Privacy Issues in Big Data and Enhanced Privacy Preservation Model

Panakal Harly Mary, Prajwal Sunil Solanki, M Lakshmi

Abstract:- Privacy is an essential factor for humans, but Big Data can expose set patterns from which trends specific to human behavior called personally identifiable information can be easily derived. We try to overcome the privacy issues in Big Data using traditional privacy preservation techniques, and K-anonymity is the most extensively used technique for preserving privacy for data publishing. This paper has investigated the privacy challenges in big data and proposes an enhanced privacy preservation model that protects the data against homogeneity and background knowledge attacks and maintains a balance between data quality and data privacy. The proposed algorithm gets executed with minimum running time. This technique will also aid data mining as it ensures data quality by reducing information loss.

Keywords: Big data, Data privacy, Data anonymization, K-Anonymization, Information loss

I. INTRODUCTION

Big data is a valuable constituent in today's world as it holds a significant role in various fields of research and management. Big data analysis leads to better decisions and strategic moves for overpowering businesses. Confidentiality, availability, integrity, and privacy are four essential aspects of Big data.

In the scenario of data publishing, we strive to minimize privacy risks and maximize data utility and data availability. Big data's privacy aspect mainly concentrates on the user's data rather than the whole data collection. Sensitive information of individual users can be protected using privacy preservation techniques. The privacy preservation models' objective is to share the data openly without revealing any individual's identity.

Data anonymization is the process in which changes are made to the data to prevent identity disclosure. It is also known as the de-identification process. Anonymization while publishing data usually refers to hide identifier attributes, i.e., attributes that are uniquely identify individuals like name, license number, voter id, passport number, etc. The purpose of the process is to avoid re-identification of the individual.

However, this process cannot prevent linking information from published data to the data obtained from external sources (i.e., linking attacks)[3]. Data privacy can get disrupted by linking attacks, background knowledge attacks, and homogeneity attacks. Sweeney et al. [12,13,14] proposed a privacy preservation model called the k-anonymization model to overcome the privacy leakage problem. K-anonymity is the most widely used method due to its simplicity, and this technique gets extensively adopted in various fields such as data mining, the medical industry, etc. K-anonymization is the process of carrying out modifications before sending the data to data analytics so that de-identification attempts will lead to K indistinguishable records. The cost of carrying out this method is significantly lesser than another anonymity technique [5]. This technique is prone to attacks such as unsorted matching, discreet, background knowledge attack, temporal attacks and homogeneity attacks [15][16] [17].

Many algorithms have been proposed [4,8,9,12,17] based on the K anonymity problem, but the information loss is considerably high.By viewing k-anonymization as a clustering problem [6], the information loss reduces without decreasing data utility. In a clustering-based k-anonymization algorithm, the set of records or data gets divided into clusters containing at least k records[6]. Our goal is to formulate an enhanced k-anonymization model based on the clustering problem with the least number of comparisons. Our model suffers minimal information loss and maintains a balance between data utility and data privacy. The paper is organized in the following order: We review the recent works related to K-anonymization and the basic concepts of anonymization from Sections 2 and 3. The most common privacy issues in Big Data are explained in Section 4. In section 5, we present our scheme along with its algorithmic description. The proposed enhanced model is evaluated, and the experimental results are described in section 6. We conclude our discussion in Section 7.

II. LITERATURE SURVEY

Researches have proposed multiple security theories and models to satisfy the three most significant aspects of Big data security confidentiality, availability, integrity, and privacy [7]. Data privacy-preservation models have been applied in various research domains such as data publishing and location-based services during the past few decades. The anonymization technique has been widely analyzed to ensure privacy preservation in Big data when dealing with quasi-identification attributes.Research shows that the majority of the commonplace k-anonymization techniques depend on speculation and concealment strategies. Wantong Zheng et al.

Manuscript received on 05 April 2021 | Revised Manuscript received on 20 April 2021 | Manuscript Accepted on 15 May 2021 | Manuscript published on 30 May 2021.

* Correspondence Author

Panakal Harly Mary*, Department of Computer Science Engineering S.R.M. Institute of Science and Technology, Chennai, India

Prajwal Sunil Solanki, Department of Computer Science Engineering S.R.M. Institute of Science and Technology, Chennai, India

Dr. M Lakshmi, Professor Department of Computer Science and Engineering S.R.M. Institute of Science and Technology, Chennai, India

[4], in their paper on K-anonymity algorithm based on improved clustering, talks about how his scheme has significantly improved data availability compared with k-member clustering, Mondrian multidimensional, and one-time k-means. However, many numerous restrictions have been identified in k-anonymization, mainly attacks such as unsorted matching, complimentary release, discreet and temporal attacks.[6]

K-anonymization gets implemented through generalization and suppression. Generalization is the procedure carried out to replace a specific value with a generalized one. Numeric values such as age and salary are generalized into intervals (e.g. [20-30]), and categorical values are generalized into a set of distinct values. Multiple generalization approaches have been suggested. In [20], a non-overlapping generalization-hierarchy is determined for each attribute of the quasi identifier.

Multiple privacy preservation models that do not depend on generalization hierarchies [17,21] have been proposed. LeFevre et al. [21] reconstructed the k-anonymity problem into a partitioning problem. In the partitioning approach, each partition possesses at least k records. Although the model appeared effective, it demanded a total order for every attribute domain and made it impractical when categorical values were involved.

Ji-Won Byun et al. [6] proposed an efficient k-anonymization model based on the clustering problem. In this approach, we find a set of clusters, also called equivalent classes, that contains at least k record each. After advancing the analysis, Wantong Zheng[1] suggested an algorithm that can reduce information loss while generating clusters. The proposed model performed positively with numerical as well as categorical attributes. There is a trade-off between data utility and data privacy; if we improve one of them, the second decreases accordingly. Therefore, it essential to maintain a balance between data privacy and data utility.

III. BASIC CONCEPTS OF ANONYMIZATION

A. Identifiers

Identifiers usually refer to any sensitive information about an individual. This type of information demands encryption to protect the data infringement. An identifier could be a credit card number, ID number, bank account details, name etc.

B. Quasi-Identifiers

Quasi-identifiers refer to non-sensitive information linked to any other external database to further investigate a particular individual's records. For example, in the case of bank account details, the IFSC code can be linked with the bank branch, while sensitive information such as debit/credit card details is sensitive.

C. Sensitive Identifiers

These identifiers demand the maximum level of discretionary means. This information usually includes mobile phone numbers, medical records etc.

D. Equivalent Classes

Every equivalent class consists of a minimum of k records. One equivalent class is the collection of all the similar valued quasi-identifiers. Raw data produces m equivalent classes after k-anonymity where m = (n/k) where n is the number of records in the data table.

E. Big Data

Big data means an extensive collection of correlated data used by the businesses according to their business model to summarize, analyze records and reform their business strategies. According to Francis Diebold, big data is the explosion in the quantity and quality of availably and potentially relevant data.

F. Data Privacy

With an expansion in the quantity of live internet audience, there is a need for every individual and businesses to raise privacy consciousness to avoid fraudulent attacks. Data Privacy is the keyword for the practices that guarantee that data are used only for its affianced needs. With as small as giving unnecessary permission to mobile application to as large as attacks on servers, a medium that enhances security by precisely detecting a malicious code is needed.

G. Data anonymization

Data anonymization, in simple terms, means to make sensitive data independent of other data stored along with it. It is a much-needed step to ensure that one single identifier's leak does not make the entire database vulnerable. Widely used Data anonymization techniques are data masking, Generalization, Data swapping, Data perturbation, and Synthetic Data. After the user grants cookies on a website, these cookies collect information about user activity. The data anonymization techniques remove the database's identifier, making it difficult to derive conclusions and enhance user experience. Thus, we cannot use anonymized data for marketing and research purposes.

H. Suppression

During suppression, quasi-identifiers get blackened by some constant alphanumeric characters like 0,*, etc., so that they cannot be linked to any external sources. Fig. 1. displays the sample dataset before generalization and suppression. The ZIP attribute from Fig 1. is suppressed in Fig. 2.

I. Generalization

Generalization is the procedure carried out to replace a specific value with a generalized one. Numeric values such as age and salary are generalized into intervals (e.g. [20-30]), and categorical values are generalized into a set of distinct values.Fig. 2. demonstrates the generalization of gender and age attribute.

ID	Gender	Age	ZIP
1	Male	27	680201
2	Female	22	680212
3	Male	23	680203
4	Male	30	680204
5	Female	28	680215
6	Female	24	680216

Fig. 1. Sample data

ID	Gender	Age	ZIP
1	Person	[26,30]	68020**
2	Person	[21,25]	68021**
3	Person	[21,25]	68020**
4	Person	[26,30]	68020**
5	Person	[26,30]	68021**
6	Person	[21,25]	68021**

Fig. 2. K-anonymous table (k=3)

J. NCP

Normalized Certainty Penalty (NCP) [22] is a standard algorithm for measuring information loss. It is also efficient and easy to use. In the outcome of our proposed model, the NCP value is converted to a percentage. The NCP value is divided by the total number of values in the dataset to estimate the NCP percentage. It is also called GCP (Global Certainty Penalty).

K. Distance and Information Loss Metric

Distance is a significant factor for achieving optimal clustering results. Hence, the distance metric should be chosen carefully. The other major factor for assessing k-anonymization is information loss. Information loss is an error condition that occurs during Big data transmission. We utilize an algorithm to determine the distance between records and between records and clusters to calculate information loss. The distance between numerical attributes cannot be applied to categorical attributes as they do not have complete ordering. We use the correlation between data semantics to measure the distance between categorical data. The equations for finding information loss are defined in [1].

IV. PRIVACY ISSUES IN BIG DATA

Privacy is the responsible factor that enables the individual to decide what data can be shared. A voluminous amount of data gets processed on the internet due to the progressive change in the internet application in different domains. If the data are in the public domain, then it is a threat to individual privacy. The information we share suffers multiple hazards such as surveillance, disclosure, discrimination, and personal embracement and abuse[19].

Many businesses analyze their customers' buying habits and suggest various products with multiple offers [18]. The suggestions are created by monitoring the customer's activity, and it is an individual privacy threat. Suppose a data holder shares the data with a third party after anonymizing it. In that case, the third party can link the data with external sources' data to identify sensitive information[3]. The disclosure of data is a severe privacy infringement. The revelation of data can also lead to discrimination. For example, when an individual's medical records get leaked, it can cause personal embarrassment and abuse.

Data analytics plays a vital role in decision-making, but the data should be kept safe from privacy threats such as homogeneity attacks, background attacks, and data infringements. Hence, privacy-preservation models in data analytics are very crucial.

V. PROPOSED SCHEME

In the scenario of data publishing, data quality is a significant factor. Anonymization problem requires minimal information loss to maintain a balance between data privacy and data utility. The clustering-based k-anonymization problem satisfies this criterion. We propose an enhanced privacy preservation model that protects the data against homogeneity and background knowledge attack, and it also maintains minimum total information loss cost. This technique will also aid data mining as it ensures data quality. The suggested algorithm makes a minimal number of comparisons to achieve the desired results.Fig. 3 illustrates the activity diagram for our model.

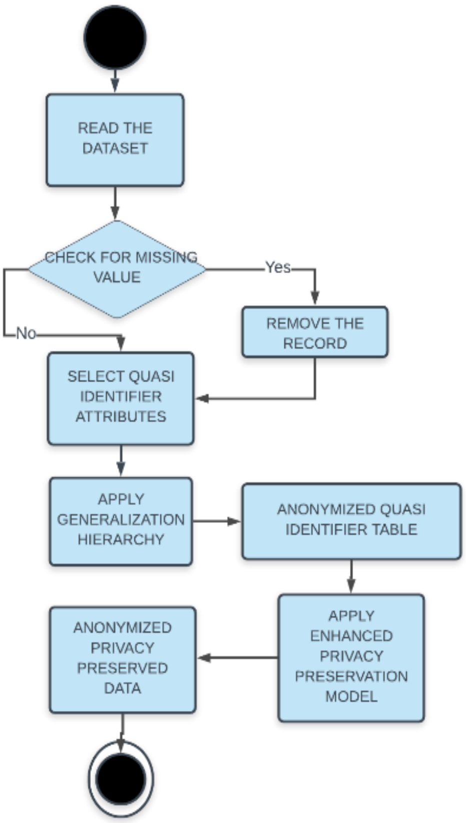
READ THE DATASET
CHECK FOR MISSING VALUE
Yes
No
REMOVE THE RECORD
SELECT QUASI IDENTIFIER ATTRIBUTES
APPLY GENERALIZATION HIERARCHY
ANONYMIZED QUASI IDENTIFIER TABLE
APPLY ENHANCED PRIVACY PRESERVATION MODEL
ANONYMIZED PRIVACY PRESERVED DATA

Fig. 3. Activity diagram for the proposed model.

A. Algorithmic Description

Our enhanced privacy preservation model follows the subsequent steps: The input to the algorithm is a set of n records S, and the privacy parameter is k. The algorithm classifies the n records from set S into different clusters or buckets. Each group will possess only k records in it.

The algorithm commences by checking whether the number of records in the set S is lesser than or equal to k. If the condition returns true, it suggests that it can produce only one cluster. It will return the set S. Until group S has records greater than or equal to the privacy parameter k, it chooses a record randomly. It adds them to the current cluster Ci. After appending the selected record to the cluster, the record gets removed from the set S.

The remaining records get examined for selecting the best record with minimum information loss. The best record selection is carried out using the divide and conquer algorithmic Paradigm. This method of selection is also called the Tournament Method. The set of remaining records in S gets divided into two collections. The array's primary collection contains the records from the 0th index till the middle of the array, and the secondary collection contains the uncommitted records.

The record index with minimum information loss from the primary collection min_index1, and the secondary collection, min_index2, is computed.

The records' information loss metric at indexes min_index1 and min_index2 get compared with the current cluster, and the record with minimum information loss gets selected as the best record. Its index gets returned to the calling function. The best record is added to the current cluster Ci and removed from S. The process gets repeated until the existing cluster has k records. Unique clusters such as Ci, Ci+1 get created until the Set S has n records greater than or equal to k. After generating all the clusters, the Set S gets examined for any spare records. If excess records are present, then the best cluster that satisfies minimal information loss is located in which the record can get added so that the data utility of the set remains secure. The best cluster is determined utilizing the exact approach used to obtain the best record previously. All the generated clusters have at least k records and get appended together to obtain the model's output.

The algorithms are as follows:

```
Algorithm 1:
Function enhanced_k_member_clustering (S, k)
Input: Set of records S and privacy preservation parameter k.
Output: Set of clusters containing at least k records
    if( size of S ≤ k )
      return S;
    end if;
    result = empty set ; r = a randomly selected record ;
    while ( size of S ≥ k )
          r = the furthest record from r;
          Remove r from set S;
          Append record r to cluster c;
          while( size of cluster c < k )
                 l = 0;
                 h = size of cluster – 1;
                 min_index = find_fittest_record(l ,h, S, c);
                 r = record at index min_index;
                 Remove the record r from set S;
                 Append record r to the cluster;
          end while;
          Append the cluster c to set outcome;
    end while;
    while (set S is not null)
             r = randomly selected record from S;
             Remove r from set S;
             l = 0;
             h = size of cluster - 1;
             min_index = find_fittest_cluster(l, h, S, c);
             c = Cluster at index min_index;
             Append the record r to the current cluster c;
    end while;
    return outcome;

End;
```

```
Algorithm 2:
Function find_fittest_record (l, h, S, c)
Input: Cluster c, set of records S, and index pointers l and h.
Output: Index of the fittest record
    if (l is equal to h)
      return l;
    end if;
    else if (h is equal to l+1)
      distance_l = IL(c ∪ {record at index l}) – IL(c);
    distance = IL(c ∪ {record at index h) – IL(c);
      If(distance_l >distance_h)
        min_index=h;
        return min_index;
      else
        min_index= l;
        return min_index;
      end if;
    else
      m = (l+h)/2;
      min_index_l= find_fittest_record(l, m, S, c);
      min_index_h= find_fittest_record(m+1, h, S, c);
      distance1= IL(c ∪ {record at index min_index_l}) – IL(c);
      distance2= IL(c ∪ {record at index min_index_h}) – IL(c);
      If (distance1 < distance2)
      return min_index_l;
      else
      return min_index_h;
      end if;
    end if;

End;
```

Published By:
Lattice Science Publication (LSP)
© Copyright: All rights reserved.

```
Algorithm 3:
Function find_fittest_cluster (l, h, r, c)
Input: Cluster c, record r from S, and index pointers l and h
Output: Index of the fittest cluster
    if (l is equal to h)
        return l;
    end if;
    else if (h is equal to l+1)
      cl = cluster at index l;
      ch = cluster at index h;
    distance_l= IL(cluster cl ∪ {r}) – IL(cl);
      distance_h= IL(cluster ch ∪ {r) – IL(ch);
    If(distance_l >distance_h)
          min_index=h;
          Return min_index;
       else
         min_index= l;
       Return min_index;
       end if;
    else
      m= (l+h)/2;
      min_index_l= find_fittest_cluster(l, m, S, c);
      min_index_h= find_fittest_cluster(m+1, h, S, c);
      cl = cluster at index min_index_l;9
      ch = cluster at index min_index_h;
    distance1= IL(cluster cl ∪ {r}) – IL(cl);
      distance2= IL(cluster ch ∪ {r) – IL(ch);
    If (distance1 < distance2)
      return min_index_l;
      else
      return min_index_h;
      end if;
    end if;

End;
```

The divide and conquer method utilized to obtain the best cluster, and the best record gets implemented with *O(n)* time complexity, where *n* is the number of remaining records or clusters, respectively.

$T(n) = T(floor(n/2)) + T(ceil(n/2)) + 2$ (1.1)
$T(2) = 1$ (1.2)
$T(1) = 0$ (1.3)
If n is a power of 2 ,then: $T(n) = 2T(n/2) + 2$ (1.4)
If the above recursion is solved,we get:
$T(n) = 3n/2 -2$ (1.5)

Hence, selection process of the best cluster and the best record performs 3n/2 -2 comparisons if n is a power of 2 as shown in (1.5). Furthermore, it performs more than 3n/2 -2 comparisons if n is not a power of 2.

Our algorithm makes the minimum number of comparisons demanded in obtaining the best record and the best cluster.

Our enhanced privacy preservation model maintains a balance between data utility and data privacy.

VI. EXPERIMENTAL RESULTS

A. Experimental Setup

The experiment was performed on a 2.80Ghz Intel Core i7 11th Gen computer having a RAM of 16.0 GB. The operating system installed on the system was Windows 10 Pro. The model was implemented using Python (2.7.18). After successful execution, the model's outcome displayed the runtime and the NCP percentage for multiple values of k. We used the Adult dataset from the UC Irvine Machine Learning Repository [10], a de facto benchmark for assessing k-anonymity algorithms' performance.

B. Results and Discussion

We summarize our proposed model's experimental outcomes based on time elapsed, performance efficiency, scalability, and data quality. The total information loss cost of the Enhanced greedy k-member is lesser than other algorithm's cost. Our proposed model runs with a minimum number of comparisons while selecting the best cluster and record compared to other algorithms. Hence, decreasing the runtime of the problem significantly. Fig. 4. Depicts the relationship between the k value and time taken to complete the process. The enhanced greedy k-member algorithm generates clusters with sizes very close to the privacy preservation parameter k.

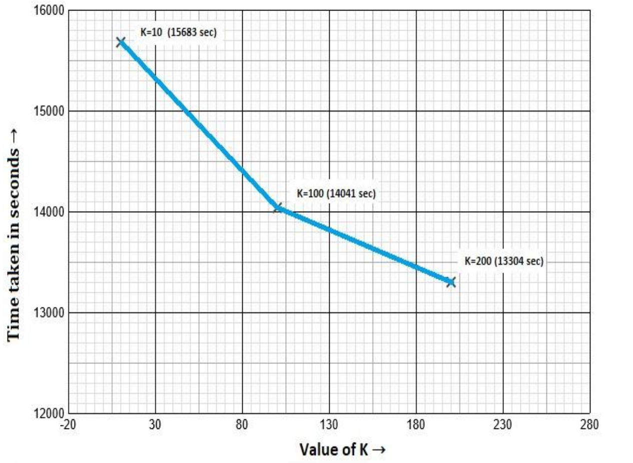

Fig. 4. K value vs Running Time

Normalized Certainty Penalty (NCP) [22] is a standard algorithm for measuring information loss. In the outcome of our proposed model, the NCP value is converted to a percentage. The NCP value is divided by the total number of values in the dataset to estimate the NCP percentage.Fig. 5 Depicts the relationship between the k value and NCP percentage. The enhanced model was run for different values of k (k=10,50,150 and 200) to measure the execution time and NCP percentage of the algorithm.

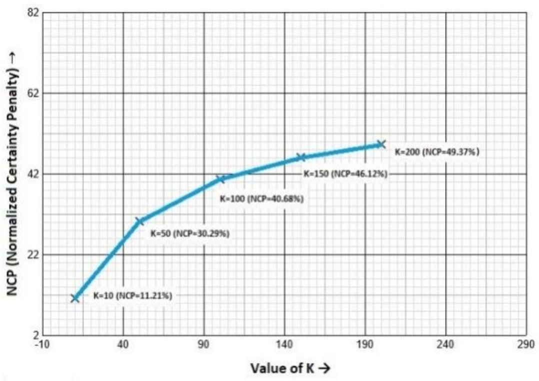

Fig. 5. K value vs NCP (Normalized Certainty Penalty)

We utilized the Adult dataset for this experiment. The enhanced privacy preservation model based on clustering offers the least Total information loss cost for any dataset size. While comparing the clustering model with the partitioning model, the former is slower than the latter. Still, the performance regarding the Total-IL metric of the clustering model is better than the partitioning model.

VII. CONCLUSION

This paper proposes an enhanced privacy preservation model for data publishing based on the k-anonymization clustering algorithm. The proposed algorithm also makes a minimal number of comparisons to achieve the desired results and balances data privacy and data quality compared to traditional anonymization algorithms. The model efficiently decreases information loss and contributes more utilizable data for data mining.

REFERENCES

1. Zheng, Wantong & Wang, Zhongyue & Lv, Tongtong & Ma, Yong & Jia, Chunfu. (2018). K-Anonymity Algorithm Based on Improved Clustering.
2. S.Subbalakshmi & K.Madhavi. "A Research on Bigdata Privacy Preservation Methods",International Journal of Recent Technology and Engineering,2019
3. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys 42(4), 14 (2010)
4. B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In International Conference on Data Engineering, 2005.
5. Samarati, P. and Sweeney, L. "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression" , February, 2017.
6. Byun, Ji-Won & Kamra, Ashish & Bertino, Elisa & Li, Ninghui. (2007). Efficient k-Anonymization Using Clustering Techniques. DASFAA 2007. LNCS. 4443. 188-200. 10.1007/978-3-540-71703-4_18.
7. A. Agarwal, A. Agarwal, The Security Risks Associated with Cloud Computing, International Journal of Computer Applications in Engineering Sciences, 2011.
8. V. S. Iyengar. Transforming data to satisfy privacy constraints. In ACM Conference on Knowledge Discovery and Data mining, 2002.
9. K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In ACM International Conference on Management of Data, 2005
10. Palanisamy, B., Liu, L., Zhou, Y., Wang, Q.: Privacy-preserving publishing of multilevel utility-controlled graph datasets. ACM Transactions on Internet Technology 18(2), 24 (2018)
11. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information. In: the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. vol. 98, p. 188. Citeseer (01 1998)
12. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: kanonymity and its enforcement through generalization and suppression. Tech. rep., Technical report, SRI International (1998)
13. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
14. Hussien, A.A., Hamza, N. and Hefny, H.A., "Attacks on Anonymization-Based privacy-preserving: A survey for data mining and data publishing", Journal of Information Security, volume 4, issue 2, 2013,pp.101–112.
15. Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. "L -diversity: privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, volume 1, Issue 1, 2007.
16. Li, N., Li, T. and Venkatasubramanian, S., "Tcloseness: Privacy beyond k-anonymity and l-diversity", ICDE IEEE 23rd International Conference on Data Engineering, 2007.
17. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In International Conference on Data Engineering, 2005.
18. Liu Y et al. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans Ind Inf. 2018
19. P. Ram Mohan Rao, S. Murali Krishna and A. P. Siva Kumar "Privacy preservation techniques in big data analytics: a survey",. J Big Data (2018)
20. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.
21. K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional kanonymity. In International Conference on Data Engineering, 2006.