Privacy Preserving Data Portals

Benjamin C. M. Fung
Simon Fraser University, Canada

INTRODUCTION

Information in a Web portal often is an integration of data collected from multiple sources. A typical example is the concept of one-stop service, for example, a single health portal provides a patient all of her/his health history, doctor's information, test results, appointment bookings, insurance, and health reports. This concept involves information sharing among multiple parties, for example, hospital, drug store, and insurance company. On the other hand, the general public, however, has growing concerns about the use of personal information. Samarati (2001) shows that linking two data sources may lead to unexpectedly revealing sensitive information of individuals. In response, new privacy acts are enforced in many countries. For example, Canada launched the Personal Information Protection and Electronic Document Act in 2001 to protect a wide spectrum of information (The House of Commons in Canada, 2000). Consequently, companies cannot indiscriminately share their private information with other parties.

A data portal provides a single access point for Web clients to retrieve data. Also, it serves a logical point to determine the trade-off between information sharing and privacy protection. Can the two goals be achieved simultaneously? This chapter formalizes this question to a problem called *secure portals integration for classification* and presents a solution for it. Consider the model in Figure 1. A hospital A and an insurance company B own different sets of attributes about the same set of individuals identified by a common key. They want to share their data via their data portals and present an integrated version in a Web portal to support decision making, such as credit limit or insurance policy approval, while satisfying two privacy requirements:

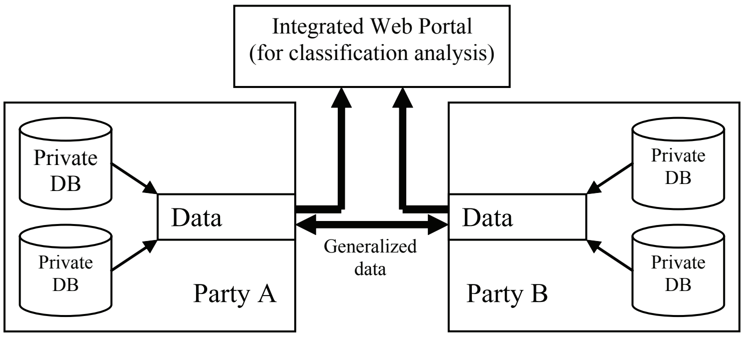
1. The final integrated table has to satisfy the k-anonymity requirement, that is, given a specified set of attributes called a *quasi-identifier* (*QID)*, each value of the QID must be shared by at least k records in the integrated table (Dalenius, 1986).
2. No party can learn more detailed information from another party other than those in the final integrated table during the process of generalization.

Simply joining their data at raw level (e.g., birthday and city) may violate the k-anonymity requirement. Therefore, data portals have to cooperate to determine a generalized version of integrated data (e.g., birth year and province) such that the generalized table remains useful for classification analysis, such as insurance plan approval. Let us first review some building blocks in the literature. Then we elaborate an algorithm, called top-down specialization for 2-party (Wang, Fung, & Dong, 2005), that studies the problem.

BACKGROUND

Privacy-preserving data mining is a study of performing a data-mining task, such as classification, association, and clustering, without violating some given privacy requirement. Recently, this topic has gained enormous attention

Figure 1. Secure portals integration for classification



in the data-mining community because the privacy issue often is an obstacle for real-life data mining and decision support systems.

Agrawal, Evfimievski, and Srikant (2000) achieved privacy on the releasing data by randomization. Randomized data are useful at the aggregated level (such as average or sum), but not at the record level.

Definition 1: k-Anonymity

Consider a person-specific table T with attributes ($D_1,\ldots,D_m$). Each D_i is either a categorical or a continuous attribute. The data owner wants to protect against linking an individual to sensitive information through some subset of attributes called a *quasi-identifier*, or *QID*. A sensitive linking occurs if some value of the QID is shared by only a small number of records in T. k-anonymity requires that each value of the QID must identify at least k records (Dalenius, 1986).

k is a threshold specified by the data owner. The larger the k, the more difficult it is to identify an individual using the QID. Typical values of k ranges from 50 to 500. Sweeney (2002) proposed an algorithm to detect the violation of a given k-anonymity requirement in a data table, and employed generalization to achieve the requirement. Generalization is replacing a specific value (e.g., city) by a consistent general value (e.g., province) according to some *taxonomy tree* in which a leaf node represents a domain value and a parent node represents a less specific value. Figure 2 shows the taxonomy trees for Sex and Education. Compared to randomization, generalization makes information less precise, but preserves the "truthfulness" of information. These works did not consider classification or a specific use of data, and used very simple heuristics to guide generalization.

Iyengar (2002) studied the anonymity problem for classification, and proposed a genetic algorithm solution to generalize and suppress a given table. The idea is encoding each state of generalization as a "chromosome" and encoding data distortion into the fitness function, and employing the genetic evolution to converge to the fittest chromosome. Wang, Yu, and Chakraborty (2004) presented an effective bottom-up approach to address the same problem, but it lacks the flexibility for handling continuous attributes. Recently, Bayardo and Agrawal (2005) proposed and evaluated an optimization algorithm for achieving k-anonymity. Fung, Wang, and Yu (2005) extended the notion of k-anonymity to a privacy requirement with multiple QIDs as follows:

Definition 2: Anonymity Requirement

Consider p quasi-identifiers $QID_1,\ldots,QID_p$ on T. $a(qid_i)$ denotes the number of records in T that share the value qid_i on QID_i. The anonymity of QID_i, denoted $A(QID_i)$, is the smallest $a(qid_i)$ for any value qid_i on QID_i. A table T satisfies the anonymity requirement $\{<QID_1, k_1>,\ldots,<QID_p, k_p>\}$ if $A(QID_i) \geq k_i$ for $1 \leq i \leq p$, where k_i is the anonymity threshold on QID_i specified by the data owner.

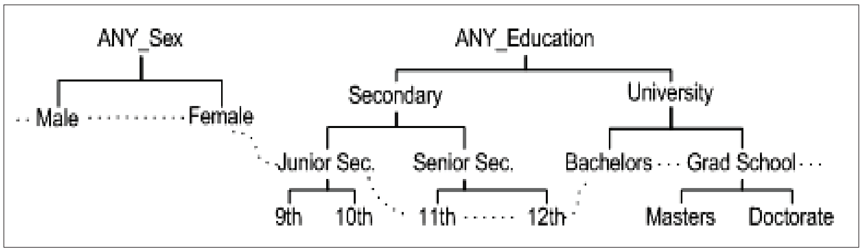
Fung et al. (2005) also presented an efficient method, called top-down specialization (TDS), for the anonymity problem for classification, with the capability to handle both categorical and continuous attributes. All these works address the anonymity problem for classification; however, they did not consider integration of private information from multiple data sources, which is the central idea in this chapter.

Many privacy-preserving algorithms for multiple data sources have been proposed in the literature. For example, secure multiparty computation (SMC) allows sharing of the computed result (i.e., the classifier in our case), but completely prohibits sharing of data (Yao, 1982). Thus, it is not applicable to our portals integration problem. Agrawal et al. (2003) and Liang and Chawathe (2004) proposed the notion of minimal information sharing for computing queries spanning private databases. Still, the shared data in these models is inadequate for classification analysis.

PORTALS INTEGRATION FOR CLASSIFICATION

Two parties want to integrate their data via their portal services to support classification analysis without revealing any sensitive information. A data portal may release data from multiple private databases. To focus on main ideas, we represent all data in $Portal_X$ as a single table T_X.

Figure 2. Taxonomy trees for Sex and Education


ANY_Sex
Male
Female
ANY_Education
Secondary
University
Junior Sec.
Senior Sec.
Bachelors
Grad School
9th
10th
11th
12th
Masters
Doctorate