A Credential Data Privacy Preserving in web Environment using Secure Data Contribution Retrieval Algorithm

Kumaran U
Research Scholar, School of Information Technology and Engineering, VIT University
Vellore, India
kumaranvit.phd@gmail.com

Neelu Khare
Assistant Professor, School of Information Technology and Engineering, VIT University
Vellore, India
neelu.khare@vit.ac.in

ABSTRACT: *Preservation of privacy is a significant aspect of data mining and as the secrecy of sensitive information must be maintained while sharing the data among different untrusted parties. There are much application is suffering from vulnerable, data leakage, data misuse, and sensitive data disclosure issues. To protect the privacy of sensitive data without losing the usability of data, various techniques have been used in privacy-preserving data mining (PPDM). However, a system is unable to maintain the privacy during online services. Some of the approaches are available to maintain the tight privacy, but they fail to minimize the execution time and error rate. The main objective of the article is to contribute and retrieve the data with minimal classification error and execution time with enhanced privacy. To overcome the issues, the paper introduces the Secure Data Contribution Retrieval algorithm to fulfill the current issues. Proposed algorithms define a privacy policy and arrange the security based on requirements. This design applies the privacy based on the compatibility of applications. This approach computes the union of private multi-datasets that each of the interacting with attributes and actors and another that tests the inclusion of an element held by one actor in a subset of another. It displays the table with hidden attributes in multiple categories wise for a user. This approach is capable of satisfying the accuracy constraints for multiple datasets. It also considers the efficient data extraction with a good ranking of attributes in tables. Based on experimental result proposed approach performs well regarding success rate, error rate and system execution time compare than existing methods.*

Keywords: Privacy Preserving in Data Mining, Web Mining, Secure Data Contribution Retrieval Algorithm, Success Rate, Error Rate, System Execution Time

Received: 17 May 2019, Revised 23 June 2019, Accepted 29 June 2019

DOI: 10.6025/jic/2019/10/3/102-110

1. Introduction

Nowadays, Privacy Preserving in Data Mining (PPDM) is playing important role data mining fields. It is the challengeable task

for the normal user to contribute and retrieve the sensitive information safely. In the current system, some technologies are available to maintain the secrecy of data. While individual services may provide interesting information/functionality alone, in most cases, users' queries require the combination of several Web services through service composition. In spite of the large body of research devoted to service composition, service composition remains a challenging task in particular regarding privacy. In a nutshell, privacy is the right of an entity to determine when, how, and to what extent it will release private information. Privacy relates to numerous domains of life and has raised particular concerns in the medical field, where personal data, increasingly being released for research, can be or have been, subject to several cases of abuse, compromising the privacy of individuals.

In literature, classification approach is utilized in many types of real-time application scenario with a respective domain. It's focused to cluster relevance data for many kinds of the business purpose for normal user. However, this method is quite dull in the terms of data extraction and cluster for application based un-serialized data. Here, existing approaches work for index designing of similar query retrieval. Here, some privacy techniques are introduced to maintain the privacy preserving for web based application. However, it only provided privacy for a limited number of data and their rights. In this scenario, data owner have to specify the service utilization and their access of data each section wise like authenticated, authorized for view general info (limited) and Authorized (fully) using DAML-S ontology. However, this approach is creating complexity to retrieve the query from a web. It may produce the error for data classification. P3P approach is introduced to apply privacy for web based application. It produced simple tools to analyze and visualize the large volume of data. However, providing effort of privacy is not sufficient for the huge amount of data. Anonymization technique worked to remove or encrypt personal or sensitive information from a given data so that the person whom the data refers to remain anonymous. Therefore Anonymization based PPDM is an approach where the identity or sensitive information about a person is hidden. However, this method suffers heavy information loss.

To overcome the current issues, the paper proposed Secure Data Contribution Retrieval algorithm for secure data contribution and retrieval from websites or online services. Proposed algorithms define a privacy policy and arrange the security based on requirements. This design applies the privacy based on the compatibility of applications. Here privacy is defined by data holder or owner based on their application requirement. This algorithm is not highly devoted for strong security in data; it also considers the how security is efficient to contribute and retrieve the data in minimal execution time and classification error. Initially, this system verifies the application compatibility. Hence, it proceeds to apply the privacy with rule-based classification approaches success rate. The systems can provide the accuracy constraints with multiple datasets. It also considers the effective information retrieval with good indexing of attributes. This design does not apply privacy on data; it also considers the types of application services. This designed utilized negotiation method to find out privacy compatibility of web based application and matching algorithm for compatibility verification. It reduces the I/O costs and efficiently exploits DBMS buffer management strategies. It operates the union of private multi-tabular data that each is interacting with attributes and actors and another that tests the inclusion of an element held by one actor in a subset of another actor. This paper contribution work follows as:

• The design proposed Secure Data Contribution Retrieval algorithm for secure data contribution and retrieval from websites or online services.

• Develop the framework to verify the privacy compatibility between application and running environment.

• Define the privacy level for application based on application requirement.

• Minimize the classification error and data retrieval time for the web-based distributed an application.

• Based on experiment result, its display that proposed approach perform well on success rate, error rate and system execution time compare than existing methods.

The rest of paper follows as Section 2 explain related work which is closest to proposed mechanism. Section 3 introduces the proposed research methodology with proposed techniques elaboration and algorithm details. Section 4 elaborated the result and discussion. Section 5 summarizes the proposed work details with future enhancements.

2. Related Work

Paper [1] developed a fast perturbation algorithm which contains tree to perform the database perturbation processes for

preventing credential information. In [2] focused on the problems of HUIM and privacy-preserving utility mining (PPUM) algorithms to mine HUIs and hide the sensitive high-utility item sets. Paper [3] focused on the significance role of Data Warehousing and Data Mining approach for business applications. It assists the organization to consolidate data from different sources. In [4] developed Privacy Preservation in Data Mining using Cluster based Greedy Method for sensitive data like medical data, government, and customer care management system. Paper [5] designed Rampart framework to protect sensitive information in mined data.

In [6] utilized data reduction algorithm CFS Subset and four different neural networking algorithms namely Multilayer Perception, Stochastic Gradient Descent, Logistic Regression and Voted Perception to reduce data set. In [7] did the comparative study of various Privacy Preserving Data mining techniques and their feature and limitations. In [8] personalized anonymization approach designed for persevering the privacy during sensitive data publication. Paper [9] designed Data distortion method for achieving privacy protection association rule mining and privacy protection. In [10] worked to maintain the privacy of distributed data with new anonymization and slicing technique. Paper achievement is to publish a Genuine or Anonymized view of integrated data, which will be immune to attacks.

Paper [11] designed a combined weight of attributes and Kong's approach together to derive the weights of attributes directly. In [12] worked a novel DMBSs technique which contains parameters to set the size of data buffer caching in a memory, and disc I/O dataset reduction. Paper [13] reviewed a various approaches which was presented by different researchers and also explained the data reduction methods to extract the essential information. Paper [14] developed a method to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing and information (i.e., the data mining results) delivering. It can help to protect sensitive information. In briefly, it performs four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. In [15] studied current scenario of privacy preserving data mining tools and techniques to develop the effective approach for PPDM.

Paper [16] develop Heuristic approach method for hiding a set of frequent items containing highly sensitive knowledge that only remove information from the transactional database with no hiding failure. In [17] introduced the reversible integer transformation method in the image processing and developed a Reversible Data Transform (RDT) algorithm that can disrupt and restore the data. RDT algorithm is used an adjustable weighting mechanism to adjust the degree of data perturbation for maximizing the flexibility of privacy-preserving. In [18] explained edge recognition technique to find low-value data, to keep input data for distribution purpose. Paper [19] developed a low-complexity robust data-dependent dimensionality reduction model for reduced-rank beam forming and steering vector estimation. In [20] designed K – anonymity method to achieve privacy in many data publishing applications. It anonymized to data utility reduction and more information loss of publishing table.

In [21] worked on Singular value decomposition based data Perturbation method to extract the original data from perturbed data. It provided perturbed data for all sample to maintain the privacy. Paper [22] developed k – anonymity with the decision tree to design pattern for medical data. It avoids linking attack in Electronic Health Record systems. In [23] developed perturbation based PDM method to data distortion and privacy maintenance. This method is designed to preserve the data without compromising sensitive information.

3. Research Methodology

The section briefly explains the proposed Secure Data Contribution Retrieval algorithm for secure data contribution and retrieval in web environments. The diagrammatical view of proposed framework is elaborated in Figure 1. This system is classified into following ways namely user interface, Input data processing, privacy definition, security compatibility identification, application & web environment matching process, result visualization. Here, data owner have to choose his application, set the privacy level. Hence proposed algorithm applies the privacy based on the application requirement. Finally, a user can search the query from the web and retrieve the result with minimal execution time and classification error rate.

3.1 Content Owner

Here, the content owner can register in a web environment and hence it will go for login after activation of owner account. Next, he/she can move to contribute the data in the web environment. Here, a content owner can provide credential information without taking any risk.

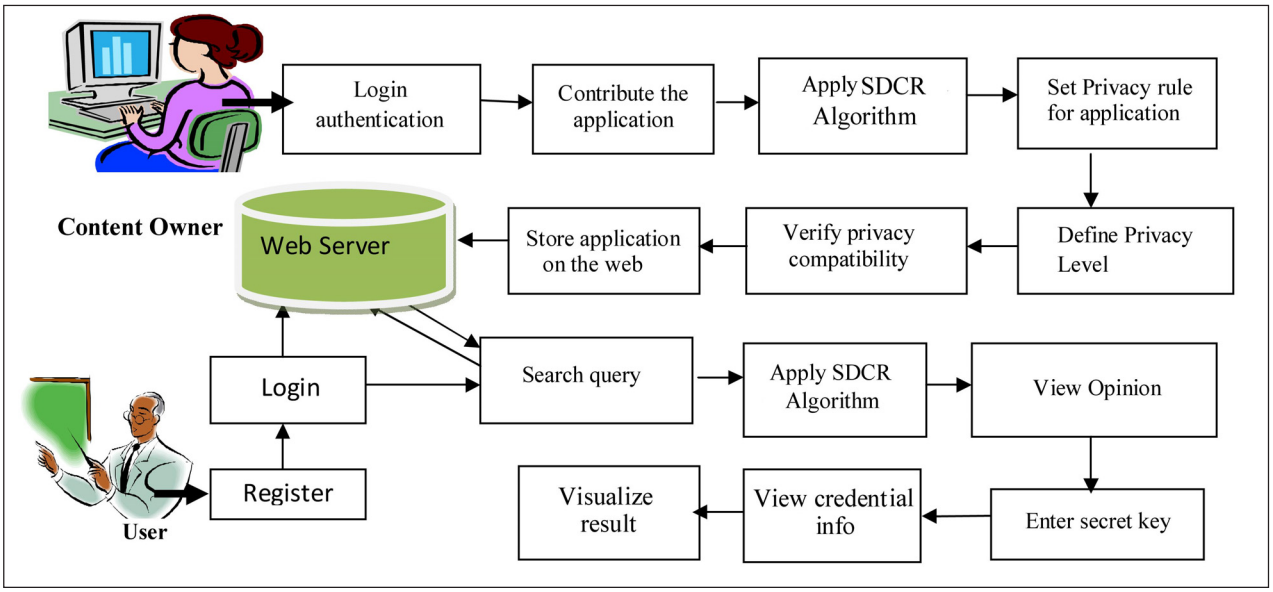

Figure 1. System architecture for Secure Data Contribution Retrieval algorithm in web environment

3.2 User

The user will create the account here to access the information from the web environment. After login, a user can search the query; view the opinion and see the credential information without data loss. However, a user needs to enter secret key to display the credential information.

3.3 Data Contribution

This module allows the content owner to contribute his/her credential information or data in a web environment. Here, a content owner can enter multiple data with multiple categories. The content owner can apply the security for his/her data based on their application requirement.

3.4 Secure Data Contribution Retrieval Algorithms

Secure Data Contribution Retrieval algorithm for secure data contribution and retrieval from websites or online services. Proposed algorithms define a privacy policy and arrange the security based on requirements. This design applies the privacy based on the compatibility of applications. Here privacy is defined by data holder or owner based on their application requirement. This algorithm is not highly devoted for strong security in data; it also considers the how security is efficient to contribute and retrieve the data in minimal execution time and classification error. Initially, this system verifies the application compatibility. Hence, it proceeds to apply the privacy with rule-based classification approaches success rate. The system can provide the accuracy constraints with multiple datasets. It also considers the effective information retrieval with good indexing of attributes. A proposed system verifies privacy compatibility between application services within a composition. It works according to the notion of privacy and cost model. A threshold is defined for application services to cater the privacy compatibility. This application is mostly utilized to get secure medical information like patient details, hospital details, lab facility and doctor visits, etc. It applies secure mining preference services based item correlations. This design develops the privacy rule for medical data to protect credential information. It contains three phases mainly the privacy policy measurement, signature, and the secure item preferences. Propose system implements a multi-agent based privacy mechanism to provide level by level privacy along with classified datasets. This algorithm is integrated with random forest algorithm to classify the credential datasets. The pseudo code of proposed algorithm is shown below.

The pseudo code Secure Data Contribution Retrieval algorithm

```
Input:  Table T with D tuples containing apparently identifiers A and Credential
        attribute C
Output: Credential Attribute      table T*
```

```
Procedure
                          Store the tuples D in table T
                            Update multiple tuples D in respective table T
                            Define the privacy rule PR
                             Collect application privacy requirement Level PL
                            Apply PR on D
                           Verify the compatibility
                          If matches then
                         Proceed to PL and Store in T
                       Else
                         Recheck the PL and apply PR
 Classify the data attribute wise
Index the attribute
Visualize the credential result CR
End
```

3.5 System Classifier and Display Results

This module displays retrieved result in multiple sets of attribute dataset along with hidden attributes tables. It's also considers ranking of attributes according to the user query. This module represents the data with multiple categories of attributes with multiple selections.

4. Result and Discussion

4.1 Experiments Setup

To compare the proposed mechanism with existing algorithm, experiments were deployed with Intel Dual Core Processor with 2 GB RAM running with windows7, JDK 1.8, Netbeans 8.0 Profiler plug-in, Apache Tomcat 8.0.3, and MYSQL 5.5 database. To evaluate proposed system with existing method we used Java base open source with Weka 3.7.2 library.

4.1.1 Data

For our experimental evaluations, proposed techniques select medical domain three different datasets namely cancer dataset with 250 records, HIV dataset 200 records and Diabetes dataset with 250 records. For retrieving the query from the centralized server used JAVA based Secure Data Contribution Retrieval system.

4.2 Result

In this phase, Secure Data Contribution Retrieval algorithm represents the mathematical model to the privacy level of data without affecting the quality and efficiency of application in the web environment. Here, it demonstrates success rate, error rate, and system execution time to evaluate the performance of proposed mechanism. It was indentified, how can achieve minimum error rate with execution time and maximum success rate.

4.2.1 Success Rate (SR)

Success rate (SR) is the probability for query retrieval is the success, means at least once a query is hit. Suppose that the queried resources are uniformly spread in the network with replication ratio R, and then SR can be evaluated as in equation (1)

$$SR = \frac{(T_N + T_P)}{(T_N + T_P + F_N + F_P)} \times 100 \quad (1)$$

Where T_N true negative, T_P is truly positive, F_N is false negative, & F_P is false positive. This derivation indicates that SR mostly

depends on some parameters condition. Based on parameters condition, this method assists to a researcher to achieve the success rate.

4.2.2 Error Rate

Error rate represents the unclassified or uncluttered data ratio during patient data retrieval from a database. Here, it represents error rate in of total clustered data in the statically way. Proposed approach counts the total number of the data and subtracts with obtained result which is indicated in equation (2).

$$ER = \frac{V_{Approx} - V_{Exact}}{V_{Exact}} \times 100 \tag{2}$$

Where V_{Approx} is the approximate or total count of patient datasets and V_{Exact} is the exact or obtained result of patient datasets.

4.2.3 System Execution Time (SET)

In this section, proposed method describes a mathematical model for system execution time in equation (3). In this step, proposed method is calculated as system execution times on the user query. Here system execution time (SET) is calculated as:

$$SET = T_{CD} \times T_{AR} \tag{3}$$

Where T_{CD} is a total number of patient dataset and T_{AR} is average retrieval time for a patient dataset.

Table 1 represents Success Rate (*SR*) in %, Error Rate (*ER*) in % and System Execution Time (*SET*) in seconds, for Cancer, HIV and Diabetes patient and we display their average values for respective parameter with respective dataset. Here, proposed mechanism Secure Data Contribution Retrieval algorithm (*SDCRA*) is compared with existing approaches namely as Perturbation [23],singular value decomposition (*SVD*) [21], Singular Value Decomposition data Perturbation (*SVD + DP*) [21], *K*-anonymity with Decision Tree (*KA + DT*) [22], on given parameters like a *SR, ER* and *SET*.

Learning Algorithms	**Cancer**			**HIV**			**Diabetes**		
	SR	ET	SET	SR	ET	SET	SR	ET	SET
Perturbation	89	20	6	91	17	4	92	13	3
SVD	85	17	9	88	15	7	90	11	6
SVD+DP	93	6	12	94	4	9	95	3	7
KA +DT	96	12	8	97	10	5	98	8	4
SDRCA	98	3	3	99	2	2	99.5	1	2

Table 1. Success Rate, Error Rate and System Execution Time of Cancer, HIV, and Diabetes Patients

Based on Figure 2 to 4 result performance, it can be said that proposed Secure Data Contribution Retrieval algorithm perform best for overall dataset namely as the medical patient dataset, Cancer, HIV, Diabetes. In details, behalf of success rate which presents a classification of data attributes where the nearest competitor was a *KA + DT* [22]. Regarding error rate which displays the accuracy attribute after parameters reductions where the closest competitor is *SVD + PD* [21]. In terms of systems execution time which performs the query retrieval from systems where Perturbation is the nearest approach [23]. Proposed approach enhances Success Rate 1.83% reduces the Error Rate 2.33% and minimizes the system execution time 2 seconds. Finally, this paper claim that proposed Secure Data Contribution Retrieval algorithm shown well on every corresponding parameter for all dataset.

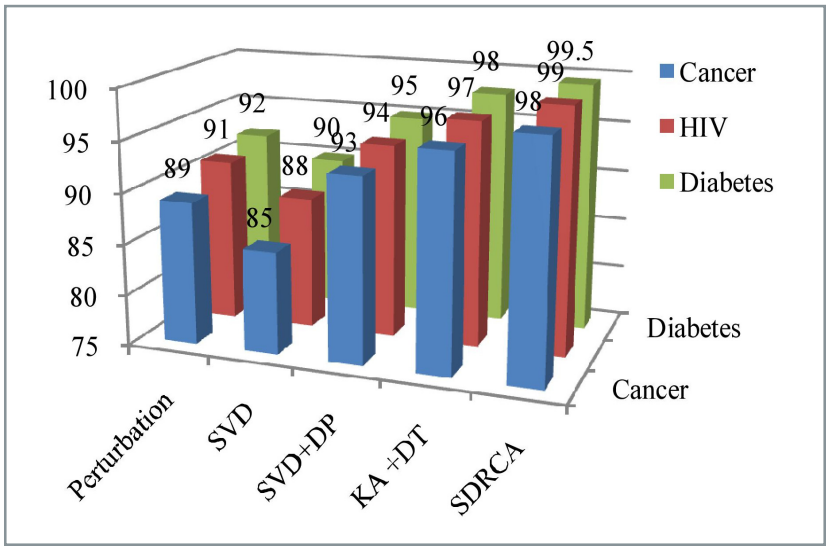

Figure 2. Success Rate for Cancer, HIV, and Diabetes Patients

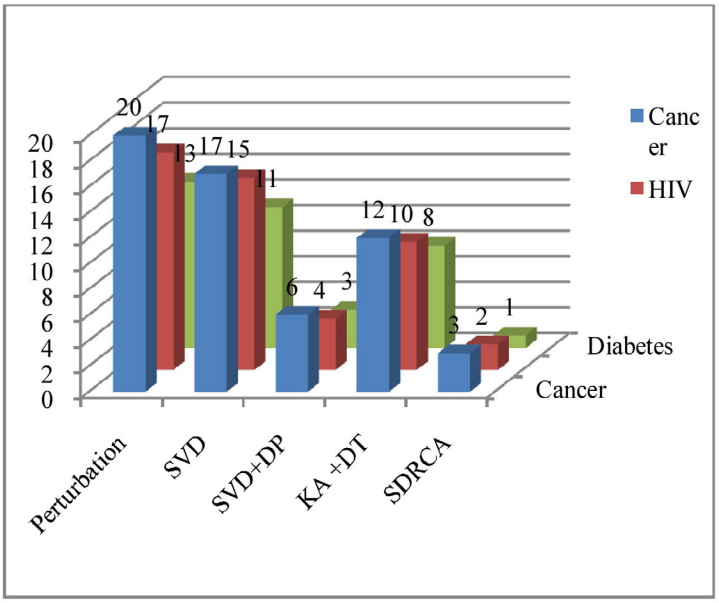

Figure 3. Error Rate (%) for Cancer, HIV, and Diabetes Patients

5. Conclusion

The paper presents Secure Data Contribution Retrieval algorithm for securing data contribution and retrieval from websites or online services. Proposed algorithms define a privacy policy and arrange the security based on requirements. This design applies the privacy based on the compatibility of applications. Here privacy is defined by data holder or owner based on their application requirement. This algorithm is not highly devoted for strong security in data; it also considers the how security is efficient to contribute and retrieve the data in minimal execution time and classification error. Initially, this system verifies the application compatibility. Hence, it proceeds to apply the privacy with rule-based classification approaches success rate. This method is capable of satisfying the accuracy constraints for multiple datasets. The proposed approach also considers the efficient information extraction with a good ranking of attributes in tables. It also reduces the I/O costs and efficiently exploits DBMS buffer management strategies. This approach performs the union of private multi-datasets that each of the interacting

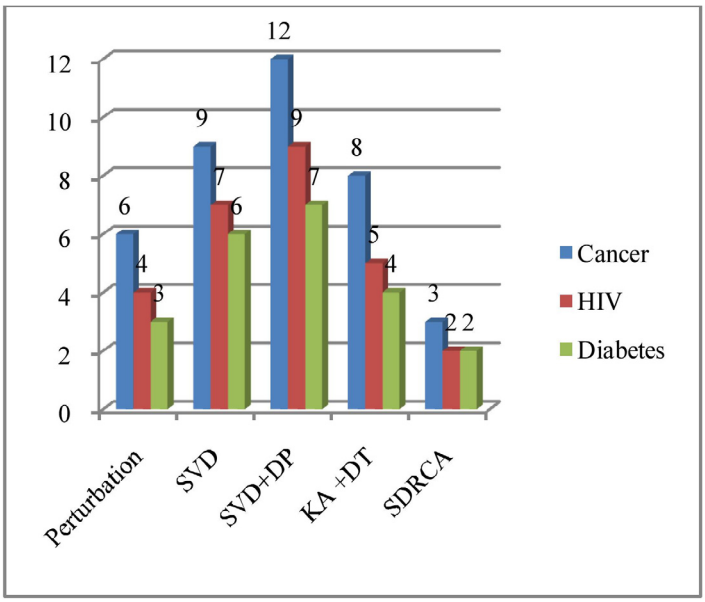

Figure 4. System Execution Time for Cancer, HIV, and Diabetes Patients

with attributes and actors and another that tests the inclusion of an element held by one actor in a subset of another. The proposed approach presents unique data extraction operation with hidden attribute table. Proposed approach enhances Success Rate 1.83% reduces the Error Rate 2.33% and minimizes the system execution time 2 seconds.

In future, this work can extended with geo-social data to maintain the privacy of location detail and user information for un-structured data.

References

[1] Yun, U., Kim J. (2015). A fast perturbation algorithm using tree structure for privacy preserving utility mining, *Expert Systems with Applications*, 42 (3) 1149-1165.

[2] Lin, J. C. W., Gan, W., Fournier-Viger, P., Yang, L., Liu, Q., Frnda, J., Voznak, M. (2016). High utility-itemset mining and privacy-preserving utility mining, *Perspectives in Science*, 7, 74-80.

[3] Joseph, M. V. (2013). Significance of Data Warehousing and Data Mining in Business Applications, *International Journal of Soft Computing and Engineering,* 3 (1) 329-333.

[4] Hariharan, R., Mahesh, C., Prasenna, P., Kumar, R. V. (2016). Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach, *Indian Journal of Science and Technology*, 9 (3) 1-8.

[5] Xu, L., Jiang, C., Chen, Y., Wang, J., Ren, Y. (2016). A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining, *Computer*, 49 (2) 54-62.

[6] Raiwani, Y. P., Panwar, S. S. (2015). Data Reduction and Neural Networking Algorithms to Improve Intrusion Detection System with NSL-KDD Dataset*, International Journal of Emerging Trends & Technology in Computer Science,* 4 (1) 219-225.

[7] Patel, A. K. (2016). A Survey: Privacy Preservation Data Mining Techniques and Geometric Transformation, *International Journal of Scientific Research in Science, Engineering, and Technology*, 2 (2) 106-111.

[8] Prakash M., Singaravel, G. (2015). An approach for prevention of privacy breach and information leakage in sensitive data mining, *Computers & Electrical Engineering*, 45, 134-140.

[9] Qi, X., Zong, M. (2012). An overview of privacy-preserving data mining. *Procedia Environmental Sciences*, 12, 341-1347.

[10] Tue, A., Priyadarshi, A. (2016). Data Mining with Big Data and Privacy Preservation, *International Journal of Advanced Research in Computer and Communication Engineering,* 5 (4).

[11] Zhang, L. (2014). A Weighted Attribute Decision Making Approach in Incomplete Soft Set. *In*: 2014 *International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC - 14)*, Atlantis Press, 1553-1556.

[12] Lee, J. H., Lee, Y. H., Park, J. S., Kim, S. H., Al-Khanjari, Z., Al-Hosni, N., Jeon, G. (2014). A Study on the Analysis of the Effectiveness according to Buffer Size of Storage, *International Journal of Software Engineering & Its Applications*, 8 (5) 1-14.

[13] Shrivastava, J., Shrivastava, N. (2014). A Review of Data Reduction/Extraction in Data mining from the Large set of Database, *International Journal of Electrical, Electronics and Computer Engineering*, 3 (2) 149-153.

[14] Xu, L., Jiang, C., Wang, J., Yuan, J., Ren, Y. (2014). Information security in big data: privacy and data mining, *IEEE Access*, 2, 1149-1176.

[15] Malik, M. B., Ghazi, M. A., Ali, R. (2012, November). Privacy-preserving data mining techniques: current scenario and prospects, *In*: IEEE *Third International Conference on Computer and Communication Technology (ICCCT),* 26-32.

[16] Mahendran, M., Sugumar, R., Anbazhagan, K., Natarajan, R. (2012). An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach, *International Journal of Advanced Research in Computer and Communication Engineering*, 1 (9) 737-744.

[17] Lin, C. Y. (2016). A reversible data transform algorithm using integer transform for privacy-preserving data mining, *Journal of Systems and Software*, 117, 104-112.

[18] Ahmadi, M., Ghaffari, H. (2014). Reducing the Size of Very Large Training Set for Support Vector Machine Classification, *International Journal of Soft Computing and Engineering (IJSCE),* 4 (5) 55-61.

[19] Li, P., Feng, J., de Lamare, R. C. (2015). Robust Rank Reduction Algorithm with Iterative Parameter Optimization and Vector Perturbation. www.mdpi.com/journal/algorithms*Algorithms*, 8 (3) 573-589.

[20] Kumar, S. N., Aparna, R. (2013). Sensitive Attributes based Privacy Preserving in Data mining using k-anonymity, *International Journal of Computer Applications*, 84 (1) 1-6.

[21] Li, G., Wang, Y. (2012). A privacy-preserving classification method based on singular value decomposition, *The International Arab Journal of Information Technology*, 9 (6) 529–534.

[22] Srivastava, A., Srivastava, G. (2015). Privacy Preserving Data Mining in Electronic Health Record using K-anonymity and Decision Tree, *International Journal of Computer Science & Engineering Technology*, 6 (7) 416-426.

[23] Patel, N., Lade S., Gupta, R. K. (2015). Quasi & Sensitive Attribute Based Perturbation Technique for Privacy Preservation, *International Journal of Advanced Research in Computer Science and Software Engineering*, 5 (11) 450-456.