ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Analyses of Algorithms and Complexity for Secure Association Rule Mining of Distributed Level Hierarchy in Web

Gulshan Shrivastava*
Department of Computer Science & Engineering,
Ambedkar Institute of Technology,
Geeta Colony, Delhi, India
gulshanstv@gmail.com

Dr. Vishal Bhatnagar
Department of Computer Science & Engineering,
Ambedkar Institute of Technology,
Geeta Colony, Delhi, India
vishalbhatnagar@yahoo.com

***Abstract*—** WWW (World Wide Web) has revolutionized the way in which people interact, carry out their works and gather information. It has proved itself to be a useful interface for its users to carry out such activities with ease. With hundreds of millions of people around the world using it a huge pile of data are collected every day. These data carries interesting insights on the way people interact with it. Web Mining is the process of using various data mining techniques to analyze and discover patterns from the data. The Web mostly contains semi-structured information. It is, however, not easy to search and extract structural data hidden in a Web page. Thus several privacy preserving techniques for association rule mining for web have also been proposed in the past few years. This paper focuses on analyses of algorithms and complexity for secure association rule mining of distributed level hierarchy in web. We also have shown that algorithm's pseudocode for easily analyzing its complexity.

Keywords - Vertical Partition, Privacy Preserving, Complexity, Pseudocode of Association Rule Mining

I. INTRODUCTION

The rapid development of computer technology, especially increased capacities and decreased costs of storage media, has led businesses to store huge amounts of external and internal information in large databases at low cost. Mining useful information and helpful knowledge from these large databases has thus evolved into an important research area. Web mining [2] [10] the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. Web mining encompasses a wide variety technique, including soft computing [12]. Web mining methodologies can generally be classified into one of three distinct categories: web usage mining, web structure mining, and web content mining.

In mathematics, computer science, and related subjects, an "algorithm" is an effective method for solving a problem expressed as a finite sequence of instructions. Algorithms are used for calculation, data processing, and many other fields. Each algorithm is a list of well-defined instructions for completing a task. Starting from an initial state, the instructions describe a computation that proceeds through a well-defined series of successive states, eventually terminating in a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate randomness. Algorithms are written in pseudocode that resembles programming languages like C and Java etc. Pseudocode is a mixture of natural language and high level programming concept that describes the main idea behind a generic implementation of a data structure or algorithm.

The rest of this paper is arranged as follows: Section 2 gives an overview about the background and related work in the area of secure association rule mining of distributed level hierarchy in web. In section 3 the details of analysis of algorithm for secure association rule mining of distributed level hierarchy in web. Section 4 results of our paper by analysis of complexity for secure association rule mining of distributed level hierarchy in web. Finally, some conclusion and prospect are put forward in Section 5.

II. BACKGROUND & RELATED WORK

Web usage mining, the art of analyzing user interactions with a web page, has been dealt by several researchers using different approaches [2]. Some researchers including [3], [6] have used classification algorithms for detecting web usage patterns. The authors [7] used similarity upper approximation clustering technique on web transactions from web log data to extract the behavior pattern of user's page visits and order of occurrence of visits.

Privacy preservation in data publishing has attracted considerable attention due to the need of several organizations to share their data without revealing information that can be traced to real person or legal entities. Privacy preservation was first studied in the relational context. In [15, 8] the authors introduce *k*-anonymity and use generalization and suppression as their two basic tools for anonymizing a dataset. [16] Proved that optimal *k*-anonymity for multidimensional QI is *NP*-hard, under both the generalization and suppression models. For the latter, they proposed an approximate algorithm that minimizes the number of suppressed values; the approximation bound is $O(k \cdot log\ k)$. [17] Improved this bound to $O(k)$, while [18] further reduced it to $O(\log k)$. *Incognito* [18] and *Mondrian* [18] guarantee *k*-anonymity for a relation table by transforming the original data using global (full-domain) and local recoding respectively. In [5] the authors provide another local recoding approach that shows superior performance to the global recoding approach of *Incognito*. A different approach is taken in [14], where the authors propose to use *natural* domain generalization hierarchies (as opposed to user-defined ones) to reduce information loss.

© 2010, IJARCS All Rights Reserved

Jyoti Pandey, et al [7] proposed data mining based service would run in background mode. The service computes the web pages likely to be requested by the user, considering their past web access log history, using association rules and thus optimizing the access time [5].

III. ANALYSIS OF ALGORITHM FOR SECURE ASSOCIATION RULE MINING OF DISTRIBUTED LEVEL HIERARCHY IN WEB

"Before there were computers, there were algorithms." - H.Cormen [8]. Now that there are computers, there are even more algorithms and algorithms lie at the heart of computing. Informally, an algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output. We can also view an algorithm as a tool for solving a well-specified computational problem. The statement of the problem specifies in general terms the desired input/output relationship. The algorithm describes a specific computational procedure for achieving that input/output relationship. For example [1], we focus on the preservation of privacy in vertical partitioned. Common source for such data are credit card log, web log etc. consider example a dataset P which contain web logs. If an attacker has background knowledge that associates queries to known user then the publication of P might lead to privacy breaches. For example, assume that attacker *Ravi* knows that the user *Ashu* was interested on *train ticket* to *Shimla*, so he have the background knowledge consisting of terms *Shimla* and *train ticket*. If P is published without any modification then the attacker can trace all record that contain both term *Shimla* and *train ticket*. If only one record exists, then he can easily get that this is the record of *Ashu*. This problem arises frequently in practice and provides fertile ground for introducing many standard design techniques and analysis tools. Pseudocode is not a computer program, but is more structured than usual prose [11] [13].

A. Expression:

a. Use standard mathematic symbols to describe numeric and boolen expression.
b. Use ← for assignment (‘ =’ in Java).
c. Use = for equality relationship (‘= =’ in Java).

B. Method Declaration:

a. Declares a new method "name" and its parameters.
Algorithm name (Parameter1, Parameter2,…)

C. Method:

a. Calls:

Object.method(args) (object is optional if it is understood).

b. Returns:

Return value (This operation returns the value specified to the method that called this one).

D. Programming Constructs:

a. Decision Structures*:* If condition then true-actions [else false-actions].
b. While-Loops*:* While condition do actions.
c. Repeat-Loops: Repeat actions until condition.
d. For-Loops*:* For variable-increment-definition do actions.
e. Array Indexing*:* A[i] represents the i^{th} cell in the array A. The cells of an n-celled array A are indexed from A [0] to A [n − 1] (consistent with Java).

To vertically partition the cluster, we follow the greedy strategy of Algorithm illustrated in [1], which is executed independently for each cluster. Thus, Algorithm takes as input a Cluster C and integer k and m. The algorithm performance a vertical partition and output a data set of secure vertical partition.

Algorithm: SECVERPART
Input: A cluster *C*, integer’s p and q
Output: A data set of secure vertical partitioning of c

i. Let T^C, set of terms of C;
ii. **for** term $t \in T$ **do**
iii. Compute the number of appearances $a(t)$;
iv. Sort T^C with decreasing $a(t)$;
v. Move all terms with $a(t) < p$ into T_T ;
vi. $v = 0$;
vii. $T_{cur} = \varnothing$; // Term which contain current set
viii. $T_{remain} = T^C - T_{term}$; // T_{remain} has the ordering of T^C
ix. **while** $T_{remain} \neq \varnothing$ **do**
x. **for** term $t \in T_{remain}$ **do**
xi. Create a chunk C using $T_{cur} \cup \{t\}$;
xii. **if** C *is* k^m *anonymous* **then** $T_{cur} = T_{cur} \cup \{t\}$;
xiii. $v + +$;
xiv. $T_v = T_{cur}$;
xv. $T_{remain} = T_{remain} - T_{cur}$;
xvi. Create record chunks $C_1, \ldots, C_v$ using $T_1, \ldots, T_v$;
xvii. Create term chunk C_T using T_T ;
xviii. return $C_1, \ldots, C_v, C_T$;

IV. ANALYSIS OF COMPLEXITY FOR SECURE ASSOCIATION RULE MINING OF DISTRIBUTED LEVEL HIERARCHY IN WEB

As the given algorithm is a logistic approach, it cannot furnish a assurance for its quality oriented result. Apart from it, its estimation involution is short. Asymptotically, the extremely valuable part of the algorithm is its perpendicular partitioning. For the generation of chunks, it requires numeration of item sets, therefore, the favorable result is a computationally intensive problem (even plain mining for common item sets has been shown to be # P complete [12]).

Moreover, to generate groups of terms to sustain their combination is a big problem. As per the size of the created chunks, the cost of each vertically partitioning cluster depends because the chunk magnitude influences the number of combination to be tested. It usually happens in worst case, as the domain of the cluster increases, size of the created chunks also increases. The average record length and the total domain of the dataset both affect the size of the cluster domain. The benefits of this anonymization algorithm are that this process maintains record of one cluster at a time. Nevertheless, the algorithm has expressed check over the size of its input despite its weighty asymptotic cost of the vertical partitioning. Since the size of the cluster does not depend on the size of the dataset |D| will yield more linearly clusters. As a result, the cost vertical partitioning phase will only increase followed by the increase in clusters. On the other hand, the horizontal

© 2010, IJARCS All Rights Reserved

partitioning phase is O ($|D|^2$), since in the beat case it demands $|D| / |C_{min}|$ partitioning (where $|C_{min}|$ is the minimum cluster size), conditioned one cluster is produced every step. Each partitioning has worth O ($|D|$) since all records might have to be tested to discover their most common item, thus the entire cost of horizontal partitioning seems to be worst case complexity O ($|D|^2$) [9].

Now we get the complexity of Algorithm for secure association rule mining of distributed level hierarchy in web that will be $O(n^2)$. In this algorithm the data set is getting partitioned with the number of appearances in that list for Example [4]:

{gaurav, kavita, gulshan, darshan, dheeraj}
{ashu, darshan, shrivastava}
{gulshan, arora, darshan}

So basically its getting determined in terms of number of appearances that is *a(t)* in this {gulshan, darshan} is the maximum occurrence, In algorithm each time it will depend upon number of terms that is n since its loop for each term is n1we have to retrieve the loop n-1 times means n(n-1) times means n^2 means $O(n^2)$ out of n terms. Here, we comparing with n-1 term with T_{remain} which is always n-1 for n^{th} term so simply our loop will iterate n (n-1) times for n is total number of terms so complexity will be $O(n^2)$.

Table 1 Comparison between Vertical and Horizontal Partition

S. No.	Vertical Partition	Horizontal Partition
1.	Asymptotically, it is extremely valuable part of algorithm.	It seems to be worst case of complexity.
2.	Vertical anonymization algorithm process maintains record of one cluster at one time.	Horizontal partitioning process is bounded to produce one cluster at every step.
3.	With comparison to other process, it is less time consuming as far as result declaration is concerned.	It takes more time to declare the desired result means it is time consuming process.

V. CONCLUSION AND FUTURE WORK

The major contributions of this paper are a privacy preserving association rule mining algorithm given a secure web mining. Our grand goal is to analyze algorithm that can be done at vertical partitioned, while respecting their privacy policies. In this paper, we analyze an algorithm and its complexity for privacy preserving in distributed level hierarchy in web.

In future we aim to improve our algorithm and implement it in real dataset. Additionally we plan to investigate how to quality of published dataset will be improved.

VI. REFERENCES

[1] Gulshan Shrivastava, Dr. Vishal Bhatnagar, "Secure Association Rule Mining of Distributed Level Hierarchy in Web" International Journal of computer Science and Engineering (IJCSE), Vol. 3, Issue 6, Pp. 2240 - 2244, 2011.

[2] Kavita Sharma, Gulshan Shrivastava, Vikas Kumar, "Web Mining: Today and Tomorrow" In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[3] Agrawal, R., Imielinski, T., and Swami, A. N., "Mining association rules between sets of items in large databases" In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216. 1993.

[4] Chen, R., Sivakumar, K., and Kargupta, H.,"Distributed Web mining using Bayesian networks from multiple data streams". In Proceedings of the IEEE International Conference on Data Mining. 2001.

[5] Verykios,S.V. Bertino,E. Fovino, I.N. Provenza,L.P. Saygin,Y. Theodoridis,Y. "State-of-the-art in Privacy Preserving Data Mining" ACM SIGMOD Vol. 33 Issue 1, March 2004.

[6] Metanat HooshSadat, Hamman W. Samuel, Sonal Patel, Osmar R. Zaïane, "Fastest Association Rule Mining Algorithm Predictor", Proceedings of The Fourth International C* Conference on Computer Science and Software Engineering (ACM), May 16-18, 2011, Pp. 43-50.

[7] Jyoti Pandey, Amit Goel, Dr. A K Sharma, " A Framework for Predictive Web Prefetching at the Proxy Level using Data Mining", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.6, June 2008.

[8] Thomas H. Cormen,Charles E. Leiserson,Ronald L. Rivest,Clifford Stein, "Introduction to Algorithms, Second Edition", McGraw-Hill 2001.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm (with discussion)". Journal of the Royal Statistical Society, B 39:1{38, 1977.

[10] Murat Kantarcioglu and Jaideep Vaidya. Architecture for privacy-preserving mining of client information." IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, volume 14, pages 37{42, Maebashi City, Japan, December 9 2002. Australian Computer Society.

[11] Yehuda Lindell and Benny Pinkas. "Privacy preserving data mining". In Advances in Cryptology CRYPTO 2000, pages 36, Springer-Verlag, August 20-24 2000.

[12] Xiaodong Lin, Chris Clifton, and Michael Zhu. "Privacy preserving clustering with distributed EM mixture modeling". Knowledge and Information Systems, 2004.

[13] H. Pang, X. Ding, and X. Xiao, "Embellishing text search queries to protect user privacy," *PVLDB*, vol. 3, no. 1, 2010.

[14] J. Cheng, A.W.-c. Fu, and J. Liu, "K-isomorphism: privacy preserving network publication against structural attacks," in *SIGMOD*, 2010.

[15] K. Liu and E. Terzi, "A framework for computing the privacy scores of users in online social networks," *Data Mining, IEEE International Conference on*, vol. 0, pp. 288–297, 2009.

[16] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," *TKDE*, vol. 16, no. 4, 2004.

[17] Backstrom, L., Dwork, C., and Kleinberg, J., "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *WWW*, 2007.

[18] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving Anonymization of Set-valued Data," *PVLDB*, vol. 1, no. 1, 2008.

© 2010, IJARCS All Rights Reserved

SHORT BIODATA OF THE AUTHOR

Gulshan Shrivastava, has obtained a degree of M.Tech. in Information Security from Ambedkar Institute of Technology, New Delhi and MBA (IT) from Punjab Technical University, Jalandhar after completing his B.Tech. & Polytechnic in Computer Science and Engineering from Hindu Society. He has rich experience in teaching the classes of Graduate and Post-Graduate in India and Abroad. He is a Sun Certified Java Programmer. He has been continuously imparting corporate training to the experienced professionals of multinational IT giants in the area of Java Programming & Information Security. He has participated in many National & International Workshop and Technical Fest. He has contributed to numerous International journal & conference publications in various areas of Computer Science. His area of interest includes Java Programming, Website Designing, Data Mining and Information Security.

Dr. Vishal Bhatnagar, Associate-Professor (CSE), has obtained his Ph.d. degree in 2010 and has done M.Tech. (IT) from Punjab University in the year 2005 and completed his B.E. (CSE) from Nagpur University in the year 1999. He has more then 13 years of experience. His area of Interest is Database and Data Mining, Data Warehouse, and application of DWDM in business domain. He joined as an Assistant Professor (CSE) in the department of Computer Science and Engineering in Ambedkar Institute of Technology, Geeta Colony, Delhi. He is currently working as an Associate Professor and HOD (CSE Deptt.) in A.I.T., New Delhi.

© 2010, IJARCS All Rights Reserved