Privacy Characterization and Quantification in Data Publishing

ABSTRACT

The increasing interest in collecting and publishing large amounts of individuals’ data to public for purposes such as medical research, market analysis and economical measures has created major privacy concerns about individual’s sensitive information. To deal with these concerns, many Privacy-Preserving Data Publishing (PPDP) techniques have been proposed in literature. However, they lack a proper privacy characterization and measurement. In this paper, we first present a novel multi-variable privacy characterization and quantification model. Based on this model, we are able to analyze the prior and posterior adversarial belief about attribute values of individuals. We can also analyze the sensitivity of any identifier in privacy characterization. Then we show that privacy should not be measured based on one metric. We demonstrate how this could result in privacy misjudgment. We propose two different metrics for quantification of privacy leakage, distribution leakage and entropy leakage. Using these metrics, we analyzed some of the most well-known PPDP techniques such as k-anonymity, l-diversity and t-closeness. Based on our framework and the proposed metrics, we can determine that all the existing PPDP schemes have limitations in privacy characterization. Our proposed privacy characterization and measurement framework contributes to better understanding and evaluation of these techniques. Thus, this paper provides a foundation for design and analysis of PPDP schemes.

**EXISTING SYSTEM**

* Privacy is an important issue when one wants to make use of data that involves individuals' sensitive information. Research on protecting the privacy of individuals and the confidentiality of data has received contributions from many fields, including computer science, statistics, economics, and social science. In the existing system, the system survey research work in privacy-preserving data publishing. This is an area that attempts to answer the problem of how an organization, such as a hospital, government agency, or insurance company, can release data to the public with violating the confidentiality of personal information.
* The system focuses on privacy criteria that provide formal safety guarantees, present algorithms that sanitize data to make it not safe for release while preserving useful information, and discuss ways of analyzing the sanitized data. Many challenges still remain. This survey provides a summary of the current state-of-the-art, based on which we expect to see advances in years to come.

**Disadvantages**

* + There is no Methods to find k-anonymity, l-diversity, and t-closeness on published data sets.
	+ There is no Data Disclosure Model.

**PROPOSED SYSTEM**

* In the proposed system, the system begins by introducing our novel data publishing framework. The proposed framework consists of two steps. First, the system models attributes in a dataset as a multi-variable model. Based on this model, the systems are able to re-define the prior and posterior adversarial belief about attribute values of individuals. Then we characterize privacy of these individuals based on the privacy risks attached with combining different attributes. This model is indeed a more precise model to describe privacy risk of publishing datasets.
* For a given dataset, before it is released, we want to determine to what extent we can achieve privacy. Therefore, the system introduces a new set of privacy quantification metrics to measure the gap between prior information belief and posterior information belief of an adversary, from both local and global perspectives. Specifically, we introduce two privacy leakage measurements: distribution leakage and entropy leakage.
* The system discusses the rationale for these two measurements and illustrates their advantages through examples. The system shows how considering only one metric ignoring the effect of the other strongly contributes to the information leakage and in turn affects the privacy. An intuitive example for this problem is reviewing a blood work. The medical status of a patient cannot be determined based on only one measure even if this particular measure is the most sensitive one.

**Advantages**

* The system model attributes in a dataset as a multi-variable model. Based on this model, the system is able to re-define the prior and posterior adversarial belief about attribute values of individuals.
* The system characterizes privacy of all data individuals based on the privacy risks attached with combining different attributes. This model is indeed a more precise model to describe privacy risk of publishing datasets.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL