International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

Data Publishing Thorough Micao Grouping Tranformation with Privacy Utility Preservation

[1]Mr. G.SRIKANTH REDDY, [2] Mr. P.GIRIDHAR

[1]M.Tech(CSE) from JAGRUTI INSTITUTE OF ENGINEERING AND TECHNOLOGY

[2] Assistant Professor, Department of Computer Science and Engineering, JAGRUTI INSTITUTE OF ENGINEERING AND TECHNOLOGY, Telangana State, India.

ABSTRACT

Now days data exchanging between two parties need secrete communication, there is a need to handle the risk of unintended information disclosure. Encrypting the data while sending is not easy job always, without revealing sensitive information about them is an important problem. K-anonymization is the most valuable method among other data protection techniques. The limitations of K- anonymity were surmount by methods like L-diversity, T-closeness, (alpha, K) anonymity; but all of these methods focus on universal approach that exerts the same amount of privacy preservation for all persons against linking attack, which result in high loss of information. Privacy was also not guaranteed 100% because of proximity and divergence attack. In this paper Our approach is to design micro data sanitization technique to preserve privacy against proximity and divergence attack and also to preserve the utility of the data for any type of mining task. The proposed approach, apply a graded grouping transformation on numerical sensitive attribute and a mapping table based transformation on categorical sensitive attribute. We conduct experiments on adult data set and compare the results of original and transformed table to show that the proposed task independent technique preserves privacy, information and utility.

Keywords : Anonymization, Data Publishing, Data utility, Privacy management.

INTRODUCTION

Anonymization means without encrypting complete data hiding or pertubarating part of data and anonymization we can apply in either using Generalization or Suppression techniques . Releasing the original data set provides the highest utility to data users but greatest disclosure risk for the subjects in the data set. On the contrary, releasing random data incurs no risk of disclosure but provides no utility. K-Anonymity, in particular, seeks to make record re-identification unfeasible by hiding each subject within a group of k subjects. To this end, k-anonymity requires each record in the anonymized data set to be indistinguishable from another k -1 records as far as the quasi-identifier attributes are concerned. Online browsing methods use a representing concept-based user profiles. The weights of the vector elements, which could be positive or negative,

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

represent the interestingness (or uninteresting nests) of the user on the concepts. MicroAggregation is Statistical Disclosure Control (SDC), also known as Statistical Disclosure Limitation (SDL), seeks to transform data in such a way that they can be publicly released whilst pre-serving data utility and statistical confidentiality, where the latter means avoiding disclosure of information that can be linked to specific individual or corporate respondent entities. When we micro-aggregate data we have to keep two goals in mind: (i) Preserving data utility. To do so, we should introduce as little noise as possible into the data i.e. we should aggregate similar elements instead of divergent ones. In the example given in Figure 1 for a security parameter k = 3, groups of three elements are built and aggregated. (ii)Protecting the privacy of the respondents. Data have to be modified to make reidentification difficult i.e. by increasing the number of aggregated elements, we in- crease data privacy. In the example given in Figure 1, after aggregating the chosen elements, it is impossible to distinguish them, so that the probability of linking any respondent is inversely proportional to the number of aggregated elements.

EXISTING SYSTEM

Service provider computing is a new computing paradigm that is built on virtualization, parallel and distributed computing, utility computing, and serviceoriented architecture. Although the great benefits brought by computing paradigm are exciting for IT companies, academic researchers, and potential users, security problems in service provider computing become serious obstacles which, without being appropriately addressed, will prevent service provider computing extensive applications and usage in the future. To achieve flexible and fine-grained access control, a number of schemes have been proposed more recently. Unfortunately, these schemes are only applicable to systems in which data owners and the service providers are within the same trusted domain. Since data owners and serviceproviders are usually not in the same trusted domain in service provider computing, a new access control scheme employing attributed encryption. The notion of attributed encryption was first introduced as a new method for fuzzy identity-based encryption. The primary drawback of the scheme in is that its threshold semantics lacks expressibility. Several efforts followed in the literature to try to solve the expressibility problem.

PROPOSED SYSTEM

We have proposed and evaluated the use of micro aggregation as a method to attain k-anonymous t-closeness. The a priori benefits of microaggregation vs generalization/ recoding and local suppression have been discussed. Global recoding may recode more than needed, whereas local recoding complicates data analysis by mixing together values corresponding to different levels of generalization. Also, recoding produces a greater loss of granularity of the data, is more affected by outliers, and changes numerical values to ranges. Regarding local suppression, it complicates data analysis with missing values and is not obvious to combine with recoding in order to decrease the amount of generalization. Microaggregation is free from all the above downsides. We have proposed and evaluated three different microaggregation based algorithms to generate kanonymous t-close data sets. The first one is a simple merging step that

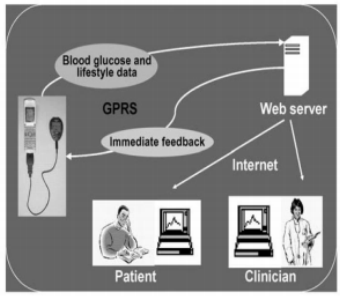
International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

can be run after any microaggregation algorithm. The other two algorithms, k-anonymity-first and T-closeness-first, take the T-closeness requirement into account at the moment of cluster formation during micro aggregation. The T-closeness-first algorithm considers tcloseness earliest and provides the best results: smallest average cluster size, smallest SSE for a given level of Tcloseness, and shortest run time (because the actual micro aggregation level is computed beforehand

4.1 Admin module Initially, the micro aggregation algorithm is run on the quasi-identifier attributes of the original data set; this step produces a k-anonymous data set. Then, clusters of Micro aggregated records are merged until t-closeness is satisfied. Selecting the cluster whose confidential attribute distribution is most different from the confidential attribute distribution in the entire data set (that is, the cluster farthest from satisfying t-closeness); and ii) merging it with the cluster closest to it in terms of quasiidentifiers.

4.2 Search module In this section that the values of the confidential attribute(s) can be ranked, that is, be ordered in some way. For numerical or categorical ordinal attributes, ranking is straight forward. Even for categorical nominal attributes, the ranking assumption is less restrictive than it appears, because the same distance metrics that are used to microaggregate this type of attributes can be used to rank them. EMD distance with respect to microaggregation. To minimize EMD between the distributions of the confidential attribute within a cluster and in the entire data set, the values of the confidential attribute in the cluster must be as spread as possible over the entire data set. Consider the case of a cluster with k records. The following proposition gives a lower bound of EMD for such a cluster.

4.3 Filtering Module In a first battery of tests we used as evaluation data the Census data set, which is usual to test privacy protection methods and contains 1,080 records with numerical attributes. Because k-anonymity and t-closeness pursue different Goals. we defined two data sets according to the correlation between the values of quasi-identifier and confidential attributes.



Identify the Patient Situation and any Modification

CONCLUSION

Algorithms such as k-anonymity and L-diversity leave all sensitive attributes intact and apply generalization and suppression to the Quasi-identifiers. The goal is to keep the data truthful and thus provide good utility for data-mining applications, while achieving less than perfect privacy. But utility is best measured by the success of data mining algorithms such as decision tree learning which take advantage of relationships between attributes. Also simple anonymization is already widely used in practice. One prime example is clinical trial studies for new drugs in the medical and pharmaceutical domain. Even though the U.S. Food and Drug Administration guidelines are known to be strict, anonymization (or de-identification) is still considered adequate in the

clinical trial setting for protecting the privacy of patients participating in the studies. Compared with other transformation techniques, anonymization is simple to carry out, as mapping objects back and forth is easy. Another advantage of anonymization is that it does not perturb data characteristics. Optimal generalization is NP hard, as well as generalization becomes complex when dimensionality of the table increases [9]. But, the proposed method is extremely efficient because of simplicity in implementation. Since the transformed table preserves the characteristics of the original table, the utility of any mining task is preserved and thereby avoiding need for developing problem specific algoritms. Being the simple procedure, transformation is done in data owners' site itself and can be supplied for any type mining task. The experiments conducted on the UCI data proved the utility and privacy of data for all typical data mining tasks. Also, the problem of proximity attack and divergence attack is solved by not forming a group or equivalence class.

REFERENCES

[1] Adam N. R., Wortmann J. C., "Security-control methods for statistical databases: A comparative study", ACM Comput. Surv 21(4), 515–556, 1989

[2] Aggarwal C. C., Yu P. S., "A Condensation approach to privacy preserving data mining", EDBT Conference, 2004

[3] Aggarwal C. C., Yu P. S., "On Variable Constraints in Privacy-Preserving Data Mining", SIAM Conference, 2005

[4] Agrawal D., Aggarwal C. C., "On the Design and Quantification of Privacy- Preserving Data Mining Algorithms", ACM PODS Conference, 2002

[5] Agrawal R., Srikant R.,"Privacy-Preserving Data Mining", ACM SIGMOD Conference, 2000

[6] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.,"Disclosure limitation of sensitive rules", Workshop on Knowledge and Data Engineering Exchange, 1999

[7] Bayardo. R. J, Rakesh Agrawal, "Data privacy through optimal k- anonymization" , ICDE, 217-228,2005

[8] Justin Brickell and Vitaly Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing", KDD conference, 2008

[9] C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, 2005

[10] I. Dinur and K. Nissim.,"Revealing information while preserving privacy", PODS, pages 202–210, 2003

[11] J. Li, Raymond chi wing wong, Ada Fu, J. pei, "Anonymization by local recoding in data with attribute hierarchical taxonomies", IEEE transaction on Knowledge and data Engg, Vol 20, No. 9, pp. 1181-1194,sep 2008

[12] Jiexing Li, Yufei Tao, Xiaokui Xiao, " Preservation of Proximity Privacy in Publishing Numerical Sensitive Data", ACM SIGMOD, 2008

[13] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing, PODS, 2005

[14] K LeFevre, David J. DeWitt, Raghu Ramakrishnan ,"Incognito: Efficient full domain k – anonymity ", SIGMOD, 49-60, 2005

[15] K. Lefevre, D. Dewatt, R. Ramakrishnana, "Workload Aware Anonymization", ACM KDDM, 2006

[16] Machanavajjhala A., Gehrke J., Kifer D.,and Venkitasubramaniam M, "l-Diversity: Privacy Beyond kAnonymity", pp.24-35, ICDE, 2006 [17] D. Martin, D.Kifer, A. Machanavajjhala, J. Gehrke, J. Halpern, " Worst-case background knowledge in privacy", ICDE, 2007