International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 2 (March-April, 2016), PP. 153-155

A SECURE DATA SHARING PLATFORM FOR OUTSOURCED ENTERPRISE STORAGE

Pranav Giri [1], Manisha Boddu [2], Ankita Bitla [3], Ritesh Bhat [4], Krupali Deshmukh [5]
Computer Engineering Department
MMIT
Pune, Maharashtra
pranavgiri070@gmail.com

***Abstract*-There are number of cloud data owners who outsource their data to third party data mining servers in order to get frequent item sets from their datasets. But the main concern arises here is the security that the client of low processing power cannot substantiate correct mining result. So we are considering the server that is probably untrusted and attempts to elude from verification by using its prior knowledge of the outsourced data. So in order to verify these kinds of servers we propose efficient probabilistic and deterministic verification approaches that analyze whether the server sends correct and complete frequent item sets of our dataset. The first approach is probabilistic approach that can catch erroneous results with high probability and second deterministic approach measures the result correctness. Here we also design efficient verification methods for both situations that the data and the mining set are updated. In order to evaluate our results we are using various datasets and developing our project using JAVA technology.**

***Index Terms*-Cloud computing, data mining-as-a-service, security, result integrity verification, Probabilistic, Deterministic.**

I. Introduction

Outsourcing data mining computations to a third-party service provider (server) offers a cost-effective solution mostly for data owners (clients) of limited resources. Such a structure introduces the data-mining-as-a-service (DMaS) paradigm. Now Cloud computing provides a natural solution for the DMaS structure. There are various active industry projects such as Google's Prediction APIs and Microsoft's Daytona project which provide cloud-based data mining as a service to clients. In our proposed work we focus on frequent item set mining as the outsourced data mining task to be carried out. Frequent item sets refer to a set of data values (e.g., product items) whose number of co-occurrences are greater than a given threshold. These FSP's i.e. frequent item sets mining has been proven important in many applications such as market data analysis and networking data study. There are various researches presented previously has shown that frequent item set mining can be computationally intensive due to the huge search space that is exponential to data size as well as the possible explosive number of discovered frequent item sets. So for those users of limited computational resources outsourcing their data for frequent item set mining to computationally powerful service providers (e.g., the cloud) is a natural solution.

In our proposed work we focus on the problem of verifying whether the server returned correct and complete frequent item sets. Here we have mention two points Correctness and Completeness. Here Correctness means that all item sets returned by the server are frequent, and completeness means that no frequent item set is missing in the returned result.

II. Literature Survey

While availing data mining as a service from cloud various data mining strategies are implemented with varying integrity issues. For the same purpose we have conducted a survey of following papers as follows,
1] Feida Zhu, Xifeng Yan, Jiawei Han, Philip S. Yu and Hong Cheng proposed "Mining colossal frequent patterns by core pattern fusion"

They studied the problem of efficient computation of a good approximation for the colossal frequent item sets in the presence of huge number of frequent patterns and proposed a model based on the concept of core pattern to evaluate the approximation quality of the mining results of Pattern-Fusion against the entire answer set. Their model also provides a general mechanism to find out the differences between two sets of frequent patterns.

2] W. K. Wong, David W. Cheung, Ben Kao, Edward Hung and Nikos Momoulis proposed "An audit environment for outsourcing of frequent item set mining"

This paper proposed a set of encryption methods for transactional databases that are suitable for outsourcing association rule mining Starting from a simple one-on-one substitution cipher which is susceptible to attacks so, they developed one-to-n item mapping scheme , which proved that the encryption technique cannot be broken by one-to-one decryption attacks and to test the security vulnerability of the proposed scheme against adversaries, they ran a generic algorithm that had background knowledge regarding the frequencies of the item set.

III. Proposed System

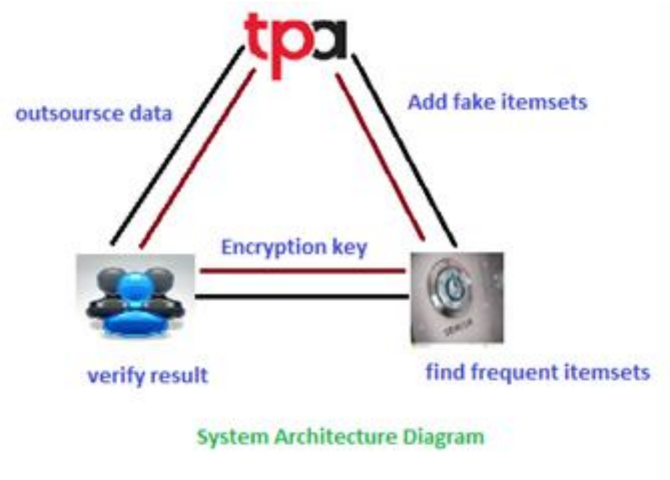
When a client outsources data through third party agency then there rises a major integrity concern that is a possibility that fake items may be included in result data set sent by server then the client accepts the wrong data without verification, to resolve this we are proposing a model for dataset verification at client side including 2 approaches as following,

1] *Probabilistic Approach:*
The key idea of our methods is to use real items to create a set of (in)frequent item sets, and use these (in)frequent item sets as proof to check the integrity of the server's mining result. We remove real items from the original dataset to construct artificial evidence infrequent

item sets (EIs), and then insert copies of items that exist in the dataset to construct artificial evidence frequent items (EFs).

2] *Deterministic Approach:*

Our deterministic approach is based on an efficient authenticated data structure. It enables proof based verification And allow access to this data to corporate network companies. We optimize the verification algorithm by reducing the number of proofs for both correctness and completeness verification. We show that a small number of proofs are sufficient to verify the correctness and completeness of a large set of frequent item sets and also catches incorrect/incomplete frequent item set mining answer with 100% probability. The key idea of our deterministic solution is to require the server to construct cryptographic evidences of the mining results. Both, definiteness and integrity of the mining results are measured against the proofs with 100% certainty.



System Architecture Diagram

Various components of our system are:

1) Outsourcing Dataset:

The data owner (client) outsources his/her dataset D, with the minimum support threshold minsup, to the service provider (server). The server performs frequent item set mining on the received dataset and returns the mining results to the client.

Finding Frequent Item set:

Frequent item set are found on server side using Prefixspan algorithm.

TPA (Third Party Application):

Sometimes the TPA adds fake item set to the original item set and sends it to the server.

Verify Frequent Item set:

Server sends frequent item set to the client using TPA, the client then verifies whether the received frequent item set is correct or not using MiniGraph approach.

IV. ALGORITHMS

In this paper we are using 2 algorithms U-Prefixspan algorithm and MiniGraph algorithm one for each approach

A. U-Prefixspan Algorithm:

In this paper we use U-Prefixspan algorithm to find frequent item sets at server side. Prefixspan mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. Moreover, prefix-projection substantially reduces the size of projected databases and leads to efficient processing. Following are the steps of this algorithm.

Seq U-prefixspan (αe, D|α),T|α)
Input: Current path αe, projected probabilistic databse D|α, element table T|α

```
VEC α e ← Φ
For each projected sequence Si|α € D|α do
Pr (Si | α e) ←0
For each instance Sij | α = Cpos, pr(Sij) > € Si|α do
Find its corresponding sequence Sij € D
If e € Sij [pos +1], len (Sij)] then
Pr (Si| α e) ← Pr (Si| α e) + Pr (Sij)
 ← min C ⩾ pos + 1 {Sij[c] =e}
Append () to Si|αe
If pr(Si|αe) > o then
Append Si|αe to VECαe
(tag,f α e)←PMFcheck(VEC α e)
If tag=true then
Output αe
T α e ← Prune (T| α , D| α e)
For each element l € T|αe do
 Seq U-Prefix span (αel, D|αe, T|αe)
```

Given a transaction dataset D that consists of n transactions, let I be the set of unique items in D. The support of the item set I _ I (denoted as supD(I)) is the number of transactions in D that contain I. An item set I is frequent if its support is no less than a support threshold minsup. Clearly the search space of all frequent item sets is exponential to the number of items in D. The (in)frequent item sets show the following monotone property. For any given infrequent item set I, any item set I s.t. I’ € I must be an infrequent item set. Similarly, for any frequent item set I, any item set I’ ⊂ I must be a frequent item set.

B. MiniGraph Algorithm:

We propose the MiniGraph approach to construct EFs.Intuitively; we pick a set of item seeds for EF construction.

STEP 1: Largest frequency infrequent 1-item sets is named as Iseed. If the item set is not present, then we select the infrequent 2-item set having the support value as the largest. Once the artificial transactions are added, Iseed is no more infrequent in the outsourced dataset.

STEP 2: Iseed is calculated from Dseed ⊆ D transactions, and then Dseed is used to construct MiniGraph G. The root of G correlates to Iseed. For every transaction in Dseed, there is an equivalent response of a non-root node in G. There is an edge from node N1 to node N2 in G if the transaction that node Nj corresponds to is the smallest superset of the transaction of node Ni. Every node is named with a number that shows the frequency of its corresponding transaction. Iseed is infrequent, thus frequency of every node in MiniGraph is less than minsup.

STEP 3: We select nodes from the second level of G as EFs. The item sets that the selected nodes correspond to are EFs. If there are not sufficient item sets to select, we add the next infrequent 1-item set of the largest frequency as another Iseed, and repeat Step 1 - 3, until we either find ` EFs or there is no infrequent 1 item set left. If it cannot return enough EFs, we will continue the same procedure from the lower levels of G. It is easy to see that the maximum number of EFs that can be constructed by MiniGraph approach is 2x - 2, where x is the number of infrequent items in D.

STEP 4: For each picked EF: (1) we compute its support f as the total frequency of all its descendants and itself.

V. CONCLUSION

In this paper, we present two integrity verification approaches for outsourced frequent item set mining. The probabilistic verification approach constructs evidence (in) frequent item sets. In particular, we remove a small set of items from the original dataset and insert a small set of artificial transactions into the dataset to construct evidence (in)frequent item sets. The deterministic approaches require the server to construct cryptographic proofs of the mining result. The correctness and completeness are measured against the proofs with 100% certainty. Our experiments show the efficiency and effectiveness of our approaches. An interesting direction to explore is to extend the model to allow the client to specify his/her verification requirements in terms of budget (possibly in monetary format) besides precision and recall threshold.

VI. ACKNOWLDGMENT

We would like to take this opportunity to express our deepest gratitude to all those who have supported us and helped us to make this paper a reality. We are highly indebted to Prof. Krupali Deshmukh for her guidance and constant supervision and also for her support in completing the paper. In conclusion, we would like to thank our parents and colleagues for their kind co-operation and investing their time, and willingly helping us out with their abilities to make this paper a reality.

REFERENCES

[1] Rakesh Agrawal and Ramakrishna Srikant. Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), pages 487–499, 1994.

[2] D. Burdick, M. Calimlim, J. Gehrke, and MAFIA: A Maximal Frequent Item set Algorithm for Transactional Databases. In Proceedings of the International Conference on Data Engineering (ICDE), 2001.

[3] Feida Zhu, Xifeng Yan, Jiawei Han, Philip S. Yu, and Hong Cheng. Mining colossal frequent patterns by core pattern fusion. In ICDE, 2007.

[4] W. K. Wong, David W. Cheung, Ben Kao, Edward Hung, and Nikos Mamoulis. An audit environment for outsourcing of frequent item set mining. In PVLDB, volume 2, pages 1162–1172, 2009.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev. in press.