INTERNATIONAL JOURNAL OF COMPUTATIONAL MATHEMATICAL IDEAS ISSN: 0974-8652

Effectively Anonymized Collection Of Tree Structured Data

Dupati Ashok Chakravarthi 1 *M. Naresh2*

Newtons Institute of Engineering

Published in Volume 13, Issue 1 Oct - Nov: 2016, Page No: 65039 to 65044

Abstract:

Anonymizing using tree structured data study about the problem of protecting privacy in the publication of set-valued data. Considering a collection of supermarket transactional data that contains detailed information about items bought together by individuals. Even after removing all personal characteristics of the buyer, which can serve as a link to his identity, thus resulting to privacy attacks from adversaries who have partial knowledge about the set. Depending upon the point of view of the adversaries. We define a new version of the k-anonymity guarantee. Our anonymization model relies on generalization instead of suppression. We develop an algorithm which find the frequent item set. The frequent-itemsets problem is that of finding sets of items that appear in (are related to) many of the same dataset.The paper proposes k(m,n)-anonymity, which guarantees that an attacker who knows up to m elements of a record and to n structural relations between the m elements will not be able to match her background knowledge to less than k matching records in the anonymized data. The anonymization procedure does not only generalize values that participate in rare item combinations but also simplifies the structure of the records. The simplification is performed by removing nodes from long paths and creating new smaller paths. In this paper introduces a new approach in k(m,n) anonymity is that the frequent item set mining of each item that has been transferred.

KEYWORDS:

Anonymity, generalization , information loss,synopsis tree, frequent item set mining .

INTRODUCTION:

In recent years the data mining community has faced a new challenge. It is now required to develop methods that restrain the power of these tools to protect the privacy of individuals. Anonymity in Data Mining [1], focus on the problem of guaranteeing privacy of data mining output. To be of any practical value, the definition of privacy must satisfy the needs of users of a reasonable application. The k-anonymity model distinguishes three entities: individuals, whose privacy needs to be protected; the database owner, who controls a table in which each row(also referred to as record or tuple) describes exactly one individual and the attacker. The k-

anonymity model makes two major assumptions: The database owner is able to separate the columns of the table into a set of quasi-identifiers. The attacker has full knowledge of the public attribute values of individuals, and no knowledge of their private data. Sequential pattern mining is a major research field in knowledge discovery and data mining. Helps in increasing availability of transaction data, it is now possible to provide new and improved services based on users' and customers' behavior. Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining[2],introduced a new approach for anonymizing sequential data by hiding infrequent, and thus potentially sensible, sub sequences. User's actions as well as customer transactions are often stored together with their timestamps, making the temporal sequentially of the events a powerful source of information. The problem of protecting privacy in the publication of set-valued data is defined in Local and global recoding methods for anonymizing set-valued data[3],Consider a collection of supermarket transactions that contains detailed information about items bought together by individuals. Even after removing all personal characteristics of the buyer, which can serve as links to his identity, the publication of such data is still subject to privacy attacks from adversaries who have partial knowledge about the set. Consider a database D, which stores information about items purchased at a supermarket by various customers. A subset of items in a transaction could play the role of the quasi-identifier for the remaining (sensitive) ones and vice-versa. Another fundamental difference is that transactions have variable length and high dimensionality, as opposed to a fixed set of relatively few attributes in relational tuples. All items can act as quasi-identifiers ,an attacker who knows them all and can link them to a specific person has nothing to learn from the original database. Her background knowledge already contains the original data. There are three classes of algorithm they are the optimal anonymization (OA) algorithm, which explores in a bottom-up fashion the lattice of all possible combinations of item generalizations, and finds the most detailed such sets of combinations that satisfy km-anonymity. The best combination is then picked, according to an information loss metric. Direct anonymization (DA) heuristic operates directly on m-sized itemsets found to violate k anonymity. There need to share person-specific records in such a way that the identities of the individuals who are the subjects of the data cannot determined, it is determined in Achieving k- Anonimity privacy protection using generalization and suppression[4].Generalization involves replacing(or recoding)a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all.

RELATED WORK:

Anonymizing Classification Data For Privacy Preserving [5] Data sharing in today's globally networked systems poses a threat to individual privacy and organizational confidentiality. First of all, knowing that the data is used for classification does not imply that the data provider knows exactly how the recipient may analyse the data. The recipient often has application-specific bias towards building the classifier. Consider the problem of publishing set-valued data, while

preserving the privacy of individuals associated to them. Local and Global Recoding Methods for Anonymizing Set-valued Data[6] study the problem of protecting privacy in the publication of set-valued data. Consider a collection of supermarket transactions that contains detailed information about items bought together by individuals. Even after removing all personal characteristics of the buyer, which can serve as links to his identity, the publication of such data is still subject to privacy attacks from adversaries who have partial knowledge about the set. A new version of the k-anonymity guarantee, the km-anonymity, to limit the effects of the data dimensionality. Unlike the k-anonymity problem in rela- tional databases there is no fixed, well-defined set of quasi-identifier attributes and sensitive data. A subset of items in a transaction could play the role of the quasi-identifier for the remaining (sensitive) ones and vice-versa. Another fundamental difference is that transactions have variable length and high dimensionality, as opposed to a fixed set of relatively few attributes in relational tuples. The concept of km-anonymity for such data and analysed the space of possible solution. k-Anonymity: A model for protecting privacy[8],k-anonymity with respect to all m subsets of the domain of the set-valued attribute can help avoiding associating the sensitive value to less than k tuples. MultiRelational k-Anonymity [7] ,k-Anonymity protects privacy by ensuring that data cannot be linked to a single individual. In a k-anonymous dataset, any identifying information occurs in at least k tuples. Much research has been done to modify a single table dataset to satisfy anonymity constraints.

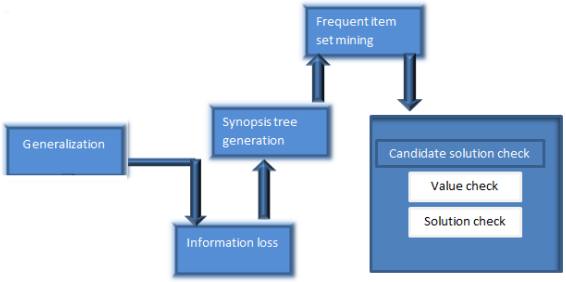
The main observation is that all clustering based anonymity algorithms make use of two basic operations on private entities: anonymization and calculation of the distance between two entities. The latter can be generally defined as the cost of the anonymization of two entities. The assumptions given in the previous section enables us to abstract private entities of multiR databases as trees where each level of a given entity tree corresponds to levels of the nested relation for a particular vip entity. To provide a formal framework for constructing and evaluating algorithms and systems that release information such that the released information limits what can be revealed about properties of the entities that are to be protected. In k-Anonymity: A model for protecting privacy [8],the data holder can identify attributes in his private data that may also appear in external information and therefore, can accurately identify quasi-identifiers. Privacy-preserving Anonymization of Set-valued Data [9], considering the problem of publishing set-valued data, while preserving the privacy of individuals associated to them. However, if the super-market decides to publish its transactions and there is only one transaction containing cheese, scissors ,and light bulb, Jim can immediately infer that this transaction corresponds to Bob and he can find out his complete shopping bag contents. A subset of items in a transaction could play the role of the quasi-identifier for the remaining (sensitive)ones and vice-versa. Another fundamental difference is that transactions have variable length and high dimensionality, opposed to a fixed set of relatively few attributes in relational tuples. Finally, considering that all items that participate in transactions take values from the same domain (i.e, complete universe of items), unlike relational data,

where different attributes of a tuple have different domains.

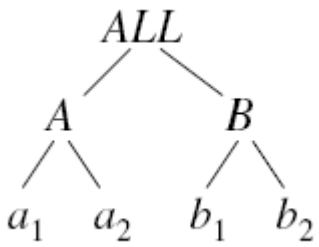
Frequent Item Set Mining Using Tree Structure :

Data anonymization technique have been proposed in order to allow processing of personal data without compromising users privacy. The problem of anonymizing tree structured data has only been addressed in the context of multirelational k-anonimity.k(m,n)-anonymity guarantees attack from the attacker.To prevent attackers who have background knowledge from associating records to individuals and provide an anonymization technique that offers protection against identity disclosure. Define the k(m,n) anonymity privacy guarantee and how it is efficient in concrete attack scenarios, here in this chose the values in a way that the background knowledge of the attacker ,both positive and negative matches atleast one record in the data set. k-anonimity guarantees the protection against identity disclosure, sensitive information may be revealed when there are many identical sensitive attribute values with in an equivalence class(attribute disclosure).k(m,n)-anonymity, which guarantees that an attacker who knows up to m elements of a record and to n structural relations between the m elements will not be able to match her background knowledge to less than k matching records in the anonymized data. The anonymization procedure does not only generalize values that participate in rare item combinations but also simplifies the structure of the records. The simplification is performed by removing nodes from long paths and creating new smaller paths.

System Architecture FIM-anonymizing using tree structured data includes generalization, information loss, synopsis tree, frequent item set mining



Generalization: Data generalization hierarchy (DGH) for every item of I. Each value of a class A is mapped to a value in the next most general level and these values can be mapped to even more general ones. All class hierarchies have a common root denoted as “*”,which means “any” value and is equivalent to suppressing the value. When a value is generalized, then all its appearances in the dataset are replaced by the new, generalized value. Moreover, when a value is generalized then all its siblings are generalized to the same item. The anonymization algorithm will identify a generalization cut C on the DGH.A generalization cut defines the generalization level for each item in the data domain I, i.e.,it defines a horizontal “cut” on the hierarchy tree



Conclusion

k-anonymity is a property processed by certain anonymized data. A release of the data with scientific guarantees that the individual who are the subjects of the data cannot before identified while data remain practically usefull. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. Frequent itemset mining (FIM) is one of the most fundamental problems in data mining. Here exploring the possibility of designing a differentially private FIM algorithm which can not only

References:

[1] Jian Xu1 Wei Wang1 Jian Pei2 Xiaoyuan Wang1 Baile Shi1 Ada Wai-CheeFu,"Utility-Based Anonymization Using Local Recoding":publication Data for Privacy Preservation,: IEEE Transactions on Knowledge and Data Engineering, VOL. 19, NO. 5, MAY 2007.

[2] Manolis Terrovitis, Panos Kalnis, "Privacy preserving Anonymization of Set-valued Data" , VLDB 08, August 24-30, 2008, Auckland, New Zealand.

[3] M. Ercan Nergiz Chris Clifton, \MultiRelational k- Anonymity": , 2007 IEEE.

[4] L. Sweeney, \k-anonymity: a model for protecting privacy ":,International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[5] Pierangela Samarati, Member, IEEE Computer Society, "Protecting Respondents Identities in Microdata Release ": , IEEE Transaction on Knowledge and Data Engineering, VOL. 13, NO. 6, Nov/Dec 2001.

[6] Arik Friedman, Ran Wol, Assaf Schuster,"Providing k- Anonymity in Data Mining".

[7] Olga Gkountouna, Student Member, IEEE and Manolis Terrovitis,,"Anonymizing Collections of Tree-Structured Data,":IEEE Transaction on Knowledge and Data Engineering, VOL. 27, NO. 8, Aug 2015.

[8] L. Sweeney, "k-anonymity: a model for protecting privacy",International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

Author 1:-

Dupati Ashok Chakravarthi
Student under M.Tech program
II year C S E (Software Engg)
ashokgnt33@gmail.com,
Newtons Institute of Engineering

Authoer 2:-

M. Naresh Associate professor
Department of CSE
Newton's institute of engineering
nareshmtech08@gmail.com