Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume I
A–Data Pre

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
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@igi-global.com
Web site: http://www.igi-global.com/reference

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.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

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

Data Confidentiality and Chase-Based Knowledge Discovery

Seunghyun Im
University of Pittsburgh at Johnstown, USA

Zbigniew W. Ras
University of North Carolina, Charlotte, USA

INTRODUCTION

This article discusses data security in Knowledge Discovery Systems (KDS). In particular, we presents the problem of confidential data reconstruction by Chase (Dardzinska and Ras, 2003c) in KDS, and discuss protection methods. In conventional database systems, data confidentiality is achieved by hiding sensitive data from unauthorized users (e.g. Data encryption or Access Control). However, hiding is not sufficient in KDS due to Chase. Chase is a generalized null value imputation algorithm that is designed to predict null or missing values, and has many application areas. For example, we can use Chase in a medical decision support system to handle difficult medical situations (e.g. dangerous invasive medical test for the patients who cannot take it). The results derived from the decision support system can help doctors diagnose and treat patients. The data approximated by Chase is particularly reliable because they reflect the actual characteristics of the data set in the information system.

Chase, however, can create data security problems if an information system contains confidential data (Im and Ras, 2005) (Im, 2006). Suppose that an attribute in an information system S contains medical information about patients; some portions of the data are not confidential while others have to be confidential. In this case, part or all of the confidential data in the attribute can be revealed by Chase using knowledge extracted at S. In other words, self-generated rules extracted from non-confidential portions of data can be used to find secret data.

Knowledge is often extracted from remote sites in a Distributed Knowledge Discovery System (DKDS) (Ras, 1994). The key concept of DKDS is to generate global knowledge through knowledge sharing. Each site in DKDS develops knowledge independently, and they are used jointly to produce global knowledge without complex data integrations. Assume that two sites S_1 and S_2 in a DKDS accept the same ontology of their attributes, and they share their knowledge in order to obtain global knowledge, and an attribute of a site S_1 in a DKDS is confidential. The confidential data in S_1 can be hidden by replacing them with null values. However, users at S_1 may treat them as missing data and reconstruct them with Chase using the knowledge extracted from S_2. A distributed medical information system is an example that an attribute is confidential for one information system while the same attribute may not be considered as secret information in another site. These examples show that hiding confidential data from an information system does not guarantee data confidentiality due to Chase, and methods that would protect against these problems are essential to build a security-aware KDS.

BACKGROUND

Data Security and Knowledge Discovery System

Security in KDS has been studied in various disciplines such as cryptography, statistics, and data mining. A well known security problem in cryptography area is how to acquire global knowledge in a distributed system while exchanging data securely. In other words, the objective is to extract global knowledge without disclosing any data stored in each local site. Proposed solutions are based primarily on the idea of secure multiparty protocol (Yao, 1996) that ensures each participant cannot learn more than its own input and outcome of

a public function. Various authors expanded the idea to build a secure data mining systems. Clifton and Kantarcioglou employed the concept to association rule mining for vertically and horizontally partitioned data (Kantarcioglou and Clifton, 2002). Du et al, (Du and Zhan, 2002) and Lindell et al, (Lindell and Pinkas, 2000) used the protocol to build a decision tree. They focused on improving the generic secure multiparty protocol for ID3 algorithm [Quinlan, 1993]. All these works have a common drawback that they require expensive encryption and decryption mechanisms. Considering that real world system often contain extremely large amount of data, performance has to be improved before we apply these algorithms. Another research area of data security in data mining is called perturbation. Dataset is perturbed (e.g. noise addition or data swapping) before its release to the public to minimize disclosure risk of confidential data, while maintaining statistical characteristics (e.g. mean and variable). Muralidhar and Sarathy (Muralidhar and Sarathy, 2003) provided a theoretical basis for data perturbation in terms of data utilization and disclosure risks. In KDD area, protection of sensitive rules with minimum side effect has been discussed by several researchers. In (Oliveira & Zaiane, 2002), authors suggested a solution to protecting sensitive association rules in the form of "sanitization process" where protection is achieved by hiding selective patterns from the frequent itemsets. There has been another interesting proposal (Saygin & Verykios & Elmagarmid, 2002) for hiding sensitive association rules. They introduced an interval of minimum support and confidence value to measure the degree of sensitive rules. The interval is specified by the user and only the rules within the interval are to be removed. In this article, we focus on data security problems in distributed knowledge sharing systems. Related works concentrated only on a standalone information system, or did not consider knowledge sharing techniques to acquire global knowledge.

Chase Algorithm

The overall steps for Chase algorithm is the following.

1. Identify all incomplete attribute values in S.
2. Extract rules from S describing these incomplete attribute values.
3. Null values in S are replaced by values (with their weights) suggested by the rules.
4. Steps 1-3 are repeated until a fixed point is reached.

More specifically, suppose that we have an incomplete information system $S = (X,A,V)$ where X is a finite set of object, A is a finite set of attributes, and V is a finite set of their values. Incomplete information system is a generalization of an information system introduced by (Pawlak, 1991). It is understood by having a set of weighted attribute values as a value of an attribute. In other words, multiple values can be assigned as an attribute value for an object with their weights (w). Assuming that a knowledge base $KB = \{(t \rightarrow v_c) \in D : c \in In(A)\}$ is a set of all classification rules extracted from S by $ERID(S, \lambda_1 \lambda_2)$, where $In(A)$ is the set of incomplete attributes in S, v_c is a value of attribute c, and λ_1, λ_2 are thresholds for minimum support and minimum confidence, correspondingly. *ERID* (Dardzinska and Ras, 2003b) is the algorithm for discovering rules from incomplete information systems, which can handle weighted attribute values. Assuming further that $Rs(x_i) \subseteq KB$ is the set of rules that all of the conditional part of the rules match with the attribute values in $x_i \in S$, and $d(x_i)$ is a null value, then, there are three cases for null value imputations (Dardzinska and Ras, 2003a, 2003c):

1. $Rs(x_i) = \Phi$. $d(x_i)$ cannot be replaced.
2. $Rs(x_i) = \{ r_1 = [t_1 \rightarrow d_1], r2 = [t_1 \rightarrow d_1], \ldots r_k = [t_k \rightarrow d_i]\}$. $d(x_i) = d_1$ because every rule predicts a single decision attribute value.
3. $Rs(x_i) = \{ r_1 = [t1 \rightarrow d_1], r2 = [t_1 \rightarrow d_2], \ldots r_k = [t_k \rightarrow d_k]\}$. Multiple values can replace $d(x_i)$.

Clearly, the weights of predicted values, which represent the strength of prediction, are 1 for case 2. For case 3, weight is calculated based on the confidence and support of rules used by Chase (Ras and Dardzinska, 2005b). Chase is an iterative process. An execution of the algorithm for all attributes in S typically generates a new information system, and the execution is repeated until it reaches a state where no improvement is achieved.

MAIN FOCUS OF THE ARTICLE

Reconstruction of Confidential Data by Chase

Suppose that an information system S is part of a knowledge discovery system (either single or distributed). Then, there are two cases in terms of the source of knowledge.

1. Knowledge is extracted from local site, S
2. Knowledge is extracted from remote site, S_i, *for* $S_i \neq S$

Let's first consider a simple example that illustrates how locally extracted rules can be used to reveal confidential data. Suppose that part of an attribute d is confidential and it is denoted as $v_{conf} = \{d(x_3), d(x_4)\}$. To protect v_{conf}, we hide them by constructing a new information system S_d, where

1. $a_{S(x)} = a_{S_d(x)}$, for any $a \in A - \{d\}, x \in X$
2. $d_{S_d(x)} = d_{S_d(x)} - v_{conf}$,

Now, we extract rules from S_d in terms of d. Assuming that a rule, $r_1 = a_1 \rightarrow d_3$ is extracted. It is applicable to objects that d_3 was hidden, meaning the conditional part of r_1 overlaps with the attribute value in the object. We can use Chase to predict the confidential data d_3. Clearly, it is possible that predicted values are not equal to the actual values or its weight is low. In general, there are three different cases.

1. $d_{S_d(x)} = d_{S(x)}$ and $w \geq \lambda$
2. $d_{S_d(x)} = d_{S(x)}$ and $w < \lambda$
3. $d_{S_d(x)} \neq d_{S(x)}$

where λ is the minimum threshold value for an information system (Ras and Dardzinska, 2005). Clearly, $d_{S_d(x)}$ in case 1 is our major concern because the confidence of approximated data is higher than λ. We do not need to take any action for case 2 and case 3.

The notion of confidential data disclosure by Chase can be extended to Distributed Knowledge Discovery System. The principal of DKDS is that each site develops knowledge independently, and the knowledge is used jointly to produce global knowledge (Ras, 1994), so that each site acquires global knowledge without implementing complex data integrations. The security problem in this environment is created by the knowledge extracted from remote sites. For example, assume that an attribute d in an information system S (See Table 1) is confidential and we hide d from S and construct $S_d = (X,A,V)$, where:

1. $a_{S(x)} = a_{S_d(x)}$, for any $a \in A - \{d\}, x \in X$
2. $d_{S_d(x)}$ is undefined, for any $x \in X$,

In this scenario, there exists no local rule describing d because d is completely hidden. Instead, rules are extracted from remote sites (e.g. r_1, r_2 in Table 2). Now, the process of missing value reconstruction is similar to that of local Chase. For example, $r_1 = b_1 \rightarrow d_1$ supports objects $\{x_1, x_3\}$, and $r_2 = a_2 \cdot b_2 \rightarrow d_2$ supports objects $\{x_4\}$. The confidential data, d_1, and d_2, can be reconstructed using these two rules.

Rules are extracted from different information systems in DKDS. Inconsistencies in semantics (if exists) have to be resolved before any null value imputation can be applied (Ras and Dardzinska, 2004a). In general, we assume that information stored in an ontology of a system (Guarino, and Giaretta, 1995),(Sowa, 1999, 2000) and in inter-ontologies among systems (if they are required and provided) are sufficient to resolve

Table 1. Information system S_d

X	a	B	c	d
x_1	a_1	b_1	c_1	hidden
x_2	a_1	b_2	c_1	hidden
x_3	a_2	b_1	c_3	hidden
x_4	a_2	b_2	c_2	hidden
x_5	a_3	b_2	c_2	hidden
x_6	a_3	b_2	c_4	hidden

Table 2. Rules in Knowledge Base

Rid	Rule	Support	confidence	Source
r_1	$b_1 \rightarrow d_1$	20%	100%	Remote
r_2	$a_2 \cdot b_2 \rightarrow d_2$	20%	100%	Remote
r_3	$c_1 \rightarrow a_1$	33%	100%	Local

inconsistencies in semantics of all sites involved in Chase.

Achieving Data Confidentiality

Clearly, additional data have to be hidden or modified to avoid reconstruction of confidential data by Chase. In particular, we need to change data from non-confidential part of the information system. An important issue in this approach is how to minimize data loss in order to preserve original data set as much as possible. The search space for finding the minimum data set is, in general, very large because of the large number of predictions made by Chase. In addition, there are multiple ways to hide data for each prediction. Several algorithms have been proposed to improve performance based on the discovery of Attribute Value Overlap (Im and Ras, 2005) and Chase Closure (Im , Ras and Dardzinska 2005a).

We can minimize data loss further by taking advantage of hierarchical attribute structure (Im , Ras and Dardzinska, 2005b). Unlike single-level attribute system, data collected with different granularity levels are assigned to an information system with their semantic relations. For example, when the age of Alice is recorded, its value can be either 20 or *young*. Assuming that exact age is sensitive and confidential, we may show her age as 'young' if revealing the value, '*young',* does not compromise her privacy. Clearly, such system provides more data to users compared to the system that has to hide data completely.

Unlike the previous example where knowledge is extracted and stored in KB before we apply a protection algorithm, some systems need to generate rules after we hide a set of data from an information system. In this case, we need to consider knowledge loss (in the form of rules). Now, the objective is that secrecy of data is being maintained while the loss of knowledge in each information systems is minimized (Ras and Dardzinska and Im, 2006). Clearly, as we start hiding additional attribute values from an information system S, we also start losing some knowledge because the data set may be different. One of the important measurements of knowledge loss can be its interestingness. The interestingness of knowledge (Silberschatz and Tuzhilin, 1996) is classified largely into two categories (Silberschatz and Tuzhilin, 1996): subjective and objective. Subjective measures are user-driven, domain-dependent. This type of measurement includes unexpectedness, novelty, actionable (Ras and Tsay, 2005) (Ras and Tsay, 2003) rules. Objective measures are data-driven and domain-independent. They evaluate rules based on statistics and structures of patterns, (e.g., support, confidence, etc).

If the KDS allows for users to use Chase with the rules generated with any support and confidence value, some of the confidential data protected by the described methods may be disclosed. This is obvious because Chase does not restrict minimum support and confidence of rules when it reconstructs null values. A naive solution to this problem is to run the algorithm with a large number of rules generated with wide range of confidence and support values. However, as we increase the size of KB, more attribute values will most likely have to be hidden. In addition, malicious users may use even lower values for rule extraction attributes, and we may end up with hiding all data. In fact, ensuring data confidentiality against all possible rules is difficult because Chase does not enforce minimum support and confidence of rules when it reconstructs missing data. Therefore, in these knowledge discovery systems, the security against Chase should aim to reduce the confidence of the reconstructed values, particularly, by meaningful rules, such as rules with high support or high confidence, instead of trying to prevent data reconstruction by all possible rules. One of the ways to protect confidential data in this environment is to find object reducts (Skowron and Rauszer 1992) and use

the reducts to remove attribute values that will more likely be used to predict confidential attribute values with high confidence.

FUTURE TRENDS

More knowledge discovery systems will use Chase as a key tool because Chase provides robust prediction of missing or unknown attribute values in knowledge discovery systems. Therefore, further research and development for data security and Chase (or privacy in data mining in general) will be conducted. This includes providing data protection algorithms for dynamic information system or improving usability of the methods for knowledge experts.

CONCLUSION

Hiding confidential data from an information system is not sufficient to provide data confidentiality against Chase in KDS. In this article, we presented the process of confidential data reconstruction by Chase, and solutions to reduce the risk. Additional data have to be hidden or modified to ensure the safekeeping of the confidential data from Chase. Data and knowledge loss have to be considered to minimize the changes to existing data and knowledge. If the set of knowledge used to Chase is not completely known, we need to hide data that will more likely be used to predict confidential data with high confidence

REFERENCES

Du, W. and Zhan, Z. (2002). Building decision tree classifier on private data. *Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining*.

Dardzinska, A., Ras, Z. (2003a). Chasing Unknown Values in Incomplete Information Systems, *Proceedings of ICDM 03 Workshop on Foundations and New Directions of Data Mining*.

Dardzinska, A., Ras, Z. (2003b). On Rules Discovery from Incomplete Information Systems, *Proceedings of ICDM 03 Workshop on Foundations and New Directions of Data Mining*.

Dardzinska, A., Ras, Z. (2003c). Rule-Based Chase Algorithm for Partially Incomplete Information Systems, *Proceedings of the Second International Workshop on Active Mining*.

Du, W. and Atallah, M. J. (2001). Secure multi-party computation problems and their applications: A review and open problems. *New Security Paradigms Workshop*

Guarino, N., Giaretta, P. (1995). Ontologies and knowledge bases, towards a terminological clarification, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*.

Im, S. (2006). Privacy aware data management and chase. *Fundamenta Informaticae, Special issue on intelligent information systems*. IOS Press.

Im, S., Ras, Z. (2005). Ensuring Data Security against Knowledge Discovery in Distributed Information System, *Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*.

Im, S., Ras, Z., Dardzinska, A. (2005a). Building A Security-Aware Query Answering System Based On Hierarchical Data Masking, *Proceedings of the ICDM Workshop on Computational Intelligence in Data Mining*.

Im, S., Ras, Z., Dardzinska, A. (2005b). SCIKD: Safeguarding Classified Information against Knowledge Discovery, *Proceedings of the ICDM Workshop on Foundations of Data Mining*

Kantarcioglou, M. and Clifton, C. (2002). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, page 24-31.

Lindell, Y. and Pinkas, B. (2000). Privacy preserving data mining. *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology*, page 36-54.

Muralidhar, K. and Sarathy, R. (1999). Security of random data perturbation methods. *ACM Trans. Database System*, 24(4), page 487-493.

Oliveira, S. R. M. and Zaiane, O. R. (2002). Privacy preserving frequent itemset mining. *Proceedings of the*

IEEE ICDM Workshop on Privacy, Security and Data Mining, page 43-54.

Pawlak, Z. (1991). Rough sets-theoretical aspects of reasoning about data, Kluwer

Quinlan, J. (1993). C4.5: Programs for machine learning.

Ras, Z. (1994). Dictionaries in a distributed knowledge-based system, *Concurrent Engineering: Research and Applications, Concurrent Technologies Corporation*

Ras, Z., Dardzinska, A. (2004a). Ontology Based Distributed Autonomous Knowledge Systems, *Information Systems International* Journal, 29(1), page 47–58.

Ras, Z., Dardzinska, A. (2005). CHASE-2: Rule based chase algorithm for information systems of type lambda, *Proceedings of the Second International Workshop on Active Mining*

Ras Z., Dardzinska, A. (2005b). Data security and null value imputation in distributed information systems. *Advances in Soft Computing*, page 133-146. Springer-Verlag.

Zbigniew Ras, A. Dardzinska, Seunghyun Im, (2006). Data Security versus Knowledge Loss, *Proceedings of the International Conference on AI*

Ras, Z. and Tsay, L. (2005). Action rules discovery: System dear2, method and experiments. *Experimental and Theoretical Artificial Intelligence*

Ras, Z. and Tsay, L.-S. (2003). Discovering extended action rules (system dear). *Proceedings of intelligent information systems*, pages 293~300.

Saygin, Y., Verykios, V., and Elmagarmid, A. (2002). Privacy preserving association rule mining. *Proceedings of the 12th International Workshop on Research Issues in Data Engineering*, page 151~158.

Skowron A and Rauszer C. (1992). The discernibility matrices and functions in information systems. Intelligent Decision Support, 11:331–362.

Silberschatz, A., Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Trans. On Knowledge And Data Engineering*, 8:970-974.

Sowa, J.F. (1999). Ontological categories, in L. Albertazzi, ed., Shapes of Forms: From Gestalt Psychology and Phenomenology to Ontology and Mathematics, Kluwer, 307-340.

Sowa, J.F. (2000). Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole Publishing Co., Pacific Grove, CA.

Yao, A. C. (1996). How to generate and exchange secrets. *Proceedings of the 27th IEEE Symposium on Founda- tions of Computer Science*, pages 162~167.

KEY TERMS

Chase: A recursive strategy applied to a database V, based on functional dependencies or rules extracted from V, by which a null value or an incomplete value in V is replaced by more complete values.

Distributed Chase: A recursive strategy applied to a database V, based on functional dependencies or rules extracted both from V and other autonomous databases, by which a null value or an incomplete value in V is replaced by more complete values. Any differences in semantics among attributes in the involved databases have to be resolved first.

Data Confidentiality: Secrecy of confidential data by hiding a set of attribute values from unauthorized users in the knowledge discovery system.

Knowledge Base: A collection of rules defined as expressions written in predicate calculus. These rules have a form of associations between conjuncts of values of attributes.

Knowledge Discovery System: A set of information systems that is designed to extract and provide patterns and rules hidden from a large quantity of data. The system can be standalone or distributed.

Ontology: An explicit formal specification of how to represent objects, concepts and other entities that are assumed to exist in some area of interest and relationships holding among them. Systems that share the same ontology are able to communicate about domain of discourse without necessarily operating on a globally shared theory. System commits to ontology if its observable actions are consistent with the definitions in the ontology.