FIM-Anonymizing Using Tree Structured Data

Rabitha T[1], Farzin Ahammed T[2]

[1] AWH Engineering College, Calicut University Department of Computer science & Engineering, Kuttikkatoor, Kozhikode, India

Abstract*: FIM-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.*

Keywords: anonymity, generalization , information loss,synopsis tree, frequent item set mining

1. Introduction

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.

2. Related Works

Literature survey deals with the related techniques which contribute to the development of Frequent item set mining method. Here present ten most important papers for developing the problem definition.

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 k^m-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. A value is replaced by a less specific, more general value that is faithful to the original. These techniques can provide results with guarantees of anonymity that are minimally distorted. Any attempt to provide anonymity protection, no matter how minor, involves modifying the data and thereby distorting its contents, so the goal is to distort minimally.

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 k^m-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 k^m-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.

Providing k-Anonymity in Data Mining [10], the one definition of privacy which has come a long way in the public arena and is accepted today by both legislators and corporations is that of k-anonymity. The guarantee given by k-anonymity is that no information can be linked to groups of less than k individuals. For the attacks against k-anonymity, even when sufficient care is taken to identify the quasi-identifier, a solution that adheres to k-anonymity can still be vulnerable to attacks. Fortunately, the attacks presented can be thwarted by due diligence to some accompanying practices. Unsorted matching attack against k-anonymity. This attack is based on the order in which tuples appear in the released table. While having maintained the use of a relational model in this discussion, and so the order of tuples cannot be assumed ,in real-world use this is often a problem. It can be corrected of course, by randomly sorting the tuples of the solution table. Otherwise , the release of a related table can leak sensitive information.

3. 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.

3.1 System Architecture

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

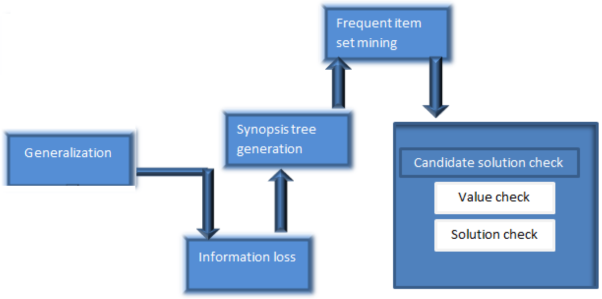

Figure 1: system architecture of FIM

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.

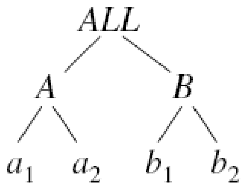

Figure 2: example for generalization

Formally, we use a generalization hierarchy for the complete domain *I* of items that may appear in a transaction. Such an exemplary hierarchy is shown in Figure 1. In this example we assume *I* = {*a*1; *a*2; *b*1; *b*2}, items *a*1, *a*2 can be generalized to *A*, items *b*1, *b*2 can be generalized to *B*, and the two classes *A*, *B* can be further generalized to *ALL*.

Formally, the set of possible transformations corresponds to the set of possible horizontal cuts of the hierarchy tree. Each such cut, defines a unique set of generalization rules. Each cut corresponds to a set of non-overlapping subtrees of the hierarchy, which altogether span the complete domain *I*. The root of each subtree correspond to the generalization rule which maps all values in its leaves to the label of the root.

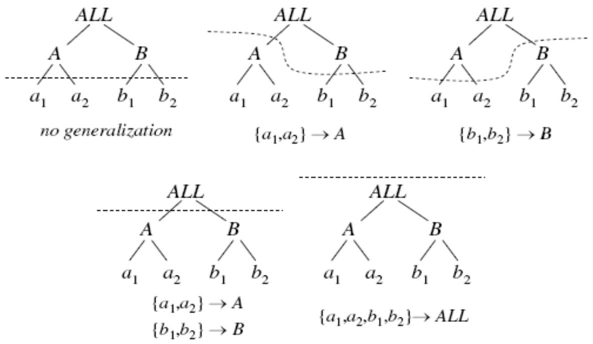

Figure 3: example for generalization cut

Information loss: The value generalization transformations distort the original data and introduce information loss to the published anonymized data**.**

Synopsis Tree: The term support refers to the number of records that contain the path. The synopsis tree facilitates deciding on the k$^{(m,n)}$anonymity of a dataset by tracing not only the support of item combinations from I, but also the support of paths that contain them. A tree structure, which is created by superimposing all records of D. Every record's root node is mapped to a single node, the root rs of the synopsis tree. All paths that appear in a record are superimposed to the synopsis tree starting from rs. Each node n has two elements:

a) A label representing the item that is mapped to it and
b) A sorted list of the ids of all the records that contain the exact path from the root to the current node.

A new entry is added to L for every generalized item gi that appears. This new entry has a list of id which is the result of the union of the id lists of all items I that are mapped to gi.We create the list of sidelinks associated with gi, as the set of all sidelinks in the entry of every item i that is mapped to gi.

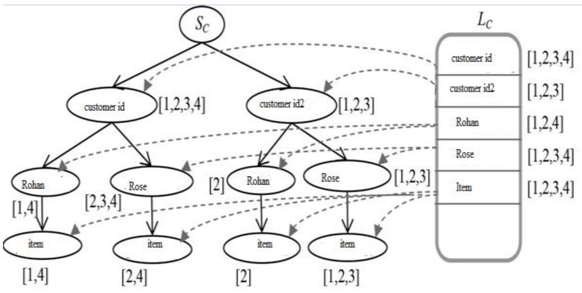

Figure 4: Synopsis tree generalization

Frequent item set mining: This paper focuses on outsourcing frequent itemset mining and examines the issue on how to protect privacy against the case where the attackers have precise knowledge on the supports of some items. The frequent-itemsets problem is that of finding sets of items that appear in (are related to) many of the same dataset.

Privacy Preserving frequent itemset mining Itemset mining FIM) is one of the most fundamental problems in data mining. In the preprocessing phase, to better improve the utility-privacy trade off ,devise a smart splitting method to transform the database. In the mining phase, a run-time estimation method is proposed to offset the information loss incurred by transaction splitting. Putting forward a dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process.

The algorithm for working of proposed system is shown below
1. Create a dataset that adds the details of id, supports.
2. Fetching out the data (supplier name, path of each node ,products details) from the data table.
3. Select the count from the data source and calculate the maximum count of each product sold out.
4. Mining the frequent item set based on the count of each product.

4. Impact and Cost

In the section we present the experimental evaluation of our algorithms. All implementations were done in C++.We implemented an algorithm for frequent item set mining,data. For the experimental evaluation of the proposed algorithms], which is a typical example of a database of customers, orders, products and suppliers, all linked via foreign keys. We parse the relational tables and use the foreign keys to create tree records that represent different individual customers. The resulting trees express the following information: each customer has made a number of orders at a particular date, each containing a number of products (items). To simplify the experimental evaluation we used only the attributes: customer nation, order price, order date, item quantity, manufacturer and brand name from the relational tables, and kept the structural relations between values implied by the schema of the database. We first created a dataset of 1M records and sampled it to create the two other ones.
We limited the fanout of the records (each customer may have up to two orders, each containing up to three items) to create a dataset where the ratio of the size of each record to the total dataset size is small (if the records are too big and too detailed compared to the size of the collection, only very low quality anonymization can be produced with any method).

5. 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 achieve high data utility and a high degree of privacy, but also offer high time efficiency.

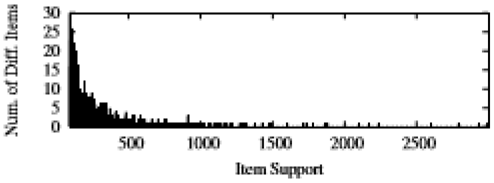

Figure 5: Analysis of Retail: Number of different items under the same support.

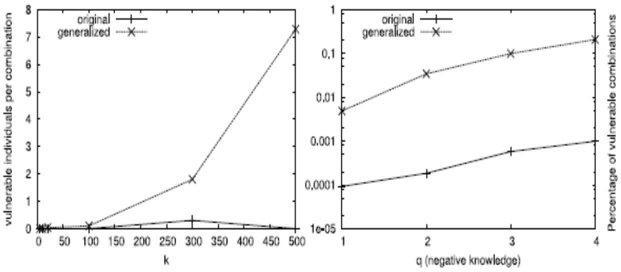

Figure 6: Impact of negative knowledge

6. Future Scope

It is possible to establish a future extension to the current work. In the current system mainly focusing on anonymizing the tree structured data. And also mining the frequent item set based on the product consumed by the customer. So this learning can be including as a future extension in to the existing system.

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.

[9] 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.

[10] Arik Friedman, Ran Wol, Assaf Schuster, “Providing k-Anonymity in Data Mining.”

[11] 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.

Author Profile

Rabitha T received the B Tech degree in Computer Science and Engineering from university of Calicut in 2014 and she is currently persuading her M-tech in Computer Science and Engineering from Calicut Universiy.

Farzin Ahammed received the B Tech and M Tech degrees in Computer Science and Engineering from Calicut University and Kerala University in 2012 and 2015, respectively. He is working with AWH Engineering College since December 2015.