Encyclopedia of Information Ethics and Security

Marian Quigley
Monash University, Australia

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 Editor: Maria Boyer
Typesetter: Sara Reed
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-pub.com
Web site: http://www.igi-pub.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.eurospanonline.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of information ethics and security / Marian Quigley, Editor.
p. cm.
Topics address a wide range of life areas affected by computer technology, including: education, the workplace, health, privacy, intellectual property, identity, computer crime, cyber terrorism, equity and access, banking, shopping, publishing, legal and political issues, censorship, artificial intelligence, the environment, communication.
Summary: "This book is an original, comprehensive reference source on ethical and security issues relating to the latest technologies. It covers a wide range of themes, including topics such as computer crime, information warfare, privacy, surveillance, intellectual property and education. It is a useful tool for students, academics, and professionals"--Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-59140-987-8 (hardcover) -- ISBN 978-1-59140-988-5 (ebook)
1. Information technology--Social aspects--Encyclopedias. 2. Information technology--Moral and ethical aspects--Encyclopedias. 3. Computer crimes--Encyclopedias. 4. Computer security--Encyclopedias. 5. Information networks--Security measures--Encyclopedias. I. Quigley, Marian.
HM851.E555 2007
174'.900403--dc22
2007007277

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

Data Security and Chase

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

Seunghyun Im
University of Pittsburgh at Johnstown, USA

INTRODUCTION

This article describes requirements and approaches necessary for ensuring data confidentiality in knowledge discovery systems. Data mining systems should provide knowledge extracted from their data which can be used to identify underlying trends and patterns, but the knowledge should not be used to compromise data confidentiality. Confidentiality for sensitive data is achieved, in general, by hiding them from unauthorized users in conventional database systems (e.g., data encryption and/or access control methods can be considered as data hiding). However, it is not sufficient to hide the confidential data in knowledge discovery systems (KDSs) due to Chase (Dardzinska & Ras, 2003a, 2003c). Chase is a missing value prediction tool enhanced by data mining technologies. For example, if an attribute is incomplete in an information system, we can use Chase to approximate the missing values to make the attribute more complete. It is also used to answer user queries containing non-local attributes (Ras & Joshi, 1997). If attributes in queries are locally unknown, we search for their definitions from KDSs and use the results to replace the non-local part of the query.

The problem of Chase with respect to data confidentiality is that it has the ability to reveal hidden data. Sensitive data may be hidden from an information system, for example, by replacing them with null values. However, any user in the KDS who has access to the knowledge base is able to reconstruct hidden data with Chase by treating them as incomplete or missing elements. For example, in a standalone information system with a partially confidential attribute, knowledge extracted from a non-confidential part can be used to reconstruct hidden (confidential) values (Im & Ras, 2005a). When a system is distributed with autonomous sites, knowledge extracted from local or remote information systems (Im, Ras, & Dardzinska, 2005b) may reveal sensitive data to be protected. Clearly, mechanisms that would protect against these vulnerabilities have to be implemented in order to build a security-aware KDS.

BACKGROUND

Security in KDS has been studied in various disciplines such as cryptography, statistics, and data mining. A well-known security problem in the 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 secure data mining systems. Clifton and Kantarcioglou (2002) employed the concept to association rule mining for vertically and horizontally partitioned data. Du and Zhan (2002) and 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 in that they require expensive encryption and decryption mechanisms. Considering that real-world systems often contain an 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. A 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 (2003) provided a theoretical basis for data perturbation in terms of data utilization and disclosure risks. In the

KDD area, protection of sensitive rules with minimum side effects has been discussed by several researchers. Oliveira and Zaiane (2002) suggested a solution to protecting sensitive association rules in the form of a "sanitization process" where protection is achieved by hiding selective patterns from the frequent itemsets. There has been another interesting proposal for hiding sensitive association rules. Saygin, Verykios, and Elmagarmid (2002) 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 data security algorithms for distributed knowledge sharing systems. Previous and related works concentrated only on a standalone information system or did not consider knowledge sharing techniques to acquire global knowledge.

CHASE ALGORITHM

Suppose that we have an incomplete information system $S = (X,A,V)$ where X is a finite set of objects, A is a finite set of attributes (functions from X into V or relations over X×V), and V is a finite set of attribute values. An 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 if their weights ($\underline{\omega}$) are larger than a minimum threshold value (λ).

Chase requires two input parameters to replaces null or missing values: (1) knowledge and (2) existing data in an information system. The main phase of the Chase algorithm for S is the following:

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

More specifically, suppose that a knowledge base $KB = \{(t \rightarrow v_c) \in D : c \in In(A)\}$ is a set of all rules extracted from S by $ERID(S, \lambda_1 \lambda_2)$, where $In(A)$ is the set of incomplete attributes in S and λ_1 λ_2 are thresholds for minimum support and minimum confidence, correspondingly. *ERID* (Dardzinska & Ras, 2003b) is the algorithm for discovering rules from incomplete information systems and used as a part of Chase. Assuming that $Rs(x_i) \subseteq KB$ is the set of rules in which all of the conditional parts of the rules match with the attribute values in $x_i \in S$, and $d(x_i)$ is a null value, there are three cases for value replacements (Dardzinska & Ras, 2003a, 2003c):

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

Clearly, the weight of predicted value is 1 for case 2. For case 3, the weight is calculated based on the confidence and support of rules used for the prediction (Ras & Dardzinska, 2005b). As defined, any predicted value with $\omega > \lambda$ is considered to be valid value.

Chase is an iterative process. An execution of the algorithm generates a new information system, and the execution is repeated until it reaches a state where no additional null value imputation is possible. Figure 1 shows the number of null value imputations at each execution for a sample data set describing the U.S. congressional voting (Hettich & Merz, 1998). In its first run, 163 null values are replaced. In the second run, 10 more attribute values are filled. The execution stops after the third iteration.

SECURITY PROBLEMS

Suppose that an information system S is part of a distributed knowledge discovery system (DKDS). 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 $S_i \in DKDS$, *for* $S_i \neq S$.

We examine confidential data disclosure for these two cases.

Figure 1. Execution of Chase algorithm

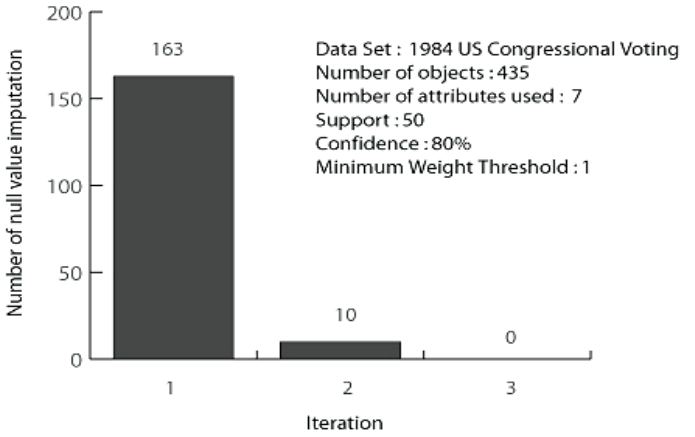

Table 1. Information system S

X	a	b	c	d
x_1	a_1	b_1	c_1	d_3
x_2	a_1	b_2	c_2	d_3
x_3	a_2	b_1	c_2	d_1
x_4	a_2	b_2	c_2	d_1
x_5	a_1	b_1	c_2	d_3
x_6	a_1	b_2	c_1	d_3

Scenario One: Disclosure by Local Knowledge

Consider a simple example that illustrates how locally discovered rules are revealing confidential data. Suppose that we have a set of confidential data v_{conf} = *{d*(x_1*), d*(x_2*), d*(x_3*), d*(x_4*)}* in *S* (see Table 1). To protect v_{conf}, we construct S_d = *(X, A, V)* as shown in Table 2. Now, we extract rules from S_d in terms of *d* using the remaining non-confidential data, as shown in Table 3. Since $r_1 = a_1 \rightarrow d_3$ is applicable to objects in $\{x_1, x_2\}$ by Chase, the execution returns the confidential data d_3 for x_1 and x_2.

It is possible that predicted values are not equal to the actual values. 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. $d_{S_d(x)} \neq d_{S(x)}$.

Clearly, confidential data can be correctly predicted in case 1. We may not need to take any action for case 2 and case 3.

Scenario Two: Disclosure by Distributed Knowledge

The principal of DKDS is that each site develops knowledge independently, and the independent knowledge is used jointly to produce global knowledge (Ras, 1994). The advantage of DKDS is that organizations are able

Table 2. Information system S_d

X	a	b	c	d
x_1	a_1	b_1	c_1	
x_2	a_1	b_2	c_2	
x_3	a_2	b_1	c_2	
x_4	a_2	b_2	c_2	
x_5	a_1	b_1	c_2	d_3
x_6	a_1	b_2	c_1	d_3

Table 3. Rules in KB

rid	rule	support	confidence	Source
r_1	$a_1 \rightarrow d_3$	20%	100%	Local

to share global knowledge without implementing complex data integrations. However, security problems may arise when knowledge is extracted from a remote site, because some of them can be used to reveal confidential data stored in local sites. For example, assume that an attribute *d* in an information system *S* (see Table 4) is confidential, and we hide *d* from *S* and construct S_d = *(X,A,V)* , where:

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

In this scenario, no local rule for *d* is generated because all attribute values in *d* are completely hidden. Instead, rules are extracted from remote sites (e.g., r_1,r_2 in Table 5). Now, disclosure risks are similar to those encountered previously. 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 object $\{x_4\}$. These confidential values can be reconstructed. In addition, we can see that a local rule $r_3 = c_1 \rightarrow a_1$ can be used to reconstruct a_1, which is involved in confidential value reconstruction by r_2. Therefore, we need to consider both local and remote rules.

Since 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 & Dardzinska, 2004a). There are two options:

1. A knowledge base at the local site is kept consistent (all inconsistencies have to be resolved before rules are stored in the knowledge base)
2. A knowledge base at the local site is inconsistent (values of the same attribute used in two rules extracted at different sites may be of different granularity levels and may have different semantics associated with them)

In general, we assume that the information stored in ontology (Guarino, 1995) and, if needed, in inter-ontologies (if they are provided) is sufficient to resolve inconsistencies in semantics of all sites involved in Chase. In other words, any meta-information in distributed information systems is described by one or more ontologies, and the inter-ontology relationships are used as a semantic bridge between autonomous information systems in order to collaborate and understand each other.

PROTECTION METHODS

Minimum Data Loss

Because some degree of data loss is almost inevitable to prevent the data reconstruction by Chase, protection algorithms should limit possible disclosure with the least amount of additional data loss to improve data

Table 4. 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 5. Rules in KB

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

availability. This task is computationally intensive because Chase often builds up reclusive predictions as shown in the previous section. Several algorithms (Im & Ras, 2005a, 2005b) have been proposed to improve performance based on the discovery of Chase closure.

When we design a protection method, one promising approach is to take advantage of hierarchical attribute structure and replace the exact value with more generalized values at a higher level in the hierarchy (Im et al., 2005c). An information system represented in hierarchical attribute structures is a way to make an information system more flexible. Unlike a single-level attribute system, data collected with different granularity levels can be assigned into an information system with their semantic relations. For example, when the age of a person is recorded, the value can be 20 or *young*. In this system, we may show that a person is 'young' instead of '20 years old' if disclosure of the value 'young' does not compromise the privacy of the person. The advantage of this approach is that users will be able to get more explicit answers for their queries.

Minimum Knowledge Loss

Another aspect of loss to be considered is knowledge loss (in the form of rules). Clearly, when we start hiding data from an information system, some knowledge will be lost because some data have to be hidden or generalized. An important issue in developing a strategy for this problem is to determine the set of significant knowledge, because the notion of significance varies considerably from site to site and is often subjective (Silberschatz & Tuzhilin, 1996). For example, some sites value common sense rules, while others are interested in surprising or actionable rules. One way to measure the significance of rules in an objective way is to check their support and confidence. Clearly, certain rules provide better guidance to the overall trends and patterns of the information system than that of possible rules, and rules with higher support also reflect the overall figure of the information system better.

FUTURE TRENDS

Much work remains in the research and development to provide data security against Chase. For example,

there is a trade-off between data security and data availability. As we add more security to confidential data against Chase, we have to hide (or modify) more data from an information system. A method of measuring the trade-off would help knowledge experts and security administrators to determine appropriate security levels for the information system. Another research direction is the study of data security in dynamic information systems. When information has to be updated, confidential data may be revealed by newly discovered rules even if protection has been applied. Integration of ontology into KDS is also important in order to more precisely identify Chase-applicable rules in a distributed knowledge discovery system.

CONCLUSION

The use of Chase may reveal confidential data in an information system. This article discussed disclosure risk and a few different approaches for reducing the risk. In general, additional data and/or knowledge loss is unavoidable to protect confidential data. If the objective is to minimize data loss, we have to hide the minimum set of additional attribute values that will prevent predictions by Chase. If an information system is represented in hierarchical attribute structures, the number of additional data hidings can be further reduced. If the objective is to achieve data confidentiality with minimum knowledge loss, the set of significant knowledge has to be determined before measuring the loss. The objective dimension of the significant rules can be measured by the most commonly used criteria, such as support and confidence of rules.

REFERENCES

Dardzinska, A., & Ras, Z. (2003a). Chasing unknown values in incomplete information systems. *Proceedings of the 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 the 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 2nd International Workshop on Active Mining.*

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

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

Hettich, C.B., & Merz, C. (1998). *UCI repository of machine learning databases.*

Im, S., & Ras, Z. (2005a). 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,* Regina, Canada.

Im, S., Ras, Z., & Dardzinska, A. (2005b). Building a security-aware query answering system based on hierarchical data masking. *Proceedings of the ICDM Workshop on Computational Intelligence in Data Mining,* Houston, TX.

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

Muralidhar, K., & Sarathy, R. (2003). A theoretical basis for perturbation methods. *Statistics and Computing,* 329-335.

Kantarcioglou, M., & 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.*

Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology.* London: Springer-Verlag.

Oliveira, S.R.M., & Zaiane, O.R. (2002). Privacy preserving frequent itemset mining. *Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining* (pp. 43-54).

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

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

Ras, Z. (1994). Dictionaries in a distributed knowledge-based system. *Concurrent engineering: Research and applications.* Pittsburgh, PA: Concurrent Technologies Corporation.

Ras, Z., & Dardzinska, A. (2004a). Ontology based distributed autonomous knowledge systems. *Information Systems International Journal, 29*(1), 47-58.

Ras, Z., & Dardzinska, A. (2004b). Query answering based on collaboration and Chase. *Proceedings of the 6th International Conference on Flexible Query Answering Systems,* Lyon, France.

Ras, Z., & Dardzinska, A. (2005). CHASE-2: Rule based Chase algorithm for information systems of type Lambda. *Proceedings of the 2nd International Workshop on Active Mining,* Maebashi City, Japan.

Ras Z., & Dardzinska, A. (2005b). Data security and null value imputation in distributed information systems. In *Advances in soft computing* (pp. 133-146). Berlin/London: Springer-Verlag.

Ras, Z., & Joshi, S. (1997). Query approximate answering system for an incomplete DKBS. *Fundamenta Informaticae Journal, 30*(3/4), 313-324.

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

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering, 8,* 970-974.

Yao, A.C. (1996). How to generate and exchange secrets. *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science.*

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 a new, more complete value.

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 a new more complete value. Any differences in semantics among attributes in the involved databases have to be resolved first.

Intelligent Query Answering System: Enhancements of a query-answering system into a sort of intelligent system (capable of being adapted or molded). Such systems should be able to interpret incorrectly posed questions and compose an answer not necessarily reflecting precisely what is directly referred to by a question, but rather reflecting what the intermediary understands to be the intention linked with the question.

Knowledge Base: A collection of rules defined as expressions written in predicate calculus. These rules have a form of association 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. The system commits to ontology if its observable actions are consistent with the definitions in the ontology.

Semantics: The meaning of expressions written in some language, as opposed to their syntax, which describes how symbols may be combined independently of their meaning.