RESEARCH ARTICLE OPEN ACCESS

Verifying Result Correctness of Outsourced Frequent Itemset in Data Mining through Probabilistic and deterministic approaches

*Dr. P. Julia Grace[1], Vanitha Suresh[2]

[1]*Assistant Professor & Research Supervisor, Department of Computer Science, JBAS College for women, Chennai.*

[2] *M.Phil., Computer Science Scholar, Mother Teresa women's University, kodaikanal.*

[3]*Assistant Professor, Department of Computer Applications, Madras Christian College, Chennai.*

*Corresponding author: *Dr. P. Julia Grace*

ABSTRACT: Cloud computing technology has enabled large organization to outsource data to a third-party service provider (server) for data mining and has provided a natural solution for the data-mining paradigms. However, outsourcing raises a serious security and privacy issue. Data outsourcing is the key task in recent days, for accessing services of the database processing. For the client with the weak computational capacity, there is no assurance or guarantee to know if the server returned correct mining results or not. This paper focuses on the specific task of frequent item set mining. The server that is potentially untrusted and tries to escape from verification by using its prior knowledge of the outsourced data is considered. An efficient probabilistic and deterministic verification approach to check whether the server has returned correct and complete frequent item sets is proposed. This proposed verification approach can catch incorrect results with high probability and the deterministic approach measures the result correctness with 100% certainty. The result shows that the proposed efficient verification method is desirable for both the cases. And the effectiveness and efficiency of our method gives secure and faster results using an extensive set of empirical results on real datasets.

Keywords: Data security, Data mining, probabilistic and deterministic approach.

Date of Submission: 27-07-2017 Date of acceptance: 14-08-2017

I. INTRODUCTION

The primary goal of this paper is to design efficient and robust integrity verification methods to catch such server that may return incorrect and incomplete frequent itemsets. The paper focuses on frequent Itemset mining as the outsourced data mining task and the problem of verifying whether the server returned correct and complete frequent itemsets. By correctness, it means that all itemsets returned by the server are frequent. By completeness, it means that no frequent Itemset is missing in the returned result. The deterministic approach is designed to catch any incorrect/incomplete frequent Itemset mining answer with cent percent probability. The key idea of our deterministic solution is to require the server to construct cryptographic proofs of the mining results. Many application areas as well as different popular Itemset mining variants are discussed. Since the focus of the paper is on finding frequent-itemsets, frequent-Itemset mining algorithms and respective optimizations is reviewed and frequent Itemset mining as the outsourced data mining task is considered. Informally, frequent itemsets refer to a set of data values (e.g., product items) whose number of co-occurrences exceeds a given threshold. Frequent Itemset mining has been proven important in many applications such as market data analysis, networking data study, and human gene association study.

II. LITERATURE SURVEY

Previous research has shown that frequent Itemset 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 itemsets. 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. Although it is advantageous to achieve sophisticated analysis on tremendous volumes of data in a cost effective way, end users hesitate to place full trust in cloud computing. This raises serious security concerns. One of the main security issues is the integrity of the mining result. There are many possible reasons for the service provider to return incorrect answers. For instance, the service provider would like to improve

its revenue by computing with less resources while charging for more. Since sometimes the mining results are so critical that it is imperative to rule out errors during the computation, it is important to provide efficient mechanisms to verify the result integrity of outsourced data mining computations.

III. CONTRIBUTIONS

Two integrity verification approaches for outsourced frequent Itemset mining are contributed. (a) The probabilistic verification approach constructs evidence in frequent itemsets. In particular, a small set of items from the original dataset is removed and insert a small set of artificial transactions into the dataset to construct evidence in frequent itemsets. (b) The deterministic approaches require the server to construct cryptographic proofs of the mining result. The correctness and completeness are measured against the proofs with far-reaching certainty. The experiments show the efficiency and effectiveness of these two approaches. The problem of verifying whether the server returned correct and complete frequent itemsets is considered. By correctness, it means that all itemsets returned by the server are frequent. By completeness, it means that no frequent Itemset is missing in the returned result. Initializing the research on integrity verification of outsourced frequent Itemset mining. Its basic idea is to insert some fake items that do not exist in the original dataset into the outsourced data. These fake items will construct a set of fake in frequent itemsets. Then by checking against the fake in- frequent itemsets, the client can verify the correctness and completeness of the mining answer by the server. Their assumption is that the server has no background knowledge of the items in the outsourced datasets, thus it has equal probability to cheat on the fake and true itemsets. Such assumption may not stand in practise, as the server may be able to possess prior knowledge of outsourced data (e.g., domain values of items and their frequency) from other sources. Apparently such server can escape easily from the verification mechanism that is based on using fake items. Moreover, the server may also be aware of the knowledge of verification techniques, and tries to escape verification by utilizing such knowledge. In particular, the following contributions are made. (1) The probabilistic approach to catch mining result that does not meet the predefined correctness/completeness requirement with high probability is designed. The key idea is to construct a set of in-frequent itemsets from real items, and use these in-frequent itemsets as evidence to check the integrity of the server's mining results.

(2) The deterministic approach to catch any incorrect/incomplete frequent Itemset mining answer with cent percent probability is designed. The key idea of our deterministic solution is to require the server to construct cryptographic proofs of the mining results. Both correctness and completeness of the mining results are measured against the proofs with cent percent certainty.

Finally for both probabilistic and deterministic approaches, the efficient methods to deal with updates on both the outsourced data and the mining setup are provided. The analytical results with extensive experiments evaluating the performance of our verification approaches is complemented. The experimental results show that the probabilistic approach can achieve the desired verification guarantee with small overhead, while our deterministic approach provides higher security guarantee with overhead more than the probabilistic approach.

IV. RELATED WORK

Data has experienced a variety of definitions, largely depending on the context of its use. For example, Information Science defines data as unprocessed information and other domains leave data as a representation of objective facts. In computer science expressions such as a data stream and packets of data are commonly used. Indeed the bandwidth of a signal is, metaphorically, the size of the pipe that the data can travel down. Other commonly encountered ways of talking about data include having sources of data or raw data. Data is placed in storage in databases, or fill a repository. It is discrete, it can pile-up, be recorded and manipulated, or captured and retrieved. Data can be mined for useful information or can extract data. Data or experience the tedium of data-entry can be looked into. Manipulation of data is possible through the interaction of these metaphors with the Thinking Is Object Manipulation metaphor and its associated mappings. It is this combination that allows us to rearrange data or to send it to another without difficulty. Yet despite this possibility of manipulation there is a limited amount of actions that can perform on data. Data is understood as discrete, atomistic, tiny packets that have no inherent structure or necessary relationship between them. In addition, data is now abundant resources in many circumstances, thus data can pile-up or "many managers find themselves drowning in data" (Mann, 2004). And since data is a substance, it can be measured, hence can have a lot of data or just a little. This substance and resource conceptualization differs crucially from a conceptualization involving larger objects such as products revealed from the apparent nonsense of expressions such as big data or small data. The conceptualizations of data as a flow in both a data stream and drowning in data occur due to our common experience of conflating a multiplicity of moving objects with a flowing substance. A real example includes watching a crowd of people from a

distance appear to behave as if they were one flowing mass. With so much data around us novel expressions such as these are likely to arise.

If data is seen as both a physical, external substance and a resource how then do the information be conceptualized? It turns out as there are many similar, but subtly different, representations with which the reason about information is described. This discussion will focus on the more recent, abstract sense of the word as distinguished from the particularistic sense, the little atoms of content, identified by Nurnberg. It is also noted that that whilst Nurnberg attributes the properties of information in the following discussion to the "reifications of the material properties of the documents that inscribe it," this does not make the metaphors used to understand information any less real to us than if they had evolved from purely embodied notions of information. Information is corpuscular, quantifiable, morselized, commoditized, objective and 'out there,' transferable, interconvertible, transparent, autonomous and measurable. It has shape and can be processed and accessed, generated and created, transmitted, stored, sent, distributed, produced and consumed, searched for, used, compressed and duplicated. Information can also be of different types with different attributes. It can be sensitive information, qualitative or quantitative information. Modern uses even extend its use to biological cells using and transmitting information, with cancers, for example, seen as spreading misinformation. It is described that the reliability and quality of information through the metaphor Attributes are Possessions.

V. METHODOLOGIES

Data Chunk Similarity and Compression: Firstly, the similarity models for our compression and clustering will be developed. The similarity model is critical and fundamental for deploying the data chunk based data compression because the similarity model is used for generating the standard data chunks.

Similarity Model for Text Data: For string type and text type similarity, a dual variable length hidden Markov model is used and updated in our work for calculating similarity between text data. Suppose there are a string pair p(str1, str2), and a time stamp series t= t1, t2, …, tn. The joint probability PR of each pair by the state time stamp series is defined. In the following, pi stands for the paper of text string with the similar time series stamp ti where i is the state.

$$\Theta = \tau\ (10) \equiv (\pi, \{\tau s\}, \{\{Os\}\})$$

The next Formula is the parameters consisting of states of the initial, transition, and output probabilities.

$$MAXt \in \tau\ (p)\ PR(p, t| \Theta)$$

Generally there are multiple state transitions that produce a given pair of strings. If the set of state sequences that produces a pair p is denoted as τ (p). Then the string similarity of the pair p is defined as the maximum alignment probability. With the transitions and states calculated with these two models can describe the string similarity. This compression technique can be considered as a type of common element similarity computation techniques.

VI. INFERENCES

The technique that is used to perform these feats in data mining is called modelling. Modelling is simply the act of building a model in one situation where one knows the answer and then applying it to another situation that you don't. For instance, if one were looking for a sunken Spanish galleon on the high seas the first thing one might do is to research the times when Spanish treasure had been found by others in the past. It is noted that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship's captains in that era. It is also noted these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand one can sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if a good model is found, treasure is found.

VII. FINAL FINDINGS

MapReduce is a framework for processing parallelizable and scalable problems across huge datasets using a large number of computers (nodes), collectively referred to as a cluster (if all nodes are on the same local network and use similar hardware) or a grid (if the nodes are shared across geographically and administratively distributed systems, and use more heterogeneous hardware). Computational processing can occur on data stored either in a file system (unstructured) or in a database (structured). MapReduce can take advantage of locality of data, processing data on or near the storage assets to reduce data transmission. "Map" function. The above idea has been implemented as working model in SBOA School and Junior College, Chennai. This module allows the staff to upload assignments, videos and text for the future usage of the students by getting an approval from the administrator. This product is run locally and cannot be accessed by everyone. It can be attached to the web server so that anyone connected to the internet can access it.

VIII. CONCLUSIONS

In this research paper, two integrity verification approaches for outsourced frequent Itemset mining are used. The probabilistic verification approach constructs evidence in-frequent itemsets. In particular, a small set of items is removed from the original dataset and insert a small set of artificial transactions into the dataset to construct evidence in-frequent itemsets. The deterministic approaches requires 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 her verification needs in terms of budget (possibly in monetary format) besides precision and recall threshold.

REFERENCES

[1]. Rakesh Agrawal and Ramakrishnan 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]. Laszlo Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy. Checking computations in polylogarithmic time. In STOC, pages 21–32, 1991.

[3]. Ran Canetti, Ben Riva, and Guy N. Rothblum. Verifiable computation with two or more clouds. In Workshop on Cryptography and Security in Clouds, 2011.

[4]. Kun-TaChuang, Jiun-LongHuang,andMing-SyanChen. Power-law relationship and self-similarity in the itemset support distribution: analysisandapplications. TheVLDBJournal,17:1121–1141,August 2008.

[5]. Rosario Gennaro, Craig Gentry, and Bryan Parno. Non-interactive verifiable computing: outsourcing computation to untrusted workers. In CRYPTO, pages 465–482, 2010.

[6]. Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Wendy Hui Wang. Privacy-preserving data mining from outsourced databases. In Computers, Privacy and Data Protection, pages 411–426. 2011.

[7]. S.Goldwasser,S.Micali,andC.Rackoff. Theknowledgecomplexity of interactive proof systems. SIAM Journal of Computing, 18:186– 208, February 1989.

[8]. Hakan Hacig¨um¨us¸, Bala Iyer, Chen Li, and Sharad Mehrotra. Executing sql over encrypted data in the database-service-provider model. In SIGMOD, pages 216–227, 2002.

[9]. Feifei Li, Marios Hadjieleftheriou, George Kollios, and Leonid Reyzin. Dynamic authenticated index structures for outsourced databases. In SIGMOD, pages 121–132, 2006.

[10]. Ruilin Liu, Hui Wang, Anna Monreale, Dino Pedreschi, Fosca Giannotti, and WengeGuo. Audio: An integrity auditing framework of outlier-mining-as-a-service systems. In ECML/PKDD, 2012.

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

Dr. P. Julia Grace. "Verifying Result Correctness of Outsourced Frequent Itemset in Data Mining through Probabilistic and deterministic approaches ." International Journal of Engineering Research and Applications (IJERA), vol. 7, no. 8, 2017, pp. 56–59.