FREQUENT ITEMSET MINING USING PFP-GROWTH VIA SMART SPLITTING

Neha V. Sonparote, Professor Vijay B. More.

Neha V. Sonparote, Dept. of computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India
Professor Vijay B. More, Dept. of computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India

---***---

Abstract - *In the context of data mining patterns searching from large databases is an interesting area. Mining or discovery of frequent pattern plays very important role in data mining. FIM is the frequent itemset mining is popular technique to discover the knowledge from given database. It finds the itemsets that appeared together in database. But it suffered from some privacy implications. Data owner impaired in computational resources can outsource to third party server. But both item and association rules considered as private property of data owner. Therefore, outsourcing data required privacy protection which noticed as a major issue in transaction splitting. Traditional methods of frequent itemset mining has problem of tradeoff between utility and privacy in designing a differentially private FIM algorithm. It does not deal with the high utility transactional itemsets and it has large time complexity. Existing methods provides the large size output combinations. According analysis of existing approaches of FIM there is requirement of a time efficient differentially private FIM algorithm.*

Key Words: — **Frequent itemset mining, differential privacy, transaction splitting, FP-Growth, FP Tree.**

1. INTRODUCTION

With increasing ability to gathered private or personal data has a major privacy issue. Privacy issues are also arises in the area of data mining i.e. frequent itemset mining. FIM is the frequent itemset mining approach widely used in many market applications. It aim to discover the frequent itemsets that bought together. It will helps to make analysis of placement of products and their marketing and many more. In this study of frequent pattern mining there are multiple techniques available for FIM which suffered from privacy issues. An algorithm for frequent itemset mining i.e. FIM accepts the input in the form of transactions form the group of individuals and generates the frequent itemsets by processing it. This processing leads privacy issues such as, after producing frequent itemsets from input transactions it is going to publish but the dataset does not leak private information about individual whose data get proceed. This is the problem associated by the reality that individuals outsourced data as well as background details or information.

Fortunately, differential privacy approach provided the guarantees that the presence of individual's data in a database does not leak about that individual information. The anonymization-based privacy models are proposed in [2], [4] on the release of data without revealing background knowledge of data. K-anonymity model which contains the formal protection to data. It is important to provide k-similar guarantees of data privacy protection. Re-identification linking approach is utilized for demonstration of re-identification on shared attributes. L-diversified does not required complete distribution of the sensitive as well as non-sensitive information attributes. It automatically covers the instant level knowledge. Traditionally different frequent itemset mining algorithms are available such as, Aprior, FP-Growth algorithms [1]. Both algorithms are most popular in the context of mining frequent itemset and patterns. Apriori algorithm utilizes BFS i.e. breadth-first search approach for candidate set generation-and-test algorithm. It required 'l' scan of the database whereas, 'l' is the length of database. In converse, FP-growth algorithm is DFS i.e. depth-first search algorithm. It required two scans of database. Hence faster than apriori. The FIM algorithm is developed on the basis of or motivating from FP-growth algorithm. FIM algorithm [10] achieves high efficiency and utilization of data and stronger degree of privacy. As FP-Growth algorithm required only two scans which limits on opportunity to re-truncate transactions during the process of mining. Transaction truncation approach is proposed in [9]. It is private frequent pattern mining approach which provides good privacy and utility. At the beginning they investigate their approach by truncating long transactions trade-off errors.

2. RELATED WORK

In 2000, J. Han, J. Pei et al [1], proposed an approach for mining FP patterns from transaction database on basis of time series databases and other many types of databases. They discussed about apriori-like candidate set generation. But the candidate set generation is very expensive, especially for long patterns. Therefore, they have proposed FP-Growth algorithm. It is extended prefix structure which is used to stored compressed and sensitive information of frequent patterns. There are three techniques represented to achieve efficiency of proposed technique as: 1. Compression of large database into small data structure therefore, cost will be reduced 2. FP based mining to neglect cost of large candidate set generation and 3. Partitioned based approach to decomposed task of mining. They have discussed and represented the performance of FP-Growth technique such as it is efficient as well as scalable for long and short FP. Also the magnitude is faster than the apriori algorithm.

In 2002, L. Sweeney, et al [2], provides solution for data re-identification. They have proposed k-anonymity model which contains the formal protection to data. It is important to provide k-similar guarantees of data privacy protection. Re-identification linking approach is utilized for demonstration of re-identification on shared attributes. In multi-level database aggregation and inferences restrict the low level classification information into higher level classification. They discussed about MDB i.e. multilevel database system which provides different security classifications and clearances. Also they represented elimination of precise inference because of functional dependencies and multi-valued dependencies. The aggregation inferences get solved by design of database but solution cannot possible practically due rich settings of data.

In 2002, J. Vaidya and C. Clifton [3], described - two-party algorithm. It is implemented for efficient frequent itemsets mining with less support level without revealing individual transaction values. To mine association rules from vertically partitioned data they proposed a privacy preserving algorithm. In this vertical partitioning is used. To detect/predict malfunctions, they gathered proprietary data collected by several parties, with a single key joining all the data sets. Ideally with the privacy constraints they would gained complete zero knowledge but it is acceptable for practical solution controlled information disclosure. They referred heterogeneous database scenario for vertical partitioning of the database between two parties A and B. In proposed framework they considered mining Boolean association rules such as, transactions are either 0 or 1. The cost of communication is depending upon number of candidate itemsets. There are several issues faced in other types of data mining including clustering, classification etc was considered by them for future work planning.

In 2006 A. Machanavajjhala et al [4] represented two simple attacks to serve the privacy problems. It has ability to discover the sensitive attributes value. In many cases attacker is familiar with the background knowledge for which author does not guarantee of privacy against background attacks. To provide detail analysis of two attacks they have proposed l-diversity mechanism and construct formal foundation of it. They provide practically efficient implementation of 'l'-diversified mechanism. They showed that how k-anonymous dataset comprises with two attacks i.e. homogeneous and background knowledge attacks. L-diversified does not required complete distribution of the sensitive as well as non-sensitive information attributes. It automatically covers the instant level knowledge. They extend their idea of managing multiple sensitive attributes in future work. They want to develop a method for continues sensitive attributes.

In 2007, W. K. Wong et al. [5], developed an effective and efficient encryption algorithm. It is specifically for applications in which owners transmit streams of transactions to SP (service provider) because it performs a single pass over the database. The proposed algorithm required nominal computational resources hence they have limited resources to maintain. Two approaches used to protect sensitive information namely, encryption and data perturbation. Encryption is function for transformation of original format into the new one whereas; data perturbation modifies the original data randomly. Mainly, they proposed and evaluate appropriate encryption techniques for outsourcing of association rules mining algorithms. For one-to-one mapping from the original set I of items to another dictionary J item mapping technique is proposed using simple encryption method. Unique itemsets are mapped using one-to-n admissible mapping algorithm whereas, complete transformation is implemented in another algorithm known as a valid and complete transformation for a transaction t.

In 2009, W. K. Wong et al. [6], introduced an approach for outsourcing of Frequent Itemset Mining. They addressed the integrity issue in the outsourcing process. They proposed and build an audit environment, which includes database transformation method and a result verification method. An artificial itemset planting (AIP) technique is the main component of proposed audit environment. They provide guarantees about the correctness of the process of verification. The problem is divided into two subparts such as: 1. Frequent itemsets computations and 2. Association rules based on the mined frequent itemsets computations. These problems are computationally inexpensive that is the reason time complexity is exponential. Security and integrity are the major issues in outsourcing which satisfactorily addressed in this paper. At very first they step towards integrity solving approach and then focused on security in outsourcing. Proposed AIP gives probabilistic guarantees that incorrect FP mining results returned by the SP will be observed by the owner with a controllably high confidence.

In 2010, R. Bhaskar, et al. [7], proposed two efficient algorithms for discovering the K most frequent patterns in a data set of sensitive records. They provided meaningful privacy guarantees in the presence of arbitrary external information. Proposed algorithms were outputting to preserve privacy. They also defined notion of utility that refines the output accuracy of private top-K pattern mining algorithms. To discover and report the patterns that occur most frequently in the data is the main goal of FIM framework. The Apriori algorithm is the successful techniques in data mining for FIM. Authors were mainly focused with mining top K itemsets from transaction data. In private FIM algorithm a natural notion of approximation for frequent itemset mining is introduced. To obtained accuracy another algorithm is presented based on Laplace Mechanism. For future work they have planned to remove dependency on the size of the universe of items by applicability of the algorithms for larger and complex data sets.

In 2012, N. Li, et al. [8], proposed PrivBasis approach. It is basically for dimension reduction. They have described algorithms for privately constructing set and then using it to identify the most frequent itemsets. To satisfy differential privacy they studied the problem of how to perform frequent itemset mining on transaction databases. The proposed approach neglects the selection of top k itemsets from huge candidate set. PrivBasis approach projecting the input dataset onto a small number of selected dimensions to meet the high dimensionality challenges. It uses several dimensions for projection and rejects any one set containing too many dimensions. For the purpose of finding the k most frequent itemsets the proposed technique enable one to select which sets of dimensions. The proposed approach deals with the curse of dimensionality in private data analysis and data anonymization.

In 2012, C. Zeng, J. F. Naughton et al[9], considered private frequent pattern mining approach. Their aim is to provide good privacy and utility. At the beginning they investigate their approach by truncating long transactions trade-off errors. Also by proposed work they used to solve the problem of mining “classical” frequent itemset. In proposed frequent itemset mining they have consider frequent 1-itemset mining. In this they represented truncation of transactions and promoting utility of 1-itemset mining. They discussed about transaction truncation using Naïve algorithm but it outputs the poor performance for two reasons such as:

1. Random Truncating: It does not differentiate between frequent and rare subsets of transactions to be truncated.
2. Propagated errors: If by mistake frequent itemset is labeled as infrequent then in-frequented superset is also computed.

They have proposed frequent β itemset Mining algorithm for random truncation approximation in the loss of information it realizes small amount of information from the database. The proposed algorithm has improved performance over apriori algorithm due to smart truncation of transaction. Basically, smart truncation is applied on original database to discover initial itemsets with the utilization of noisy results. Quantification of information loss by smart truncating is focused in future work.

In 2015, S. Su, S. Xu [10], proposed private FP-Growth (PFP-growth) algorithm to address the challenging issues of privacy in outsourcing transaction database. Proposed approach consists of two phases such as, preprocessing and mining phase. In preprocessing, database transformation limit to the length of transaction, it is irrelevant to the user specified threshold and it required only onetime database scan. Whereas, in mining phase, user specified threshold and transformed database is privately extracted for frequent itemsets.

3. SYSTEM ARCHITECTURE

We proposed FP-Growth partitioning-based, depth-first search algorithm. It adopts a divide-and-conquer manner to decompose the mining task into many smaller tasks for finding frequent item sets in conditional pattern bases. FP-growth constructs a FP-Tree, FP tree for the database. For the frequent items in each transaction, they are arranged according to the order of HT and inserted into FP-Tree as a branch.

It consists of two phases:

1. Preprocessing of dataset
2. Mining Phase

Following figure represents the system architecture and as per designing system flow is also given:

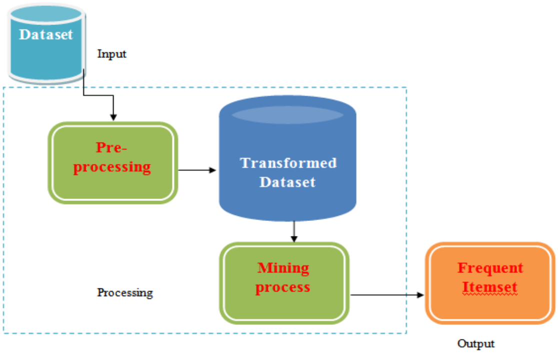

Fig. -1: System architecture

4. CONCLUSION

In this review paper we have studied some existing techniques of frequent itemset mining. FIM is the process of extracting itemsets from large database. Several methods are available for mining frequent itemsets such as algorithm, Boolean association rules, which suffers from privacy issues. Also some techniques such as, PrivBasis has the problem of high dimensionality. Security and integrity are the major issues in outsourcing database. From literature survey analysis, we analyzed that there is need of a time efficient differentially private FIM algorithm. Lastly, we conclude that PFP-growth algorithm can be a better solution which has ability to dynamically minimize the amount of noise added to guarantee privacy during the mining process and it is differentially private technique.

REFERENCES

[1] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1–12.

[2] L. Sweeney, “k-anonymity: A model for protecting privacy,” Int. J. Uncertainty Fuzziness Knowl.-Base Syst., vol. 10, no. 5, pp. 557–570, 2002.

[3] J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 639–644.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in Proc. 22nd Int. Conf. Data Eng., 2006, p. 24.

[5] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, “Security in outsourcing of association rule mining,” in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 111–122

[6] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, “An audit environment for outsourcing of frequent itemset mining,” Proc. VLDB Endowment, vol. 2, no. 1, pp. 1162–1173, 2009.

[7] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, “Discovering frequent patterns in sensitive data,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp. 503–512.

[8] N. Li, W. Qardaji, D. Su, and J. Cao, “Privbasis: Frequent itemset mining with differential privacy,” Proc. VLDB Endowment, vol. 5, no. 11, pp. 1340–1351, 2012.

[9] C. Zeng, J. F. Naughton, and J.-Y. Cai, “On differentially private frequent itemset mining,” Proc. VLDB Endowment, vol. 6, no. 1, pp. 25–36, 2012.

[10] S. Su, S. Xu, X. Cheng, Z. Li, and F. Yang, “Differentially Private Frequent Itemset Mining via Transaction Splitting,” in IEEE Trans. Knowl. Data Eng., vol. 27, no. 7, pp., July. 2015.

BIOGRAPHIES

Sonaparote Neha V. has completed Bachelor’s Degree from Late PVG College of Engg. Nashik. Now pursuing ME Degree in Computer Engineering from MET’s Institute of Engineering, BKC, Nashik.

Prof. More Vijay B. is working in MET's Institute of Engineering, BKC Nashik. He has published/presented research papers in National and International Journals/Conferences. His area of interest is in Data mining.