Chapter 48
Scalable l-Diversity:
An Extension to Scalable k-Anonymity for Privacy Preserving Big Data Publishing

Udai Pratap Rao
Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India

Brijesh B. Mehta
Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India

Nikhil Kumar
Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India

ABSTRACT

Privacy preserving data publishing is one of the most demanding research areas in the recent few years. There are more than billions of devices capable to collect the data from various sources. To preserve the privacy while publishing data, algorithms for equivalence class generation and scalable anonymization with k-anonymity and l-diversity using MapReduce programming paradigm are proposed in this article. Equivalence class generation algorithms divide the datasets into equivalence classes for Scalable k-Anonymity (SKA) and Scalable l-Diversity (SLD) separately. These equivalence classes are finally fed to the anonymization algorithm that calculates the Gross Cost Penalty (GCP) for the complete dataset. The value of GCP gives information loss in input dataset after anonymization.

INTRODUCTION

The success and failure of any organizations highly depend on the analysis of their business/transaction data. But size of such data is in massive form; hence, it cannot be analyzed by traditional analytical methods. To analyze and handle these data, distributed environment such as MapReduce framework

DOI: 10.4018/978-1-7998-8954-0.ch048

(Dean & Ghemawat, 2008) is required, in which this large volume of data can be distributed over many distributed systems to process and analyze it. Almost every organization used to make their business data public for the use of researchers. This business/transaction data contains private information of their customers, so organizations need to anonymize their data before publishing it publicly.

Preserving privacy as well as to keep the high utility of data is a big challenge in order to publish the big data because data are collected from different sources which may leads to privacy issues (Wu, Zhu, Wu & Ding, 2014; Mehta & Rao, 2016). The anonymization of data will reduce the utility of underlying data. Preserving the privacy of an individual in order to publish the big data with high utility is a challenging task and can be considered as open research problem. In this paper, the discussion about two privacy model *k*-anonymity and *l*-diversity is given and further we propose scalable algorithms for *k*-anonymity and *l*-diversity. The results of SKA and SLD for different value of *k* and *l* for large dataset are also compared.

Privacy Models

Mehta, Rao, Kumar & Gadekula (2016) discussed about the different privacy models and concluded that for big data *k*-anonymity and *l*-diversity are more suitable to preserve the privacy. As *l*-diversity is an extension to *k*-anonymity, first *k*-anonymization need to be applied on dataset and then it can be *l*-diversified. In both the approaches, attributes of dataset are categorizes into four types: Personal Information Identifier (PII), Quasi Identifier (QID), Sensitive Attribute (SA), and Non-sensitive attribute. PII uniquely identifies the individuals so this attribute is removed from the published table. QID is a collection of one or more attribute which alone cannot identify the data owner but its combination with publically available dataset may reveal the identity and sensitive value of the individual. The attribute which data owner do not want to disclose is known as sensitive attribute. Apart from PII, QID and SA all other attributes are called Non sensitive attribute. Now discussion about *k*-anonymity and *l*-diversity is given one by one. Table 1 is an example of the patients published data. In the dataset, UID is PII; Sex, ZIP Code and Age are QIDs; and Disease is SA.

Table 1. Original patient data

S#	UID	Sex	ZIP Code	Age	Disease
1	728953467896	M	852219	34	HIV
2	786545678901	M	852227	32	Flu
3	456732190876	M	855007	43	Flu
4	678904523679	M	855010	49	Malaria
5	890567432673	F	853457	54	Cancer
6	976543097645	F	853401	51	Cancer