Research Journal of Applied Sciences 11 (2): 27-34, 2016
ISSN: 1815-932X

A Technique of Data Privacy Preservation in Deploying Third Party Mining Tools over the Cloud Using SVD and LSA

Yousra Abdul Alsahib S. Aldeen,
Mazleena Salleh and Mohammad Abdur Razzaque
Department of Computer Science, University Technology Malaysia,
Johor, Skudai, Malaysia

Abstract: In these days, information sharing as a crucial part appears in our vision, bringing about a bulk of discussions about methods and techniques of privacy preserving for data mining which are regarded as strong guarantee to avoid information disclosure and protect individuals' privacy. k-anonymity has been proposed for privacy preserving for data mining and publishing which can prevent linkage attacks by the means of anonymity operation such as generalization and suppression. Manyanonymity algorithms have been utilized for achieving k-anonymity. Here, there is need to discover the relationships between the quasi identifier and other attributes that lead to disclosure the sensitive information. The main goal of this study is to discover the attributes with high variance which lead to disclosure the sensitive information to apply anonymity method on them. While the attributes with low variance, they can consider as quasi-identifier. This study proposed a technique based on transfer the conceptualization of the data base table into another domain which maintains the privacy and reduces the loss of information by decomposing the table using the Singular Value Decomposition (SVD) and revealing latent semantic relationships among attributes in the semantic space using Latent Semantic Analysis (LSA). This technique is the innovative in term of preventing more smart attack which tries to build linkages and binding across distributed data bases over the cloud.

Key words: SVD, LSA, privacy preserving, k-anonymity, latent semantic

INTRODUCTION

In recent years, cloud computing has become as a very popular paradigm for hosting and delivering services over the internet. Cloud computing is a model for allowing convenient on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Yang and Ma, 2013).

Due to the emerging of cloud based computation systems, large number of organizations starts their migration to that environment where information systems are built based on ready to go services provided over the cloud which reduces the cost and increase the reliability of these systems. Data mining tools are one of the significant services offered over the cloud which carry the knowledge and experience to reveal information from the data published by organizations; this has introduced new risk factor due to the exposing private data over the cloud (Xu *et al*., 2014). Data mining is a set of automated techniques used to extract hidden or buried information from large databases. In recent years, several successful applications in data mining have been reported from varied sectors such as marketing, finance, medical diagnosis, banking, manufacturing and telecommunication (Weiss, 2009). The benefits of using data mining tools can reveal invaluable knowledge that was unknown to the data holder before hand. The extracted knowledge patterns can provide insight to the data holders as well as be invaluable in tasks such as decision making and strategic business planning (Wu, 2004). But, at the sametime, serious concerns have grown over individual privacy in data collection, processing and mining (Jain *et al*., 2013a, b).

Data privacy is the main concern of organizations which publish data over the cloud, thus, a lot of efforts have been targeting this issue and many generalization and suppression techniques have been introduced. The comparison between the performance of data mining methodologies in discovering and retrieving knowledge from a repository of data and the privacy of this data stored within the repository is the starting point in investigating privacy in cloud computing applications

Corresponding Author: Yousra Abdul Alsahib S. Aldeen, Department of Computer Science, University Technology Malaysia, Johor, Skudai, Malaysia

(Fung *et al*., 2010). Mining data is crucial to retrieve knowledge from data and discover patterns to connect correlated profiles (Han *et al*., 2012); this can be highly observed in the facebook and other social applications where people are introduced to each other based on mining their profiles. Linking attack is used to bind general information to the sensitive information and eventually violate the privacy; this attack mainly works on quasi-identifier from two or more data table (Pan and Chen, 2012).

k-anonymityis the most common approach in privacy preservation and the most effective defend against linking attack where data variables with specific attributes (i.e., quasi-identifiers) are published. Quasi-identifiers are variables that carry potential information and at the same time do not refer to the sensitive information embedded with the table. k-anonymity (i.e., where k is an integer) is the approach that published record is anonymous with k-1 other records. Basically, three types of variables are there in a dataset table as it is presented in the Table 1 where quasi-identifier is the variable that can be published over the cloud if the condition of k-anonymity is maintained.

More advanced categorization for the attributes of a Table 1 is presented by Xu *et al*. (2014) where attributes are divided into four categories:

- Identifier can uniquely represent an individual
- Quasi-identifiers is a specific sequence of attributes in the table that malicious attackers can take advantage of these attributes to break the privacy by correlating released dataset with other dataset that has been already acquired
- Non-quasi attributes have less effect on data processing
- Sensitive attribute are the private information which it should be kept private and not reached by the mining tools

Anyway, three types of linkage attack are observed which are: record linkage, attribute linkage and table linkage; these attacks will be presented in the research and methodologies of defense are also presented (Sweeney, 2002).

Table 1: Types of attributes

	Quasi-identifier			
Key attribute names	DOB	Gender	Zip code	Sensitive attribute Disease
Andre	1/21/76	Male	53715	Heart disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53715	Bronchitis
Dan	1/21/76	Male	53715	Broken arm
Ellen	4/13/86	Female	53715	Flu
Eric	2/28/76	Female	53715	Hang nail

LITERATURE REVIEW

Privacy preservation on data in cloud based anonymity method has been extensively investigated and rich gain has been made by research communities.

A new approach for anonymizing data in a way that fulfils data publishers' utility demand and afford low information loss have been designed by Loukides and Gkoulalas-Divanis (2012). They presented an accurate information loss measure and an efficient anonymization algorithm that discovers a huge part of the problem space. Presented expanded meanings of k-anonymity and used them to confirm that a assumed data mining model does not disclosure the k-anonymity of the individuals represented in the learning examples. Their expansion offers a tool that measures the volume of anonymity reserved during data mining. They presented that their model could be used to different data mining problems such as classification association rule mining and clustering.

k-anonymity could be joined with data mining for preserving the individuality of the respondents to whom the data being extracted mention. Have defined the possible risks to k-anonymity that could rise from implementing mining on a collection of data and considered two main approaches to combine k- anonymity in data mining. They have also presented various methods that can be used to detect k-anonymity violations and thus remove them in association rule mining and classification mining (He *et al*., 2012). Proposed an algorithm which is based on clustering to produce a utility-friendly anonymized version of micro data. Their extensive performance study showed that their methods outperform the non-homogeneous technique where the size of QI-attribute is >3. They proposed a clustering-based k-Anonymity algorithm which achieves k-anonymity through clustering. Wide experiments on actual data sets are also directed, viewing that the utility has enhanced by their approach (Soodejani *et al*., 2012). Employed a version of the chase, called standard chase which put some restrictions on the dependencies and constrains such as being positive and conjunctive. Investigating the applicability of other versions of the chase in the method can be studied further. The anonymity principle of their method has some similarities to the L-diversity privacy model. Investigating other privacy models such as t-closeness, provide a stronger privacy model for the proposed method which can be valuable.

A new definition of k-anonymity for personal sequential data which provides an effective privacy

protection model is introduced and a method that transforms sequential datasets into a k-anonymous form is presented while preserving the utility of data with reference to a variety of analytical properties (Monreale *et al*., 2014). Through, a wide set of experiments with various real-life sequential datasets, they demonstrated that the suggested method substantially protect consecutive pattern mining results both in terms of number of mined patterns and their support; results are extremely interesting in the case of dense datasets (Cacheda *et al*., 2011) have discovered the weaknesses of many algorithms in mining information from user side view. They have also assured that SVD-based techniques have good results compare to other methods. Clustering and classification algorithms are considered a very significant part in providing low-power and low sampling rate wireless ECG systems by Balouchestani *et al*. (2014). They proved that clustering and classification sufer from drawbacks which are including that they coudnt be ready in real time and also they need higher computational costs. These disadvantages have motivated them to improve new enhanced clustering algorithm based on K-Singular Value Decomposition (K-SVD) approach. Their proposed algorithm is better than existing algorithms by attaining a classification accuracy of 99.3% (Jain *et al*., 2013a) explained the matrix decomposition is considered big part in privacy preserving in data mining decision tree. They suggested approach that contents the utility based anonymization code that vital information is endangered from being suppressed. Similarly, weights assumed to attributes recover clustering and give the aptitude to control the generalization's depth. In their proposed algorithm, attributes are grouped according to their distance difference similarity by clustering the data set using decision tree classification. Their proposed algorithm satisfies the utility based anonymization principle that crucial information is protected from being suppressed. But, their algorithm is limited to only classification and clustering algorithms. clustering algorithm which is one of the data mining algorithms and it is the procedure of grouping a set of data objects into numerous groups or clusters so as to objects within a cluster have high similarity but are very dissimilar to objects in other clusters.

Compare our work to previous works, it is used SVD to discover the attributes with high variance to apply the anonymity method on them while the attributes with low variance can publish as quasi identifier. This technique offers a good trade-off between privacy and utility of data. Moreover, it reduces dimensions space as well as it is able to discover dependence relationships between the structures of dataset, so it makes easy to deal with big data. Despite the fact that SVD comes out with dimensionality reduction, there is another beneficial outcome which is knowledge condensing vectors; this is the vector that results on maximum knowledge about the records. In data mining application, the initial matrix is an array of objects and attributes. The number of rows, n of the matrix is typically very large in the range. The number of columns, m is also large10-10. However, this is large enough for many of the difficulties of working in high dimension to play a significant role.

CHALLENGES IN PREVIOUS PRIVACY PRESERVATION TECHNIQUES

As mentioned in related research there are many challenges that are facing the conventional methodologies of privacy preservation for data published for mining; this is due the huge expansion in number of resources that contain information about the published data (e.g., hospital patients might subscribed to social sites like facebook or forums). It can be summarized in two main points including:

Quasi-identifiers are masked through different techniques (i.e., generalization and suppression) but smart attacks still can reveal latent semantic relationships for quasi-identifiers with other published data bases. Let:

Repository = $\{R_1, R_2, ..., R_K\}$ is a k-resources for information (i.e., Repository could be Web which contains profiles from Twitter, Facebook, iCloud,... and others).

$$\exists_{table} \exists_{var1, var2, var3 \in table} KeyAttributes(var1) \wedge QueasiIdentifiers(var2) \wedge SensitiveInfo(var3)$$

$$\exists_{table} \exists!_{function} Mapping(QuasiIdentifiers(X), KeyAttributes(Y)) \wedge Probability(function) \leq threshold \rightarrow PrivacyPreserved(table)$$

$$\exists_{table} \forall_{x \in table} Published(Repository, x) \wedge Size(Repository) \geq k \rightarrow Increased(Probability(function))$$

Adding more records to the published quasi-identifiers adds more potential for having latent semantic relationships among table attributes where attributes are connected through the key attributes in the latent semantic space.

DISCOVERING AND MASKING QUASI-IDENTIFIER

In this study, an innovative technique to discover Quasi-identifiers is presented where the table is decomposed into three matrices ($U\Sigma V^T$). Each matrix characterizes the original matrix without a direct

identification to the contents. Thus, database tables are represented by attributes they contain. Table attribute is conceptualized as the projection of human interpretation to domain semantic; this is very crucial in smart attack where attacker can reveal the latent relationship among quasi-identifiers and other published tables.

Suppose T a table of $\{A_1, A_2, ..., A_m\}$ then it can be expressed this table in term of semantic using the following equaiotn:

$$T = \sum_{i=1}^{N} A_i.\vec{v}_i \quad (1)$$

Where:

T = A table

A_i = The ith attribute composing that table

$\vec{v}_i$ = The semantic meaning of the attribute

Let, quasi be a set $\{A_1, A_2, ..., A_m\}$ such that $\subseteq T$, then it can be expressed quasi using the following equation:

$$quasi = \sum_{i=1}^{m} A_i.\vec{v}_j \quad (2)$$

where, quasi is the quasi-identifier semantic domain for table (T), A_j is the jth quasi-identifier within the table and $\vec{v}_i$ is the semantic unit vector. Anyway, quasi-identifiers are a subset of the original table (i.e., database table) as following equation:

$$quasi \subset T \quad (3)$$

Now, a technique has to study the variance resultant due changes in quasi-identifiers over the entire table; this is done by using the awesome statistical tool of SVD through which a table T is represented as following equation:

$$T = U\Sigma V^T = \sum_{i=1}^{N} \sigma_i u_i v_i^T \quad (4)$$

$$\rightarrow T = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \sigma_3 u_3 v_3^T + ,..., + \sigma_N u_N v_N^T \quad (5)$$

σ_i is the variance and it is equal to $\sqrt{\lambda}$, $\sum$ is a diagonal matrix of σ_i and reflects the variance of the latent semantic in attributes domain (i.e., table attributes include quasi, key-attributes and sensitive information) and the semantic in the record domain (i.e., the combination of attributes composing each row of the table). Despite the fact that SVD comes out with dimensionality reduction, there is another beneficial outcome which is knowledge condensing vectors; this is the vector that results on maximum knowledge about the records. In data mining application, the initial matrix is an array of objects and attributes. The number of rows, n of the matrix is typically very large in the range 10^3-10^9. The number of columns, m is also large 10-10^4. However, this is large enough for many of the difficulties of working in high dimension to play a significant role.

DATASET DESCRIPTION

The present study is based on bank direct marketing data set which is collected from different web sources of the University of California at Irvine (UCI) Machine Learning Repository. The data set is used to implement and estimate the performances of proposed approach for privacy preserving data mining using anonymization technique. This data set have been collected and arranged by Moro *et al*. (2011) and also utilized by Elsalamony (2014). The data is associated to different marketing campaigns of a Portuguese banking institution based on phone calls. Regularly, more than one contact has required with one client in order to analyse if the product (bank term deposit) has been (or not) subscribed. There are two types of datasets:

- Bank-full.csv with all examples, ordered by date (from May 2008 to November 2010)
- Bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv

The bank direct marketing data set includes (300) number of data samples with 17 attributes without any missing values (Asuncion and Newman, 2007). The properties of data set are comprised of two different kinds: nominal and numeral attributes as given in Table 2.

Table 2: Attributes description

Attributes	Kind of attributes	Attributes design
Age	Numerical	Numerical
Job	Categorical	Admin, unknown, unemployed, management, house main, entrepreneur, student, blue collar, self employed, retired, technician, services
Marital	Categorical	Married, divorced (widowed), single
Education	Categorical	Unknown, secondary, primary, tertiary
Default	Binary	Yes, No
Balance	Numeric	Numeric
Housing	Binary	Yes, No
Loan	Binary	Yes, No.
Contact	Categorical	Unknown, telephone, cellular
Day	Numeric	Numeric
Month	Categorical	0-12
Duration	Numeric	Numeric
Campaign	Numeric	Numeric
Pdays	Numeric	Contacted, numeric
Pervious	Numeric	Numeric
Poutcome	Categorical	Unknown, failure, success, other
Output	Binary	Yes, No

Three types of attributes are depicted in Table 2 such as numerical (which includes age, balance, day, duration, campaign, pdays and previous), categorical (consists of job, marital, education, contact, month, poutcome) and binary categories which includes the attributes as yes or no in their classes (for example default, housing, loan, output) (Elsalamony, 2014).

PRIVACY PRESERVING BASED SVD MODEL

A privacy preserving technique for data mining for distributed datasets on cloud computing which has been designed as levels. Moreover, the description includes demonstration of the effect and the interaction between the levels. In the first level, the datasets are analyzed to discover depth relationships among attributes (key attributes, quasi-identifier and sensitive attributes) composing the dataset and to discover the variance of attributes in dataset due the change in quasi-identifier that lead to disclosure the sensitive information. To discover the variance resultant due changes in quasi-identifier in the dataset, SVD and LSA Methods are applied. This level is adopted by the owner of data sets or honest third party. In second level, after classification these attributes depending on the variance between these attributes in level one. The attributes that have lowest variance is the highest candidate to be a quasi-identifier. While the attributes that have highest variance, they should be anonymized. These attributes of highest variance and sensitive attributes are anonymized based on enhancing of k-Anonymity Method.

Attributes variance in dataset: As mentioned above, the data set is prepared for analysing. The study applied the pseudo code to apply SVD.

Coding scheme: The data set is depended in this study are described in dataset description, it is the bank dataset. To start mining data for the bank dataset, coding scheme is needed to translate string values to scalar as the following as shown in Table 3 and 4.

Result analyses: In order find the variance between the above attributes, it is needed to analyse the data set. SVD is used to decompose and to detect the variance of these attributes. SVD is used to study the effect of every one of these attributes on other attributes. Figure 1 is shown the analysis of first 100 attributes vectors.

From this Fig. 1, it is obvious that 'loan' and 'contact' are having the less variance in the first 100 attribute vectors within the dataset. Thus, they are highly nominated to be the quasi-identifier for the table but with minimum valuable information gained when mining this table. The same result was achieved even for 300 attribute vectors as shown in Fig. 2.

Table 3: Coding scheme

I	ID	Values
1	Age	Neumeric
2	job	Admin: 0, unknown: 1, unemployed: 2, management: 3, housemain: 4, entrepreneur: 5, student: 6, blueCollar: 7, selfEmployed: 8, retired: 9, technician: 10, services: 11
3	Marital	Married: 0, divorced (widowed): 1, single: 2
4	Education	Unknown: 0, secondary: 1, primary: 2, tertiary: 3
5	Default	Yes: 0, no: 1
6	Balance	Numeric
7	Housing loan	Yes: 1, no: 0
8	Personal loan	Yes: 1, no: 0
9	Contact	Unknown: 0, telephone: 1, cellular: 2
10	Day	Numeric
11	Month	0-12
12	Duration	Numeric
13	Campaign	Numeric
14	pDays	-1: not contacted, numeric
15	previous	Numeric
16	poutcome	Unknown: 0, failure: 1, success: 2, other: 3
17	Subscribed	Yes: 1, No: 0

Table 4: Examples on coding scheme for attributes' vectors

I	Attribute vector	Coding
1	30;"unemployed";"married";"primary";"no";1787;"no";"no";"cellular";19;"oct";79;1;-1;0;"unknown";"no"	30,2,0,2,1,1787,0,0,2,19,10,79,1,-1,0,0,0
2	33;"services";"married";"secondary";"no";4789;"yes";"yes";"cellular";11;"may";220;1;339;4;"failure";"no"	33,11,0,1,1,4789,1,1,2,11,5,220,1,339,4,1,0
3	35;"management";"single";"tertiary";"no";1350;"yes";"no";"cellular";16;"apr";185;1;330;1;"failure";"no"	35,3,2,3,1,1350,1,0,2,16,4,185,1,330,1,1,0
4	30;"management";"married";"tertiary";"no";1476;"yes";"yes";"unknown";3;"jun";199;4;-1;0;"unknown";"no"	30,3,0,3,1,1476,1,1,0,3,1,199,4,-1,0,0,0
5	59;"blue-collar";"married";"secondary";"no";0;"yes";"no";"unknown";5;"may";226;1;-1;0;"unknown";"no"	59,7,0,1,1,0,1,0,0,5,5,226,1,-1,0,0,0

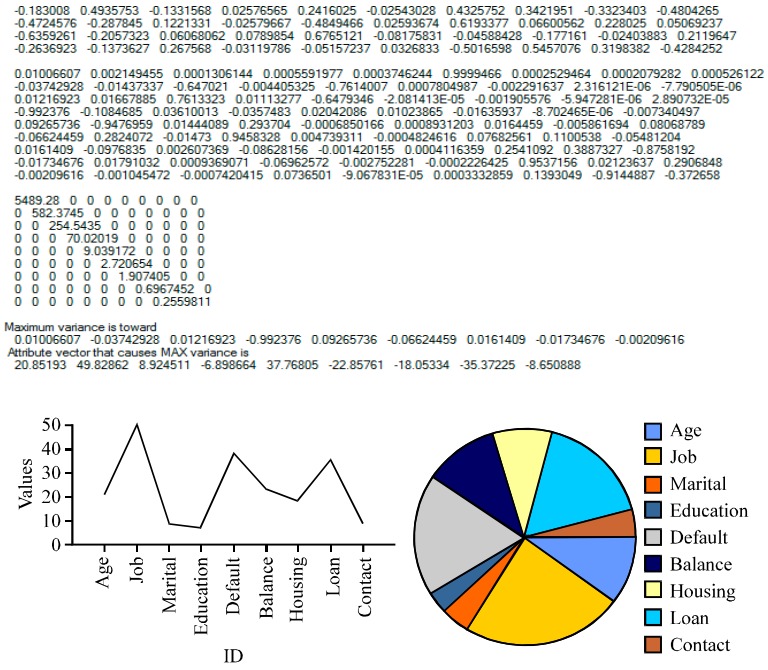

Fig. 1: Analyses of first 100 attribute vector

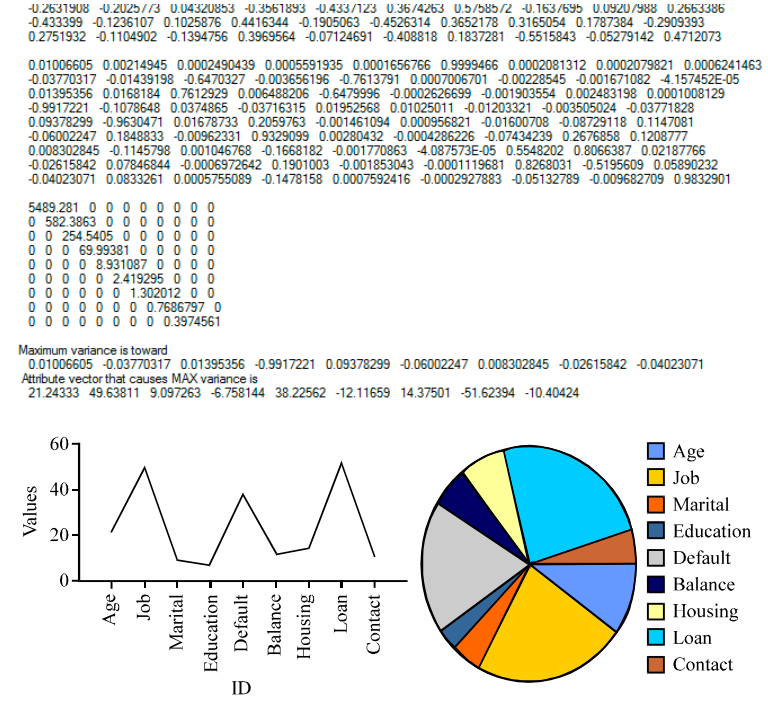

Fig. 2: Analysis for the 300 attribute vector within the dataset

CONTRIBUTION AND FUTURE WORK

In this study, an innovative approach of data analysis is introduced to capture semantic latent relationships among quasi-identifiers and the key attributes; this approach goes beyond the traditional methodologies that have been exploited to discover quasi-identifiers which preserve the anonymity for certain

database. Database tables nowadays hold an intensive amount of semantic especially with the advent of the cloud repositories; hence, the traditional methodologies are not enough to preserve the privacy of the published data bases where linkage attacks can reduce the anonymity factor by binding quasi-identifier of the published data with other related data base tables that have a correlation with the quasi-identifier in the semantic space. Our research, discover depth relationships among attributes composing the table and later on the variance in data table due the change in quasi-identifiers to asses assess the generalization or suppression methodology used to establish the k-anonymity in that table. In the future research, we will preserve privacy by developing k-anonymity.

CONCLUSION

The aim of this study is to transfer the conceptualization of the data base table into another domain which maintains the privacy and reduces the loss of information. Basically, the idea is to decompose the table using the Singular Value Decomposition (SVD) into key-attribute and sensitive information as an orthogonal matrix; this in a side and represent the quasi-identifiers as another orthogonal matrix in the other hand. SVD will handle the mapping between these two matrices. Latent Semantic Analysis (LSA) will reveal latent semantic relationships among attributes in the semantic space (Landauer *et al*., 1998); this technique is innovative in term of preventing more smart attack which tries to build linkages and binding across distributed data bases over the cloud. The main goal of this study is to discover the attributes with high variance which lead to disclosure the sensitive information to be anonymities. While the attributes with low variance, they can consider as quasi identifier.

ACKNOWLEDGEMENT

The dataset of this study was supported by University of California at Irvine (UCI) Machine Learning Repository.

REFERENCES

Asuncion, A. and D.J. Newman, 2007. UCI machine learning repository. University of California, Department of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Balouchestani, M., L. Sugavaneswaran and S. Krishnan, 2014. Advanced K-means clustering algorithm for large ECG data sets based on K-SVD approach. Proceedings of the 9th International Symposium on Communication Systems, Networks and Digital Signal Processing, July 23-25, 2014, Manchester, pp: 177-182.

Cacheda, F., V. Carneiro, D. Fernandez and V. Formoso, 2011. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. ACM Trans. Web, 5: 2-33.

Elsalamony, H.A., 2014. Bank direct marketing analysis of data mining techniques. Int. J. Comput. Applic., 85: 12-22.

Fung, B.C.M., K. Wang, R. Chen and P.S. Yu, 2010. Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv., Vol. 42. 10.1145/1749603.1749605.

Han, J.W., M. Kamber and J. Pei, 2012. Data Mining Concepts and Techniques. 3rd Edn., China Machine Press, Beijing, China, pp: 327-390.

He, X., H.H. Chen, Y. Chen, Y. Dong, P. Wang and Z. Huang, 2012. Clustering-Based k-Anonymity. In: Advances in Knowledge Discovery and Data Mining, Tan, P.N., S. Chawla, C.K. Ho and J. Bailey (Eds.). Springer, Heidelberg, Germany, ISBN-13: 978-3-642-30217-6, pp: 405-417.

Jain, P., P. Tapashetti, A.S. Umesh and S. Sharma, 2013a. Privacy preserving processing of high dimensional data classification based on sample selection and singular value decomposition. Proceedings of the International Conference on Control, Automation, Robotics and Embedded Systems, December 16-18, 2013, Jabalpur, pp: 1-5.

Jain, P., N. Pathak, P. Tapashetti and A.S. Umesh, 2013b. Privacy preserving processing of data decision tree based on sample selection and singular value decomposition. Proceedings of the 9th International Conference on Information Assurance and Security, December 4-6, 2013, Gammarth, pp: 91-95.

Landauer, T.K., P.W. Foltz and D. Laham, 1998. An introduction to latent semantic analysis. Discourse Process., 25: 259-284.

Loukides, G. and A. Gkoulalas-Divanis, 2012. Utility-preserving transaction data anonymization with low information loss. Expert Syst. Applic., 39: 9764-9777.

Monreale, A. and D. Pedreschi, R.G. Pensa and F. Pinelli, 2014. Anonymity preserving sequential pattern mining. Artif. Intell. Law, 22: 141-173.

Moro, S., P. Cortez and R. Laureano, 2011. Using data mining for bank direct marketing: An application of the CRISP-DM methodology. Proceedings of the European Simulation and Modelling Conference, October 24-26, 2011, Guimaraes, Portugal.

Pan, Y. and T. Chen, 2012. Research on privacy preserving on k-anonymity. J. Software, 7: 1649-1656.

Soodejani, A.T., M.A. Hadavi and R. Jalili, 2012. k-Anonymity-Based Horizontal Fragmentation to Preserve Privacy in Data Outsourcing. In: Data and Applications Security and Privacy XXVI: 26th Annual IFIP WG 11.3 Conference, DBSec 2012, Paris, France, July 11-13, 2012. Proceedings, Cuppens-Boulahia, N., F. Cuppens and J. Garcia-Alfaro (Eds.). Springer, Heidelberg, Germany, ISBN-13: 978-3-642-31540-4, pp: 263-273.

Sweeney, L., 2002. k-Anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst., 10: 557-570.

Weiss, G.M., 2009. Data Mining in the Real World: Experiences, Challenges and Recommendations. CSREA Press, Nevada, USA., pp: 124-130.

Wu, X., 2004. Data mining: Artificial intelligence in data analysis. Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology, September 20-24, 2004, Beijing, China.

Xu, Y., T. Ma, M. Tang and W. Tian, 2014. A survey of privacy preserving data publishing using generalization and suppression. Applied Math. Inform. Sci., 8: 1103-1116.

Yang, Y. and M. Ma, 2013. Proceedings of the 2nd International Conference on Green Communications and Networks 2012 (GCN 2012): Volume 3. Springer Science and Business Media, New York, USA., ISBN-13: 9783642354700, pp: 79-84.