Chapter 16
A Secure Protocol for High-Dimensional Big Data Providing Data Privacy

Anitha J.
Dayananda Sagar Academy of Technology and Management, India

Prasad S. P.
Dayananda Sagar Academy of Technology and Management, India

ABSTRACT

Due to recent technological development, a huge amount of data generated by social networking, sensor networks, internet, etc., adds more challenges when performing data storage and processing tasks. During PPDP, the collected data may contain sensitive information about the data owner. Directly releasing this for further processing may violate the privacy of the data owner, hence data modification is needed so that it does not disclose any personal information. The existing techniques of data anonymization have a fixed scheme with a small number of dimensions. There are various types of attacks on the privacy of data like linkage attack, homogeneity attack, and background knowledge attack. To provide an effective technique in big data to maintain data privacy and prevent linkage attacks, this paper proposes a privacy preserving protocol, UNION, for a multi-party data provider. Experiments show that this technique provides a better data utility to handle high dimensional data, and scalability with respect to the data size compared with existing anonymization techniques.

INTRODUCTION

The term big data is defined as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis. Based on this definition, the properties of big data are reflected as volume, velocity, variety, veracity and value. Volume refers to the amount of data generated. With the emergence of social networking, there is dramatic increase in the size of the data. The rate at which new data are generated is often characterized as velocity. A big data may contain text, audio, image, or video

DOI: 10.4018/978-1-5225-9611-0.ch016

etc. This diversity of data is denoted by variety. Veracity refers to the data that are generated uncertain in nature. It is hard to know which information is accurate and which is out of date. Finally the Value of data is valuable for society or not.

The life cycle of the big data has various phases like data generation, data storage and data processing. In data generation phase, large, diverse and complex data are generated by human and machine. Usually, the data generated is large, diverse and complex. Therefore, it is hard for traditional systems to handle them. The data generated are normally associated with a specific domain such as business, Internet, research, etc. Data storage phase refers to storing and managing large data sets. A data storage system consists of two parts namely hardware infrastructure and data management. Hardware infrastructure is utilizing information and communications technology (ICT) resources for various tasks. Data management refers to the set of software deployed on top of hardware infrastructure to manage and query large scale data sets. It should also provide several interfaces to interact with and analyze stored data. In data processing phase, various computations and transformations takes place on data set.

Data processing phase is the process of data collection, data transmission, pre-processing and data extraction. Data collection is needed because data may be coming from different diverse sources i.e., sites that contains text, images and videos. In data transmission phase, after collecting raw data from a specific data production environment, a high speed transmission mechanism to transmit data into a proper storage for various types of analytic applications. The pre-processing phase aims at removing meaningless and redundant parts of the data so that more storage space could be saved. Finally in data extraction phase only useful information are retrieved from data sets.

The excessive data and domain specific analytical methods are used by many application to derive meaningful information. Although different fields in data analytics require different data characteristics, few of these fields may leverage similar underlying technology to inspect, transform and model data to extract value from it. Emerging data analytics research can be classified into the following six technical areas: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics (Xu et al., 2014).

Data generation can be classified into active data generation and passive data generation. Active data generation means that the data owner is willing to provide the data to a third party, while passive data generation refers to the situations that the data are generated by data owner's online activity (e.g., browsing) and the data owner may not even be aware of that the data are being collected by a third party. The major challenge for data owner is that how can he protect his data from any third party who may be willing to collect them. The data owner wants to hide his personal and sensitive information as much as possible and is concerned about how much control he could have over the information.

The data processing phase includes Privacy Preserving Data Publishing (PPDP). During PPDP, the collected data may contain sensitive information about the data owner. Directly releasing the information for further processing may violate the privacy of the data owner, hence data modification is needed in such a way that it does not disclose any personal information about the owner. On the other hand, the modified data should still be useful, not to violate the original purpose of data publishing. The privacy and utility of data are inversely related to each other. Intrusion Detection Scheme (IDS) schemes have been implemented in wired and semi-wired networks. These systems look for certain misbehavior patterns in the network which would give a whiff of a malicious act and thereby trigger attack mitigating mechanism. Many studies have been conducted to modify the data before publishing or storing them for further processing.