PRIVACY PRESERVING ASSOCIATION RULE MINING FROM HIGHLY SECURED OUTSOURCED DATABASES

K. Bhuvaneswari
Research Scholar, Department of Computer Science, PRIST University, Thanjavur, India

K. Saravanan
DEAN, Faculty of Computer Science, PRIST University, Thanjavur, India

ABSTRACT

Database outsourcing is becoming more commercial in the recent distributed and parallel systems. This paper considers Association Rule Mining, Frequent Itemset Mining and Privacy Preserving Mining. There is always a controversy between security and the flexibility. For better mining approaches flexibility is more needed from the Database servers but it increases its security risks on distributed network. In current trend of system setup on distributed clouds the Database servers are separated from the service providing web servers. It also extends the resource of the web servers where they can access more than one Database to analyze and retrieve results. At this scenario the web server acts as the intermediate between the Database servers and the client applications. It is responsibility of the web server to preserve privacy of both client and the Database server. This paper concentrates on both client side and Database server side privacy by introducing the algorithms No-Cache Rules Mining on client and Encrypted Database Access on Database server to preserve both client and server privacy. By internally it follows the traditional Association Rule Mining, Frequent Itemset Mining but in different manner.

Key words: Data Mining, Knowledge Discovery, Frequent Itemset Mining, Encrypted Database Access and No-Cache Rules Mining.


http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=8&IType=4
1. INTRODUCTION

1.1. Data mining
Data mining becoming popular technology and becoming wide range applied technology all over the computing world. With introduction of large number of social Medias and blogs there are large number of data’s gathered with can be analyzed to get some valuable results. Using the gathered data many useful things can be retrieved so every piece of data cannot be negligible. Data Mining is the field which is used for the knowledge discovery and it is the attempt to answer the need for the current scenario. With respect to the statistical methods mining technique is used for the analysis of statistical data. There can be various types of report analyzed in single set of data, for that a specific algorithm should be constructed. There are plenty of data mining algorithms available in the world of data engineering, but the improvement on those algorithms is an ongoing process to meet our requirements. There are many security issues may occurs in the process of data mining if the data is outsourced. The need for privacy is sometimes due to law or can be motivated by business interests. The data may be collected in many ways one from the database, from the data center or from the outsourced databases. A key utility of large databases today is research, whether it be scientific, or economic and market oriented. Thus, for example, the medical field has much to gain by pooling data for research as can even competing businesses with mutual interests. There is always be a controversy between the flexibility and security in the data mining. So steps should be taken to solve such an problems.

1.2. Knowledge Discovery and Data Mining
Knowledge discovery in data mining makes an mining algorithm to think by itself on the way in analyzing the future datum. For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on a quarterly basis. Knowledge discovery in Data mining used to make a computer algorithm to decide to act for a situation based on the data collected previously. This can be used to develop Artificial Intelligence applications for general purpose usage. The knowledge discovery can be used to apply in any kind of data like structured, semi-structured and unstructured data even on the moving datas over the networks. The implementation of the knowledge discovery makes the data mining technique to compute faster with the predetermined results. The data from all the fields like finance, marketing, health-care, hospital can be used for specific or separated purpose with the implementation of the knowledge engineering. It can be clearly visualized that the previous methods are not suitable for the current scenario of data mining. Classic frequent itemset mining and association rule mining algorithms, such as Apriori, Eclat and FP-growth, were designed for a centralized database setting where the raw data is stored in the central site for mining. Privacy concerns were not considered in this setting. Vaidya et al and Kantarciooglu et al. are the first to identify and address privacy issues in horizontally / vertically partitioned databases. In recent times number of privacy preserving mining solutions proposed to give importance to the data privacy. In their settings, there are multiple data owners wishing to learn association rules or frequent itemsets from their joint data. Most of the data owners are doubted to send the raw data to central domain for the security and privacy reasons. If each and every owner has some rows in joint database then it is known as horizontally partitioned database. There is the database model named vertically partitioned when every data owner has some columns in the joint databases. The privacy consideration is taken on such databases.
2. RELATED WORK

The review of related literature of the proposed work refers to data mining in a secured and effective way. Simon Fong et al. described the feature selection is designed particularly for mining streaming data on the fly, by using accelerated particle swarm optimization (APSO) type of swarm search that achieves enhanced analytical accuracy within reasonable processing time\(^1\). Once the features are extracted against faulted and unfaulted situations for each-phase, the data set is built to train the decision tree (DT), which is validated on the unseen data set for fault detection in the micro grid. Role mining has been extensively used to automatically generate roles for role-based access control. Nevertheless, the two core problems in role mining, role minimization and edge concentration, are both NP-hard. Gokilam et al. describe the algorithm processed to find the initial centroid of the cluster by arranging the data elements in sorting order then subtracting first element from the last element, resulting value are divided by number of cluster that values are taken intervals and form the cluster then use arithmetic mean to find initial centroid\(^2\).

Equivalence feature vector and matrix are defined firstly to update approximations of DTRS in different levels of granularity. To reduce the data set dimensionality, this model base the approach on the Principal Component Analysis (PCA), which is the statistical tool most commonly used in factorial analysis. Debi Prasad Mishra et al. proposed a Multi-Context Trajectory Embedding Model, called MC-TEM, to explore contexts in a systematic way. MC-TEM is developed in the distributed representation learning framework, and it is flexible to characterize various kinds of useful contexts for different applications\(^2\). Geosocial networks like Yelp and Foursquare have been rapidly growing and accumulating plenty of data such as social links between users, user check-ins to venues, venue geographical locations, venue categories, and user textual comments on venues. Lijun Dong et al. presents a data-mining-based intelligent differential relaying scheme for transmission lines, including flexible ac transmission system device, such as unified power flow controller (UPFC) and wind farms\(^3\).

Chuan Luo et al. described Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. This Work relaxes this assumption and expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM)\(^4\). Foued Saadaoui et al. proposed a novel methodology based on particle swarm optimization for storage systems sizing in wind power integrated microgrids. Traditional parallel algorithms for mining frequent itemsets aim to balance load by equally partitioning data among a group of computing nodes\(^5\). Xiao Zhang et al. described to address this problem by developing a data partitioning approach called FiDoop-DP using the MapReduce programming model\(^6\). The overarching goal of FiDoop-DP is to boost the performance of parallel Frequent Itemset Mining on Hadoop clusters. At the heart of FiDoop-DP is the Voronoi diagram-based data partitioning technique, which exploits correlations among transactions. Specifically, the spatiotemporal correlations among historical spectrum sensing data are exploited to form prior knowledge of channel availability probability, and Bayesian inference is used to derive posterior probability of channel availability\(^7\).

The algorithm name colored network-based model (CNBM) for characterizing economic behaviors, social relationships and the IATs between taxpayers, and generating a Taxpayer Interest Interacted Network (TPIIN)\(^8\). Xin-Lin Huang et al. proposed a two-stage methodology for online identification of power system dynamic signature using phasor measurement unit (PMU) measurements and data mining. This paper proposes an approach based on geometric data perturbation and data mining service-oriented framework\(^14\). The approach has three unique features compared to the existing approaches: 1) with geometric data perturbation, these protocols can work for many existing popular data mining algorithms, while most of other approaches are only designed for a particular mining algorithm; 2) both the two major factors:
data utility and privacy guarantee are well preserved, compared to other perturbation-based approaches; and 3) two of the three proposed protocols also have great scalability in terms of the number of participants, while many existing cryptographic approaches consider only two or a few more participants\cite{17}. Feng Tian et al. described the methodology have exploited sensors that are available on these devices, and have identified frequent behavioral patterns with a temporal granularity, which has been inspired by the way individuals segment time into events. These patterns are helpful to both end-users and third parties who provide services based on this information. This scalability makes analysis on resource constrained and small devices such as smart watches feasible\cite{15}.

Yaping Li et al. proposed the sparse computation for the K nearest neighbors algorithm (KNN), for graph-based machine learning techniques of supervised normalized cut and K-supervised normalized cut (SNC and KSNC) and for support vector machines with radial basis function kernels (SVM), on real-world classification problems\cite{11}. Joobin Gharibshsh et al. described the efficient probabilistic and deterministic verification approaches to check whether the server has returned correct and complete frequent itemsets\cite{18}. Their probabilistic approach can catch incorrect results with high probability, while our deterministic approach measures the result correctness with 100% certainty.

3. PROBLEM DEFINITION

One of the major challenges in Rule Caching is to develop a programming language for its software development. On one hand, this programming language should be sufficiently flexible and rich to allow new network applications, but on the other hand, existing caching techniques are relies on the client based cache process that leads to insecurity for the clients. The proposed idea is taken by two works intention to provide high level of security with minimum time consumption of data transfer. Primary work is No-Cache Rules Mining Find all set of items (itemsets) with minimum support. Generate all the rules with minimum confidence from the frequent itemsets. Then all the users’ details are maintained by server. Only server is keeps the Database not in client Sides. Secondary Work is Encrypted Database Used for Authorized client only access and retrieve the data. First client enter theirs query then server clarify searchable client is authorized are unauthorized. Last, client receives the matching database collection from the server and decrypts them with the key. Provide high level of security with minimum time consumption of data.

4. PROPOSED METHODOLOGY

4.1. SYSTEM MODEL

Following architecture diagram shows the environment setup of the distributed outsourced database data grid.
Figure 1 shows that the data owner and client are considered connected to each other when they each have a connected WiFi. Every client can search the query, these files are ready to be extracted for processing. After this extraction the data moved to webserver. At this point, each device can obtain input and output streams and data transfer can begin.

4.2. NO-CACHE RULES MINING

The proposed idea is taken by two works. The first work takes on No-Cache Rules Mining where the web server does not allow the database server are itself to create caches on the client side application mostly on web servers. But it is difficult to provide better accuracy on search without the client side cache. To avoid it the No-Cache Rules Mining uses session based analysis on user session to retrieve user details. These details are given by the servers not by the client on network so the client details are totally hidden from the server. Association Rules Mining is applied on the retrieved result from the multiple databases to ensure the accuracy.

ALGORITHM

**input:** search query q, where q ∈ {q1, q2, q3, .., qn}
**output:** r = {r1, r2, r3, .., rn}, where r is the result

**step 1:** get input query q

**step 2:** get user parameters p, l, id, b, ip
where p-platform, l-location, id-userid, b-browser, ip-internet protocol address

**step 3:** store p, l, id, b, ip in S, where S is session

**step 4:** get S from {S1, S2, S3, .., Sn} set of Sessions

**step 5:** extract p, l, id, b, ip from S

**step 6:** for query using q, p, l, id, b, ip

**step 7:** perform encrypted database access

**step 8:** retrieve {r1, r2, r3, .., rn}
4.3. NO-CACHE RULES BY SESSION CACHE

No-Cache selects a set of important rules from among the rules given by the controller to be cached in the Client Application, while redirecting the cache misses to the software switches. The output is a prioritized list of rules to store in the Client Application. The objective is to maximize the sum of the weights for accuracy that “user-agent” in the Client Application, while processing “user-agent” packets according to the semantics of the original prioritized list. As the packet traffic distribution of rules change over time, the set of cached rules chosen by our caching algorithms also change over time. This would mean periodically updating the Client Application with a new version of the policy cache. Session Cache has the ability to retrieve cached rules from the session data to handle and serve a client connected to the server. Through this the cost can be minimized, power consumption is lowered and security is improved along with the performance. The rule caching is done between the peer to peer where the peer receives the rule form the centralized server.

4.4. ENCRYPTED DATABASE ACCESS

Secondly this proposed work concentrates on the Encrypted Database Access on Outsourced database. The database providers assign accessible user id and password for the database consumer that is web servers. Now a day's database servers are not only secured with the user id and password, the data tables are preserved with the encrypted manner. This means that no unauthorized persons can access the database from nowhere. Even if they access they couldn't get the complete database. Applying Knowledge Discovery and Data Mining (KDDM) on encrypted database is the challenging one. For access and authentication the database server should send and validate the decryption key to the web servers. But Encrypted database Access avoid this procedures and follow difference methodologies. Instead of getting the data and decrypting it the web server encrypt the query and send to the database server. These methodologies are a light weight procedure which can save plentiful time on data retrieval.

ALGORITHM

input: query Q, where Q consist of q, p, l, id, b, ip
output: R ∈ [R1,R2,...,Rn] ∈ {r1,r2,r3,...,rn}

step 1: get user parameters q, p, l, id, b, ip
step 2: prepare Q, where Q is prepared query statement
step 3: encrypt Q→ EQ, EQ - encrypted Query
step 4: request(q=EQ)
step 5: get request(q← EQ)
step 6: parse(EQ)
step 7: get uid← id
step 8: retrieve uk← get(uid)
step 9: decrypt EQ(uk)→ Q
step 10: if(validate(Q))
retrive R← {r1,r2,r3,...,rn}
return R
step 11: else
return R← {empty}
5. RESULT AND DISCUSSION

To evaluate performance of the proposed algorithm the following parameters computation time, encryption time, key validation time and data retrieval time. The comparison is done with these parameters for existing and proposed methodology. The proposed experimental setup has number of virtual servers. The mining Servers, Authentication Server, Account Management Server, Key Management Server and File Server. They are connected in parallel to form a distributed cloud environment. All the servers are controlled by the centralized data mining servers. The test cloud server has the connection limitations of 8192. That means 8192 concurrent users can be in session at a time with the server. The key management server has the capability to store millions of key and file attributes similarly the account management server can manage millions of user account. The file server is limited with the file storage size of 1024 GB.

As the running time of association rule mining is only slightly higher than that of frequent Itemset mining. Only the results of association rule mining are presented. Show our solutions’ running time at the end (i.e. running time of mining) and data owner side (i.e. running time of pre-processing and decrypting) separately. As expected, our solutions are not as efficient as the most efficient algorithms / solutions of low privacy levels. However, they achieve a higher privacy level with an acceptable running time. Compared with the fastest algorithm’s running time, the cloud’s running time is about one order higher for most cases, while data owner’s running time is very low. This is the classic trade-off between privacy preserving and efficiency.

5.1. Performance analysis

The following graph shows the performance of the proposed system with the given resources. The performance criteria can be calculated for various configurations of environment setups.

The below graph 1 shows the user and time taken to complete the file upload and sharing it. The graph clearly explains that the server is capable of handling 2000 number of user at a time and they can upload file of size 1 GB within 16 milliseconds.

Evaluate the computational complexity, communication complexity and storage cost of our association rule mining and frequent Itemset mining solutions. In the evaluation, choose one of solutions and classic no-cache rules mining algorithms as the baseline. The former is chosen because it and our solutions achieve the same privacy level. In contrast, other solutions achieve lower privacy levels. Classic algorithms are chosen as baselines because they are the most efficient known solutions.

![Graph 1](image.png)

**Graph 1** Time taken for outsourced databases vs no. of users on server
5.2. Security analysis

Table 1 shows the prevented option is used to analysis the security level of various outsourced models.

<table>
<thead>
<tr>
<th>Outsourced Models</th>
<th>Prevented Intrusion Attack</th>
<th>Prevented Session Hijacking</th>
<th>Prevented Query Injection</th>
<th>Privacy presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third party auditing</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Triple Encryption Model</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Hybrid Encryption Model</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Our Scheme</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Yes – possible prevented option in various techniques
No – Not possible prevented option in various techniques

Privacy preserving association rule mining (our scheme) only have all prevented option like i) prevented intrusion attack ii) prevented session hijacking iii) prevented query injection iv) privacy presentation. Other techniques have only a specific prevented option so, the proposed model has the highest security compared with other techniques (third party auditing, triple encryption model, hybrid encryption model).

Unlike other system out proposed model have the capabilities to prevent intrusion and session hijacking attacks by validating user attribute. The query injection attack is prevented by validating the file attribute on the request along with the user attribute. As the key & file attribute and file and user attribute are stored in different servers the privacy of the user is preserved securely.

6. CONCLUSION

In this paper, there are two new generation algorithms No-Cache Rule Mining and Encrypted Database Access is proposed to provide security for both the client and the Data Owner in the distributed outsourced database network. The No-Cache Rule Mining is a developmental procedure for preserving the client personal details from the outsourced data owners. The Data owners can only get the information what the web server provided within the session details. Encrypted Database Access is an innovative mining security for Outsourced Database Servers, where the web server can get the valid result only if it can send the valid encrypted query. Instead using the time consuming process of encrypting and decrypting results the security is preserved by allowing web server to encrypt the query with valid key provided to the server. Thus the proposed methodology shows that the Time Efficiency, Accuracy/Relevancy are improved and privacy is preserved on the network.
REFERENCE


[2] A Combined Wavelet and Data-Mining Based Intelligent Protection Scheme for Microgrid Debi Prasad Mishra, SubhransuRanjanSamantaray, Senior Member, IEEE, and GezaJoos, Fellow, IEEE, 2015

[3] A Data-Centric Approach to Quality Estimation of Role Mining Results Lijun Dong, Kui Wu, Senior Member, IEEE, and Guoming Tang, Student Member, IEEE, 2016


[5] A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining FouedSaâdaoui, Pierre R. Bertrand, Gil Boudet, KarineRouffiac, FrédéricDutheil, and Alain Chamoux, 2014

[6] A General Multi-Context Embedding Model for Mining Human Trajectory Data Ningnan Zhou, Student Member, IEEE, Wayne Xin Zhao, Member, IEEE, Xiao Zhang, Ji-Rong Wen, Senior Member, IEEE, and Shan Wang Member, IEEE, 2013

[7] A Random Decision Tree Framework for Privacy-Preserving Data Mining JaideepVaidya, Senior Member, IEEE, BasitShafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, and David Lorenzi, 2014


[10] Detecting and Removing Web Application Vulnerabilities with Static Analysis and Data Mining Ibéria Medeiros, NunoNeves, Member, IEEE, and Miguel Correia, Senior Member, IEEE, 2015


[13] FiDoop-DP: Data Partitioning in Frequent Itemset Mining on Hadoop Clusters YalingXun, Jifu Zhang, Xiao Qin, Senior Member, IEEE, and Xujun Zhao, 2016


[16] Online Identification of Power System Dynamic Signature Using PMU Measurements and Data Mining, TingyanGuo, Student Member, IEEE, and Jovica V. Milanović, Fellow, IEEE, 2015

[17] Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation Keke Chen, Member, IEEE, and Ling Liu, Senior Member, IEEE, 2009


[21] Sparse Computation for Large-Scale Data Mining, Dorit S. Hochbaum, Philipp Baumann, 2016
