ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

A Survey on Different Techniques for Verification of Frequent Itemset Mining in Cloud

Pandit Jyotikumari[1], Yash Singh[1], Shailesh Mahanta[1], Aditi Kalia[2], D.A. Phalke[2]

B.E. Student, Dept. of Computer Science, DYPCOE, Savitribai Phule Pune University, Pune, India. [1]

Assistant Professor, Dept. of Computer Science, DYPCOE, Savitribai Phule Pune University, Pune, India. [2]

ABSTRACT*:* Small companies and research teams that are dealing with colossal amounts of data have started relying on third party service providers (server) for providing them with cloud computing and data mining services. Outsourcing data to these service providers is effective and helpful but it also raises serious chances regarding inauthentic results returned by these servers to their respective clients. Results by these servers may potentially be very damaging for clients business or research. To avoid these deplorable conditions form taking place,the returned results by the servers need to be checked for any inauthenticity. This paper analyzes different methods and techniques for verification of results received by the server.

KEYWORDS*:* Cloud computing; data mining as a service; frequent item set mining; result integrity verification.

I. INTRODUCTION

Cloud computing means cloud which provide computing services which is hosted by other bodies. Nowadays, cloud computing has become more powerful then it was before. Frequent itemset mining has been proven important in many applications such as networking data study, human gene association and market data analysis study. Many analysis has been done that show frequent item set mining are often computationally intensive, attributable to the massive search area that's exponential to information size further because the doable explosive variety of discovered frequent item sets. Therefore, for those clients of limited computational resources, outsourcing frequent itemset mining to computationally powerful service providers (e.g., the cloud) is a natural solution.

There are a lot of companies in the world which take innumerable benefits from data mining. Some of these small sized companies lack financial resources for mining data themselves. Hence, these companies depend upon some third party service providers for mining.

These service providers offer an opportunity for small-scale businesses to take advantage of cloud computing services at a very reasonable price. These services however come with a few share of disadvantages, one of them being dishonest service providers. Clients might not even be aware that they are getting incorrect results by the service providers. This irresponsibility by third party service providers can be costly for client’s business or research. These inauthentic results have the potential to divert a client’s work in a wrong direction.

Hence, it is very important for client’s to verify the results they receive from the server. But again there are different problems like the verification cost should be less. It should not take considerably less time than actually computing the results. The server should not gain any inside information from the outsourced data. This paper analysis different techniques for verification of outsourced data. It also gives a better approach for finding trusted server.

Our end goal is to help clients verify the authenticity of the cloud servers and possibly making them much more vigilant in selecting their cloud service providers.

DOI: 10.15680/IJIRCCE.2017. 0503126

II. RELATED WORK

A perennial problem of dishonest clients, end users who modify their client software to return plausible results without performing any actual work[1]. Users commit such fraud even when the only motive is to higher their ranking on a website listing. Many projects deal with such fraud via redundancy the same work unit is sent to several clients and the results are compared for consistency[2].Apart from wasting resources, this provides little defence against colluding users. A related fear plagues cloud computing, where businesses buy computing time from a service, rather than purchasing, provisioning, and maintaining their own computing resources. Sometimes the applications outsourced to the cloud are so critical that it is imperative to rule out accidental errors during the computation[8]. Moreover, in such arrangements, the business providing the computing services may have a strong financial incentive to return incorrect answers, if such answers require less work and are unlikely to be detected by the client[2].The proliferation of mobile devices, such as smart phones and net books, provides yet another venue in which a computationally not so strong device would like to be able to outsource a task for computation, e.g., a cryptographic operation or a photo manipulation task , to a third party and yet obtain a strong assurance that the result returned is correct[9].Given the fact that many data mining applications (e.g., fraud detection and business intelligence) are so critical, it is important to provide efficient and practical methods to enable computationally-weak clients to verify the *result* integrity of data mining computations that are outsourced to a potentially dishonest service provider[10]. In this paper, we focus on verification of frequent itemset mining, a popular and important data mining problem

III. LITERATURE SURVEY

The requirement of creating endless amount of data and finding meaningful patterns from them has increased the necessity of data mining. Outsourcing data for mining to a third party server offers a financially effective alternative, particularly for clients with less computation power. It presents the data-mining-as-an administration (DMaS) worldview. It condenses on regular item set mining as the outsourced data mining undertaking. Frequent item set mining has been used in numerous applications for example; advertise data analysis, organizing data study, and human quality affiliation think about. Past research has shown that authenticity of frequent item set mining results can be done in many possible ways like construction of proofs. In this manner, for those customers that have less computational powers can outsource data to computationally effective third party server. Some past survey for related work has been done here:

In [1], authors have focused on frequent itemset mining as outsourced data mining. This paper deals with verifying the integrity of the outsourced frequent itemset mining. The proposed methodology here is adding some fake items to the original dataset that is outsourced. Adding fake items leads to construction of fake infrequent itemsets. Once the fake frequent itemsets are constructed the clients can check against these itemsets and verify the completeness and correctness of the mining result returned by the server. However, there is an assumption considered in this paper. The assumption is that the server should not have any background knowledge of the items in the outsourced data this gives the server a fair chance to either return true itemsets or return fake itemsets. Firstly, a probabilistic approach is proposed that catches mining result that do not return correct and complete results with high probability. In this approach a set of infrequent itemsets are constructed from real items. This infrequent itemsets are used as affirmation to verify the authentication of the server's mining result. MiniGraph approach is proposed by authors to construct Evidence Frequent Itemsets(EFs*)* and Evidence Infrequent itemsets *(*EIs).After construction of *EFs* and *EIs* robustness of the verification approach based on these two evidence itemsets. This is termed by the authors as Robustness Analysis. Secondly, a deterministic approach is proposed to catch any inauthentic mining result with 100% probability. The deterministic approach suggests the server to construct cryptographic proofs of the mining answers. Thirdly, for both approaches an effective approach is provided that deals with the updates on the mining setup as well as the outsourced data. Both of the approaches, the probabilistic approach and the deterministic approach have fair amount of advantages over one another. The probabilistic approach can attain the required verification assurance with small overhead. While the deterministic approach provides higher overhead than probabilistic approach, it also provides higher security.

In [2], authors have prescribed a cloud server is capable of giving inauthentic results also, verifiable computation is introduced to handle this issue by providing clients with proofs of the output with regards of dynamically chosen inputs. Data which is outsourced, say ‘F’ is computed on some random inputs. The returned output is not only expected to return the correct value but also to generate proof that randomly chosen data were indeed used to generate the results. However, the computation effort of verifying the proof should be less than calculating the actual outsourced data for output. The above-proposed system requires the client to preprocess the outsourced data for computing some auxiliary information which is expected to be very time consuming. Though preprocessing has to be done only once, it is advised by the publishers to use some trusted outsourcing service provider for the job. The approach doesn’t have to deal with preprocessing making it not at all mandatory to run the outsourced data on an already known verified and trusted server for the first time. We have the freedom to upload and verify the returned results on any server anytime.

In [3], authors have researched the pattern fusion algorithm for finding approximation to the colossal patterns. Colossal patterns are longer sequence with larger support set. Mining algorithms like FP Close, LCM2 and TFP (top-k) failed to complete execution of colossal patterns as these algorithms are stuck in mid-sized patterns. It has been seen that even closed and maximal frequent item set of microarray and frequent graph pattern are vast in size. Hence algorithms like FP Close, LCM2 and TFP(top-k) fails to mine complete frequent item set. Pattern fusion algorithm is able to find a result set, which is a very close approximation of the complete set of colossal patterns. Pattern fusion algorithm has the capability of finding shortcuts in search space, which helps pattern fusion algorithm to reach colossal patterns more quickly. Pattern fusion algorithm eliminates the drawbacks of breadth-first and depth-first search by traversing the tree in bounded breadth way. Pattern fusion algorithm generates the colossal patterns directly in mining process. Pattern fusion algorithm works in two phases. Firstly, initial pool is constructed which contains frequent item set of small size (say 3). These set of frequent set is complete. Secondly, the pattern fusion algorithm works in iterative manner. Pattern fusion algorithm stops when support set become zero for every super pattern. These patterns have a very high probability to be a descendant of the colossal pattern. These all patterns are fused together to generate a larger descendant of the complete set. Lastly quality evaluation model is proposed were this pattern fusion an approximation algorithm results are been tested. This model tests the quality of approximation solution against complete solution. Quality evaluation model can be used to test any approximation algorithm results.

In [4], authors expressed Practical Delegation of computation using multiple servers. Data mining is done on two or more cloud service provider. Out of these service providers one cloud service provider should be honest. Client can be oblivious to the fact that which server is honest. Client asks for mined results from two or more servers. If the results returned are same then the servers are honest. If the results seem to contradict each other, Referred Delegation of Computation (RDoC) protocol is used for determining malicious servers from honest servers. The limitations of the above proposed system is that client has to rely on multiple servers. And as these service providers are not free, the cost of verification increases inevitably. Furthermore, the client should have at least one honest server for giving authentic results also raises the inefficiency of the system. Our proposed system doesn’t deal with multiple servers. Thus the cost of system reduces drastically making the system economically feasible. The verification is only done on one server; therefore the necessity of a server being honest is completely eradicated.

In [5], authors have carried out a formal inquiry on the concept of Frequent Item set Mining techniques and its applicability on the Map-Reduce platform. This paper also shows the quantifiable methods and algorithms being used. The use of these methods considers customer behavioral analysis of buying products. The authors have also discussed about the existing system i.e. Apriori algorithm and its serious scalability problems and have proposed two new algorithms, i.e., Dis t-Eclat and Big-FIM for improving the efficiency of the application. The Dist-Eclat algorithm focuses on speed by using a simple load balancing scheme based on K-FIs, whereas Big-FIM focuses on mining very large databases by utilizing a hybrid approach. The authors have also studied several techniques for balancing the load of the K-FIs and discussed how using 3-FIs in combination with a basic Round-Robin allocation scheme resulted in a good workload distribution. The authors have also taken into account the customer behavior of buying products and ways to improve the retailer business performance by considering the Market Basket Analysis.

In [6], authors overviewed some of the existing frequent mining algorithms. It provides with the preliminaries of basic concepts about frequent pattern tree (fp-tree) and gives a cue to the recent developments in this particular area. The focus is on the recent fp-tree modifications and some other new techniques other than apriori, also various existing mining algorithms for frequent item sets are discussed in this paper. Among the methods discussed for frequent item set mining the Linear Prefix (LP-Tree) is found to be the simplest method for mining. This paper also gives an efficient method for reducing the search space when creating and mining the frequent patterns by considering k-item set at a time. It also provides a detailed analysis of some of the existing frequent pattern mining algorithms. Here the analysis also reveals that every method/algorithm has its set of advantages and disadvantages. The researchers can improve the efficiency by contributing some new techniques with the existing new methods. These algorithms can be reformed in a constructive way to lower the runtime and memory usage.

In [7], authors have put forward an efficient algorithm for mining association rules in considerably large databases and instigate the issue of mining a large collection of basket data type transaction for the association rules between sets of items with the least specified confidence and give an efficient algorithm. The project offers an efficient generation for large item sets by hash methods. It also gives an effective reduction on itemset scan and the option of reducing the number of databases scan required. The proposed hash and division based technique is highly efficient. The generation of candidate large item sets which are generated by HD, is lesser compared to many other methods like Apriori algorithm, DHP algorithm and the DIC algorithm. An extensive simulation study was conducted to evaluate the performance of the proposed algorithm. The simulation results show that the proposed technique is efficient than any existing algorithms. Simulation results show that the proposed algorithm is usually 15 to 38 percent better than DHP and also HD performs much better than the Apriori algorithm and the DIC algorithm.

IV. GOALS

Our main goal is to define a verification technique, which can verify completeness and correctness of the returned results. To achieve a cryptographic approach that doesn’t allow the database to be tampered by a server or a client. Another goal is to identify whether the servers providing services can be trusted or not. Dishonest servers are to be avoided by the clients in future.

V. IDENTIFIED CHALLENGES

The client should be able to perform the verification process in an efficient manner. The client shouldn’t be performing the frequent itemset mining task all again. Another challenge in verification are correctness and completeness. Correctness means that item which are present in frequent itemset mining results are actually frequent and there support counts are correct. Completeness means all the items that are actually frequent are included in the frequent itemset mining results. No frequent item is missed by the server performing task is checked. The verification process is performed on small numbers of item hence it is very important to provide correctness and completeness guarantee.

VI. GENERALIZED APPROACH

The system contains three modules:

1. **Evident Itemsets Generation**: Infrequent itemsets from frequent itemsets are selected. These infrequent itemsets are added in the outsourced data to avoid cheating at server’s end.
2. **Cryptography Proof Generation**: The proof generation algorithm constructs item- based inverted index E1 and Merle hash tree (T) of E. Then it keeps the root element value with itself and outsources the data for mining. When server receives the task of mining it performs the task and also constructs the proofs for every item that belongs to frequent item sets and also constructs the proof for every infrequent itemset.
3. **Verification:** The server returns the proof along with the mining result. The proof assures that the data hasn’t been tempered with and neither did server cheat on the task.

VII. CONCLUSION AND FUTURE WORK

In this paper, surveys of different papers are done which show different techniques of finding frequent item set mining and verification techniques. There are many approaches, which can help small-scale firm to outsource their data and gain valuable information with security. Since a single computer may require a lot of processing time for verification of the returned result, distributed networking can be fairly easily be implemented for verification of the frequent itemsets. Also multiple mining can be achieved. The areas where the task verification can be really helpful are linear programming, bioinformatics.

REFERENCES

1. Boxiang Dong, Ruilin Liu, Hui (Wendy) Wang, ”Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-mining-as-a-service Paradigm” IEEE transaction on services computing, Vol.9 pp.1939-1374, 2015.
2. Rosario Gennaro ,Craig Gentry and Bryan Parno,”Non-interactive Verifiable Computing: Outsourcing Computation to Untrusted Workers”,CRYPTO'10 Proceedings of the 30th annual conference on Advances in cryptology, pp.465-482, 2010.
3. Feida Zhu, Xifeng Yan, Jiawei Han, Philip S. Yu, Hong Cheng, ”Mining Colossal Frequent Patterns by Core Pattern Fusion”, Data Engineering, 2007. ICDE 2007, IEEE 23rd International Conference, pp.1063-6328, 2007.
4. Bryan Parno, Mariana Raykova and Vino Vaikuntanathan ,”How to Delegate and Verify in Public: Verifiable Computation from Attribute-Based Encryption” , TCC'12 Proceedings of the 9th international conference on Theory of Cryptography, pp.422-439, 2012.
5. Boxiang Dong, Ruilin Liu, Wendy Hui Wang,”Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee”, Data Mining (ICDM), 2013 IEEE 13th International Conference on , pp.1550-4786, 2013.
6. Kiran Chavan, Priyanka Kulkarni, Pooja Ghodekar, ”Frequent itemset mining for Big data”, International Conference on Green Computing and Internet of Things (ICGCIoT), pp.1365 – 1368, 2016.
7. O.Jamsheela, Raju G., “Frequent itemset mining algorithms: A literature survey”, Advance Computing Conference (IACC), 2015 IEEE International, pp.1099 – 1104, 2015.
8. Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Wendy Hui Wang,”Privacy-preserving data mining from outsourced databases”, Computers, Privacy and Data Protection: an Element of Choice, pp.411–426, 2011.
9. J.Wang, J. Han, and J. Pei, “Closet+: Searching for thebest strategies for mining frequent closed itemsets”,9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,InKDD’03, pp.236–245, 2003.
10. F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. Zaki,”CARPENTER: Finding closed patterns in long biological datasets”,Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, In KDD’03, pp.637–642, 2003.

BIOGRAPHY

Pandit Jyotikumari is a student in the Computer Science Department, D.Y. Patil College of Engineering, Akurdi, Pune -411044, India Affiliated to Savitribai Phule Pune University, Pune, Maharashtra state, India -411007. She is currently pursuing B.E. (Computer) Degree. Her research interests are cloud computing and data mining.

Yash Singh is a student in the Computer Science Department, D.Y. Patil College of Engineering, Akurdi, Pune -411044, India Affiliated to Savitribai Phule Pune University, Pune, Maharashtra state, India -411007. He is currently pursuing B.E. (Computer) Degree. His research interests are cloud computing and data mining.

Shailesh Mahanta is a student in the Computer Science Department, D.Y. Patil College of Engineering, Akurdi, Pune -411044, India Affiliated to Savitribai Phule Pune University, Pune, Maharashtra state, India -411007. He is currently pursuing B.E. (Computer) Degree. His research interests are cloud computing and data mining.

Mrs. Aditi Kalia is presently working as Assistant professor at Department of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Pune-411044, India Affiliated to Savitribai Phule Pune University, Pune, Maharashtra state, India -411007.
She has received M.E. degree in computers. Her area of interest is data mining and networking.

Mrs. D.A Phalke is presently working asAssistant Professor at Department of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Pune-411044, India Affiliated to Savitribai Phule Pune University, Pune, Maharashtra state, India -411007. She has received M.E. degree in computers and currently pursuing Ph.D. Her area of interest is data mining and information retrieval.