BEST: International Journal of Management Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348–0513, ISSN (E): 2454–471X
Vol. 7, Issue 8, Aug 2019, 19–22

VALIDATING THE OUTSOURCED RESULTS OF FREQUENT ITEM SET MINING AS A SERVICE

M. KANIMOZHI[1] & M. AARTHI[2]

[1]Department of Computer Science, Prist University, Thanjavur, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science, Prist University, Thanjavur, Tamilnadu, India

ABSTRACT

"Cloud Computing" is playing a vital role by outsourcing data which is being stored in cloud server to 'n' number of third-party providers. The volume of information which is being exchanged between providers is charged and data-owner and service provider are getting benefited. Outsourcing will always become a big challenge; because nowadays data is shared between systems are attacked by man in the middle. In this paper, we had proposed certain techniques to validate whether the server had returned right mining result or not and we also concentrate particular on task of regular item-set mining. Un-trusted server which tries to elude from authentication, proposes probabilistic-validation and deterministic-validation method to validate whether server has returned right and complete results as recurrent item-sets. The proposed probabilistic-validation method is used to filter in-correct results returned from cloud server with high expectation, while our deterministic-validation method measures results with 98% accuracy. The obtained result shows accuracy of our proposed methods using comprehensive set of actual results on live data-set.

KEYWORDS*:* Item-set Mining, Outsourcing Item-set, Data-Mining as a Service, Outsourced Data-Securities, & Validation Results

1. INTRODUCTION

Generating very huge quantities of data-set is technically challenge able task for effective data-set mining. Outsourcing data-set mining computations to service provider acts as server which can be cost-effective one for data-owners who operates on very small number of limited resources. This problem in future establishes a concept called "Cloud Based Data Mining as a Service" (CBDMaS). For example: Amazon, Google, Microsoft, etc. are providing cloud based data mining as a service directly to consumers. Current journal mainly concentrates on recurrent item-set mining for outsourced data via third-party service provider. Recurrent item-set refers to a set of data-values whose co-occurrences exceed configured thres-hold value. Recurrent item-set mining are implemented and used in various software applications such as marketing survey analysis, network traffic analysis, heath care data analysis, etc.

Recent researches in recurrent item-set mining shows that it's computationally rigorous. Due to huge search space and estimated number of located recurrent item-sets. Users those who have limited number of computational resource will outsource recurrent item-set mining via third-party providers. Most prominent security issues are trustworthiness of data-mining results. Service providers will always try to maximize their revenue by computing with very less number of resources instead of billing more cost. It's important to validate liability, security and trustworthiness of outsourced data-mining computations by using any efficient mechanisms.

Current journal had addressed the problem of validating results which returned from the cloud-server are accurate and perfect recurrent item-sets. Inserting small amount of dummy items into outsourced data which does not prevail in the original data-set. Verifying right and perfect recurrent item-set mining by client, then an acknowledgement will be sent to cloud server. This situation confirms that cloud server does not have knowledge of recurrent items; it has equal possibility to cheat on fraud and true item-sets. From other sources server able to possess prior understanding of outsourced data-set. Moreover, cloud server may be aware of all validation techniques & try to elude from validation by utilizing such understanding.

First and foremost goal is implementing an accurate and strong trustworthiness validation method to catch the server which returns's wrong and partial recurrent item-sets. First approach is probabilistic-validation approach which is used to separately filter mining result which did not met pre-defined wrong/partial requirement with high probability and second approach is deterministic approach which is used to filter any wrong/partial recurrent item set mining answer with 98% probability. Both probabilistic-validation and deterministic-validation approaches provide accurate method to handle with updates on outsourced data and mining setup. The obtained result shows probabilistic-validation approach can achieve desired guaranteed validation, while our deterministic-validation approach provides guaranteed higher security more than probabilistic-validation approach.

2. GROUNDWORK

2.1. Recurrent Item-set Mining

Given large data-set 'T' which consist 'n' number business process logs; let 'Q' set of distinct item's in data-set 'T'? An item-set 'Q' is recurrent if support is not less than support threshold mins-up. Search-space of all recurrent item-sets is exponential to the number of items in 'T'.

2.2. Settings of Outsourcing

Data owner outsources data-set 'T', with minimum support thres-hold mins up, to third-party provider. Cloud server performs recurrent item-set extraction on received data-set and returns extracted results to service provider. Privacy-preserving recurrent item-set-mining algorithm used to encrypt data-set. Cloud server will respond with accurate item-sets in encrypt format, so that man in the middle attack cannot be performed.

2.3. Untrusted Server

There are 2 different types of un-trusted servers that may respond with more results. First server possesses backstage understanding outsourced dataset which includes domain items & their rate of occurrence information. Second server aware of frequency information of both items and business process logs, details of validation steps. We had implemented various validation approaches to capture two type of server.

3. PROBABILISTIC APPROACH

The proposed methodology used to build unusual item-sets from real-time items and use unusual item-sets as evidence to check trustworthiness of extraction results. We had removed real-items from original data-set to build artificial evidence of un-usual item-set & insert copies of items that does not exist in data-set to build artificial evidence of recurrent items.

4. DETERMINISTIC APPROACH

Our deterministic approach uses efficient validated data-structure, which built up on standard “Merkle Trees” and “Bilinear-map Accumulators” & enables proof based validation scheme. Validation algorithm optimized by minimizing number of proofs for right & whole validation scheme. Small number of proof is sufficient to validate right & whole large set of recurrent item-sets.

5. RESULTS AND DISCUSSIONS

5.1. Probabilistic and Deterministic Approaches

We had experimented and compared performance of proposed probabilistic-validation and deterministic-validation approaches. Table: 1 shows comparison results on ‘T3’ data-set of various settings. We had selected only error ratios of 1%, and vary probabilistic-validation guarantee thres-hold from 90% to 100%. If probability is equal to 100% which corresponds to proposed deterministic-validation approach. Table: 1 shows details of comparison results. Deterministic-validation approach brings high overhead to server side than probabilistic-validation approach.

Table 1: Time Performance: Deterministic vs. Probabilistic Approaches

Error ratio	Integrity Prob.	Type	Client	Server	
			Verify	Proof prep.	Mining
1%	90%	R	N/A	0	0.042
		M	1.433	0	1024.53
1%	95%	R	N/A	0	0.042
		M	0.945	0	1204.59
1%	99%	R	N/A	0	0.042
		M	0.689	0	1498.67
1%	100%	R	0.000628	1660.12	0.5707
		M	0.4123	2785.6	0.5707

5.2. Comparison with Existing Work

We had proved that evidence patterns constructed by en-coding approach are identified even by an attacker without prior understanding of data. Our probabilistic-validation approach is robust against attacks and it is more efficient. Existing approaches shows that it takes 1 second to generate one evidence pattern, while proposed approach takes only 900 seconds to generate 7900 evidence item-sets overall. Goodrich et al., proposed an efficient crypto-graphic based approach to validate result trustworthiness of web-content searching by using same set intersection validation protocol. Time taken on server to build crypto-graphic proof for a query that involves 2 terms is between 0.45 to 0.9 seconds. Proposed deterministic-validation approach requires 0.6 seconds to build crypto-graphic proof for an item-set of length 2 at an average.

6. CONCLUSIONS

We had proposed 2 different trustworthy validation approaches for outsourced recurrent item-set mining. First one is probabilistic approach & second deterministic-validation approach. Probabilistic validation approach constructs proofs of unusual item-sets. We had removed small amount of items from original dataset and inserted small amount of artificial transactions into data-set to build evidence unusual item-sets. Deterministic-validation approach requires cloud server to build crypto-graphic proofs of mining result. Right & whole outsourced recurrent item-set mining are measured against crypto-graphic proofs with 98% accuracy. Results obtained show accuracy & effectiveness of proposed approaches.

REFERENCES

1. 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.
2. Ranade, R., & Kanwar, K. (2014). Examining synergistic effects of TDZ and TIBA on adventitious shoot induction in Dianthus caryophyllus L. leaf explants. Int J Agric Sci Res, 4(2), 17–26.
3. Bryan Parno, Mariana Raykova, and Vinod Vaikuntanathan. How to delegate and verify in public: verifiable computation from attribute based encryption. In TCC, 2012.
4. Michael T. Goodrich, Charalampos Papamanthou, Duy Nguyen, Roberto Tamassia, Cristina Videira Lopes, Olga Ohrimenko and Nikos Triandopoulos Efficient Verification of Web-Content Searching Through Authenticated Web Crawlers in PVLDB, volume 5, pages 920–931, 2012.
5. Devi, S. V. S. G. (2014). A survey on distributed data mining and its trends. International Journal of Research in Engineering & Technology (IJRET), 2(3), 107–120.
6. Siavosh Benabbas, Rosario Gennaro, and Yevgeniy Vahlis. Verifiable delegation of computation over large datasets. In CRYPTO, 2011.
7. Ran Canetti, Ben Riva, and Guy N. Rothblum. Verifiable computation with two or more clouds. In Workshop on Cryptography and Security in Clouds, 2011.
8. 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.
9. Charalampos Papamanthou, Roberto Tamassia, and Nikos Triandopoulos. Optimal verification of operations on dynamic sets. In CRYPTO, 2011.
10. Dario Fiore and Rosario Gennaro. Publicly verifiable delegation of large polynomials and matrix computations, with applications. In CCS, 2012.
11. Ran Canetti, Ben Riva, and Guy N. Rothblum. Practical delegation of computation using multiple servers. In CCS, 2011.
12. Haque, S. A. A. The Heterogeneous Realm of South Asian Literature in Determining the Public Sphere of Partition.
13. Srinath Setty, Andrew J. Blumberg, and Michael Walfish. Toward practical and unconditional verification of remote computations. In HotOS, 2011.
14. Ian Molloy, Ninghui Li, and Tiancheng Li. On the (in)security and (im)practicality of outsourcing precise association rule mining. In ICDM, pages 872–877, 2009.
15. Yadav, S., & Sharma, G. Improvisation of Data Mining Techniques in Cancer Site Among Various Patients Using Market Basket Analysis Algorithm.
16. Charalampos Papamanthou, Roberto Tamassia, and Nikos Triandopoulos. Authenticated hash tables. In CCS, pages 437–448, 2008.