Frequent Itemsets Mining With Differential Privacy Over Large-Scale Data

Veeravenkata Sai Kumar Nimmala, Smt.K.R.Rajeswari, Sri.V.Bhaskara Murthy
MCA Student, Assistant Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

ABSTRACT Frequent itemsets mining with differential privacy refers to the problem of mining all frequent itemsets whosesupports are above a given threshold in a given transactional dataset, with the constraint that the mined results should not break the privacy of any single transaction. Current solutions for this problem cannot well balance efficiency, privacy and data utility over large scaled data. Toward this end, we propose an efficient, differential private frequent itemsets mining algorithm over large scale data. Based on the ideas of sampling and transaction truncation using length constraints, our algorithm reduces the computation intensity, reduces mining sensitivity, and thus improves data utility given a fixed privacy budget. Experimental results show that our algorithm achieves betterperformance than prior approaches on multiple datasets.

I. INTRODUCTION

In recent years, with the explosive growth of data and the rapid development of information technology, various industries have accumulated large amounts of data through various channels. To discover useful knowledge from large amounts of data for upper-layer applications (e.g. business decisions, potential customer analysis, etc.), data mining [1]–[9] has been developed rapidly. It has produced a positive impact in many areas such as business and medical care. Along with the great benefits of these advances, the large amount of data also contains privacy sensitive information, which may be leaked if not well managed. For instance, smartphone applications are recording the whereabouts of users through GPS sensors and are transferring the data to their servers. Medical records are also storing potential relationships between diseases and a variety of data. Mining on user location data or medical record data both provide invaluable information; however, they may also leak user privacy. Thus mining knowledge under confident privacy guarantees is highly expected. This paper investigates how to mine frequent item sets with privacy guarantee for big data. We consider the following application scenario. A company (such as information consulting firm) has a large-scale dataset. The company would like to make the dataset public and therefore allow the public to execute frequent item sets mining for getting cooperation or profits. But due to privacy considerations, the company cannot provide the original dataset directly. Therefore, privacy mechanisms are needed to process the data, which is the focus of this paper.

To ensure privacy of data mining, traditional methods are based on k-anonymity and its extended models [10]–[16].These methods require certain assumptions; it is difficult to protect privacy when the assumptions are violated. The insufficiency of k-anonymity and its extended models is that there is no strict definition of the attack model, and that the knowledge of the attacker cannot be

quantitatively defined. To pursue strict privacy analysis, Dwork proposed a strong privacy protection model called differential privacy [17]. This privacy definition features independence of background knowledge of the attacker and proves very useful.

Frequent pattern mining with privacy protection has also received extensive attention. As preliminary methods [18]–[24], these works have provided a lot of contributions in this area. But with the advance of research, these privacy method shave not been able to provide effective privacy. In order to overcome these difficulties, researches began to focus on the differential privacy protection framework [25]–[31]. Although guaranteeing privacy temporary, however, the balance between privacy and utility of frequent item sets mining results needs to be further pursued.

In this paper, we propose a novel differential private frequent item sets mining algorithm for big data by merging the ideas of [27], [30], which has better performance due to the new sampling and better truncation techniques. We build our algorithm on FP-Tree for frequent item sets mining. In order to solve the problem of building FP-Tree with large-scale data, we first use the sampling idea to obtain representative data to mine potential closed frequent item sets, which are later used to find the final frequent items in the large-scale data. In addition, we employ the length constraint strategy to solve the problem of high global sensitivity. Specifically, we use string matching ideas to discover the most similar string in the source dataset, and implement transaction truncation for achieving the lowest information loss. We finally add the Laplace noise for frequent item sets to ensure privacy guarantees.

A few challenges exist: First, how to design a sampling method to control the sampling error? We use the central limit theorem to calculate a reasonable sample size to control the error range. After obtaining the sample size, the dataset is randomly sampled using a data analysis toolkit. The second challenge is how to design a good string matching method to truncate the transaction without losing information as far as possible? We match the potential item sets in the sample data to find the most similar items and then merge them with the most frequent items until the maximum length constraint is reached.

II. EXISTING SYSTEM

Explosive growth of data and the rapid development of information technology, various industries have accumulated large amounts of data through various channels. To discover useful knowledge from largeamounts of data for upper-layer applications (e.g. business decisions, potential customer analysis, etc.), data mining has been developed rapidly. It has produced a positive impact in many areas such as business and medical care. Along with the great benefits of these advances, the largeamount of data also contains privacy sensitive information, which may be leaked if not well managed. For instance. Medical records are also storing potential relationshipsbetween diseases and a variety of data. Mining on user location data or medical record data both provide invaluable information; however, they may also leak user privacy. The company would like to make the dataset public and therefore allow the public to execute frequent itemsets mining for getting cooperation or profits. But due to privacy considerations, the company cannot provide the original dataset directly. Therefore, privacy mechanisms are needed to process the data.

III. PROPOSED SYSTEM

We propose a novel differential private frequent itemsets mining algorithm for big data by merging theideas, which has better performance

due to the new sampling and better truncation techniques. We build our algorithm on FP-Tree for frequent itemsets mining. In order to solve the problem of building FP-Tree with large-scale data, we first use the sampling idea to obtain representative data to mine potential closed frequent itemsets, which are later used to find the final frequent items in the large-scale data.

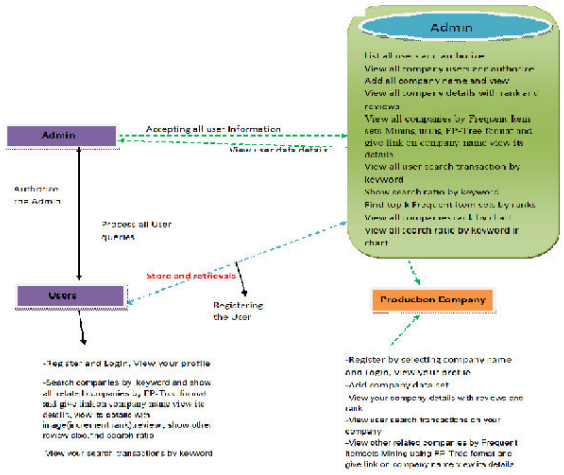
Future Work

DPFIM Data Partition Frequent Itemset Mining, which merges the ideas of, but employs a different (better) truncationscheme and boosts computation efficiency using both sampling and truncation. Compared with previous work using random truncation, our new string similarity-matching-based truncation mechanism has better performance than previous work, which is because string-similarity-matching-based truncation preserves more useful frequent itemset candidates.

Algorithm

Newly proposed algorithm, called DPFIM, which merges the ideas of, but employs adifferent(better) truncation scheme and boosts computation efficiency using both sampling and truncation. Compared with previous work using random truncation, our new stringsimilarity-matching-based truncation mechanism has better performance than previous work, which is because string-similarity-matching-based truncation preserves more useful frequent itemset candidates. The experimental results also confirm the better performance. The algorithm is differentially private; it takes a threshold value and outputs the frequent itemsets with support at least. The basic idea is as follows: first, compute a noisy support for the threshold,then truncate the original database noisily, finally construct a noisy FP-Tree for mining frequent itemsets.

IV. ARCHITECTURE DIAGRAM



V. IMPLEMENTATION

Admin

In this module, the Cloud has to login by using valid user name and password. After login successful he can do some operations such as List all users and authorize, View all company users and authorize Add all company name and view, View all company details with rank and reviews,View all companies by Frequent Item sets Mining using FP-Tree format and give link on company name view its details, View all user search transaction by keyword, Show search ratio by keyword, Find top k Frequent item sets by ranks View all companies rank by chart, View all search ratio by keyword in chart

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Production Company

In this module, there are n numbers of Owners are present. Owner should register before doing

any operations. Once registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful Owner will do some operations like View your profile, Add company data set, View your company details with reviews and rank, View user search transactions on your company, View other related companies by Frequent Itemsets Mining using FP-Tree format and give link on company name view its details

Users

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like View your profile, Search companies by keyword and show all related companies by FP-Tree format and give link on company name view its details, view its details with image(increment rank),review , show other review also, find search ratio, View your search transactions by keyword

VI. CONCLUSION

We propose a novel differentially private algorithm for frequent itemsets mining. The algorithm featuresbetter data utility and better computation efficiency. Various experimental evaluations validate that the proposed algorithm has high F-Score and low relative error. That fine-tuned parameters lead to better differentially private frequent itemsets mining algorithms with regard to data utility.

REFERENCES

[1] Z. John Lu, “The elements of statistical learning: data mining, inference, and prediction,” Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 173, no. 3, pp. 693–694, 2010.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI magazine, vol. 17, no. 3, p. 37,1996.

[3] H. Yang, K. Huang, I. King, and M. R. Lyu, “Localized support vector regression for time series prediction,” Neuro computing, vol. 72, no. 10-12, pp. 2659–2669, 2009.

[4] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, pp. 601–618, Nov 2010.

[5] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[6] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, “Robust semi-supervised subspace clustering via non-negative low-rank representation,” IEEE Transactions on Cybernetics, vol. 46, pp. 1828–1838, Aug 2016.

[7] M. Pe˜na, F. Biscarri, J. I. Guerrero, I. Monedero, and C. Le´on, “Rulebasedsystem to detect energy efficiency anomalies in smart buildings, a data mining approach,” Expert Systems with Applications, vol. 56,pp. 242–255, 2016.

[8] Y. Guo, F. Wang, B. Chen, and J. Xin, “Robust echo state networks based on correntropy induced loss function,” Neuro computing, vol. 267,pp. 295–303, 2017.

[9] H. Lim and H.-J. Kim, “Item recommendation using tag emotion in social cataloging services,” Expert Systems with Applications, vol. 89,pp. 179–187, 2017.

[10] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, “(_, k)-anonymity:an enhanced k-anonymity model for privacy preserving data

publishing,"in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 754–759, ACM, 2006.
[11] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,vol. 10, no. 05, pp. 557–570, 2002.
[12] S. Latanya, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 571–588, 2002.
[13] A. Meyerson and R. Williams, "On the complexity of optimal kanonymity,"in Proceedings of the Twenty-third ACM SIGMOD-SIGACTSIGARTSymposium on Principles of Database Systems, pp. 223–228,ACM, 2004.
[14] Y. Zhang, J. Zhou, F. Chen, L. Y. Zhang, K. Wong, X. He, and D. Xiao, "Embedding cryptographic features in compressive sensing, "Neuro computing, vol. 205, pp. 472–480, 2016.
[15] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramani"l-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, p. 3, 2007.
[16] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyondk-anonymity and l-diversity," in IEEE 23rd International Conference onData Engineering, pp. 106–115, IEEE, 2007.
[17] C. Dwork, "Differential privacy," in Encyclopedia of Cryptography and Security, pp. 338–340, Springer, 2011.
[18] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 9, pp. 1026–1037,2004.
[19] J. Vaidya and C. Clifton, "Privacy preserving association rule mining invertically partitioned data," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639–644, ACM, 2002.
[20] Z. Teng and W. Du, "A hybrid multi-group approach for privacy preserving data mining," Knowledge And Information Systems, vol. 19,no. 2, pp. 133–157, 2009.