Encyclopedia of Portal Technologies and Applications

Arthur Tatnall
Victoria University, Australia

Volume II
M–Z

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Acquisitions Editor: Kristin Klinger
Development Editor: Kristin Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Sara Reed
Assistant Managing Editor: Diane Huskinson
Copy Editors: Lanette Ehrhardt, April Schmidt, Katie Smalley, Angela Thor, and Larissa Vinci
Typesetter: Diane Huskinson and Laurie Ridge
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.info-sci-ref.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanonline.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of portal technologies and applications / Arthur Tatnall, editor.
p. cm.
Summary: "This book offers complete coverage of the nature, characteristics, advantages, limitations, design, and evolution of Web portals. Other topics include semantic portals, philosophical portal issues, and personal portals. This authoritative encyclopedia encompasses the economics of setting up and using personal portals, knowledge management, strategic planning, user acceptance, security and the law"--Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-59140-989-2 (hardcover) -- ISBN 978-1-59140-990-8 (ebook)
1. Web portals--Encyclopedias. 2. World Wide Web--Encyclopedias. 3. Knowledge management--Encyclopedias. 4. Online information services--Encyclopedias. 5. Computer network resources--Encyclopedias. I. Tatnall, Arthur.
ZA4201.E53 2007
025.0403--dc22
2007007262

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

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:

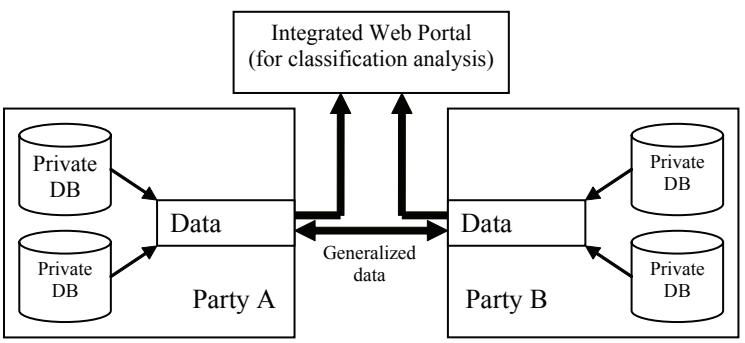
1. The final integrated table has to satisfy the k-anonymity requirement, that is, given a specified set of attributes called a *quasi-identifier* (*QID)*, each value of the QID must be shared by at least k records in the integrated table (Dalenius, 1986).
2. No party can learn more detailed information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

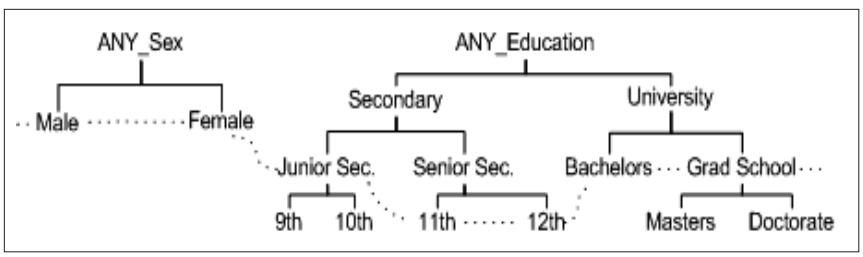
Fung et al. (2005) also presented an efficient method, called top-down specialization (TDS), for the anonymity problem for classification, with the capability to handle both categorical and continuous attributes. All these works address the anonymity problem for classification; however, they did not 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



Definition 3: Secure Portals Integration for Classification

Given two private tables T_A and T_B owned by $Portal_A$ and $Portal_B$ respectively, a joint anonymity requirement $\{<QID_1, k_1>,\ldots,<QID_p, k_p>\}$, and a taxonomy tree for each categorical attribute in QID_i, the secure data integration is to produce a generalized integrated table T such that (1) T satisfies the joint anonymity requirement, (2) T contains as much information as possible for classification, (3) each portal learns nothing from another portal more specific than what is in the final generalized T.

Example 1

Consider the data in Table 1 and the taxonomy trees in Figure 2. $Portal_A$ owns T_A(SSN, sex, class) and $Portal_B$ owns T_B(SSN, education, age, class). Each row represents one or more original records and class contains the distribution of class labels Y and N. After integrating the two tables (by matching the SSN field), the "female doctorate" on (sex, education) becomes unique; therefore, vulnerable to be linked to sensitive information such as age. To protect against such linking, we can generalize master's and doctorate to grad school so that this individual becomes one of many female doctorates. No information is lost for classification analysis because all masters' and doctorates in Table 1 have the same value Y on class. In other words, class does not depend on the distinction of master's and doctorate.

A *cut* of the taxonomy tree for an attribute D_j, denoted Cut_j, contains exactly one value on each root-to-leaf path. The dashed line in Figure 2 represents some cuts on sex and education. We want to find a *solution cut* $\grave{E}Cut_j$ such that the generalized T represented by $\grave{E}Cut_j$ satisfies the anonymity requirement and preserves quality structure for classification. An insight from (Fung et al., 2005) suggested that these two goals are indeed dealing with two types of information: The classification goal requires extracting general structures that capture patterns while the privacy goal requires masking sensitive information, usually specific descriptions that identify individuals. If generalization is performed "carefully," identifying information can be masked while the patterns for classification can be preserved.

Table 1. Raw tables

Shared Attributes		$Portal_A$	$Portal_B$	
SSN	**Class**	**Sex**	**Education**	**Age**
1-3	0Y3N	M	9th	30
4-7	0Y4N	M	10th	32
8-12	2Y3N	M	11th	35
13-16	3Y1N	F	12th	37
17-22	4Y2N	F	Bachelor's	42
23-25	3Y0N	F	Bachelor's	44
26-28	3Y0N	M	Master's	44
29-31	3Y0N	F	Master's	44
32-33	2Y0N	M	Doctorate	44
34	1Y0N	F	Doctorate	44

An Unsecured Solution: Integrate-then-Generalize

An unsecured solution is to first join T_A and T_B into a single table T and then generalize T using the top-down specialization (or TDS) method (Fung et al., 2005). Although this method fails to satisfy requirement (3) in Definition 3, it does satisfy requirements (1) and (2). Here, we first describe TDS; then a secured solution will be discussed next.

TDS is a method proposed for k-anonymizing a single table T for classification analysis. Initially, all attributes in QIDs are generalized to the top-most value and Cut_j contains the top-most value for each attribute D_j. $\grave{E}Cut_j$ represents a set of candidates for specialization. In each iteration, the algorithm selects the specialization w having the highest Score from $\grave{E}Cut_j$, performs the specialization on w in the table, and updates the Score(x) of the affected x in $\grave{E}Cut_j$. Let w → child(w) denote a specialization, where w is parent value and child(w) is a set of child values of w. To specialize a categorical value, a parent value is replaced by its child values according to some given taxonomy tree. To specialize a continuous value, a taxonomy tree is grown at runtime, where each node represents an interval, and each nonleaf node has two subintervals representing some "optimal" binary split of the parent interval. The algorithm keeps pushing $\grave{E}Cut_j$ downwards and terminates if further specialization would lead to violation of the anonymity requirement.

Example 2

Consider Table 1 with QID={Sex, Education, Age}. Initially, every value in QID is generalized to the top-most value. $\grave{E}Cut_j$ = {Any_Sex, Any_Education, [30-44]}. Then compute a Score for each candidate in $\grave{E}Cut_j$. Suppose the winning specialization is ANY_Education → {Secondary, University}. We perform this specialization by replacing every value ANY_Education in the table by either Secondary or University based on the raw value in a data record. Finally, we update $\grave{E}Cut_j$ = {Any_Sex, Secondary, University, [30-44]} and update the Scores for the affected candidates in $\grave{E}Cut_j$.

Algorithm 1. TDS2P for $Portal_B$

```
1:  Initialize Tg to include one record containing top most values;
2:  Initialize ∪Cutj to include only top most values;
3:  while there is some candidate in ∪Cutj do
4:      Find the local candidate x having the highest Score(x);
5:      Communicate Score(x) with PortalA to find the winner;
6:      if the winner w is local then
7:          Specialize w on Tg;
8:          Instruct PortalA to specialize w;
9:      else
10:         Wait for the instruction from PortalA;
11:         Specialize w on Tg using the instruction;
12:     end if
13:     Replace w with child(w) in the local copy of ∪Cutj;
14:     Update Score(x) for candidates x in ∪Cutj;
15: end while
16: return Tg and ∪Cutj;
```

A Secured Solution: TDS for Two Parties

Consider two tables, T_A and T_B, with a common key owned by $Portal_A$ and $Portal_B$ respectively. Each portal keeps a copy of the current $\grave{E}Cut_j$ and generalized joined table, denoted T_g. The nature of the top-down specialization approach implies that T_g is more general than the final answer; so requirement (3) in Definition 3 is satisfied. In each iteration, the two portals cooperate to perform the same specialization with the highest Score, as discussed in TDS. Algorithm 1 describes the procedure at $Portal_B$ (same for $Portal_A$).

Example 3

Consider the same procedure illustrated in Example 2, but the data is partitioned into two tables. Initially, both portals generalize their values to the top most values. $Portal_B$ finds the local best candidate and communicates with $Portal_A$ to identify the overall winning specialization. Suppose the winner is ANY_Education → {Secondary, University}. $Portal_B$ performs this specialization on its copy of $\grave{E}Cut_j$ and T_g. This means specializing records with SSN=1-16 to Secondary, and specializing records with SSN=17-34 to University. Since $Portal_A$ does not have the attribute Education, $Portal_B$ needs to instruct $Portal_A$ how to partition these records in terms of SSNs.

TDS2P has the following practical features:

- **Information vs. Privacy:** Both information and privacy are considered at each specialization. This notion is captured by the Score function, which aims at maximizing the information gain and minimizing the privacy loss.
- **Handling both Categorical and Continuous Attributes:** TDS2P can generalize categorical attributes according to some user-specified taxonomy trees and dynamically grow taxonomy trees at runtime for continuous attributes.
- **Efficiency and Scalability:** In each iteration, a key operation is updating the Scores of the affected candidates in $\grave{E}Cut_j$. In general, this requires accessing data records. TDS2P incrementally maintains some "count statistics" to eliminate the expensive data access.
- **Anytime Solution:** User may step through each specialization to determine a desired trade-off between accuracy and privacy, stop at any time, and produce a table satisfying the anonymity requirement. The bottom-up generalization method, such as Wang et al. (2004), does not support this feature.

Evaluation of TDS2P

The TDS2P algorithm was experimentally evaluated in Fung et al. (2005) and Wang et al. (2005). To illustrate the impacts of generalization on the classification analysis, we compared the classification error on the original data table to the classification error on the generalized (i.e., k-anonymized) data table, and examined with different classifiers. The difference between the two classification errors is small, suggesting that accurate classification and privacy protection can coexist. Typically, there were redundant (classification) structures in the data. If generalization eliminated some structures, other previously unused structures took over the classification task.

Experiments show that the top-down specialization approach is significantly more efficient and scalable than

Iyengar's (2002) genetic approach. TDS2P took only 20 seconds to generalize the data, including reading data records from disk and writing the generalized data to disk, in a multiportal environment. Iyengar reported that his method requires 18 hours to transform the same dataset for a single data source. Also, Iyengar's solution is not suitable for the problem of secure portals integration. Moreover, TDS2P is scalable for handling large data sets by maintaining count statistics instead of scanning raw records. On an enlarged dataset, TDS2P can generalize 200K records within several minutes. (See Fung et al., 2005, and Wang et al., 2005 for details.)

FUTURE TRENDS

In September 2004, the Department of Homeland Security received $9 million grants to foster and evaluate uses of "state-of-the-market" information technology that will improve information sharing and integration among the network of security agencies (The United States Department of Homeland Security, 2004). On the other hand, several surveys indicate that the public feels an increased sense of intrusion and loss of privacy (Gatehouse, 2005). A future trend in enterprise information systems is considering privacy protection as a fundamental requirement. Data portal serves a logical point for determining an appropriate trade-off between privacy protection and information analysis.

Dynamic data types, such as stream data and multimedia data, become very popular in many portal applications, for example, security, monitoring, stocks trading, and fraud detection systems. Many new data analysis algorithms were invented to handle these data types. It would be challenging, but potentially beneficial, to design these systems with the consideration of privacy preservation.

CONCLUSION

We studied secure portals integration for the purpose of joint classification analysis, formalized this problem as achieving the k-anonymity on the integrated data without revealing more detailed information in this process, presented a solution, and briefly evaluated the impacts of generalization on classification quality, efficiency, and scalability. Compared to classic secure multiparty computation, a unique feature of TDS2P is to allow data sharing instead of only result sharing. This feature is important for online data analysis in portal environment where user interaction usually leads to better results. Being able to share data across portals would permit such exploratory data analysis and explanation of results.

REFERENCES

Agrawal, R., Evfimievski, A., & Srikant, R. (2003). Information sharing across private databases. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data* (pp. 86-97). San Diego, CA.

Agrawal, R., & Srikant, R. (2000). Privacy preserving data mining. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (pp. 439-450). Dallas, TX.

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *Proceedings of the 21st IEEE International Conference on Data Engineering* (pp. 217-228). Tokyo, Japan.

Dalenius, T. (1986). Finding a needle in a haystack—or identifying anonymous census record. *Journal of Official Statistics, 2*, 329-336.

Fung, B. C. M., Wang, K., & Yu, P. S. (2005). Top-down specialization for information and privacy preservation. *Proceedings of the 21st IEEE International Conference on Data Engineering* (pp. 205-216). Tokyo, Japan.

Gatehouse, J. (2005). You are exposed. *Maclean's*, November 21, 26-29.

The House of Commons in Canada. (2000). *The personal information protection and electronic documents act*. Retrieved February 21, 2006, fromk http://www.privcom.gc.ca

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data mining* (pp. 279-288). Edmonton, AB, Canada.

Liang, G., & Chawathe, S. S. (2004). Privacy-preserving inter-database operations. In *Proceedings of the 2004 Symposium on Intelligence and Security Informatics* (pp. 66-82). Tucson, AZ.

Samarati, P. (2001) Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge Engineering, 13*(6), 1010-1027.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10*, 571-588.

The United States Department of Homeland Security. (2004). *Department of Homeland Security announces $9 million in information technology grants*. Retrieved February 21, 2006, from http://www.dhs.gov/dhspublic/display?content=4022

Wang, K., Fung, B. C. M., & Dong, G. (2005). Integrating private databases for data analysis. In *Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics* (pp. 171-182). Atlanta, GA.

Wang, K., Yu, P. S., & Chakraborty, S. (2004). Bottom-up generalization: A data mining solution to privacy protection. In *Proceedings of the 4th IEEE International Conference on Data Mining* (pp. 249-256). Brighton, UK.

Yao, A. C. (1982). Protocols for secure computations. In *Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science* (Vol. 12, pp. 160-164).

KEY TERMS

Data Portal: A Web service that provides an access point for Web clients (or other Web services) to retrieve information from a data owner.

K-Anonymity Requirement: Given a specified subset of attributes called a *quasi-identifier*, the k-anonymity requirement requires each value of the quasi-identifier must identify at least k records. The larger the k, the more difficult it is to identify an individual using the quasi-identifier.

Privacy-Preserving Data Mining: A study of achieving some data mining tasks, such as classification, association, and clustering without revealing any sensitive information of the individuals' in the analyzed dataset. The definition of privacy constraint varies in different problems.

Quasi-Identifier (QID): A quasi-identifier is a set of attributes $(A_1,\ldots,A_j)$ whose release must be controlled according to a specified k-anonymity privacy requirement.

Secure Multiparty Computation: A cryptographic protocol among a set of data owners, where some of the inputs needed for computing a function have to be hidden from parties other than the original owner.

Secure Portals Integration: Given two private tables, T_A and T_B, owned by $Portal_A$ and $Portal_B$, respectively, a joint anonymity requirement $\{<QID_1, k_1>,\ldots,<QID_p, k_p>\}$, the secure portals integration is to produce a generalized integrated table T such that (1) T satisfies the joint anonymity requirement, (2) each portal learns nothing about the other portal more specific than what is in the final generalized T.

Secure Portals Integration for Classification: Extending the definition of Secure Portals Integration, the generalized integrated table T has to contain as much information as possible for classification analysis.

Taxonomy Tree: A leaf node represents a domain value and a parent node represents a less specific value. Generalization and specialization replaces record values according to some taxonomy trees.