ..., A_k$ satisfies ($\max\{\varepsilon_i\}$)-differential privacy.

When mining frequent itemsets, in each iteration, adding or removing an input transaction in D can, in the worst case, affect the support of all candidate sequences by one, so the sensitivity of the algorithm is the number of candidate itemsets. Since the noise injected is $Lap(\eta / \varepsilon)$, differential privacy is satisfied in every iteration. Moreover, according to the composition property, as the privacy budget we used in each iteration is the same ε, we can come to the conclusion that our algorithm satisfies ε-differential privacy.

Algorithm 1 Candidate Pruning-based FIM

```
Input: $D, \lambda, \gamma, \delta, \varepsilon$
Output: $Frequent\ itemsets$
1. $\zeta \leftarrow (\gamma, \delta) - approximation\ of\ database\ D$
2. $L_1 \leftarrow I$, $k \leftarrow 2$
3. $while\ L_{k\text{-}1} \neq \varnothing$
```

```
    $C_k \leftarrow \{m \cup \{n\} \mid m \in L_{k-1} \wedge n \notin m\} - \{x \mid \{y \mid y \subseteq x \wedge |y| = k-1\} \not\subset L_{k-1}\}$
    $for\ itemsets\ i \in C_k$
        $if\ \ i \notin \zeta$
            $remove\ i\ from\ C_k$
    $for\ transactions\ t \in D$
        $C_t \leftarrow \{i \mid x \in C_k \wedge i \subseteq t\}$
        $for\ itemsets\ i \in C_t$
            $count[i]{+}{+}$
    $for\ itemsets\ i \in C_k$
        $count[\mathrm{i}] \leftarrow count[\mathrm{i}] + Lap(|C_k| / \varepsilon)$
    $L_k \leftarrow \{i \mid i \in C_k \wedge count[i] \ge \lambda * |D|\}$
    $k{+}{+}$
Return: $\bigcup_k L_k$
```

5. Experiments

In this section, we experimentally evaluate the effectiveness of our algorithm on two real data sets, and the summary of those data sets is given in Table 2.

Table 2. Data Set Description

Data Set	$\lvert D \rvert$	$\lvert I \rvert$
BMS-WebView-1(WV1)	59602	497
BMS-WebView-2(WV2)	77512	3340

5.1. Experimental Settings

We implemented our algorithm in C++, and all experiments are performed on an Intel Xeon 2.27GHz CPU with 8GB RAM. We use the F-score defined in [14] as the metric. As a reference point, we compare our algorithm with algorithm proposed in [14] denoted as TLT and PrivBasis proposed in [13].

Definition 5 (F-score): Suppose R and $\tilde{R}$ represent the set of correct frequent itemsets and the set of frequent itemsets generated by differential private FIM respectively. F-score is defined as:

$$F - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

where $precision = \frac{|\mathrm{R} \cap \tilde{R}|}{|\tilde{R}|}$ and $recall = \frac{|\mathrm{R} \cap \tilde{R}|}{|\mathrm{R}|}$.

The privacy budget is set to 1. Due to the fact that the performance of our algorithm is related to the parameter (γ,δ) used in preprocessing phase, we tune these parameters dynamically according to the utility of result. Because of space limitation, meanwhile, for the sake of convenience in comparison with other algorithms, we only show the relationship between threshold and F-score. In addition, we run all algorithms ten times to get the average results of them.

5.2. Competing Algorithms

Figure 1 and Figure 2 compare the utility of PrivBasis [13], TLT[14] and our algorithm in terms of F-score. Owning to the fact that PrivBasis is designed for top-k frequent itemsets mining, we compare our algorithm with it by setting k to be the number of frequent itemsets for the given threshold.

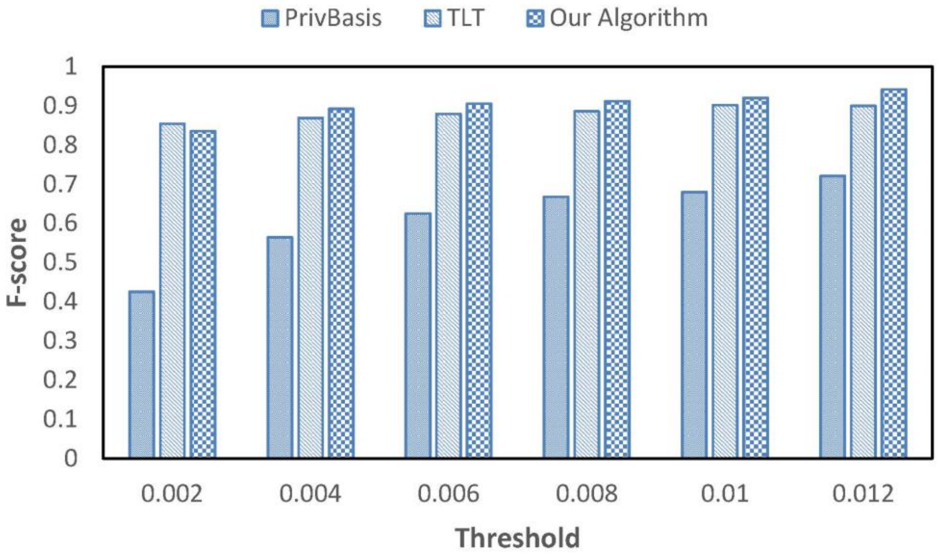

Figure 1. Frequent Itemsets Mining on WV1

As we can see from them, our algorithm outperforms TLT and PrivBasis overall. PrivBasis has poor performance, and we guess the reason for this probably is PrivBasis is designed for top-k itemsets mining. Although top-k mining problem can be easily transformed to frequent itemsets mining, the performance of PrivBasis becomes disappointed when k is large. This may explain why the F-score is increased as the threshold increases.

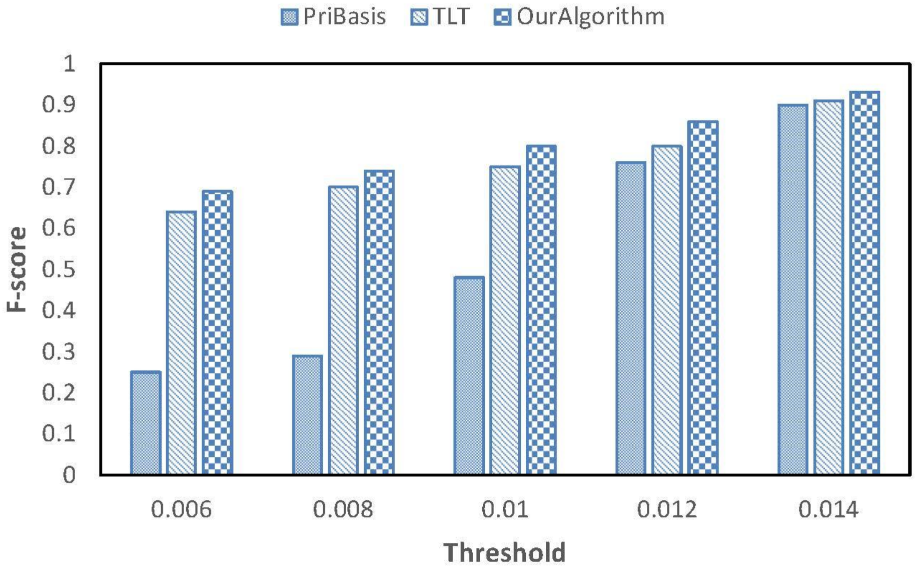

Figure 2. Frequent Itemsets Mining on WV2

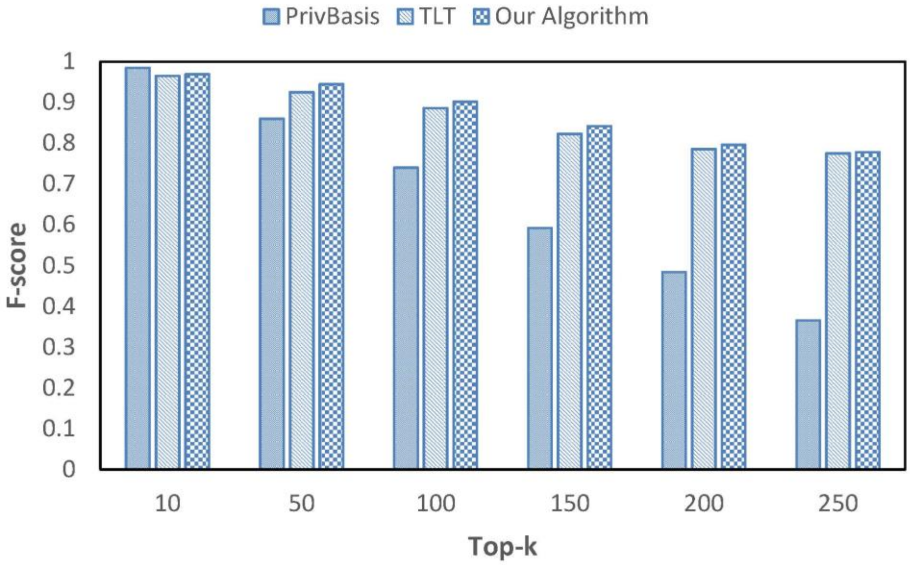

Figure 3. Top-k Mining on WV1

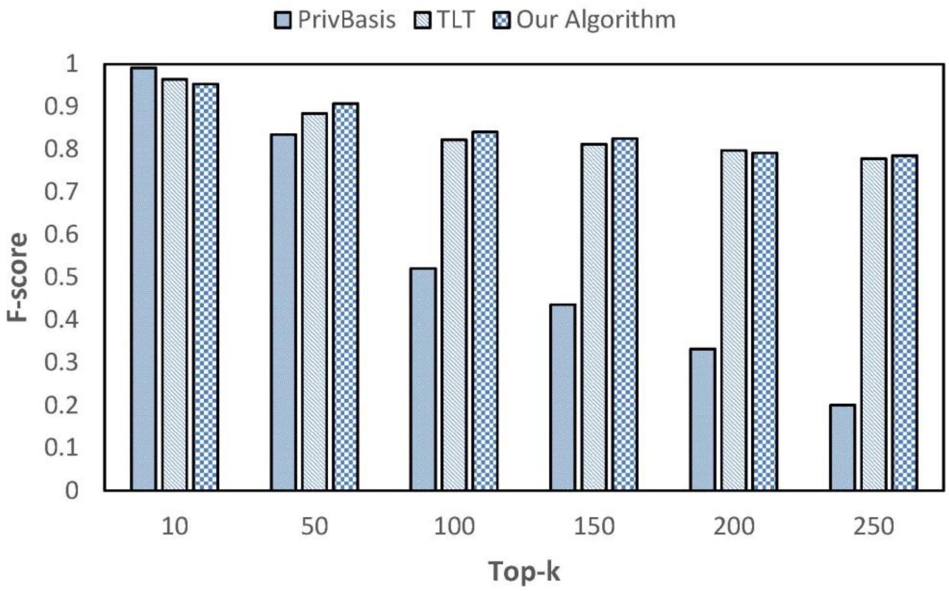

Figure 4. Top-k Mining on WV2

Our algorithm has little advantages over TLT as its F-score is higher than TLT's overall. This could be explained by the fact that our algorithm injects less noise into the support of itemsets, so the accuracy of generated result is better. This could illustrate that our method is effective to improve the utility of result.

We also modify our algorithm and TLT to compare them with PrivBasis for top-k frequent itemsets mining. Top-k frequent itemsets mining and frequent itemsets mining have some internal relationship. As mentioned in many works, they can be easily transformed to each other. As shown in Figure 3 and Figure 4, in general, our algorithm achieves better result than the other two. We observe that the performance of PrivBasis becomes worse as k increases. One reason for this is PriBasis is not an ideal choice for large k. Another reason is, as described in [14], PrivBasis mistakes some frequent itemsets as infrequent ones. Therefore, TLT and our algorithm have a more stable performance than PrivBasis. Experiment on top-k frequent itemsets mining also illustrates that, to some degree, our algorithm is effective to improve the utility of result.

6. Conclusion

In this paper, we focus on improving the utility of result by reducing the noise scale injected. The key observation is that pruning the number of unpromising candidate items can effectively reduce noise added in differential privacy mechanism, which can bring about a better tradeoff between utility and privacy of the result. As the noise scale is proportional to size of candidate frequent itemsets, we seek for an effective way to shrink candidate set by removing unpromising itemsets from the candidate set. Previous works mostly focus on using new structure to reduce the noise scale while guaranteeing the privacy requirement. However, we aim at solving the problem from the bottom. So we propose a novel 2-stage mechanism to mine frequent itemsets under differential privacy model. We use a progressive sampling method to generate a super set of frequent itemsets to shrink the size of candidate set, after which the noise scale injected will be greatly lowered. our algorithm would increase the utility of result while satisfying differential privacy.

Acknowledgements

This work is supported by the National Science Foundation for Distinguished Young Scholars of China under grant no. 61225010, NSFC under grant nos. of 61370198, 61370199 and 61300187, The Fundamental Research Funds for the Central Universities under grant nos. of 3132014215, 3132014208.

References

[1] A. Rakesh, T. Imieliński and A. Swami, “Mining association rules between sets of items in large databases”, ACM SIGMOD Record, vol. 22, no. 2, **(1993)**.

[2] G. Bart and M. J. Zaki, “Advances in frequent itemset mining implementations: report on FIMI'03”, ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, **(2004)**.

[3] A. Eytan, D. S. Weld, B. N. Bershad and S. S. Gribble, “Why we search: visualizing and predicting user behavior”, Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, **(2007)**.

[4] B. Sergey, R. Motwani, and C. Silverstein, “Beyond market baskets: Generalizing association rules to correlations”, ACM SIGMOD Record, vol. 26, no. 2, **(1997)**.

[5] X. Wu and J. Ling, “Design and Realization of the Improved Sunday Pattern Matching Algorithm”, Journal Harbin University of Science and Technology, vol. 6, no. 52, **(2013)**.

[6] N. G. Niklas, A. Bate, J. Hopstadius, K. Star and I. R. Edwards, “Temporal pattern discovery for trends and transient effects: its application to patient records”, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, USA, **(2008)**.

[7] H. Vagelis, “ed. Information discovery on electronic health records”, CRC Press, Boca Raton, **(2009)**.

[8] D. Cynthia, F. McSherry, K. Nissim and A. Smith, “Calibrating noise to sensitivity in private data analysis”, Theory of Cryptography Conference, New York, USA, **(2006)**.

[9] D. Cynthia, “Differential privacy”, Encyclopedia of Cryptography and Security, Springer, New York **(2011)**.

[10] R. Matteo and E. Upfal, “Mining frequent itemsets through progressive sampling with Rademacher averages”, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, **(2015)**.

[11] B. Raghav, S. Laxman, A. Smith and A. Thakurta, “Discovering frequent patterns in sensitive data”, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, **(2010)**.

[12] M. Frank and K. Talwar, “Mechanism design via differential privacy”, Foundations of Computer Science, Rhode Island, USA, **(2007)**.

[13] L. Ninghui, W. Qardaji, D. Su and J. Cao, “PrivBasis: frequent itemset mining with differential privacy”, Proceedings of the VLDB Endowment, vol. 5, no. 11, **(2012)**.

[14] Z. Chen, J. F. Naughton and J. Y. Cai, “On differentially private frequent itemset mining”, Proceedings of the VLDB Endowment, vol. 6, no. 1, **(2012)**.

[15] X. Shengzhi, S. Su, X. Cheng, Z. Li and L. Xiong, “Differentially Private Frequent Sequence Mining via Sampling-based Candidate Pruning”, International Conference on Data Engineering, Seoul, Korea, **(2015)**.

[16] A. Rakesh and R. Srikant, “Fast algorithms for mining association rules”, Proceedings of 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, **(1994)**.

[17] D. M. Frank, “Privacy integrated queries: an extensible platform for privacy-preserving data analysis”, Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, New York, USA, **(2009)**.

Authors

Yangyang Xu, He received the bachelor's degree from the Dalian Maritime University, China, in 2014. Currently, he is working toward the master's degree at the School of Information Science and Technology, Dalian Maritime University. His main research interests include data privacy protection and data mining.

ZhaoBin Liu, He is a Professor in School of Information Science and Technology, Dalian Maritime University, China. He received his Ph.D. in Computer Science from Huazhong University of Science and Technology in China in 2004. His research areas include Cloud computing, Big data, Computer Networks and Embedded Systems. He has more than 60 publications in international journals, conference proceedings as well as book chapters, and has successfully coordinated several research projects funded by various funding agencies across China.

Zhonglian Hu, The main research interests of him are differential privacy and set theory. In 2014, he received the bachelor's degree from the Dalian Maritime University, China. At present, he is studying for master's degree at the School of Information Science and Technology, Dalian Maritime University.