Efficient and robust integrity verification methods of frequent itemset mining using data mining as a service

Miss.Chaudhari Bhagyashri N[1], Prof.A.N.Nawathe[2],

[1]*Department of Computer Engineering, Amrutvahini College of Engineering,Sangamner,Savitribai Phule Pune University, Maharastra,India*
[2]*Associate Professor, Amrutvahini College of Engineering, Sangamner,Savitribai Phule Pune University, Maharastra,India*

Abstract - *Now a days, cloud computing familiarizes the computing technique in which is data outsource to third-party service provider for data mining process. Outsourcing, however, set a major security issue: how can the client of delicate computational power verify that the server returned correct mining result? And give to respective result. To aim on the specific work of frequent itemset mining in data mining. To consider the server that is possible unauthorized and tries to escape from verification by using its important knowledge of the outsourced data. To Proposed effective probabilistic and deterministic verification approaches to check whether the server has returned correct and complete frequent itemsets. Our probabilistic approach can catch incorrect results with high probability, while our deterministic technique compute the result of correctness with 100 percent certainty. To developed effective verification methods for both cases that the data and the mining setup are updated. To analyze the systematic and efficiency of our methods using an extensive set of empirical results on real datasets.*

Key Words: Cloud computing, data mining as a service, security, result integrity verification, efficiency.

1. INTRODUCTION

Existing system are the nearby ours. It has been proven that the evidence patterns constructed by the encoding technique in can be analysed even by an attacker without advance knowledge of the data. To state that our probabilistic verification method is long lasting against the attack. Our probabilistic approach is more effective. Display that it may take 2 seconds to create one evidence pattern, while our approach only takes 600 seconds to create 6900 evidence item sets. To Proposed an well planned cryptographic point of view to verify the result integrity of web-content searching by using the same set intersection verification protocol as ours. It state that the time spent on the server to make the evidence for a query that involves two terms. Our deterministic method requires seconds to make the evidence for an item set of length average, which is differentially to the performance. Proposed a set intersection verification protocol to prove that E is the correct intersection of system. Whether the coefficients are calculate correctly by the server. Any given accumulate value is indeed calculated from the original dataset. When satisfies the subset condition by using satisfies the intersection completeness condition. To exclude the details of evidence of construction and verification due to restricted area.

2. PROBLEM STATEMENT

The Problem is to determine how to handle an efficient probabilistic and deterministic verification approaches to check whether the server has returned correct and complete frequent itemset.

3. REVIEW OF LITERATURE

R. Agrawal and R. Srikant, [2] To state that the problem of locating association rules between items in a large database of sales transactions. To present two new algorithms for solving this problem that are basically different from the known algorithms. Empirical execution display that these algorithms efficiency the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. To show how the best quality of the two proposed algorithms can be integrated into a hybrid algorithm, called AprioriHybrid. Scale-up experiments show that prioriHybrid scales linearly with the number of communication. AprioriHybrid also has superior scale-up properties with respect to the communication size and the number of items in the database.

L. Babai, L. Fortnow, L. A. Levin, and M. Szegedy, [3]Inspire by Manuel Blum's concept of instance checking, To consider new, very fast and generic system checking of operation. Our results exploit recent advances in interactive evidence system protocol [LFKN92], [Sha92], and especially the M IP = N EXP protocol from [BFL91]. To show that every nondeterministic calculating task S(x; y), defined as a polynomial time relation between the instance x, representing the input and output integrated, and the witness y can be updated to a task S 0 such that: (i) the same instances remain granted; (ii) each instance/witness pair becomes analysable in polylogarithmic Monte Carlo time; and (iii) a witness satisfying S 0 can be calculated in polynomial time from a witness satisfying S. Here the occurrence and the description of S have to be provided in error-correcting code (since the checker will not notice slight changes). A updating of the M IP evidence was required to achieve polynomial time in (iii); the earlier mechanism yields N O (loglogN) time only. This output becomes significant if software and hardware ability are regarded as a important cost factor. The polylogarithmic checker is the only part of the system that needs to be reliability; it can be hard wired. The checker is tiny and so probably can be processed and checked o-line at a moderate cost. In this setup, a single reliable PC can monitor the work of a herd of supercomputers accessing with manageable extremely powerful but unreliable software and unproven hardware. In another contribution, to display that in polynomial time, every formal mathematical evidence can be transformed into a transparent evidence, i.e. a evidence verifiable in polylogarithmic Monte Carlo time, assuming the candidate" is given in error-correcting code. In fact, for any " > 0, we can transform any evidence P in time kP k1+" into a transparent evidence verifiable in Monte Carlo time (log kP k)O(1=").As a by-data, to developed a binary error correcting code with very efficient error-correction. The code transforms messages of length N into code words of length N 1+"; and for strings within 10per of a valid password, it allows to retrieve any bit of the unique password within that distance in polylogarithmic ((log N)O(1=")) time.

K.-T. Chuang, J.-L. Huang, and M.-S. Chen, [5]to distinguish and investigate that the power-law relationship and the self-comparative marvel show up in the itemset bolster circulation. The itemset bolster dispersion alludes to the conveyance of the tally of itemsets versus their backings. Investigating the qualities of these normal marvels is helpful to numerous applications, for example, giving the course of tuning the execution of the continuous itemset mining. Be that as it may, because of the unstable number of itemsets, it is restrictively costly to recover loads of itemsets before we distinguish the attributes of the itemset bolster dissemination in focused information. Thusly, we additionally propose a substantial and financially savvy calculation, called calculation PPL, to concentrate qualities of the itemset bolster conveyance. Besides, to completely investigate the upsides of our revelation, we additionally propose novel components with the assistance of PPL to take care of two critical issues: (1) deciding an unpretentious parameter for mining rough successive itemsets over information streams; and (2) deciding the adequate example measure for mining incessant examples. As approved in our test comes about, PPL can proficiently and unequivocally recognize the attributes of the itemset bolster appropriation in different genuine information. Likewise, experimental reviews additionally exhibit that our systems for those two testing issues are in requests of extent superior to anything past works, demonstrating the conspicuous preferred standpoint of PPL to be a vital preprocessing implies for mining applications.

R. Gennaro, C. Gentry, and B. Parno, [6] Undeniable Computation empowers a computationally frail customer to ”outsource” the calculation of a capacity F on different sources of info x1,...,xk to at least one specialists. The specialists give back the after effect of the capacity assessment, e.g., yi= F(xi), and also a proof that the calculation of F was completed accurately on the given esteem xi. The confirmation of the evidence ought to require considerably less computational exertion than registering F(xi) starting with no outside help. To display a convention that permits the specialist to give back a computationally-solid, non-intelligent confirmation that can be checked in O(m) time, where m is the bit-length of the yield of F. The convention requires a one-time pre-handling stage by the customer which takes O(|C|) time, where C is the littlest Boolean circuit figuring F. Our plan likewise gives information and yield security to the customer, implying that the laborers don't take in any data about the xi or yi values.

F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and W. Hui Wang,[7]to state that Provided by improvements, for example, distributed computing, there has been extensive late enthusiasm for the worldview of information mining-as-administration: an organization (information proprietor) ailing in ability or computational assets can outsource its mining needs to an outsider specialist organization (server). Be that as it may, both the outsourced database and the information extricate from it by information mining are viewed as private property of the information proprietor. To secure corporate protection, the information proprietor changes its information and boats it to the server, sends mining questions to the server, and recoups the genuine examples from the separated examples got from the server. In this paper, we concentrate the issue of outsourcing an information mining undertaking inside a corporate security protecting system. To propose a plan for security saving outsourced mining which offers a formal insurance against data updation , and demonstrate that the information proprietor can recuperate the right information mining comes about productively.

4. EXISTING SYSTEM

Existing work are the closest to ours. It has been proven that the evidence patterns constructed by the encoding method in can be identified even by an attacker without knowledge of the data. We argue that our probabilistic verification approach is robust against the attack. Our probabilistic approach is more efficient. Shows that it may take 2 seconds to generate one evidence pattern, while our method only takes 600 seconds to generate 6900 evidence itemsets. To Propose an efficient cryptographic approach to verify the result integrity of web-content searching by using the same set intersection verification protocol as ours. It shows that the time spent on the server to construct the proof for a query that involves two terms. Our deterministic approach requires seconds to construct the proof for an item set of length average, which is comparable to the performance.

5. PROPOSED SYSTEM

Proposed a set intersection verification protocol to verify that E is the correct intersection of protocol. Whether the coefficients are computed correctly by the server. Whether any given accumulation value is indeed calculated from the original dataset. Whether satisfies the subset condition by using whether satisfies the intersection completeness condition. To omit the details of proof construction and verification due to limited space.

6. MATHEMATICAL MODEL

Input :
V(Z) = {v1,v2,v3....vn}
C(Z) = {c1,c2,c3....cn}
P(Z) = {p1,p2,p3,...pn}
A(Z) = {a1,a2,a3.....an}
where ,
V= Number of Verifier
C= Number of Customer
P= Number of Product
A= Number of Attacker
U= Number of User
Output :
The union of two sets are
C(Z) ∪V (Z) :

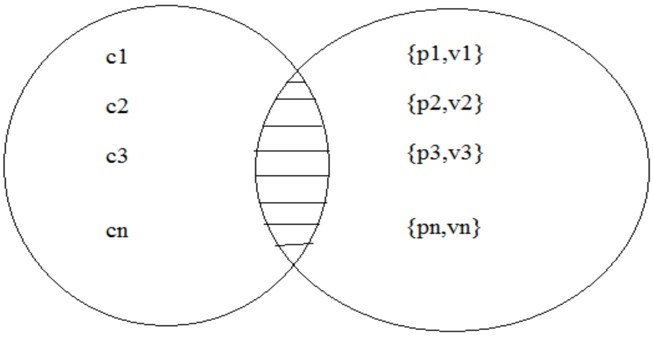

Fig -1(a)

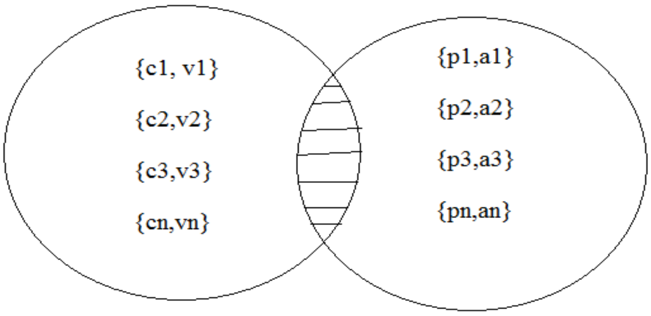

Fig -2 (b)

Above fig shows verification of product as well as attacker product
1. Fig (a) shows combination of customer with respective product.
2. Fig (b) shows attack on product.

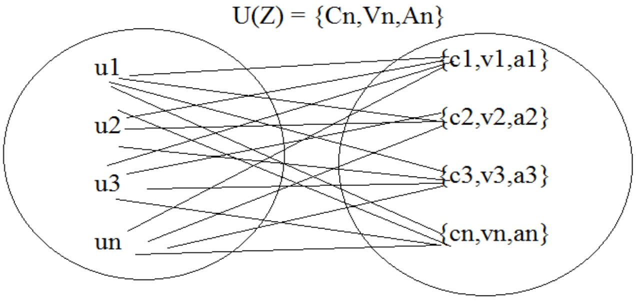

Fig -3(c)

Success Condition :
Properly Verify Product.
Failure Condition :
Before Attack not manage.

7. ALGORITHM

7.1 Apriori algorithm:

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules. Apriori uses a ”bottom up” approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data.)

7.2 Clustering algorithm:

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Clustering: unsupervised classification: no predefined classes.

7.3 Polynomial Time Algorithm:

An algorithm that is guaranteed to terminate within a number of steps which is a polynomial function of the size of the problem. See also computational time complexity. Search the data without loss of time to provide out stream for the process.

8. SYSTEM ARCHITECTURE

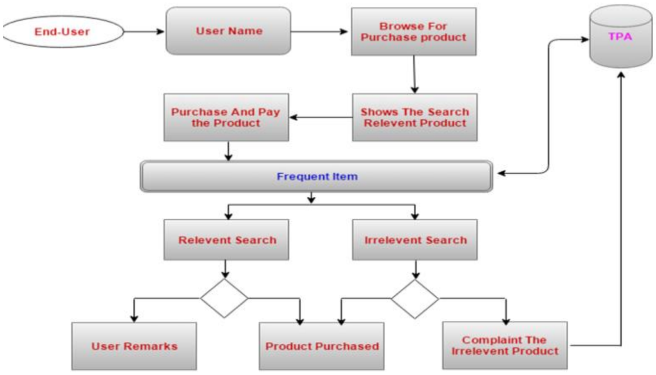
End-User
User Name
Browse For Purchase product
TPA
Purchase And Pay the Product
Shows The Search Relevent Product
Frequent Item
Relevent Search
Irrelevent Search
User Remarks
Product Purchased
Complaint The Irrelevent Product

Fig -4: System Architecture

Module:

1. Product Upload
2. Product Search
3. Auditing.

1.Product Upload:

The admin wants to upload new product to the cloud, it needs to verify the validity of the cloud and recover the real secret key. We show the time for these two processes happened in different time periods. They only happen in the time periods when the client needs to upload new product to the cloud. Furthermore, the work for verifying the correctness of the can fully be done by the cloud.

2. Product Search:

To can consider the dishonest cloud server as a suspect, the data user as a search data to the server .If the server show the search relevant data. Then the user select and buying the product. After continue the relevant product to show the user side. If the user want buying the product and complaint the irrelevant product. The product search based on index based .The cloud provide the data based on index terms. The relevant product specified for the user frequently buying product of services

3. Auditing:

Public auditing schemes mainly focus on the delegation of auditing tasks to a third party auditor (TPA) so that the overhead on clients can be offloaded as much as possible. However, such models have not seriously considered the fairness problem as they usually assume an honest owner against an untrusted CSP. Since the TPA acts on behalf of the owner, then to what extent could the CSP trust the auditing result? What if the owner and TPA collude together against an honest CSP for a financial. In this sense, such models reduce the practicality and applicability of auditing schemes. Tpa check the user remarks of the product to be verify. Then the product to be removed from the list based on number of user putting the negative comments of the products.

9. SYSTEM ANALYSIS

Performance Measurement

Discounted cumulative gain (DCG) is a measure of attacker quality. In information retrieval, it is often used to measure effectiveness of algorithms or related applications. Using a graded relevance scale of attacker product result set, DCG measures the usefulness, or gain, of a based on its position in the result list. The gain is accumulated from the top of the result list to the bottom with the gain of each result discounted at lower ranks.

Two assumptions are made in using DCG and its related measures.

1) Highly Attacker User are more useful when appearing earlier in a Black result list

2) Highly Revocation Process are more useful than marginally Un-Revoke Process, which are in turn more useful than Unrevocation. DCG originates from an earlier, more primitive, measure called Cumulative Gain.

Cumulative Gain:

Cumulative Gain(CG) is the predecessor of DCG and does not include the position of a result in the consideration of the usefulness of a result set. In this way, it is the sum of the graded relevance values of all results in a Revocation result list.

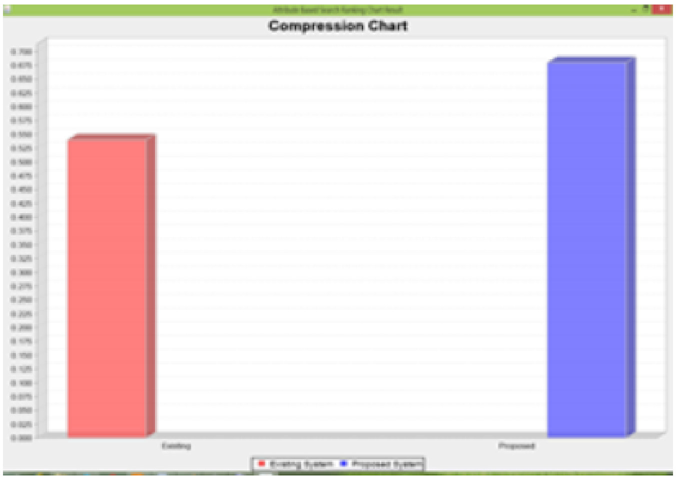

Chart -1: DCG of Proposed VS Existing System

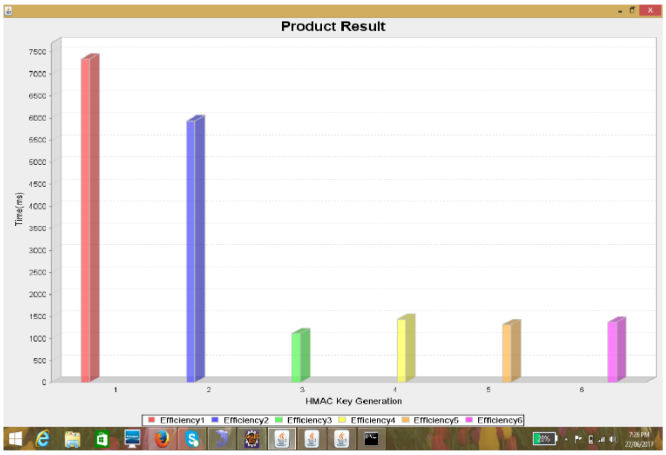

Chart -2: Efficiency of Product

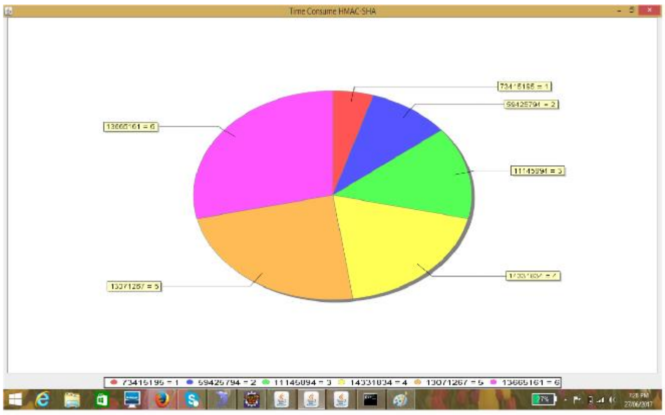

Chart -3: Time Consume HMAC-SHA

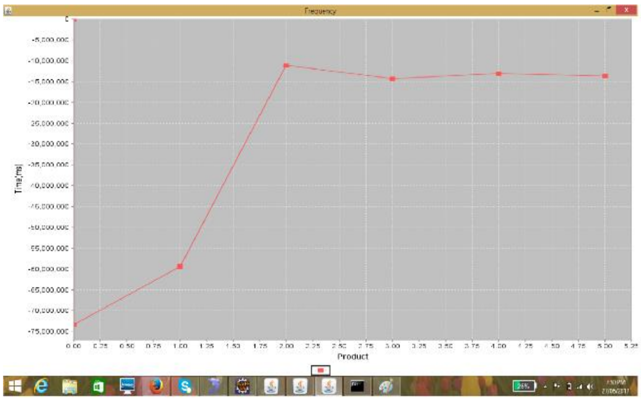

Chart -4: Frequency of Product

Chart.2 shows efficiency of each product based on initial time of entering data of product and final time of generating the HMAC key.

Efficiency=End Time-start Time/Total Time

Chart.3 shows the time consume for HMAC key in millisecond.

Chart.4 shows the frequency of each product increase entering of new multiple product.

10. SOFTWARE REQUIREMENTS

Operating system : Windows Family
Coding Language : JAVA
Data Base : MySql(Front Controller)
Front end : JSP,HTML
Back end : java(Servlet Classes)
Scripting Language : JavaScript
Style sheet : CSS JDK : 1.8
Server : Apache Tomcat 8.0

11. EXPERIMENT SETUP

In this the Structure consist of technologies like JAVA, HTML, CSS, Java script. For back end MySQL is used. Hence before investigational set up Software like Eclipse, Tomcat is predictable to be installed on server. User must have basic windows Family, good browser to view the results. Supervised Dataset or Un-Supervised dataset is used for testing in MySQL is tested

12. RESULT TABLES

Table -1: Result Table

Sr. No.	Existing System(DCG)	Proposed System(DCG)
1	0.45	0.78

13. CONCLUSIONS

To Present two integrity verification approaches for outsourced frequent itemset mining. The probabilistic verification approach constructs evidence (in)frequent itemsets. 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 itemsets. The deterministic approaches requires the server to construct cryptographicproofs of the mining result. The correctness and completeness are measured against the proofs with 100per 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.

ACKNOWLEDGEMENT

It gives me an immense pleasure to express my sincere and heartiest gratitude towards my guide Prof. A. N. Nawathe for her guidance, encouragement, moral support and affection during the course of my work. I am especially appreciative of her willingness to listen and guide me to find the best solution, regardless of the challenge. This work is also the outcome of the blessing guidance and support of my parents and family members and friends. I am also thankful to all who have contributed indirectly and materially in words and deeds for completion of this work.

REFERENCES

[1] Boxing Dong, Ruilin Liu, and Hui (Wendy) Wang,”Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-Mining-As-a-Service Paradigm,” IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 9, NO. 1, JANUARY/FEBRUARY 2016

[2] R. Agrawal and R. Srikant, ”Fast algorithms for mining association rules in large databases,” in Proc. 20th Int. Conf. Very Large Data Bases, pp. 487499, 1994.

[3] L. Babai, L. Fortnow, L. A. Levin, and M. Szegedy,”Checking computations in polylogarithmic time,” in Proc. 23rd Annu. ACM Symp. Theory Comput. 1991, pp. 2132.

[4] R. Canetti, B. Riva, and G. N. Rothblum,”Verifiable computation with two or more clouds,”in Proc. Workshop Cryptography Security Clouds, 2011.

[5] K.-T. Chuang, J.-L. Huang, and M.-S. Chen,”Power-law relationship and Self-similarity in the itemset support distribution: Analysis and applications,” VLDB J., vol. 17, pp. 11211141, Aug. 2008.

[6] R. Gennaro, C. Gentry, and B. Parno, ”Non-interactive verifiable computing: Outsourcing computation to untrusted workers,”in Proc. 30th Annu. Conf. Adv. Cryptol., 2010, pp. 465482.

[7] F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and W. Hui Wang,”Privacy-preserving data mining from outsourced databases,”in Proc. 3rd Int. Conf. Comput, Privacy Data Protection,2011, pp. 411426.

[8] Y.Gu, A.Lo and I. S. Goldwasser, S. Micali, and C. Rackoff,”The knowledge complexity of interactive proof systems,”SIAM J. Comput., vol. 18, pp. 186208, Feb. 1989.

[9] C.L.Wang and Y.S.H. Hacigum u s, B. Iyer, C. Li, and S. Mehrotra,”Executing SQL over encrypted data in the database-service-provider model,” in Proc. ACM SIGMOD Int. Conf. Manag. Data, 2002, pp. 216227.

[10] F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin,”Dynamic authenticated index structures for outsourced databases,” in Proc. ACM SIGMOD Int. Conf. Manag. Data, 2006, pp. 121132

[11] R. Liu, H. Wang, A. Monreale, D. Pedreschi, F. Giannotti, and WengeGuo, ”Audio: An integrity auditing framework of Outlierminingas-a-service systems,” in Proc. Eur. Conf. Mach. Learning Knowl. Discovery Databases, 2012, pp. 118.

[12] L.Molloy, N. Li and T. Li, “On the (In)security and (Im)practicality of outsourcing precise association rule mining,” in Proc. IEEE 9th Int. Conf. Data Mining, 2009, pp. 872–877.

[13] F. Zhu, X. Yan, J. Han, P. S. Yu, and H. Cheng, “Mining colossal frequent patterns by core pattern fusion,” in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 706–715.

[14] R. Canetti, B. Riva, and G. N. Rothblum, “Practical delegation of computation using multiple servers,” in Proc. 18th ACM Conf. Comput. Commun. Security, 2011, pp. 445–454.

[15] C. Papamanthou, R. Tamassia, and N. Triandopoulos, “Optimal verification of operations on dynamic sets,” in Proc. 31st Annu. Cryptol. Conf. Adv. Cryptol., 2011, pp. 91–110.

BIOGRAPHIES

Miss Chaudhari Bhagyashri N Perusing maters in Computer Engineering at AVCOE, Sangamner.She. She has received his Bachelors in Computer Engineering from AVCOE, Sangamner.She is member of Association of Computer Machinery (ACM).

Prof. A. N. Nawathe is Associate Professor in Amrutvahini College of Engineering, Sangamner.She is having 18 years teaching experiences in Computer Engineering. She is doing PhD degree in Data Mining.