Chapter 14
Database Anonymization Techniques with Focus on Uncertainty and Multi-Sensitive Attributes

B. K. Tripathy
VIT University, India

ABSTRACT

Publication of Data owned by various organizations for scientific research has the danger of sensitive information of respondents being disclosed. The policy of removal or encryption of identifiers cannot avoid the leakage of information through quasi-identifiers. So, several anonymization techniques like k-anonymity, l-diversity, and t-closeness have been proposed. However, uncertainty in data cannot be handled by these algorithms. One solution to this is to develop anonymization algorithms by using rough set based clustering algorithms like MMR, MMeR, SDR, SSDR, and MADE at the clustering stage of existing algorithms. Some of these algorithms handle both numerical and categorical data. In this chapter, the author addresses the database anonymization problem and briefly discusses k-anonymization methods. The primary focus is on the algorithms dealing with l-diversity of databases having single or multi-sensitive attributes. The author also proposes certain algorithms to deal with anonymization of databases with involved uncertainty. Also, the aim is to draw attention of researchers towards the various open problems in this direction.

INTRODUCTION

There is a constant demand on organizations to publish micro data (i.e. data published in its raw, non-aggregated form) in their electronic form for a variety of purposes including demographic and public health research. To protect the anonymity of the entities, called the respondents, data holders often remove or encrypt explicit identifiers. However, de-identifying data provides no guarantee of anonymity as released information often contains other attributes called quasi identifiers, which can be linked to publicly available information for re-identifying data respondents, thus leaking information that was not intended for disclosure. The process of transforming a database into a suitable form before its release such that the disclosure of sensitive attribute values of respondents can be

DOI: 10.4018/978-1-4666-2518-1.ch014

avoided is called anonymization. Two common approaches to anonymise databases have been suppression and generalization. In suppression a value is not released at all. In the process of generalization, the quasi-identifier values are replaced by values which are less specific but semantically consistent. Also, the notion of generalization is enhanced by imposing on each value generalization hierarchy a new maximal element, atop old maximal element. As a result of generalization, more records have the same set of quasi-identifier values. Such a set of records are said to form a cluster or sometimes an equivalence class. The large amount of information easily accessible today, together with the increased computational power available to the attackers, makes linking attacks a serious problem (He at al, 2009). The information disclosure has been identified to be of two types. These are; identity disclosure and attribute disclosure. In identity disclosure an individual is linked to a tuple in the database and so the information available is supposed to belong to that individual. Attribute disclosure occurs when additional information about an individual is obtained, which were not present at the time of release of the data table. It is worth noting that identity disclosure leads to attribute disclosure. However, attribute disclosure may not necessarily need identity disclosure.

LITERATURE SURVEY

To handle linking disclosure while preserving the integrity of the released data, Samarati and Sweeney proposed the concept of k-anonymity (Samarati et al, 1998). In this approach, data privacy is guaranteed by ensuring that any record in the released data is indistinguishable from at least (k-1) other records with respect a set of attributes called the quasi-identifiers. In later years it was further expanded by Sweeney (1998, 2002a, 2002b) to the context of table releases. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. Although the idea of k-anonymity is conceptually straightforward, the computational complexity of finding an optimal solution for the k-anonymity problem has been shown to be NP-hard (Meyerson et al, 2004), even when one considers only the technique of suppression of values (Agrawal et al, 2005; Chiu et al, 2007). In order to obtain k-anonymity, several algorithms have been introduced in recent times (R.Agrawal et al.,2005; Bayardo et al, 2005; Byun et al,2007, LeFevre et al, 2005; Li et al, 2006; Lin et al, 2008; Nergiz et al, 2007; Ng et al 2009; Samarati et al, 2007; Sweeny, 2002). The basic idea in most of these algorithms is that k-anonymization problem can be viewed as a clustering problem. Intuitively, the k-anonymity requirement can be naturally transformed into a clustering problem where we want to find a set of clusters, each of which contains at least k records. In order to maximize data quality, we also want the records in a cluster to be similar to each other as much as possible. This ensures that less distortion is required when the records in a cluster are modified to have the same quasi-identifier value. Some significant contributions in the devise of k-anonymization algorithms are as follows.

The k-anonymity requirement is typically enforced through generalization, where real values are replaced with "less specific but semantically consistent values (Samarati et al (1998)". Given a domain, there are various ways to generalize the values in the domain. Typically numerical values are generalized into intervals and categorical values are generalized into a set of distinct values or a single value that represents such a set. Many times a non-overlapping generalization hierarchy is first defined for each attribute of quasi-identifier. Then an algorithm tries to find an optimal (or good) solution which is allowed by such generalization hierarchies. Although this leads to much more flexible generalization, possible generalizations are still limited by the imposed generalization hierarchies.