Encryption Scheme for Privacy Preserving Mining Association Rules in Outsourced Database

N. PRANEETHA REDDY [1], Dr. P. CHITTIBABU[2]

[1]M.Tech. ANNAMACHARYA P.G COLLEGE OF COMPUTER STUDIES,Rajampeta, Kadapa

[2]Professor, ANNAMACHARYA P.G COLLEGE OF COMPUTER STUDIES,Rajampeta, Kadapa

Abstract:

With the rapid development of data mining analysis tools now a days it is penetration of data mining and analysis within different fields for disciplines, security and providing security at mining activities. The pattern mining in large database provides introduction of new and novel algorithms in data mining technology for providing secured pattern mining at outsourced or remote servers. The aim of project is to provide privacy preserving in data mining pattern analysis results and maintaining the inferences which secures the private information about organization or individuals firms. In older data mining analysis of patterns, Fast Distributed Mining algorithm is used. The Fast Distributed Mining Algorithm is invention of new protocol with set of mining rules for secure mining of data in horizontally distributed databases. But the protocol was not able to secure distributed versions of database, because fast distributed mining algorithm. To overcome the problem proposes another new technique called secure computations with Multi party computations control model on background knowledge. The goal of the proposed system is to use mining and cryptography as joint technique for faster and secure mining computations. The protocol also provides all the resource usage and requirements under control.

Keywords

Frequent, Associations, Data Mining.

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 has been developed rapidly. It has produced a positive impact in many areas such as business 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, smart phone applications are recording the whereabouts of users through GPS sensors and are transferring the data to their servers. Business records are also storing potential relationships between diseases and a variety of data. Mining on user location data or Business record data both provide invaluable information; however, they may also leak user privacy. Thus mining knowledge under confident privacy guarantees is highly expected. This project investigates how to mine frequent itemsets with privacy guarantee for big data. 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 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, Various records are also storing potential relationships between diseases and a variety of data. Mining on user location data or various 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 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. Propose a novel differential private frequent item sets mining algorithm for big data by merging the ideas, 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.

II. Related Work

High utility thing sets (HUIs) mining is a rising subject in information mining, which alludes to finding all thing sets having an utility meeting a client determined least utility edge min_util. Notwithstanding, setting min_util suitably is a troublesome issue for clients. As a rule, finding a fitting least utility edge by experimentation is a monotonous procedure for clients. In the event that min_util is set too low, an excessive number of HUIs will be produced, which may bring about the mining procedure to be exceptionally wasteful. Then again, if min_util is set too high, it is likely that no HUIs will be found. In this paper, we address the above issues by proposing another structure for top-k high utility thing set mining, where k is the coveted number of HUIs to be mined. Two sorts of proficient calculations named TKU (mining Top-K Utility thing sets) and TKO (mining Top-K utility thing sets in one stage) are proposed for mining such thing sets without the need to set min_util. We give an auxiliary examination of the two calculations with talks on their preferences and restrictions. Exact assessments on both genuine and manufactured datasets demonstrate that the execution of the proposed calculations is near that of the ideal instance of best in class utility mining calculations.

1. Mining top-k sequential rules

Mining sequential rules requires specifying parameters that are often difficult to set (the minimal confidence and minimal support). Depending on the choice of these parameters, current algorithms can become very slow and generate an extremely large amount of results or generate too few results, omitting valuable information. This is a serious problem because in practice users have limited resources for analyzing the results and thus are often only interested in discovering a certain amount of results, and fine-tuning the parameters can be very time-consuming. In this paper, we address this problem by proposing TopSeqRules, an efficient algorithm for mining the top-k sequential rules from sequence databases, where k is the number of sequential rules to be found and is set by the user. Experimental results on real-life datasets show that the algorithm has excellent performance and scalability.

2. Mining high utility mobile sequential patterns in mobile commerce environments

Mining user behaviors in mobile environments is an emerging and important topic in data mining fields. Previous researches have combined moving paths and purchase transactions to find mobile sequential patterns. However, these patterns cannot reflect actual profits of items in transaction databases. In this work, we explore a new problem of mining high utility mobile sequential patterns by integrating mobile data mining with utility mining. To the best of our knowledge, this is the first work that combines mobility patterns with high utility patterns to find high utility mobile sequential patterns, which are mobile sequential patterns with their utilities. Two tree-based methods are proposed for mining high utility mobile sequential patterns. A series of analyses on the performance of the two algorithms are conducted through experimental evaluations. The results show that the proposed algorithms deliver better performance than the state-of-the-art one under various conditions.

3. Discovering high utility item sets with multiple minimum supports

Generally, association rule mining uses only a single minimum support threshold for the whole database. This model implicitly assumes that all items in the database have the same nature. In real applications, however, each item can have different nature such as medical datasets which contain information of both diseases and symptoms or status related to the diseases. Therefore, association rule mining needs to consider multiple minimum supports. Association rule mining with multiple minimum supports discovers all item rules by reflecting their characteristics. Although this model can identify meaningful association rules including rare item rules, not only the importance of items such as fatality rate of diseases but also attribute of items such as duration of symptoms are not considered since it treats each item with equal importance and represents the occurrences of items in transactions as binary values. In this paper, we propose a novel tree structure, called MHU-Tree (Multiple item supports with High Utility Tree), which is constructed with a single scan. Moreover, we propose an algorithm, named MHU-Growth (Multiple item supports with High Utility Growth), for mining high utility itemsets with multiple minimum supports. Experimental results show that MHU-Growth outperforms the previous algorithm on both real and synthetic datasets, and can discover useful rules from a medical dataset.

4. Top-k high utility pattern mining with effective threshold raising Strategies

In pattern mining, users generally set a minimum threshold to find useful patterns from databases. As a result, patterns with higher values than the user-given threshold are discovered. However, it is hard for the users to determine an appropriate minimum threshold. The reason for this is that they cannot predict the exact number of patterns mined by the threshold and control the mining result precisely, which can lead to performance degradation. To address this issue, top-k mining has been proposed for discovering patterns from ones with the highest value to ones with the kth highest value with setting the desired number of patterns, k. Top-k utility mining has emerged to consider characteristics of real-world databases such as relative importance of items and item quantities with the advantages of top-k mining. Although a relevant algorithm has been suggested in recent years, it generates a huge number of candidate patterns, which results in an enormous amount of execution time. In this paper, we propose an efficient algorithm for mining top-khigh utility patterns with highly decreased candidates. For this purpose, we develop three strategies that can reduce the search space by raising a minimum threshold effectively in the construction of a global tree, where they utilize exact and pre-evaluated utilities of itemsets. Moreover, we suggest a strategy to identify actual top-k high utility patterns from candidates with the exact and pre-calculated utilities. Comprehensive experimental results on both real and synthetic datasets show that our algorithm with the strategies outperforms state-of-the-art methods.

5. Mining top-k high utility item sets

Mining high utility itemsets from databases is an emerging topic in data mining, which refers to the discovery of itemsets with utilities higher than a user-specified minimum utility threshold min_util. Although several studies have been carried out on this topic, setting an appropriate minimum utility threshold is a difficult problem for users. If min_util is set too low, too many high utility itemsets will be generated, which may cause the mining algorithms to become inefficient or even run out of memory. On the other hand, if min_util is set too high, no high utility itemset will be found. Setting appropriate minimum utility thresholds by trial and error is a tedious process for users. In this paper, we address this problem by proposing a new framework named top-k high utility itemset mining, where k is the desired number of high utility itemsets to be mined. An efficient algorithm named TKU (Top-K Utility itemsets mining) is proposed for mining such itemsets without setting min_util. Several features were designed in TKU to solve the new challenges raised in this problem, like the absence of anti-monotone property and the requirement of lossless results. Moreover, TKU incorporates several novel strategies for pruning the search space to achieve high efficiency. Results on real and synthetic datasets show that TKU has excellent performance and scalability.

6. TOP-COP: Mining TOP-K strongly correlated pairs in large databases

Recently, there has been considerable interest in computing strongly correlated pairs in large databases. Most previous studies require the specification of a minimum correlation threshold to perform the computation. However, it may be difficult for users to provide an appropriate threshold in practice, since different data sets typically have different characteristics. To this end, we propose an alternative task: mining the top-k strongly correlated pairs. In this paper, we identify a 2-D monotone property of an upper bound of Pearson's correlation coefficient and develop an efficient algorithm, called TOP-COP to exploit this property to effectively prune many pairs even without computing their correlation coefficients. Our experimental results show that the TOP-COP algorithm can be

orders of magnitude faster than brute-force alternatives for mining the top-k strongly correlated pairs.

7. Hyper clique pattern discovery

Existing algorithms for mining association patterns often rely on the support-based pruning strategy to prune a combinatorial search space. However, this strategy is not effective for discovering potentially interesting patterns at low levels of support. Also, it tends to generate too many spurious patterns involving items which are from different support levels and are poorly correlated. In this paper, we present a framework for mining highly-correlated association patterns called hyperclique patterns. In this framework, an objective measure called h-confidence is applied to discover hyperclique patterns. We prove that the items in a hyperclique pattern have a guaranteed level of global pairwise similarity to one another as measured by the cosine similarity (uncentered Pearson's correlation coefficient). Also, we show that the h-confidence measure satisfies a cross-support property which can help efficiently eliminate spurious patterns involving items with substantially different support levels. Indeed, this cross-support property is not limited to h-confidence and can be generalized to some other association measures.

8. A fast perturbation algorithm using tree structure for privacy preserving utility mining

As one of the important approaches in privacy preserving data mining, privacy preserving utility mining has been studied to find more meaningful results while database privacy is ensured and to improve algorithm efficiency by integrating fundamental utility pattern mining and privacy preserving data mining methods. However, its previous approaches require a significant amount of time to protect the privacy of data holders because they conduct database scanning operations excessively many times until all important information is hidden. Moreover, as the size of a given database becomes larger and a user-specified minimum utility threshold becomes lower, their performance degradation may be so uncontrollable that they cannot operate normally. To solve this problem, we propose a fast perturbation algorithm based on a tree structure which more quickly performs database perturbation processes for preventing sensitive information from being exposed. We also present extensive experimental results between our proposed method and state-of-the-art algorithms using both real and synthetic datasets. They show the proposed method has not only outstanding privacy preservation performance that is comparable to the previous ones but also 5–10 times faster runtime than that of the existing approaches on average. In addition, the proposed algorithm guarantees better scalability than that of the latest ones with respect to databases with the characteristics of gradually increasing attributes and transactions.

9. Incremental high utility pattern mining with static and dynamic databases

Pattern mining is a data mining technique used for discovering significant patterns and has been applied to various applications such as disease analysis in medical databases and decision making in business. Frequent pattern mining based on item frequencies is the most fundamental topic in the pattern mining field. However, it is difficult to discover the important patterns on the basis of only frequencies since characteristics of real-world databases such as relative importance of items and non-binary transactions are not reflected. In this regard, utility pattern mining has been considered as an emergent research topic that deals with the characteristics. In real-world applications, meanwhile newly generated data by continuous operation or data in other databases for integration analysis can be gradually added to the current database. To efficiently deal with both existing and new data as a database, it is necessary to reflect increased data to previous analysis results without analyzing the whole database again. In this paper, we propose an algorithm called HUPID-Growth (High Utility Patterns in Incremental Databases Growth) for mining high utility patterns in incremental databases.

10. Mining top-k high utility sequential patterns

High utility sequential pattern mining is an emerging topic in the data mining community. Compared to the classic frequent sequence mining, the utility framework provides more informative and actionable knowledge since the utility of a sequence indicates business value and impact. However, the introduction of "utility" makes the problem fundamentally different from the frequency-based pattern mining framework and brings about dramatic challenges. Although the existing high utility sequential pattern mining algorithms can discover all the patterns satisfying a given minimum utility, it is often difficult for users to set a proper minimum utility. A too small value may produce thousands of patterns, whereas a too big one may lead to no findings. In this paper, we propose a novel framework called top-k high utility sequential pattern mining to tackle this critical problem. Accordingly, an efficient algorithm, Top-k high Utility Sequence (TUS for short) mining, is designed to identify top-k high utility sequential patterns without minimum utility. In addition, three effective features are introduced to handle the efficiency problem, including two strategies for raising the threshold and one pruning for filtering unpromising items. Our experiments are conducted on both synthetic and real datasets. The results show that TUS incorporating the efficiency-enhanced strategies demonstrates impressive performance without missing any high utility sequential patterns.

III. Modules

Admin Module

This module is used to view user details. Admin is used to view the item set based on the user processing details using association role with Apriori algorithm.

- Activities of Admin Module
- Admin Login
- Add Company
- Add Product
- Order Management
- Frequent Item Set Mining
- Computations
- Reports
- Graphs

User Module

In this module, privacy preserving data mining provides two related operations. In the first operation the data owner and the data miner are two different entities, and another, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold. In the first operation, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonymizing the data prior to its release. The main approach in this context is to apply data perturbation. He perturbed data can be used to infer general trends in the data, without revealing original record information. In the second operation, the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners.

- User Registration
- User Login
- User Catalog View

- User Shopping transactions
- User Mail/Inbox
- User Orders

Frequent Item set Mining

The task of frequent item sets mining is to find all item sets that have support greater than a given threshold. Frequent item sets is employed for finding association rules for a group of data items. Association rules show correlational relations of different items, which have numerous practical application. Association rule generation is usually split up into two separate steps:

1. A minimum support threshold is applied to find all frequent item sets in a database.
2. A minimum confidence constraint is applied to these frequent item sets in order to form rules.

Mining Phase

Given the large-scale dataset, we first sample the dataset and then compute the closed frequent item sets in the smaller sample using a traditional frequent item sets mining algorithm. We later estimate the length distribution of the sampled dataset and obtain the maximum length constraint, which is later used to shrink the dataset. Some elements out of the closed frequent item sets are removed from the source dataset if their supports are below the support threshold.

IV. IMPLEMENTATION

Implementation is the process of assuring the information system which is operational and then allowing user take its operation for its operations for use and evaluation. Implementation includes the following activities. Obtaining and installing the system hardware. Installing the system and making it run on its intended hardware. Providing users access to the system. Creating and updating the database. Train the users on the new system. Documenting the system for its users and for those who will be responsible for maintaining it in future. Making arrangements to support the user as the system is used. Transferring ongoing responsibilities for system from its developer to the operations or maintenance.

Results

V. Conclusion

The project proposed designing of new web based Secure Multiparty Computations Mining algorithm with pair wise learning methodology. The proposed protocol performs secure mining of association rules in portioned horizontally distributed databases and improves mining activities in significant way providing privacy and efficiency. The main ingredients of proposed protocol is providing a novel secure multi-party protocol for computing private subsets with high intermediate

security in distributed database. Another extensible ingredient is that the protocol has been tested in large real time databases giving positive result. In the project various concepts of data mining has been used implemented around for short time with a innovative computing technology and software of the new generation in the current decade providing a effective tool data mining analysis. As Data mining is a powerful tool for ethical considerations and providing summarized reports used for decision making. To provide and ensure the integrity of its use, and therefore the confidence of the users, the proposed research must adequately adjust and regulate itself concerning privacy issues.

Future Enhancements

As the project is to cover distributed databases, but the project has covered distributed databases of same platform. In our future enhancements we design secure multiparty computations for multi platform distributed databases with different operating systems and different RDBMS applications, and the project also to be extended for Mobile Apps, as the next generation performs all their Net activities through tabs and mobiles.

References

[1] Gentry C, Halevi S. Implementing gentry's fully-homomorphic encryption scheme. in: Advances in Cryptology–EUROCRYPT 2011. Berlin, Heidelberg: Springer press, pp. 129-148, 2011.

[2] Brakerski Z, Vaikuntanathan V. Efficient fully homomorphic encryption from (standard) LWE.in: Proceeding of IEEE Symposium on Foundations of Computer Science. California, USA: IEEE press, pp. 97-106, Oct. 2011.

[3] Qihua Wang, HongxiaJin. "Data leakage mitigation for discertionary access control in collaboration clouds".the 16th ACM Symposium on Access Control Models and Technologies (SACMAT), pp.103-122, Jun. 2011.

[4] Adam Skillen and Mohammad Mannan.On Implementing Deniable Storage Encryption for Mobile Devices.the 20th Annual Network and Distributed System Security Symposium (NDSS), Feb. 2013.

[5] Wang W, Li Z, Owens R, et al. Secure and efficient access to outsourced data. in: Proceedings of the 2009 ACM workshop on Cloud computing security. Chicago, USA: ACM pp. 55-66, 2009.

[6] Maheshwari U, Vingralek R, Shapiro W. How to build a trusted database system on untrusted storage.in: Proceedings of the 4th conference on Symposium on Operating System Design & Implementation-Volume 4. USENIX Association, pp. 10-12, 2000.

[7] Kan Yang, XiaohuaJia, KuiRen: Attribute-based fine-grained access control with efficient revocation in cloud storage systems. ASIACCS 2013, pp. 523-528, 2013.

[8] Crampton J, Martin K, Wild P. On key assignment for hierarchical access control.in: Computer Security Foundations Workshop. IEEE press, pp. 14-111, 2006.

[9] Shi E, Bethencourt J, Chan T H H, et al. Multi-dimensional range query over encrypted data. in: Proceedings of Symposium on Security and Privacy (SP), IEEE press, 2007. 350- 364

[10] Cong Wang, KuiRen, Shucheng Yu, and KarthikMahendraRajeUrs.Achieving Usable and Privacy-assured Similarity Search over Outsourced Cloud Data. IEEE INFOCOM 2012, Orlando, Florida, March 25-30, 2012

[11] Yu S., Wang C., Ren K., Lou W. Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing. INFOCOM 2010, pp. 534-542, 2010

[12] Kan Yang, XiaohuaJia, KuiRen, Bo Zhang, RuitaoXie: DACMACS: Effective Data Access Control for Multiauthority Cloud Storage Systems. IEEE Transactions on Information Forensics and Security, Vol. 8, No. 11, pp.1790-1801, 2013.

[13] Stehlé D, Steinfeld R. Faster fully homomorphic encryption. in: Proceedings of 16th International Conference on the Theory and Application of Cryptology and Information Security. Singapore: Springer press, pp.377-394, 2010.

[14] Junzuo Lai, Robert H. Deng ,Yingjiu Li ,et al. Fully secure keypolicy attribute-based encryption with constant-size ciphertexts and fast decryption. In: Proceedings of the 9th ACM symposium on Information, Computer and Communications Security (ASIACCS), pp. 239-248, Jun. 2014.

[15] Bethencourt J, Sahai A, Waters B. Ciphertext-policy attribute based encryption in: Proceedings of the 2007 IEEE Symposium on Security and Privacy (SP). Washington, USA: IEEE Computer Society, pp. 321-334, 2007.

[16] Liang Xiaohui, Cao Zhenfu, Lin Huang, et al. Attribute based proxy re-encryption with delegating capabilities. in: Proceedings of the 4th International Symposium on Information, Computer and Communications Security. New
York, NY, USA: ACM press, pp. 276-286, 2009.

[17] Pirretti M, Traynor P, McDaniel P, et al. Secure atrribute-based systems. in: Proceedings of the 13th ACM Conference on Computer and Communications Security. New York, USA: ACM press, pp. 99-112, 2006.

[18] Yu S., Wang C., Ren K., et al. Attribute based data sharing with attribute revocation. in: Proceedings of the 5th International Symposium on Information, Computer and Communications Security (ASIACCS), New York, USA: ACM press pp. 261-270, 2010.

[19] Sandhu R S, Coyne E J, Feinstein H L, et al. Role-based access control models. Computer, 29(2): 38-47, 1996.

[20] Tian X X, Wang X L, Zhou A Y. DSP RE-Encryption: A flexible mechanism for access control enforcement management in DaaS. in: Proceedings of IEEE International Conference on Cloud Computing. IEEE press, pp.25-32, 2009

[21] Di Vimercati S D C, Foresti S, Jajodia S, et al. Over-encryption: management of access control evolution on outsourced data. in: Proceedings of the 33rd international conference on Very large data bases. Vienna, Austria: ACM, pp. 123-134, 2007.

[22] Kan Yang, XiaohuaJia, KuiRen, RuitaoXie, Liusheng Huang: Enabling efficient access control with dynamic policy updating for big data in the cloud. INFOCOM 2014, pp.2013-2021, 2014

[23] Jia W, Zhu H, Cao Z, et al. SDSM: a secure data service mechanism in mobile cloud computing. in: Proceedings of 30th IEEE International Conference on Computer Communications. Shanghai, China: IEEE, pp. 1060-1065, 2011.

[24] D. Huang, X. Zhang, M. Kang, and J. Luo. Mobicloud: A secure mobile cloud framework for pervasive mobile computing and communication. in: Proceedings of 5th IEEE International Symposium on Service-Oriented System Engineering. Nanjing, China: IEEE, pp. 90-98, 2010.

[25] Benjamin Livshits, Jaeyeon Jung. Automatic Mediation of Privacy-Sensitive Resource Access in Smartphone Applications.USENIX Security, pp.113-130, Aug. 2013.