A Comprehensive Survey of Privacy Preserving Algorithm of Association Rule Mining in Centralized Database

Archana Tomar
M.Tech, Scholar, Computer Science
LNCT, Bhopal, India

Prof. Vineet Richhariya
Head, Computer Science
LNCT, Bhopal, India

Prof. R.K. Pandey
UIT, Bhopal, India

ABSTRACT

The recent advancement in data mining technology to analyze vast amount of data has played an important role in several areas of Business processing. Data mining also opens new threats to privacy and information security if not done or used properly. The main problem is that from non-sensitive data, one is able to infer sensitive information, including personal information, fact or even patterns which are generated by any algorithm of data mining. In order to focusing on privacy preserving association rule mining, the simplistic solution to address the problem of privacy is presented. The solution is to survey different aspects which are discussed in the several research papers and after analyzing those research papers conclude a new solution which is best in efficiency and performance. Before analyzing the algorithms, the data structure of database and sensitive association rule mining set have been analyzed to build the more effective model.

Keywords

Data Mining, Association Rule Mining, Privacy Preserving

1. INTRODUCTION

In 1993[1], the association rule mining has received a great deal of attention. It is still one of most popular pattern-discovery methods in the field of data mining. Various proposals and algorithms have been designed for it in recent years. Simultaneity, Data mining algorithms are analyzed for the side-effects which incur in data privacy. Thus, several privacy-preserving techniques for association rule mining have also been proposed in the past few years. Various proposals and algorithms have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution.

Data mining technology can analyze massive data. Although it plays vital role in many domains, if it is used improperly it can also cause some new problem of information security. There are some new problems in the application of data mining recently.By studying deep in some special algorithms with association rule mining, some techniques also can be applied to other data mining computations, such as decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks etc.

Fast increasing of a series of digitized data causes people of the world attend the privacy problem of information more and more. Because the data mining technology of traditional centralized database must collect all the data together to process, it will cause the individual information abused or misused easily. Therefore more and more people will not to provide individual privacy data and suspect the using of data mining. Some people mine the privacy information pattern of the database owner from the original data. It has harmed the database owner's benefit. In order to solve the privacy preserving problem of association rule in centralized database, before publishing database we should hide the privacy or the sensitive information pattern of the database owner including the sensitive association rule information [2, 3]. Usually we use disturbing data method to change the data of original database to hide association rule. But the data disturbance may generate some information pattern that is not existed at all or reduce the accuracy of the original database. Before executing the privacy preserving algorithm, we should analyze the information pattern of association rule and the data structure of the database and find the preferred plan to keep the balance between the accuracy of database information and the privacy of sensitive information.

However, data mining also brings some problems. For example, credit card centres may intentionally or unconsciously make sensitive information of clients leak while mining relating information of clients. With the Internet popularity, because more and more information can be obtained in electronic form, that people have their own privacy confidential is becoming increasingly urgent. According to statistics, even if privacy protection measures, about one-fifth of Internet users don’t like to provide their own information to the Web site and more than the half investigators only in good privacy-preserving measures are willing to provide their own information to the Web site. Among the potential consumers shopping in internet browser, there are almost half who gave up the hope for internet shopping because of worrying about no protection of their privacy. Therefore, how to ensure personal privacy in data mining has become a need to be addressed. The service overview is shown in fig1.

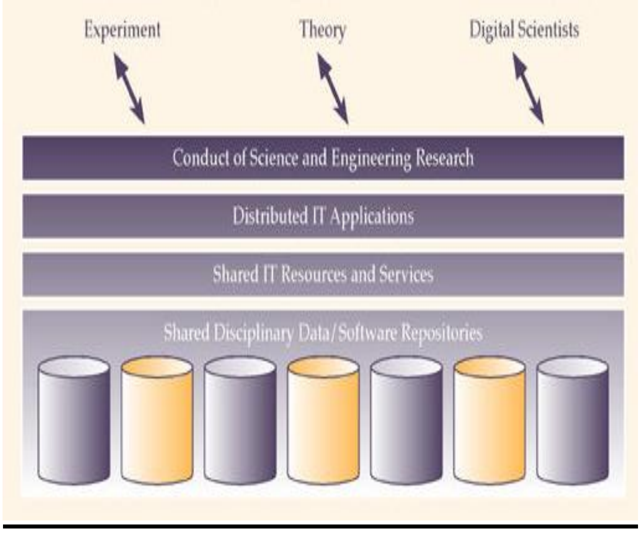
Experiment
Theory
Digital Scientists
Conduct of Science and Engineering Research
Distributed IT Applications
Shared IT Resources and Services
Shared Disciplinary Data/Software Repositories

Fig1 Service Overview

We provide here an overview of privacy preserving association rule mining. The rest of this paper is arranged as follows: Section 2 introduces Association rule mining strategies; Section 3 describe about Privacy Preserving Algorithm; Section 4 shows the evolution and recent scenario; Section 5 describes the challenges. Section 6 describes Conclusion and prospect.

2. ASSOCIATION RULE MINING

The association rule mining can be conceptualized as follows [4]: Let I= {i1,i2,…,im} be the set of all items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that T is subset of I. Each transaction is associated with an identifier, called TID. Let X be a set of items. A transaction is said to contain X if and only if X is subset of T. An association rule is an implication of the form X is subset of Y, where X is subset of T, Y is subset of T, X∩Y=NULL, the support s and confidence c of the rule X is subset of Y are defined as: s=Count(X)/|D|,c= Count(X is subset of Y)/ Count(X).

A set of items is referred as an itemset. An itemset that contains k items is a k-itemset. The support count of an itemset is the number of transactions containing the itemset. An itemset is frequent if its support count is not less than the minimum support count.Rules with the support more than a minimum support threshold (smin) and the confidence more than a minimum confidence threshold (cmin) are called strong. Association rule mining is a two-step process: (1) Finding all frequent itemsets; (2) Generating strong association rules from the frequent itemsets.

The purpose of privacy preserving is to discover accurate patterns without precise access to the original data. The algorithm of association rule mining is to mine the association rule based on the given minimal support and minimal confidence. Therefore, the most direct method to hide association rule is to reduce the support or confidence of the association rule below the minimal support of minimal confidence. With regard to association rule mining, the proposed methodology that is effective at hiding sensitive rules is implemented mainly by depressing the support and confidence. The existing tree algorithms, D_CONF1, D_CONF2 and D_SUPP, are simply introduced in [5], which are to hide the sensitive association rule all by reducing the support or confidence.

2.1 Privacy Preserving Algorithm

A lot of implementations of the confidentiality of data and knowledge are applied in association rule mining process. According to privacy protection technologies, at present, privacy preserving association rule mining algorithms commonly can be divided into three categories [6].

Heuristic-Based Techniques Heuristic-based techniques are to resolve how to select the appropriate data sets for data modification. Since the optimal selective data modification or sanitization is an NP-Hard problem, heuristics can be used to address the complexity issues. The methods of Heuristic-based modification include perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise), and blocking, which is the replacement of an existing attribute value with a "?". There is a basic principle of choosing the transaction or the item of itemset to be modified that we should reduce the influence of the original database as far as possible. Those related works are given below.

1) Data Perturbation-Based Association Rule

The algorithms can be described as the following one. Let D be the source database, R be a set of significant association rules that can be mined from D, and let Rh be a set of rules in R.How can we transform database D into a database D', the released database, so that all rules in R can still be mined from D', except for the rules in Rh. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. Therefore, the key question of this algorithm is how to put D into D' with the use of heuristic thought.

A subsequent work described in [7] extends the sanitization of sensitive large itemsets to the sanitization of sensitive rules. The work in [8] aims at balancing between privacy and disclosure of information by trying to minimize the impact on sanitized transactions or else to minimize the accidentally hidden and ghost rules. The utility in this work is measured as the number of non-sensitive rules that were hidden based on the side-effects of the data modification process. Wang et al. propose a matrix based sanitization approach to hide the sensitive patterns in [9]. It is the first paper to involve the consideration of avoiding the Forward-Inference Attacks [10], which can also be avoided in the sanitized database generated by our sanitization process. Oliveira et al. propose a novel method to modify databases for hiding sensitive patterns in [11]. Multiplying the original database by a sanitization matrix yields a sanitized database with private content. The method can avoid the question of the Forward-Inference Attacks.

This paper in [12] describes a technique that uses a queue and a random number generator to generate the items so that each item has an approximately equal frequency of being added to transactions. And the work avoids the question that it is hard for [13, 14] to utilize existing tools for association rule mining.

2) Data Blocking-Based Association Rule

The approach of blocking is implemented by reducing the degree of support and confidence of the sensitive association rules. That is by replacing certain attributes of some data items with a question mark or a true value. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lies below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated. Yucel Saygin et al. first apply blocking to the association rule confusion, which has been presented in [15, 16].

a) Replacement-Based Techniques

After original data is replaced the value of some data with the unknown value, the support and confidence of sensitive association rules will not be able to determine, which may be a range of arbitrary values. The paper in [17] discusses specific examples with the use of an uncertain symbol used in association rule mining, in which case the support and confidence interval are used to support and confidence interval to replace.

b) Anonymity Techniques

Agrawal et al. improve on the distribution reconstruction technique presented in [18] by using the Expectation Maximization (EM) method. The authors claim that EM is more effective than the currently available technique in terms of the level of information loss. Finally, they propose novel metrics for the quantication and measurement of privacy preserving data mining algorithms.

The paper in [19] presents a new generalization framework on the concept of personalized anonymity in order to perform the minimum generalization for satisfying everybody's requirements, the core of personalized anonymity is the concept of personalized anonymity. It provides privacy protection of different size for the records of data table. The paper in [20] proposes a personalized anonymity model on the
base of (α,k)-anonymization model in order to resolve the problem of privacy self management and proposes corresponding anonymity method by using local recoding and
sensitive attribute generalization. Although these personalized generalization approaches are flexible, the definitions of sensitive attributes are the same with other approaches. Thus, dynamic specifying sensitive information needs be future researched.

We should choose the algorithm according to the different situation that can reduce the influence for the original database as far as possible. D_CONF1 algorithm will increase one transaction item when it circulates one time. Usually there are much data and many association rules while it processing the problem of privacy preserving. Frequent using of D_CONF1 algorithm will increase the quantity of data and generate some association rules that do not exist. It has influenced the accuracy of database. If there are some important items that cannot be modified or deleted, D_CONF1 algorithm is suitable for this situation. The front parts of D_CONF2 and D_SUPP are same. D_CONF2 algorithm can only select sacrifice item in back-end itemset and D_SUPP algorithm can select sacrifice item in whole generated itemset. So if the influence for the original database of selecting sacrifice item in front-end is smallest we can choose D_CONF2 algorithm. If the influence for the original database of selecting sacrifice item in back-end is smaller, we can choose algorithm through comparing the efficiency, the quantity and importance of selected sacrifice item and support of D_CONF2 and D_SUPP algorithm.

We should analyze the transaction set of the original database and the sensitive association rule set to be hidden and find the relation of them. So we could select the sensitive transaction and sacrifice item more efficient. The modified data will be fewer and the influence for the original database will be smaller.

3. EVOLUTION AND RECENT SCENARIO

Recently some studies on data privacy are being conducted [21, 22, 23, 24, 25, 26]. The idea is to replace classical model of secure computation with special algorithms [27], which represent from disclosing private data and keep a functionality simultaneously. Most of methods were based on a secure mathematical operations, like secure size of intersection [21], secure sum of sets [21],secure sum [21].Data mining with preserving data privacy only concerns mining from distributed data. In case of association rule mining, data can be partitioned and distributed horizontally and vertically. For horizontally partitioned data, HPSU algorithm was introduced. It uses secure sum of sets, secure sum.For vertically partitioned data, VPSI algorithm can be used, which utilizes secure sum of sets and secure size of set intersection.

For vertically partitioned data there is another approach, which is based on secure scalar product, but it has a considerable drawback, which limits its application to only two data sources. The two, early mentioned algorithms, can mine from several data sources. It is also feasible to increase a level of data privacy, by generalization of the data. This approach assumes smaller harm by giving away generalized data than detailed data. Data can be generalized in several manners. In our study we use VPSI algorithm and aggregation as a manner of generalization.

Agrawal et al. in [27] first proposed the method of distribution reconstruction on numeric data which is disturbed by Bayesian algorithm in 2000.

Then, Dakshi and Charu in [28] improve the work over the Bayesian-based reconstruction procedure by using an Expectation Maximization (EM) algorithm for distribution reconstruction.

The work presented in [29] and [30] deals with binary and categorical data in the context of association rule mining. Both papers consider randomization techniques that offer privacy while they maintain high utility for the data set.

Agrawal et al. in 2002 proposed a privacy protection approach on reconstruction-based association rule to deal with categorical data, which is uniform randomization [31]. This approach for randomizing transactions would be to generalize Warner's "randomized response" method. Before sending a transaction to the server, the client takes each item and with probability *p* replaces it by a new item not originally present in this transaction. This process is called uniform randomization.The algorithm is applied in categorical data and the key is to mining the frequent itemsets.

Shariq J. et al. in [32] present a scheme called MASK, which attempts to simultaneously provide a high degree of privacy to the user and retain a high degree of accuracy in the mining results. Its scheme is based on a simple probabilistic distortion of binary data, employing random numbers generated from a pre-defined distribution function. And These works in [33,34] based on the "select-a-size" and "cut-andpaste" random transform operation to hide the original data set method, and then convert the transformed data into project itemsets

The security of the scalar product protocol is based on the inability of either side to solve k equations in more than k unknowns. Some of the unknowns are randomly chosen, and
can safely be assumed as private. A similar approach has been proposed by Ioannidis et al. who present an extremely efficient and sufficiently secure protocol for computing the dot-product of two vectors by using linear algebraic techniques and demonstrate superior performance in terms of computational overhead, numerical stability, and security by using analytical as well as experimental results [35].

Another way for computing the support count utilizes the secure size of set intersection method described in [36]. If the
transactions are vertically partitioned across the sites, this problem can be solved by generating and computing a set of independent linear equations [37]. The work in [38] develops a log-linear model approach for strictly vertically partitioned databases and a more general secure logistic regression for problems involving partially overlapping data bases with measurement error.

In addition, privacy-preserving algorithms relying on other means of data mining are proposed one after another. For example, GENG Bo[39] etc. developed privacy preserving technique in multi-temporal sequence rule mining, which employs one simple untrusted third-party algorithm to solve the problem of calculating the frequency of temporal sequence rule by multiple partners together.

In 2008, Shaofei Wu et al. [40] proposed a new algorithm for balance privacy preserving and knowledge discovery in association rule mining. The solution is to implement a filter after the mining phase to weed out or hide the restricted discovered association rules. Before implementing the algorithms, the data structure of database and sensitive association rule mining set have been analyzed to build the more effective model.

In 2009, Yongcheng Luo et al. [41] proposed a privacy preserving association rule mining into three categories: heuristic-based techniques,reconstruction-based techniques, cryptography-based techniques. Finally, they conclude further research directions of privacy preserving algorithms of association rule mining by analyzing the existing work.

In 2009, Jie Liu et al. [42] proposed the algorithm which is designed to solve the shortage of low privacy protection of the geometric transform algorithm. The algorithm first gives four parameters, corresponding to the probability of four different types of geometric transformations. According to the various random number generated, different geometric transformation method is selected, which serves the dual effect of privacy protection.

In 2010, Brian, C.S. Loh et al. [43] proposed a framework involves several components designed to anonymize data while preserving meaningful or actionable patterns that can be discovered after mining. In contrast with existing works for traditional data-mining, this framework integrates domain ontology knowledge during DGH creation to retain value meanings after anonymization. In addition, users can implement constraints based on their mining tasks thereby controlling how data generalization is performed. Finally, attribute correlations are calculated to ensure preservation of important features. Preliminary experiments show that an ontology-based DGH manages to preserve semantic meaning after attribute generalization. Also, using Chi-Square as a correlation measure can possibly improve attribute selection before generalization.

In 2010, Chirag N. Modi et al.[44] proposed an algorithm provides privacy and security against involving parties and other parties (adversaries) who can reveal information by reading unsecured channel between involving parties.

In 2010, Wang Yan et al.[45] proposed a privacy preserving association rule mining algorithm based on SRRCR is presented, which can achieve significant improvements in terms of privacy and efficiency. Finally, they present experimental results that validate the algorithms by applying it on real datasets.

4. CHALLENGES

The influence of data mining technology diversity makes privacy-preserving data mining methods variable. But some basic data privacy-preserving theories and application still need detailing. All in all, except considering the betterment of dependence, accuracy and expansibility of privacypreserving,the privacy-preserving research yet needs to seek breakthrough in the following aspects.

1) The perfection of formalized model of security requirement. Formalized methods including modeling, proof, analysis and expression of application logic, are the important basis of security theory and application. At present, it couldn’t strictly prove with formalized method the security intensity of one system. Data mining faces considerable data, while SMC needs plentiful calculation and message exchange of many times among participants, which has low efficiency. So while preserving privacy, how makes designed algorithm bear suitable calculation efficiency and transmission spending is one of key problems needing to be solved. If different security requirement can be dealt with by different levels and processed strict formalized definition, computing efficiency would be improved greatly so as to decrease transmission spending.

2) The general privacy-preserving way research of data mining. At present, scholars have had definite study in privacy and security question of some certain data mining algorithms, but the kinds of data mining algorithms are various, so general privacy-preserving technology necessarily becomes the future research trend.

5. CONCLUSION AND PROSPECT

We present a classification and an extended description and clustering of various algorithms of association rule mining. The work presents in here, which indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users.

At present, privacy preserving is at the stage of development. Many privacy preserving algorithms of association rule mining are proposed, however, privacy preserving technology needs to be further researched because of the complexity of the privacy problem. We conclude three research directions of privacy preserving association rule mining algorithms by analyzing the existing work in the future.

- The research of personalized privacy preserving association rule mining will become the issue. How to perform the minimum generalization for satisfying everyone’s requirements and retain the largest amount of information from the micro data need to be further researched.

- How to improve the efficiency of implementation and ensure available of the result in order to meet the various requirements. Thus, some better algorithms need to be proposed and verified by experiments.

- The research of application-oriented strategy for privacy protection. Each end-user may have different privacy concern when sharing the data. Thus, it is necessary to investigate the end-user-oriented privacy preserving data mining.

6. REFERENCES

[1] Agrawal R, Imielinski T, Swami A, “Mining association rules between sets of items in large databasesWashington, DC, 1993, pp.207–216.

[2] Chen M S, Yu P S. Data Mining:An Overview from a Database Perspective [J]. IEEE Trans on Knowledge and Data Engineering, 2004,8(6) :866-883.

[3] M K Reiter. Crowds:Anonymity for Web Transactions[J]. The ACM Transactions on Information and System Security,2005.

[4] R. Agrawal, R. Srikant, “Fast algorithms for mining association rules,”In: Proc. 20th Int’l Conf. Very Large Data Bases, 1994.

[5] Shaofei Wu, Hui Wang, “Research On The Privacy Preserving Algorithm Of Association Rule Mining

InCentralized Database,” International Symposiums on Information Processing, 2008.

[6] Vassilios S. Verykios, Elisa Bertino， et al., “State-of-the-art in Privacy Preserving Data Mining,” March 2004, pp.50-57.

[7] Elena Dasseni, Vassilios S. Verykios, Ahmed K.Elmagarmid, and Elisa Bertino, “Hiding Association Rules by using Confidence and Support,” 2001.

[8] Stanley R. M. Oliveira and Osmar R. Zaiane, “Privacy preserving frequent itemset mining, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining 2002.

[9] E.T. Wang, G. Lee, Y.T. Lin, “A novel method for protecting sensitive knowledge in association rules mining,” 2005.

[10] E.T. Wang, G. Lee, “An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining,” 2008.

[11] S.R.M. Oliveira, O.R. Zaıane, Y. Saygin, “Secure association rule sharing, advances in knowledge discovery and data mining, in: PAKDD2004.

[12] Jun Lin Lin, Yung Wei Cheng, “Privacy preserving itemset mining 2009.

[13] Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J., “Privacy preserving mining of association rules,” 2002.

[14] Rizvi S J, Haritsa J R., “Maintaining data privacy in association rule mining,” August 2002.

[15] Yucel Saygin, Vassilios Verykios, and Chris Clifton, “Using unknowns to prevent discovery of association rules,” 2001.

[16] Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, “Privacy preserving association rule mining,” 2002.

[15] L. Sweeney, “k-anonymity: a model for protecting privacy”,

2002.

[17] Xiao X, Tao Y, “Personalized privacy preservation”, 2006.

[18] LIU Ming, Xiaojun Ye, “Personalized K-anonymity”, Computer Engineering and Design, Jan.2008.

[19] Rakesh Agrawal and Ramakrishnan Srikant, “Privacy-preserving data mining 2000.

[20] Dakshi Agrawal and Charu C. Aggarwal, 2001.

[21] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya,Xiaodong Lin, Michael Y. Zhu, 2002.

[22] Chris Clinton, „Privacy Preserving Distributed Data Mining”, 2001.

[23] Murat Kantarcioglu, Chris Clinton, IEEE 2003.

[24] Rakesh Agrawal, Ramakrishnan Srikant, “Privacy- Preserving Data Mining”, 2000.

[25] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis, “State-of-the-art in Privacy Preserving Data Mining”, 2004.

[26] Shipra Agrawal, Vijay Krishnan, Jayant Haritsa, “On Addressing Efficiency Concerns in Privacy Preserving Data Mining”,DB/0310038.

[27] Rakesh Agrawal and Ramakrishnan Srikant, “Privacy-preserving data mining,” 2000.

[28] Dakshi Agrawal and Charu C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms 2001.

[29] Rizvi S J, Haritsa J R., “Maintaining data privacy in association rule mining,” August 2002.

[30] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, 2002.

[31] Evfimievski A,Srikant R,Agrawal R, et al. 2002.

[32] Rizvi S J, Haritsa J R., “Maintaining data privacy in association rule mining,” In: Proceedings of t he 28th International Conference on Very Large Data Bases , Hong Kong , China , August 2002.

[33] A. Evfimievski, J. Gehrke and R. Srikant, “Limiting Privacy Breaches in Privacy Preserving Data Mining 2003.

[34] A. Sanil, A. Karr, X. Lin, and J. Reiter, “Privacy preserving regression modelling via distributed computation,” 2004 .

[35]Ioannidis, I.; Grama, A, Atallah, M., “A secure protocol for computing dot-products in clustered and distributed environments,” 2002.

[36] Chris Clifton, Murat Kantarcioglou, XiadongLin, and Michael Y. Zhu, “Tools for privacy preserving distributed data mining 2002.

[37] Vaidya, J. & Clifton, C.W., “Privacy preserving association rule mining in vertically partitioned data,” July 2002.

[38] ZongBo Shang; Hamerlinck, J.D., “Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases,”2007. Seventh IEEE

[39] GENG Bo,ZHONG Hong,PENG Jun,WANG Da-gang Temporal Rule Distribution Mining of Privacy-preserving□ 2008.

[40] Shaofei Wu and Hui Wang ,IEEE International Symposiums on Information Processing,2008.

[41] Yongcheng Luo,Yan Zhao and Jiajin Le ,Second International Symposium on Electronic Commerce and Security,2009 IEEE .

[42] Jie Liu and Yifeng XU,2009 Fourth International Conference on Internet Computing for Science and Engineering,IEEE.

[43] Brian, C.S. Loh and Patrick, H.H. Then,2010 Second International Symposium on Data, Privacy, and E-Commerce,IEEE.

[44] Chirag N. Modi, Udai Pratap Rao and Dhiren R. Patel,2010 International Conference on Advances in Communication, Network, and Computing,IEEE.

[45] Wang Yan,Le Jiajin and Huang Dongmei,2010 International Conference on Web Information Systems and Mining,IEEE.