International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 1, January - February 2017 **ISSN 2278-6856**

A Survey on Verification of Outsourced Data Mining Computation

Umesh D Borse[1], Dr.Lalit V Patil[2]

[1] Department of Information Technology Smt. Kashibai Navale College of Engineering,(SKNCOE) Pune, India.

[2] Department of Information Technology Smt. Kashibai Navale College of Engineering (SKNCOE) Pune, India.

Abstract

Cloud computing is a computing where data is processed using computing resources (e.g. Server). A Client of weak computational power outsourced their data to cloud (e.g. Server) for computation. This introduces Data mining as Service paradigm. Outsourcing data to the server facing a critical problem of verification. Whether the server returned correct and complete computation result to the client. Other issue is related to the security or privacy of outsourced data for computation. Then another issue is related to revenue generation by cloud server. Most of the previous research paper focus on different algorithms associated with frequent itemset mining and data security. Work on result integrity verification is rare as compared to the data security and efficient mining algorithms. This paper focuses on the problem of result verification of outsourced computation of frequent itemset. By generating power set, frequent and infrequent itemset and by creating M tree of a frequent itemset, the result verification of frequent itemset mining computation is done. This paper put forward an idea in the end on verification of outsourced frequent itemset mining computation by analyzing numerous work but different authors in coming Section. The actual study reveals the efficiency and effectiveness of the verification approaches.

Keywords: Data mining as a service, Cloud computing, Frequent itemset mining, Result verification.

1. INTRODUCTION

Cloud computing is an Internet-based computing which provides shared computer processing resources and data to computers and other devices. On user demand access to shared computing resources (e.g., Computer Networks, Storage, Servers, Applications, and Services) which can be rapidly provisioned and released with minimal management effort. Cloud computing provides various capabilities to store and process users data in third-party data centers; Third party data centers may be located far from the user in the distance from across a city or across the world. In the simple word, cloud computing are nothing but storing and accessing the stored data and programs using the internet instead of computer's physical drive.

Data mining is the process of analyzing the data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of the analytical tools for analyzing data. It permits users to analyze data from many different angles, differentiate it, and summarize the relationships identified. Technically process of data mining used for finding patterns or co-relationship among dozens of fields in large relational databases. A general definition of frequent itemset mining is finding a frequent occurrence of itemset from a collection of 'n' no of itemset where support for that itemset is greater than or equal to the minimum support threshold. An itemset is nothing but a collection of one or more items. Item can be anything. For e.g. Milk, Butter, Eggs, Coke, etc.

Cloud computing, an emerging trend of provisioning scalable computing services, provides the opportunity that data mining is offered as an outsourced service[1]. Frequent itemset mining plays important roles in data analysis from a large database and many other significant data mining task[6]. Frequent itemset mining has been proven essential in many applications such as market data analysis, networking data study, and human gene association study[1].A frequent itemset mining algorithm runs on a set of text documents produced over a social network can display the central topic of discussion and pattern of usage of words in discussion threads and blogs[6]. But frequent itemset mining is very expensive computation task because of exponentially large data size. There for the weak computational power systems (Client) outsource their data to the computationally more powerful systems (Server).It has been shown that outsourcing data to a service provider brings several benefits to the data owner such as cost relief and a less commitment to storage and computational resources. This introduces the Data-Mining as Service (DMaS) paradigm[1]. Cloud computing provides a natural solution for the Data Mining as Service paradigm. In the cloud computing era, a better solution is to outsource the computations to a cloud service provider.

In addition to saving on overhead and labor costs, Most of the client outsources data to improved efficiency, greater productivity and the opportunity to focus on core products and functions of the business. Outsourcing is a practice used by different companies to reduce costs by transferring portions of work to outside suppliers rather than completing it internally. Outsourcing is an effective cost-saving strategy when used properly.

Although outsourcing data to the third party (server) is a viable option to the data owners (client), but client hesitate to place trust on cloud computing. Outsourcing data to the cloud raises few issues 1) First issue is about the integrity of the computed results[1]. There are many possible reasons for the cloud to cheat with computation result. The correctness integrity of mining results can be corrupted if the service provider is with random fault or not honest (e.g., lazy, malicious, etc.).Or Computation result can be damaged if the service provider is honest but makes mistakes in the mining process. 2) The second issue is related to the security or privacy of outsourced data for computation if service provider contaminates or tampers computational result[5] or server can access to valuable data of the owner and may learn or disclose sensitive information from it.3) The third issue is related to revenue generation by cloud server[2]. A cloud would like to improve its revenue by investing fewer resources while charging for more. A cloud server may provide some fake results instead of being spending its resources in computing the correct ones.

The existing research works contribute to practical data security by using different encryption scheme and mapping scheme on data content[7]. Most of the previous research paper focuses on various algorithms which are efficiently used in frequent itemset mining. Integrity verification outsourcing frequent itemset mining is rare and challenging. Out of these above mentioned issues this paper focus on the first issue i.e. result integrity verification of outsourced computation of frequent itemset mining.

This article has been classified as follows Section 2 is dedicated to related work, Section 3 Related work, Section 4 Acknowledgement and Section 5 Conclude efforts of the work.

2. LITERATURE SURVEY

This section of literature survey eventually reveals some facts based on thought analysis of many authors work as follows.
Boxiang Dong, Ruilin Liu[1] Proposes an idea of efficient probabilistic and deterministic verification approaches to verify whether the server has returned correct and complete frequent itemsets. In proposed word probabilistic approach is designed to catch mining result that does not meet a predefined requirement and Deterministic approach aims to find frequent itemset mining result are incomplete and incorrect with 100% probability. The basic idea is to create frequent and infrequent itemsets. Those infrequent itemsets are used as evidence to check server mining result. In this proposed

Table 1: Literature Survey on Verification of Outsourced Computation

Sr. No	Title of Paper	Methodology	Advantages	Disadvantages	Techniques Used
[1]	Trust But Verify: Verifying Result Correctness of Out-sourced Frequent Itemset Mining in Data-mining-as-a-service Paradigm	Propose efficient probabilistic and deterministic verification approaches	Better performance with respect to time to construct evidence and proof of itemset	Proposed work is limited to small set of itemset	Bilinear Pairing and Markel Hash Tree
[2]	Practical Delegation of Computation using Multiple Servers	Propose a protocol for any efficiently computable function and An adaptation of the protocol for the X 86 computation models.	Overhead of the protocol itself is lower	Limitation of using different Operating systems	Markel Hash Tree and X86 computational model.

[3]	How to Delegate and Verify in Public Verifiable Computation from Attribute-Based Encryption	Construct a Verifiable Computation scheme with a public delegation and public verifiability from any ABE scheme.	Allows arbitrary parties to submit inputs for delegation And verify the correctness of the results returned.	Requires running an expensive preprocessing to generate a secret key and evaluation key	Attribute-based Encryption technique
[4]	Verifiable Computation over Large Database with Incremental Updates.	Propose the notion of a verifiable database with incremental updates that can lead to huge efficiency gain.	The time cost is very low as compared to the existing scheme.	Not be applicable for the insertion operation due to the vector commitment.	Bilinear Pairing, VDB Scheme, Vector Commitment Scheme.
[5]	An Audit Environment for Outsourcing of Frequent Itemset Mining	Propose and develop an audit environment, which consists of a database transformation method and a result verification method.	A malicious miner cannot gain from performing any malicious actions and thus the returned mining result is both correct and complete with high confidence.	The artificial database takes more time to generate as an original database is large.	Artificial Itemset Planting

work no fake itemset are used to check mining result. Bilinear Pairing and Markel Hash Tree Technique are utilized for these approaches. This proposed methodology has better performance with respect to time to construct evidence and proof of itemset and with the limitation that proposed work is limited to a small set of an itemset.

Ran Canetti, Ben Riva, Guy N. Rothblum[2] Introduces a protocol for any efficiently computable function and an adaptation of the protocol for the X 86 computation models and a prototype implementation, called Quin. As key concept of this methodology is outsource the computation to N number of cloud and client takes number of return output out of that majority of same return output is the correct result of the outsourced computation. The client asks for the result of the function f(x) from two (or more) cloud servers. In case they make contradictory claims about f(x), the client engages in a protocol with each of the servers, at the end of which the client can efficiently determine the true claim as long as there is at least one honest server. In this protocol client searches for inconsistencies between the intermediate states of the two server's computations. In this Markel Hash Tree and X 86 computational models is used for computation. It having an advantage of an overhead of the protocol is itself low. This proposed methodology is having a limitation of using the different operating system.

Bryan Parno, Mariana Raykova, Vinod Vaikuntanathan[3] proposes public verifiable computation in two directions one is Public Delegation and other one is Public Verifiability, in which Public Delegation allows arbitrary parties to submit inputs for delegation, and Public Verifiability allows arbitrary parties (and not just the delegator) to verify the correctness of the results returned by the worker. The verifiable computation schemes with the public delegation and public verifiability have advantage that this methodology allows arbitrary parties to submit inputs for delegation, and verify the correctness of the results returned. Disadvantage of the proposed system is it requires running an expensive preprocessing to generate a secret key and evaluation key before delegation.

Xiaofeng Chen, Jin Li, JianWeng[4] introduce the notion of verifiable database with incremental updates (Inc-VDB).The update algorithm in Inc-VDB is an incremental one, i.e., the client can efficiently compute the new cipher text and the updated tokens with previous values, rather than from scratch. Thus, Inc-VDB schemes can lead to huge efficiency gain when the database undergoes frequent while small modifications. Also propose a general Inc-VDB framework by incorporating the primitive of vector commitment and the encrypt-then-incremental MAC mode of encryption. Besides, the proposed Inc-VDB scheme supports the public verifiability. Also introduce a new property called accountability for VDB schemes. That is, after the client detected the tampering of the server, the client should be able to provide a proof to convince the judge of the facts. Author proves that the proposed Inc-VDB scheme satisfies the property of accountability. The proposed methodology

is used Bilinear Pairing, Verifiable Database, and Vector commitment scheme. This proposed methodology provides an advantage that the time cost is minimal as compared to the existing scheme. But system has limitation that it not applicable to the insertion operation due to the vector commitment.

W. K. Wong David W. Cheung Edward Hung[5] Proposed and developed an audit environment, which consists of a database transformation method and a result verification method. The Figure 1 shows the architecture diagram of an audit environment. The proposed work is used for a database transformation method and a result verification method. The primary component verification environment is an Artificial Itemset Planting (AIP) technique. The main component of this audit environment is an artificial itemset planting technique. The approach to solve the integrity problem is to construct an audit environment. Essentially, an audit environment consists of (i) a set of transformation methods that transform a database T to another database U, based on which the service provider will mine and return a mining result R; (ii) a set of verification methods that take R as an input and return a deduction of whether R is correct and complete; (iii) auxiliary data that assist the verification methods. An interesting property of this approach is that the audit environment forms a standalone system. It is self-contained in the sense that the verification process can be done entirely by using only the auxiliary data that are included in the environment. In other words, the original database need not be accessed during verification. A malicious miner cannot benefit from performing any malicious actions and thus the returned mining result is both correct and complete with a high confidence, but artificial database takes more time to generate as an original database is large.

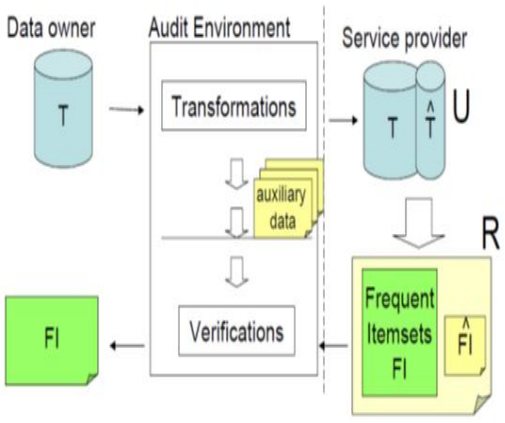

Figure 1: The architecture of the scheme [5]

3. RELATED WORK

In [7] author focuses on the integrity and verification issue in Uncertain Frequent Itemset mining problem during the outsourcing process, i.e., how the client verifies the mining results. The author specifically explores and extends the existing work on deterministic frequent itemset out-sourcing verification to the uncertain scenario. For this purpose, the author extends the existing outsourcing frequent itemset mining work to uncertain area w.r.t. the two popular uncertain frequent itemset definition criteria and the approximate uncertain frequent itemset mining methods. Specifically, authors construct and improve the basic/enhanced verification scheme with such different uncertain frequent itemset definition respectively.

To verify the outsourced computation[8] proposed different idea that client can use two or more cloud for the computation instead of using a single cloud for computation. If the client wants better assurance of the integrity of outsourced computation, then he can use more cloud for the guarantee. Minimum three clouds must be utilized for computation to get the better assurance of the integrity of computation. Client outsourced his data to the three different clouds for computation, now all cloud returns the output of computation out of that majority of the expected output to be his correct output. The disadvantage of this proposed idea is client have to pay more if the number of clouds used more. □

An author first introduces the rational lazy-and-partially-dishonest workers[9] in the outsourcing computation model. A new fair conditional payment scheme proposed for outsourcing computation that is only based on traditional electronic cash systems. The proposed construction uses a semi-trusted third party to get the fairness and efficiency. However, the third party is only involved in the protocol in case of any disputes. Compared with the other existing solutions the proposed solution is much more efficient. This advanced work has a disadvantage that Third Party is not fully trusted, and may collude with one party to obtain profits at the expense of the other party.

4. PROPOSED SYSTEM

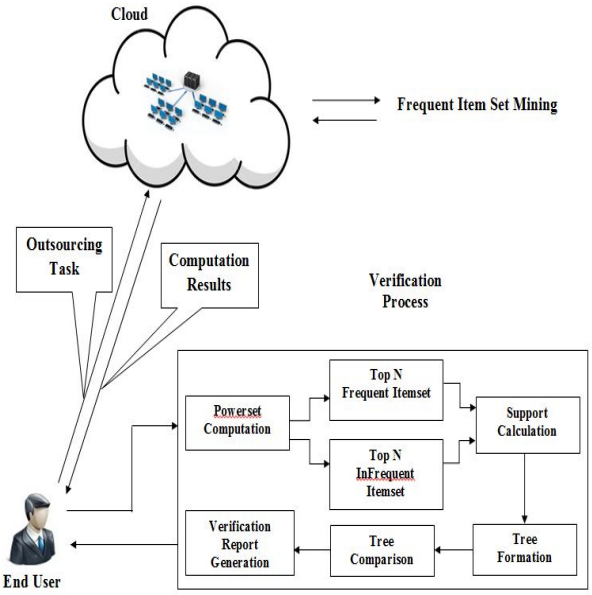

Figure 2: Proposed system architecture for verification of outsourced data mining computation.

The proposed work put forward an idea of verification of outsourced data mining computation of frequent itemset mining. The figure 2 shows proposed system architecture of outsourced data mining computation. The proposed system architecture describes the process of verification of outsourced data mining computation of frequent itemset. In which end user outsourced the data to the cloud for frequent itemset mining. Then cloud returns the computation result to the end user. The end user generates Power set from the small set of the outsourced data. Then compute for Top N frequent and Top N infrequent itemset by considering threshold value. Calculate the support for the Top N frequent and Top N infrequent itemset. The M tree algorithm is used to form tree on the basis of the support calculated. Same support value itemsets are clustered at node end. The same tree is formed for the return result from the cloud. Then end user compares both the trees to verify the return result received from the cloud. If the both tree matched then returned mining result from the cloud is correct else return result is incorrect. At the end verification report is generated for the end user.

5. CONCLUSION

The presented paper is literature review on different methodology used for the verification of frequent itemset mining computation. This presents the survey on different research paper, finding advantages, disadvantages and different issues related to verification. Simple problem definition for research work has been derived from above survey. An observed conclusion is need of enriched framework for the verification. The future scope is to design and implement a framework for verification of outsourced data mining computation of frequent itemset.

ACKNOWLEDGEMENTS

I am extremely thankful to my guide Dr. L. V. Patil for suggesting the topic for survey and providing all the assistance needed to complete the work. He inspired me greatly to work in this area. His guidance and discussions with him are valuable in realization of this survey.

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 Transactions 10.1109/ TSC.2015. 24363872015.

[2] Ran Canetti, Ben Riva, Guy N. Rothblum, "Practical Delegation of Computation using Multiple Servers" ACM 978-1-4503-0948-6/11/10 October 17–21, 2011.

[3] Bryan Parno, Mariana Raykova, Vinod Vaikunta – nathan, "How to Delegate and Verify in Public Verifiable Computation from Attribute-Based Encryption" International Association for Cryptologic Research LNCS 7194, pp. 422–439, 2012.

[4] Xiaofeng Chen, Jin Li, JianWeng, Jianfeng Ma, and Wenjing Lou, "Verifiable Computation over Large Database with Incremental Updates" European Symposium on Research in Computer Security (ESORICS 2014), LNCS8712, Springer, pp. 148-162, 2014.

[5] W. K. Wong, David W. Cheung, Edward Hung, "An Audit Environment for Outsourcing of Frequent Itemset Mining" VLDB Endowment, August 2428, ACM 2009.

[6] Zahra Farzanyar, Nick Cercone, "Efficient Mining of Frequent Itemsets in Social Network Data based on MapReduce Framework" ACM 978-1-4503-2240-9 /13/08 August 25-29, 2013.

[7] Qiwei Lu, Wenchao Huang, Yan Xiong, and Xudong Gong, "Integrity Verification for Outsourcing Uncertain Frequent Itemset Mining" arXiv:1307. 2991v2 [cs.DB] 12Jul2013.

[8] Ran Canetti, Ben Riva, Guy Rothblum, "Verifiable Computation with Two or More Clouds" Workshop on Cryptography and Security in Clouds, 2011.

[9] Xiaofeng Chen, Jin Li, and Willy Susilo, Senior Member, IEEE "Efficient Fair Conditional Payments for Outsourcing Computations" IEEE Transactions on Information Forensics and Security, Vol. 7, No. 6, December 2012.