

An Approach Towards the Secure Preserving of Sensitive Information

¹P.Anusha, ²K. Sudheer Kumar

^{1,2}Dept. of CSE, DRK College of Engineering & Technology, Hyderabad, Andhra Pradesh, India

Abstract

Privacy preserving data publishing is a promising research area since individual privacy has become a major concern. In many data analysis applications a major issue is privacy preservation. Privacy-preserving techniques are often required to reduce the possibility of identifying sensitive information about individuals, when a data set is released to other parties for data analysis. Sensitive information can be a specific location of an individual for spatial data. The study is made on protecting the privacy of persons and data has received from many fields. The information that a user browses certain websites may be considered sensitive in web surfing. We review the research work in privacy-preserving data publishing. In this paper, we describe how an organization, without violating the confidentiality of personal information, can release the data to the public is discussed. However, the simplest solution to protect the sensitive information is not to disclose the information but it will delay the data analysis process. The data must be disclosed under the government regulations in some applications. On the other hand, if the data owner first modifies the data then the modified data can guarantee privacy, such that the modified data retains enough utility and can be released to other parties safely. In this paper, we examine how the data owner can modify the data and how the modified data can preserve privacy and protect sensitive information.

Keywords

Privacy Preservation, Sensitive Information, Data Publishing, Data Analysis

1. Introduction

Many organizations use the data mining techniques to gain the characteristics and the behaviors of the customers. The process of extracting useful, interesting, and previously unknown information from large data sets is Data mining. The availability of high quality data and effective information sharing depends on data mining. An environment that facilitates large-scale data mining and data analysis is the collection of digital information (i.e. by governments, corporations and individuals). On the other hand, the public has an increased sense of privacy invasion due to the increased level of security after revolutionary attacks. However, the public has acquired a negative impression that data mining is a technique for privacy intrusion since data mining is often a key component of many security systems, monitoring and enterprise information systems and surveillance systems. The main obstacle to the advancement of technology is the lack of trust in data mining. Many techniques are proposed for protecting individual privacy and sensitive information in order to avoid the obstacles. To restrict the types of publishable data and agreements on the use and storage of sensitive data the current privacy protection practice primarily depends up on the policies and guidelines. The constraint of this approach is that the data is distorted extremely or it requires a trust level that is impractically high in many data-sharing scenarios. The sensitive data will not be carelessly misplaced and end up in the wrong hands. We describe the data collection and publishing Figure1.

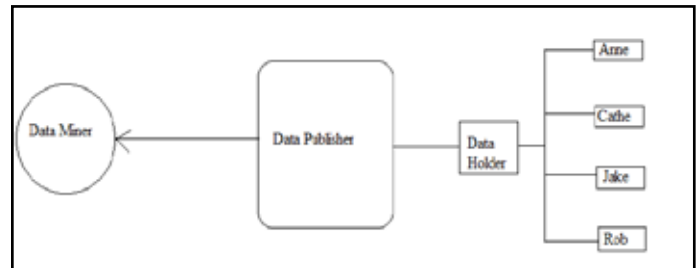


Fig. 1: Data Collection and Publisher

The data holder collects data from record owners (e.g., Anne and Rob) in the data collection phase. The data holder releases the collected data to a data miner or the public in the data publishing phase, called the data recipient who will then conduct data mining on the published data. There are two models of data holders i.e. the un-trusted model in which the data holder is not trusted and attempt to identify sensitive information from record owners (example Anne and Jake). To collect records anonymously from their owners without revealing the owners' identity the various cryptographic, anonymous communications and statistical methods were proposed. In the trusted model, the data holder is trustworthy and record owners are willing to provide their personal information to the data holder. However, the trust is not transitive to the data recipient. In this paper, we assume the trusted model of data holders and consider privacy issues in the data publishing phase. Every data publishing scenario in practice has its own assumptions and requirements on the data holder, the data recipients, and the data publishing purpose. The following are different assumptions and properties that may be adopted in practical data publishing depending on the application. To perform data mining on behalf of the data recipient the data holder is not required to have the knowledge. After receiving the data from the data holder any data mining activities can be performed by the data recipient. Sometimes, the data holder does not even know who the recipients are at the time of publication. For example, publishing patient records in hospitals on the internet. The hospitals do not know who the recipients are and how the recipients will use the data. The hospital publishes patient records because it is required by regulations, or because it supports general medical research, not because the hospital needs the result of data mining. Therefore, it is not possible to expect the data holder to anonymizing the data for publication. In other scenarios, the data holder is interested in the data mining result, but lacks the in-house expertise. The data holder could release a customized data set that preserves specific types of patterns for such a data mining task in order to improve the quality of the data mining result. The data mining activities are performed by the data recipient but not by the data holder. The data recipient could be an adversary. In PPDP, one assumption is that the data recipient could also be an adversary. For example, the data recipient, says a drug research company, is a trustworthy entity however; it is difficult to guarantee every staff in the company to be trustworthy as well. The authorized and trustworthy recipients are given the private key for accessing the clear text and this

assumption makes the PPDP problems and solutions to be very different from the encryption and cryptographic approaches. A major challenge in PPDP is to preserve both the privacy and information utility in the anonymous data. This requirement is more rigid than publishing data mining results. For example, in the case of the Netflix data release, useful information may be some type of associations of movie ratings. Netflix decided to publish data records instead of publishing such associations because the participants, with data records, have greater flexibility to perform required analysis and data exploration, such as mining patterns in one partition but not in other partitions, visualizing the transactions containing a specific pattern, trying different modeling methods and parameters, and so forth. The assumption of publishing data, not data mining results, is also closely related to the assumption of a non-expert data holder. In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life. Consider the example of patient records. The pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug. It is difficult to deploy data mining results in the real world if a published record does not correspond to an existing patient in real life. Randomized and synthetic data do not meet this requirement. Also, randomized and synthetic data become meaningless to data recipients who may want to manually interpret individual data records. Encryption is another commonly employed technique for privacy protection. Although an encrypted record corresponds to a real life patient, the encryption hides the semantics required for acting on the represented patient. Encryption aims to prevent an unauthorized party from accessing the data, but enable an authorized party to have full access to the data. In PPDP, it is the authorized party who may also play the role of the adversary with the goal of inferring sensitive information from the data received. Thus, encryption may not be directly applicable to some PPDP problems. In this paper we discuss about the developed algorithms which are evaluated in terms of the level of privacy protection, data quality anonymization, that are applicable to real life databases, efficiency and scalability. Anonymizing classification is a fundamental problem in data analysis. To access large collection of data a classifier is required. A threat may pose to an individual's privacy by releasing person-specific data. A useful k-anonymization model used to combat the linking attacks which are discussed above is anonymizing the quasi-identifier, denoted T , so that at least k released records match each value combination of T . By masking specific values to general values the anonymization is done, preferably in a way to minimize the distortion to the data. The first contribution of this paper is to efficiently identify a k-anonymous solution which preserves the classification structure. By the considering the both information and privacy the search is done by detailing the level of information in a top-down manner. This top-down refinement (TDR) algorithmic framework is highly efficient and natural for handling different types of attributes. This approach exploits by the noise and redundant structures in the data for achieving both a privacy goal and a classification goal. The privacy concern related to the input of data mining methods is addressed by the Confidence bounding which is the first contribution, but the output of data mining methods could also cause privacy threats. Although it can be used to infer sensitive properties about record holders the output is an aggregate pattern, it is not intended to identify a record holder. To preserve the information for a wanted data analysis and to limit the usefulness of unwanted inferences that may be derived from the release of data are the two main goals.

If sensitive values of a group of individuals can be confidently inferred by some linking attributes from the released data then a sensitive inference occurs. The identities of record holders and some released data records are prior to k-anonymity which focuses on limiting the associations.

II. Advantages

To improve the current state of the art, we introduce a novel data anonymization technique called slicing.

Based on the privacy requirement of k-diversity we show that slicing can be effectively used for preventing attribute disclosure.

III. Conclusion

This paper we have discussed about the overview of privacy preserving data publishing. We have shown how it overcomes the limitations of generalization and preserves the better utility while protecting against privacy threats. We discussed how it is used to prevent attribute disclosure and membership disclosure. Before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization this is the general proposed method. However, in this paper, we have described how an organization, without violating the confidentiality of personal information, can release the data to the public is discussed.

References

- [1] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.
- [2] A. Blum, C. Dwork, F. McSherry, K. Nissim, "Practical Privacy: The SULQ Framework", Proc. CM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.
- [3] J. Brickell, V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.
- [4] B.-C. Chen, K. LeFevre, R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.
- [5] H. Cramér, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.
- [6] I. Dinur, K. Nissim, "Revealing Information while Preserving Privacy", Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.
- [7] C. Dwork, "Differential Privacy", Proc. Int'l Colloquium Automata, Languages and Programming (ICALP), pp. 1-12, 2006.
- [8] C. Dwork, "Differential Privacy: A Survey of Results", Proc. Fifth Int'l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008.
- [9] C. Dwork, F. McSherry, K. Nissim, A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis", Proc. Theory of Cryptography Conf. (TCC), pp. 265-284, 2006.
- [10] J.H. Friedman, J.L. Bentley, R.A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time", ACM Trans. Math. Software, Vol. 3, No. 3, pp. 209-226, 1977.
- [11] B.C.M. Fung, K. Wang, P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation", Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.
- [12] G. Ghinita, Y. Tao, P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data", Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

- [13] Y. He, J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.
- [14] A. Inan, M. Kantarcioglu, E. Bertino, "Using Anonymized Data for Classification", Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 429-440, 2009.
- [15] L. Kaufman, P. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley & Sons, 1990.
- [16] D. Kifer, J. Gehrke, "Injecting Utility into Anonymized Data Sets", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 217-228, 2006.
- [17] N. Koudas, D. Srivastava, T. Yu, Q. Zhang, "Aggregate Query Answering on Anonymized Tables", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.
- [18] K. LeFevre, D. DeWitt, R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 49-60, 2005.
- [19] K. LeFevre, D. DeWitt, R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity", Proc. Int'l Conf. Data Eng. (ICDE), pp. 25, 2006.
- [20] K. LeFevre, D. DeWitt, R. Ramakrishnan, "Workload-Aware Anonymization", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 277-286, 2006.
- [21] N. Li, T. Li, S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and '-Diversity", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.
- [22] T. Li, N. Li, "Injector: Mining Background Knowledge for Data Anonymization", Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 446-455, 2008.
- [23] T. Li, N. Li, "On the Tradeoff between Privacy and Utility in Data Publishing", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.
- [24] T. Li, N. Li, J. Zhang, "Modeling and Integrating Background Knowledge in Data Anonymization", Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 6-17, 2009.
- [25] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkatasubramanian, "Diversity: Privacy Beyond k-Anonymity", Proc. Int'l Conf. Data Eng. (ICDE), p. 24, 2006.
- [26] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J.Y. Halpern, "Worst-Case Background Knowledge for Privacy-Preserving Data Publishing", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.
- [27] M.E. Nergiz, M. Atzori, C. Clifton, "Hiding the Presence of Individuals from Shared Databases", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 665-676, 2007.
- [28] P. Samarati, "Protecting Respondent's Privacy in Microdata Release", IEEE Trans. Knowledge and Data Eng., Vol. 13, No. 6, pp. 1010-1027, Nov./Dec. 2001.
- [29] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, Vol. 10, No. 6, pp. 571-588, 2002.
- [30] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, Vol. 10, No. 5, pp. 557-570, 2002.
- [31] M. Terrovitis, N. Mamoulis, P. Kalnis, "Privacy-Preserving Anonymization of Set-Valued Data", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 115-125, 2008.
- [32] R.C.-W. Wong, A.W.-C. Fu, K. Wang, J. Pei, "Minimality Attack in Privacy Preserving Data Publishing", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 543-554, 2007.
- [33] R.C.-W. Wong, J. Li, A.W.-C. Fu, K. Wang, "(_k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 754-759, 2006.
- [34] X. Xiao, Y. Tao, "Anatomy: Simple and Effective Privacy Preservation", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.
- [35] Y. Xu, K. Wang, A.W.-C. Fu, P.S. Yu, "Anonymizing Transaction Databases for Publication", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008.
- [36] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A.W.-C. Fu, "Utility-Based Anonymization Using Local Recoding", Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 785-790, 2006.



P. Anusha has completed B.Tech (IT) from Nagarjuna Institute of Technology Sciences and pursuing M.Tech (C.S.E) in DRK college of Engineering and Technology, JNTUH, Hyderabad. Her main research interest includes Data Mining, Databases.



K. Sudheer Kumar is working as an Assistant Professor in DRK College of Engineering and Technology JNTU, Hyderabad, Andhra Pradesh, India. He has received M.Tech (CSE) degree and M. Sc (Systems Theory & Computer Modeling). His main research interest includes Operating Systems, Computer Modeling and Advanced Computer Architecture.