ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

A Survey on k-Anonymity Generalization Algorithms

Kavitha S[1], Sivaraman E[2], Raja Vadhana P[3]

PG Scholar, Computer Science and Engineering, Dr. N. G. P Institute of Technology, Coimbatore, India [1,3]

Assistant Professor, Computer Science and Engineering, Dr. N. G. P Institute of Technology, Coimbatore, India [2]

Abstract: The issue of data privacy is at the forefront of everybody's mind. Media commercials advertise security merchandise and news programs oftentimes describe the most recent knowledge breach. Public perception aside, any organization incorporates a legal obligation to make sure that the privacy of their workers is protected. Laws compel some knowledge from being employed for secondary reasons aside from the aim that it absolutely was originally collected. We can't collect data on the health of your workers. Also, we can't share sure knowledge with third parties. Within the world of cloud computing, we currently have a third party operational and managing our infrastructure. By it's terribly nature, that supplier can have access to our data. Hence data anonymity techniques are used to share the public data such as medical records by preserving the data privacy. In this paper we discuss on various the k-anonymity generalization algorithms used for privacy preserving.

Keywords: anonymity, data privacy, generalization, privacy preservation, data publishing

I. INTRODUCTION

Today's databases contain a lot of sensitive personal data. So it's crucial to design information systems which might limit the revealing of personal data. As an example, think about a hospital that maintains patient records. The hospital desires to disclose data to a company in such some way that the company cannot infer that patients have that diseases. One technique to formally specify privacy policies is to specific sensitive data as queries and enforces excellent privacy, an awfully sturdy notion of privacy that guarantees that the other question answered by the information won't disclose any data regarding the sensitive data.

A. Privacy Preserving Data Publishing

Personal records of people are progressively being collected by numerous government and company establishments for the needs of data analysis. The data analysis is facilitated by these organizations to publish "sufficiently private" ideas over this information that are collected.Privacy could be a double edged brand -- there ought to be enough privacy to make sure that sensitive data concerning the people isn't disclosed by the views and at a similar time there ought to be enough information to perform the analysis. Moreover, an adversary who needs to collect sensitive data from the revealed views sometimes has some information concerning the people within the information. The main objective is to convert the original information into some anonymous type to stop from inferring its record owners' sensitive data as discussed in [1].

B. Data Anonymization

Data anonymization is the process of removing personally identifiable information from data sets, to make the people anonymous about whom the data describe. It permits the transfer of knowledge across a boundary, like between two departments inside centre or between two agencies, whereas reducing the danger of inadvertent revealing, and in bound environments in an exceedingly manner that permits analysis and analytics post-anonymization. This technique is used in enterprises to increase the security of the data while allowing the data to be analysed and used. It changes the data that will be used or published in order to prevent the identification of key information. Data anonymization techniques such as k-anonymity, l-diversity and t-closeness are widespread.

1) k-Anonymity: The basic plan of k-anonymity is to shield a dataset against re-identification by generalizing the attributes that might be utilized in a linkage attack (quasi identifiers). A information set is taken into account k-anonymous if every information item can't be distinguished from a minimum of k-1 alternative data things.

2) l-Diversity: l-diversity could be a variety of cluster based mostly anonymization that's wont to preserve privacy in knowledge sets by reducing the coarseness of a knowledge representation. This reduction may be a trade off that ends up in some loss of effectiveness of knowledge management or mining algorithms so as to achieve some privacy. The l-diversity model is associate degree extension of the k-anonymity model that reduces the roughness of information illustration victimization techniques as well as generalization and suppression specified any given record maps onto a minimum of k alternative records within the data [2].

3) t-Closeness: t-closeness could be a an additional refinement of l-diversity cluster based mostly anonymization that's accustomed preserve privacy in knowledge sets by reducing the coarseness of an information representation. t-closeness could be a an extra refinement of l-diversity cluster primarily based anonymization that's wont to preserve privacy in knowledge sets by reducing the coarseness of an information illustration. This reduction could be a trade off

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

that ends up in some loss of effectiveness of knowledge management or mining algorithms so as to realize some privacy [3].

II. K-ANONYMITY

k-Anonymity could be a formal model of privacy created by L.Sweeny [4]. The goal is to form every record indistinguishable from an outlined variety (k) records if tries area unit created to spot the information. A set of information is k-anonymized if, for any record with a given set of attributes, there square measure at least k-1 alternative records that match these attributes. The attributes can be any of the following types.

TABLE I
K-ANONYMITY ATTRIBUTES

Attributes	Description	Example
Explicit_identifier	Set of attributes	Name, Id
Quasi_identifier	Potentially identify record owners	Age, Sex, Zip
Sensitive attributes	Person's sensitive information that cannot revealed	Salary, Disease

The implementation of k-anonymity needs the preliminary identification of the quasi-identifier. The quasi-identifier relies on the external data available to the recipient, because it determines the linking ability (not all possible external tables area unit accessible to each potential knowledge recipient); and different quasi-identifiers will doubtless exist for a given table [5].

Example:

If the above mentioned table is to be anonymized with Anonymization Level (AL) set to 2 and the set of Quasi-identifiers as QI = {AGE, SEX, ZIP, PHONE}.Sensitive attribute = {SALARY}. The quasi-identifiers and sensitive attributes are identified by the organization according to their rules and regulations.

TABLE II
TABLE TO BE ANONYMIZED

ID	Age	Sex	Zip	Phone	Salary (in Rs.)
1	24	M	641015	9994258665	78000
2	23	F	641254	9994158624	45000
3	45	M	610002	8975864121	85000
4	34	M	623410	7456812312	20000

TABLE II
ANONYMIZED TABLE

ID	Age	Sex	Zip	Phone	Salary (in Rs.)
*	20-50	ANY	641***	999*******	78000
*	20-50	ANY	641***	999*******	45000
*	20-50	ANY	612***	897*******	85000
*	20-50	ANY	623***	745*******	20000

This anonymization can be done by *Generalization* and *Suppression*

A. Generalization

Generalization is the process of converting a value into a less specific general term. For ex, "Male" and "Female" can be generalized to "Any". At the following levels [6, 7] generalization techniques can be applied.

1) Attribute (AG): Generalization is performed at the column level; All the values in the column are generalized at a generalization step.

2) Cell (CG): Generalization can also be performed on a single cell; finally a generalized table might contain, for a specific column and values at different levels of generalization.

B. Suppression

Suppression consists in preventing sensitive data by removing it. Suppression can be applied at the level of single cell, entire tuple, or entire column, allows reducing the amount of generalization to be imposed to achieve k-anonymity [6].

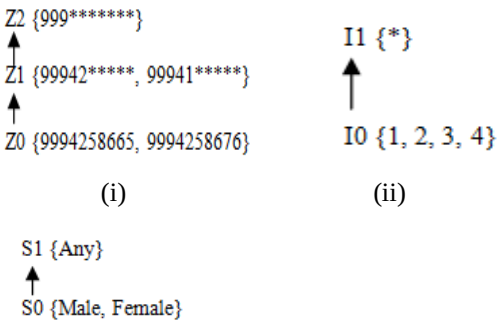
1) Tuple (TS): Suppression is performed at row level; suppression operation removes whole tuple

2) Attribute (AS): Suppression is performed at column level; suppression operation hides all the values of a column.

3) Cell (CS): Suppression is performed at single cell level; finally k-anonymized table might wipe out only certain cells of a given tuple/attribute.

III. GENERALIZATION ALGORITHMS

Samarati and Sweeney [4, 7, 8 and 9] formulated the idea of k-anonymization using generalization and suppression. The hierarchy of generalization is defined to be a set of domains that is ordered by the relationship <D. Consider the hierarchy as a chain of nodes, and if there is an edge from Di to Dj , It is said to be the direct generalization from Dj to Di. The generalization relationship is transitive, if Di <D Dj and Dj <D Dk, then Di <D Dk. We can say domain Dk an implied generalization of Di. In a domain hierarchy chain paths correspond to implied generalizations whereas edges correspond to direct generalizations. Fig. 1(i,ii,iii) shows domain generalization hierarchies for the Phone, ID and Sex attributes [10].



(i) (ii)

(iii)

Fig.1 Generalization Hierarchy

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

The generalization technique can be applied with the following algorithms.

A. Top Down Specialization

Top Down Specialization is an iterative method that ranges within the taxonomy trees of attributes from the topmost domain values. Each and every round of iteration consists of three main steps, namely, first, identifying the most effective specialization, second, performing specialization and finally, change values of the search metric for succeeding round.

Benjamin C. M. Fung, Ke Wang, Philip S. Yu [11]:
In this Paper the top down specialization algorithm is implemented for the generalization which is enforced by specializing or particularisation the amount of data in an exceedingly top-down manner until a minimum privacy demand is profaned. For handling the categorical and continuous attributes the top-down specialization approach is the feasible way. By minimizing the privacy specification and maximising the data utilization the top-down approach uses iterative method to convert the general information into special information. Multiple anonymity issues can be handled with this approach.

Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen[12]:
In this paper, they used TDS to measure the drawback of large-scale data anonymization, and introduced a new method called Two-Phase TDS approach which uses Map and Reduce phase on cloud. In this approach first data partition is done and the anonymization is done in parallel as initial state then further intermediate results are produced. In the second section the intermediate results are incorporated with additional anonymization for providing k-anonymous data sets which are consistent. Data anonymization is creatively applied on cloud using MapReduce and it is implemented in the way to produce a highly climbable specialization result. In the experimental results, the scalability and efficiency of large scale data sets has been improved in the Two Phase TDS compared to the centralized TDS.

N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee [13]:
In this paper two algorithms are proposed to overcome the issues in centralized and distributed anonymization for health institutes in order to support the data processing. The privacy and information requirements of BTS motivated to the development of LKC-privacy model for relational data that are high-dimensional and according to the BTS' need on privacy and information, they developed algorithms that can accommodate their requirements. The Privacy-Preserving Data Mining (PPDM) is different from the proposed solution because the fact is they allow data mining result sharing instead of data sharing. The essential requirement of BTS it requires the flexibility for performing different data analysis tasks. In health care sectors the proposed solutions serve as a model for data sharing.

Fung BCM, Wang K, Yu PS [14]:
They experimented to verify several claims about the proposed TDR method. First, TDR masks a given table to satisfy the anonymity requirements without sacrificing the usefulness to classification. Second, the efficiency of TDR is measured with previously reported genetic algorithm in [15]. Third, It is not necessarily translate the optimal k-anonymization [16] [17] into the optimality of classification. The better anonymization solution for classification is provided by the proposed TDR. Fourth, with large data sets and requirements of complex anonymity, the proposed TDR scales well. They experimented preserve both information utilization and individual's privacy effectively. They conferred a top-down approach that uses iterative method to convert the general information into special information which is guided by increasing the trade-off between information and anonymity.

B. Bottom Up Generalization

Bottom-up generalization is an iterative method from data processing to generalize the information. It is difficult to link to alternative sources even though the generalized data remains helpful for classification. The generalization house is mere by a data structure of generalizations. A key is at each iteration the best generalization is distinguished to climb up the hierarchy.

Ke Wang, Philip S. Yu, Sourav Chakraborty [18]:
To discover helpful patterns, they explained another optimistic use of the data mining technology even if we mask private information. The bottom-up generalization converts the specific data to less specific but semantically consistent data for privacy preservation and also they focused on two main problems, scalability and quality. The scalability problem was addressed by a unique data structure to focus on pretty good generalizations. The same quality is achieved by the proposed system however far better measurability compared to existing solutions. Our current algorithm has the likelihood of obtaining stuck at a neighbourhood optimum by greedily hill climbs to a k-anonymity state.

Jian Xu, Wei Wang1 Jian Pei, Xiaoyuan Wang, Baile Shi, Ada Wai-Chee Fu [19]:
The bottom-up approach is also known as greedy technique. In every iteration, it merges groups specified the resulted weighted certainty penalty is locally decreased. In one iteration, the cluster becomes larger than k, if one cluster is unified with multiple groups. If a group has over 2k tuples, then the cluster is to be split in order to avoid the over-generalization. The resulted table should be within the warranted, in which every cluster has up to (2k − 1) tuples. This paper usually win higher anonymization in quality than the Multidimensional methodology, the progressive approach by using both bottom-up methodology and also the top-down methodology in which the intensive experiments of real data sets and artificial data sets show that, in terms of utility and sharpness.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Compared to bottom-up method, the top-down method is better.

Tiancheng Li, Ninghui Li [20]:
For locating best anonymization, they presented a strategy called bottom-up search. Once the worth of k is small this strategy works significantly well. They showed the practicability through experiments on real census data for this approach. To find the optimal solution for small k values very quickly, the bottom up approach works efficiently and when k increases, the running time of a generalization scheme increases.

IV. CONCLUSION

In this paper we discussed about the Privacy preserving data publishing and data anonymization. We also discussed about various anonymization techniques and mainly focused on k-anonymity which comprises of both generalization and suppression. The last part is about the generalization algorithms and its implementation for protecting the privacy of data used mainly for data analysis.

REFERENCES

[1] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Devel opments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.
[2] Machanavajjhala, J. Gehrke, and D. Kifer, "LDiversity: Privacy beyond K-Anonymity", in Proc. of the IEEE ICDE (2006), pp. 24.
[3] N. Li, T. Li, and S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", in Proc. Of the IEEE ICDE (2007), pp. 106–115.
[4] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, ppp. 557-570, 2002.
[5] Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity", Springer US, Advances in Information Security (2007)
[6] Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity Data Mining: A Survey", Springer US, Advances in Information Security (2007)
[7] Latanya Sweeney, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 10 Issue 5, October 2002, Pages 571 – 588.
[8] P. Samarati, "Protecting respondants' identities in microdata release", IEEE Transactions on Knowledge and Data Engineering, 13(6), November/December 2001.
[9] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.
[10] Kristen LeFevre David J. DeWitt Raghu Ramakrishnan University of Wisconsin Madison 1210 West Dayton St. Madison, WI 53706, "Incognito: Efficient FullDomain KAnonymity", SIGMOD 2005 June 1416, 2005, Baltimore, Maryland, USA ACM 1595930604/05/06
[11] Benjamin C. M. Fung, Ke Wang, Philip S. Yu, "Top-Down Specialization for Information and Privacy Preservation", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), 1084-4627/05 $20.00 © 2005 IEEE
[12] Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, Member, IEEE, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 2, February 2014
[13] N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data" , ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4,Article 18, 2010.
[14] Fung BCM, Wang K, Yu PS (2007), "Anonymizing classification data for privacy preservation", IEEE TKDE 19(5):711–725.
[15] V. S. Iyengar, "Transforming data to satisfy privacy constraints" , in Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, July 2002, pp. 279.288.
[16] Brodsky A, Farkas C, Jajodia S (2000), "Secure databases: Constraints, inference channels, and monitoring disclosures", IEEE Transactions on Knowledge and Data Engineering 12:900–919.
[17] Farkas C, Jajodia S (2003) , "The inference problem: A survey", ACM SIGKDD Explorations Newsletter 4(2):6–11
[18] Ke Wang, Philip S. Yu, Sourav Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection", Fourth IEEE International Conference on Data Mining, 2004. ICDM '04.. Pages 249 - 256
[19] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. Fu. , "Utility-based anonymization using local recoding" , In ACM SIGKDD, 2006.
[20] Tiancheng Li, Ninghui Li, "Optimal k-Anonymity with Flexible Generalization Schemes through Bottom-up Searching", ICDM Workshops 2006. Sixth IEEE International Conference on Data Mining Workshops, 2006,pages: 518 - 523 .

BIOGRAPHIES

Kavitha S is currently pursuing her Masters in Computer Science and Engineering in Dr. N. G. P Institute of Technology, Anna University, Coimbatore, Tamil Nadu, India. She was a Software Trainer in NIIT Pvt Ltd. in the year 2011. She worked as a Computer Instructor in Kendriya Vidyalaya, Coimbatore for elementary board for the year 2013. She is specialized in .Net Framework and other areas of interest are Data Mining and Cloud Computing.

Sivaraman E is currently working as Assistant Professor in the department of Computer Science and Engineering at Dr. N. G. P Institute of Technology, Coimbatore and also pursuing his Ph.D (part time) from Bharathiar University, Coimbatore. He is a Microsoft Certified Professional and EMC^2 Proven Professional. He is life member in Computer Society of India and Indian Society for Technical Education. His area of research includes Cloud computing and Big Data Analytics.

Raja Vadhana P is currently pursuing her Masters under the department of Computer Science and Engineering in Dr. N. G. P Institute of Technology, Anna University, Coimbatore, Tamil Nadu, India. She was a Software Engineer and Associate Consultant in HCL Technologies Ltd. from 2008 to 2011. She worked as an Associate Consultant with Larsen and Toubro Infotech Ltd. under Enterprise Application Integration for the year 2011. She is specialized in BFSI domain and has extensive experience in Service Oriented Architecture technologies like TIBCO, ORACLE BPM and IBM BPM.