ABSTRACT

Privacy is an important issue to be taken into consideration when one wants to use the information, especially at a time when the sensitive information is to be disclosed. In order to protect sensitive information application of privacy preserving data mining techniques are employed. This thesis describes the research issues in publication of confidential information. The primary focus is on the data custodians, for public analysts, researchers and anyone who wants to use the sensitive data. The overall intention is to use information in evaluating the economic model, identifying social trends and the state-of-the-art surveillance in various fields. Sensitive information is included in personal information such as medical records, salary, etc., so that information cannot be readily disclosed. One way to solve this problem is to sign user non-disclosure agreements. Besides this, information cannot be protected against theft even when the victim receives reasonable alerts. So, it is essential to explore the technical solution, which conceals information before its release. Releasing sensitive information like social security numbers by administration, banks and healthcare providers of government agencies and non-government organizations has increased gradually over the past few decades. Although, it has long been realized that both the data protection and conversion need to be stored, it has recently become clear that the data must be adequately protected from unauthorized disclosure when testing new applications related to the data.

In this thesis two extended algorithms are proposed, namely frequency based top-down partitioning algorithm and frequency based bottom-up partitioning algorithm that can efficiently anonymize the microdata with reduced information loss in the presence of unsafe Full Functional

Dependencies (FFDs). The proposed top-down approach, considers the frequency distribution of sensitive values to act as a deciding factor for choosing split points during partitioning. The proposed frequency-based bottom up partitioning, is carried out by considering the frequency of occurrence of sensitive attribute values which substantially minimizes the information loss and thereby increase the performance of a conventional bottom-up approach.

This thesis investigates the privacy-preserving issue and publishes the microdata which considers both Conditional Functional Dependencies (CFDs) and FFDs as the opponents' knowledge. A Compact Frequent Pattern Development Section Algorithm (CFPGBS) is proposed to discover at least CFDs from a dataset when its range is wide. Algorithm formulates (d,l)-privacy model to protect the confidentiality of information created by CFDs, FFDs. An automatic compact repeat pattern is generated for mounting the best CFD patterns that are arranged in the Compact Frequency Pattern Growth Branch Algorithm (CFPGBS) for efficient pattern mining. Also, a small pattern tree is generated, which captures the CFD pattern information of the insertion phase and provides good pattern-matching performance enhancement. Construction of primary partition for the proposed frequency-based bottom-up method is performed by the frequency-distribution of the Log-Skew-Normal Alpha-Power Distribution (LSKNAPD) Frequency Distribution Function. A wide set of experimental results show that the proposed frequency based approaches are efficient when compared to conventional approaches.