基于粗糙集理论的数据库推理控制

张增军[1]，李向阳[2]，肖军模[2]

(1. 空军装备研究院通信所 北京 海淀区 100085； 2. 解放军理工大学通信工程学院 南京 210007)

【摘要】提出了一种基于粗糙集理论的数据库推理泄漏控制方法。该方法采用粗糙集理论技术，分析、提取出数据库中敏感和非敏感数据之间的联系，根据这些联系和各个属性的重要性，在防止敏感数据被推理泄漏的前提下，对数据库系统返回给普通用户的数据动态地做最小修改。实验结果表明该方法可扩展性强，在保证较高的数据可用性的同时提高了数据库的安全性。

关 键 词 数据的可用性； 推理控制； 粗糙集理论； 可扩展性； 安全数据库
中图分类号 TP309 文献标识码 A

Inference Control of Databases Based on Rough Set Theory

ZHANG Zeng-jun[1]，LI Xiang-yang[2]，XIAO Jun-mo[2]

(1. The Communication Institute, Air Force Equipment Academy Haidian Biejing 100085;
2. Institute of Communication Engineering, PLA University of Science and Technology Nanjing 210007)

Abstract This paper describes an approach to controlling inference disclosure of secure relation databases based on Rough Set Theory (RST). We analyze and discover all the relation between non-sensitive and sensitive data with RST. Then according to the importance of these relations and attributes, data of database queried by generic users are modified dynamically and most parsimoniously with preventing sensitive data being inferred from non-sensitive data. Experimental result shows that the approach is scalable and preserves high availability of data while improving security of inference control.

Key words availability of data; inference control; rough set theory; scalability; secure databases

推理泄漏问题是数据库安全方面研究的重点和难点之一，很多文献对此做了研究[1-2]，并认为是个NP-hard问题[2]。随着新的知识发现技术的发展，推理泄漏问题再次面临新的挑战[3-4]。

在数据库中，数据库模式上的约束关系，最终将体现在数据之间的联系上[1]，而且非敏感数据和敏感数据之间通常会存在间接的、不明显的联系。普通权限的用户可能使用这些联系，综合已知的非敏感数据推理得到敏感信息。因此，高级别授权用户如数据库管理员，应该获得并根据这些联系对数据库实施一定的控制，以阻止推理泄漏的发生。实施控制的方法主要包括修改属性(值)的安全级别或普通用户在数据库中可见的数据。但前一种会引起新的推理泄漏通道的产生[2]；后一种需要建立普通用户的数据库视图，缺乏灵活性。数据的可用性是控制方法的主要性能指标之一。目前的一些方法中，都是用各个属性之间的依赖概率来描述数据中的联系[3-4]，并需要很多先验知识，难以准确描述和有效地实施推理泄漏的控制。

本文引入了粗糙集理论(Rough Set Theory，RST)来分析存在于数据库中敏感和非敏感数据之间确定性推理规则，不需要额外的先验知识。依据获得的规则和各个属性对规则重要性的高低，动态地对发布给普通用户的数据做最小的修改，确保阻止推理泄漏的发生，提高了推理控制的灵活性和数据的可用性。

1 粗糙集理论的决策规则获取

粗糙集是一种发现、表示和分析数据内在规律的技术。

定义 1[5] 在粗糙集理论中，一个决策表由$T=(U, A, C, D)$表示。其中U是对象的非空有限集合，称为论域； A是属性的非空集合，即对属性a：$U\rightarrow V_a$，其中V_a被称作属性a的值集；集合$V=\bigcup_{a\in A} V_a$称为属性集A的值区域；$C, D \subseteq A$是属性集A的两个子集，$C\cap D=\Phi$，C为条件属性，D为决策属性。

收稿日期：2005 – 01 – 19
基金项目：国家自然科学基金资助项目(69931040)；江苏省基金资助项目(BK2004015)
作者简介：张增军(1976 –)，男，博士，主要从事网络信息安全、数据库安全方面的研究.

在决策表中，将通过约去过剩的条件属性，利用最少的属性提供同样多的决策信息，以简化决策算法。这个简化过程包括属性的约简和决策规则的约简。在C中找出约简是一个NP-hard问题，但存在快速计算约简的启发性方法[5]。对于决策表的属性约简P可以得到形为$\alpha\rightarrow\beta$的最小决策规则集，其中$\alpha=\bigcap_{a\in P}(a=a(x))$，$\beta=\bigcap_{d\in D}(d=d(x))$，$x\in U$。

定义 2[5]　关于决策规则$\alpha\rightarrow\beta$的约简，是指规则$\alpha'\rightarrow\beta$，使得$\alpha\rightarrow\beta$蕴含$\alpha'\rightarrow\beta$，并且α'具有最小的属性值集合。

命题 1　对于决策表$T=(U, A, C, D)$，由属性和规则的约简得到最小决策规则集(P,Q)中，对于任一规则$\alpha\rightarrow\beta$，用α/a(其中$a\in\alpha$)表示从α中移去$a=a(x)$所剩余的公式，则规则$\alpha/a\rightarrow\beta$在(P,Q)中不成立。

证明　用反证法。假设$\alpha/a\rightarrow\beta$在(P,Q)中成立，则在T中，规则$\alpha\rightarrow\beta$和$\alpha/a\rightarrow\beta$同时成立。因为对于两条规则的前件α和α/a，存在$\alpha\subset\alpha/a$，而后件相等均为β，所以$\alpha\rightarrow\beta$蕴含$\alpha/a\rightarrow\beta$，即条件$a=a(x)$在$\alpha\rightarrow\beta$中是不必要的，根据定义2，α不是最小的，因此规则$\alpha\rightarrow\beta$不是约简的，即$\alpha\rightarrow\beta\notin(P,Q)$，和题设相矛盾，故命题1成立。　　证毕。

利用命题1，将规则中任一条件属性值隐藏，则将使该规则失去决策意义，不能表达任何推理知识。另外，为了衡量一个规则中每个条件属性对规则的重要程度和规则本身的重要程度，需要计算条件属性的重要性和规则的支持度。

定义 3　设$Y\subseteq D$是决策属性，规则$\alpha\rightarrow\beta$中的条件属性$a(a\in P\subseteq C)$对规则的重要性定义为$\mathrm{sig}_a^Y(x)=\mathrm{card}([x]_{\mathrm{ind}(a)}\cap[x]_{\mathrm{ind}(Y)})/\mathrm{card}([x]_{\mathrm{ind}(Y)})$，其中card表示集合的元素数；$x\in U$。

定义 4　规则$\alpha\rightarrow\beta$的支持度：$\mathrm{supp}(\alpha\rightarrow\beta)=\mathrm{card}(\{x|\ x\in U$且$x$满足$\alpha$和$\beta\})/\mathrm{card}(U)$。

2　运用粗糙集理论获取推理规则

本文假定数据库中包含一系列相关的关系表，其中的敏感数据是以属性为单位标记敏感和非敏感信息的。同时整个数据库中已形成了一个全体关系样本表(Universal Relation Paradigm)，它包含数据库中的所有属性。这一过程的实现在文献[6]中做了深入的研究，本文不再赘述。

将数据库的全体关系样本表看成是一个决策表$T_{\mathrm{Td}}=(U, A, C, D)$，其中的所有的属性列组成属性集$A$，每一行代表一个对象，所有的对象组成论域空间$U$。关系表第$x$行第$p$列取值为$p(x)$，作为属性$p$的一个信息值，所有的$p(x)$组成$V$。令$D$表示普通用户不能访问的敏感属性集，且$C=A-D$。

本文利用粗糙集理论技术，获得T_{Td}的最小决策规则集(P, Q)作为推理规则集。由于决策表T_{Td}的属性约简可能存在多个，在获取推理规则集时，需要考虑所有约简产生的最小决策集。算法描述如下：

(1) 指定决策属性$D'(D'\subseteq D)$，令$T_{\mathrm{Td}'}=(U, A, C, D')$；(2) 对$T_{\mathrm{Td}'}$进行属性的约简，即找到使得$\mathrm{ind}(P, D)=\mathrm{ind}(C, D)$成立的最小属性集$P$，约简方案集$\{P_n\}=\{P\}$。若约简$P$不唯一，则将所有约简属性集加入$\{P_n\}$；(3) 若$\{P_n\}=\Phi$，则算法结束；否则，选取约简$P_i\in\{P_n\}$，形成规则集$(P_i,Q_i)$，并令$\{P_n\}=\{P_n\}-\{P_i\}$；(4) 将$(P_i,Q_i)$中的决策规则逐条进行规则的约简，并计算每条规则的支持度。若该规则的支持度$\mathrm{supp}(r)<S_{\mathrm{STH}}$，其中$S_{\mathrm{STH}}$是管理员指定的一个支持度阈值，则从$(P_i,Q_i)$中删除该规则$r$；(5) 将约简后的$(P_i,Q_i)$加入$(P,Q)$中，转(3)。

因为在安全数据库中，敏感信息可以是不同的属性值，也可以是一个关系(即属性之间的联系)。因此，在T_{Td}中利用粗糙集提取推理规则时，决策属性应该是可变的，对不同的决策属性应重复使用该算法。该算法同样可以用于敏感数据是以元素为单位设定的数据库，改为以敏感数据元素所在的属性为决策属性，只取和该元素有关的规则，其余同上即可得推理规则集。

在获得的推理规则中，大部分都对应着现实中的一些知识或经验，有些虽然没有明显的实际意义，但它们是在特定的数据集中所表现出来的，可能蕴含未发掘的必然性，因此这些规则都应加以考虑。可以通过改变S_{STH}的值，忽略部分偶然性的规则，以提高数据的可用性(Availability of Data)。

3　利用推理规则进行推理控制

对于推理规则集(P,Q)中的任意一条规则$\alpha\rightarrow\beta$，其中$a\in\alpha$表示普通用户能直接访问的属性，属于非敏感信息；$d\in\beta$表示只有高授权用户才能访问的属性，属于敏感信息。利用这些规则，普通用户可以根据非敏感信息推理出敏感信息。为了控制此类问题的发生，根据命题1，利用(P,Q)动态地修改发布给普通用户的数据，若记录集满足某一规则，则隐藏该规则某一个条件对应的属性值。但是在有些情况下，隐藏某个属性值后，用户仍能根据规则中其他属性值，以较大可信度推理出敏感数据。为了减小这种不确定推理泄漏的可能性，本文提出了隐藏对规则重要性最大的一个条件属性值的办法。

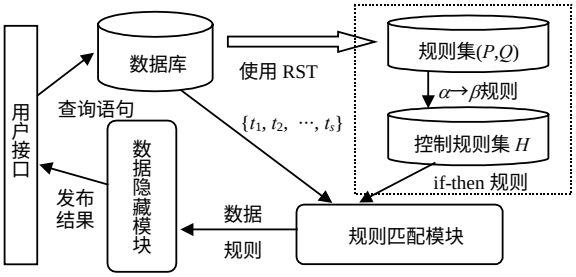

图1 推理控制过程图

如图1所示推理控制过程分为两步：1) 将从数据库中得到的(P,Q)中规则转化为控制规则(由图1中虚线框中的内容表示)；2) 对用户的查询进行动态控制，具体过程如下：

(1) 对于(P,Q)中的每一条推理规则$\alpha\rightarrow\beta$，即$\bigcap_{i=1}^{n}(a_i=a_i(x))\rightarrow\bigcap_{j=1}^{\mathrm{card}(D')}(d_i=d_i(x))$，其中$0<n\leq\mathrm{card}(C)$；$x\in U$；$D'\subseteq D$。① 根据定义3计算每个条件属性$a_i$对规则的重要性$\mathrm{sig}_{a_i}^{D'}(x)$，求$\mathrm{sig}(a)=\max(\mathrm{sig}_{a_i}^{D'}(x))$ $(0<i\leq n)$，表示属性a是条件属性a_i $(0<i\leq n)$中对规则重要性最大的；② 将规则改写为if-then形式的控制规则：if $a_1=a_1(x)$ and $a_2(x)$ and … and $a_n=a_n(x)$ then HIDE a，并加入到控制规则库H中。

这一步属于控制过程的准备阶段，若数据库数据不发生变化，只需进行一次。

(2) 用户发出查询“select Y from R where W”，数据库系统返回元组集$\{t_1,t_2,\cdots,t_s\}$，具有属性集Y，并且满足条件W。① 对所有$t\in\{t_1,t_2,\cdots,t_s\}$检查是否存在控制规则$r_s$ $(r_s\in H)$：if $a_1=a_1(x)$ and … and $a_n=a_n(x)$ then HIDE a，使得t和r_s匹配，且未被执行过。若不存在，则将$\{t_1,t_2,\cdots,t_s\}$发布给用户，控制过程结束；否则，执行②；② 根据r_s，将$a(t)$隐藏，用“—”表示，转①。

4 实验及分析

本文的试验数据库数据采用文献[4]中的实例数据(合并数据库(Combined Database))。粗糙集的分析工具是由挪威大科技大学计算机和信息科学系的知识系统研究小组(Knowledge System Group)开发的软件包ROSSETA。该软件是一个图形界面下的粗糙集分析软件，可计算出决策表的属性约简和最小决策规则集，并给出每条规则的支持度。

同文献[4]，将艾滋病的诊断结果，即属性AIDS作为敏感信息。为便于比较，设阈值参数S_{STH}为0。将属性AIDS作为决策属性，其他属性为条件属性。用ROSSETA可直接生成包含多个约简方案的最小决策规则集合。再按照推理控制过程，计算每个规则中重要性最大的条件属性，生成if-then控制规则集。然后根据这些规则对数据库表进行隐藏修改及计算修改率(数据库数据被隐藏的数量/整个数据库数据的数量)，并将其和文献[4]的方法产生的结果作了比较。另外，为了测试本文算法的性能，在保持文献[4]给出的属性之间的依赖概率的情况下，随机产生了记录，并按照本文的方法，计算出了需要隐藏的数据的数量和修改率。每种情况实验10次，求平均值。实验结果如表1所示。

表1 推理控制方法对数据库数据修改量的比较

	在文献[4]数据库条件下		在不同记录数的情况下/个			
	文献[4]的方法	本文的方法	40	100	300	500
修改的数据量/个	6	4	16.7	39.4	118.6	199.7
修改率	0.067	0.044	0.040	0.044	0.043	0.044

从表1中可以看出本文方法：

(1) 提高了数据库数据的可用性。在相同条件下，本文提出的方法和文献[4]中的方法相比，对数据库数据的修改减少达到了33%。这是由于利用粗糙集理论获得的决策规则集是最小冗余的，而且根据这些规则对用户获得的数据库记录做最小的动态修改，因此需要隐藏的非敏感信息量明显减少，从而提高了数据库数据的可用性。

(2) 提高了数据库数据的安全性。粗糙集理论作为一种分析数据内部关联关系和特征的技术，对于一个决策表，可以获取条件属性对决策属性的所有分类规则[5]。本文将数据库作为决策表，采用粗糙集理论技术进行分析，能够考虑到表中数据所有的约简情况，获取所有非敏感数据和敏感数据之间的确定性决策规则，针对这些规则进行推理泄漏控制，可以极大提高数据库的安全性。

(3) 具有较强的可扩展性。通常数据库所描述的现实对象间的关系是确定，反映在数据中，蕴含的推理关系也是确定的。因此即使数据库中数据大幅度增长，利用本文方法获得的数据间推理规则集也将是相对固定的。可见该方法具有良好可扩展性，可以适应于大规模数据库中推理泄露的控制。但该方法目前仅考虑数据量或属性单维增加情况下的可扩展性，而对于更为复杂的数据量和属性同时增加情况下的可扩展性有待进一步验证。

(下转第537页)

的ARM)进行优化。

4 结 束 语

本文通过前向安全来实现基于身份的非交互式密钥更新，方案简单，只需用户自身来更新密钥，适合于对安全性要求不高的环境。还可通过入侵弹性实现基于身份的非交互式密钥更新，该方案提供更强的安全性，但需要引入与网络无连接的、计算资源和存储资源受到限制的私有设备来帮助更新密钥，需要进一步的研究。

参 考 文 献

[1] SHAMIR A. Identity-based cryptosystems and signature schemes[C]//Proc. of Crypto'84, LNCS 196. Berlin: Springer-Verlag, 1985: 47-53.

[2] BONEH D, FRANKLIN M. Identity-based encryption from the Weil pairing[C]//Proc. of Crypto'01, LNCS 2139. Berlin: Springer-Verlag, 2001: 213-229.

[3] CANNETI R, HALEVI S, KATZ J. A forward secure public key encryption scheme[C]//Proc. of Eurocrypt'03, LNCS 2656.Berlin: Springer-Verlag, 2003: 255-271.

[4] DODIS Y, KATZ J, XU S. Key-insulated public key cryptosystems[C]//Proc. of Eurocrypt'02, LNCS 2332. Berlin: Springer-Verlag, 2002: 65-82.

[5] DODIS Y, FRANKLIN M, KATZ J. Intrusion-resilient public-key encryption[C]//Proc. of CT-RSA'03, LNCS 2612, Berlin: Springer-Verlag, 2003:19-32.

[6] FUJISAKI E, OKAMOTO T. Secure integration of asymmetric and symmetric encryption schemes[C]//Proc. of Crypto'99, LNCS 1666. Berlin: Springer-Verlag, 1999: 537-554.

[7] 毛文波. 现代密码学: 理论与实践[M]. 北京: 电子工业出版社, 2004.

编 辑 黄 莘

(上接第530页)

5 结 论

本文将粗糙集理论引入数据库的推理泄漏控制，可以提取出数据库中非敏感和敏感数据之间蕴含的确定性推理规则。根据这些规则和属性的重要程度，在保证数据库推理泄漏控制的前提下，实现发布给用户数据量修改的最小化。同现有方法相比，该方法计算量小，可扩展性强，在保证大规模数据库可用的同时，增加了数据库的安全性。

参 考 文 献

[1] MARKS D. Inference in MLS database system[J]. IEEE Trans Knowledge and Data Eng, 1996, 8(1): 46-55.

[2] DAWSON S, VIMERCATI S D C. Minimal data upgrading to prevent inference and association attacks[C]// In: Proceedings of the Eighteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Pennsylvania: ACM Press, 1999: 114-125.

[3] SHAFER G. Detecting inference attacks using association rules[J/OL]. http://www. glennshafer. com/courses/ downloads/ raman.pdf, 2004-04-18.

[4] CHANG L, MOSKOWITZ S. A study of inference problem in distributed database systems[C]//In: Proceedings of IFIP Data Security and Applications. Cambridge, UK: Cambridge Uni. Press, 2002: 229-243.

[5] 刘 清. Rough集及Rough推理[M]. 北京: 科学出版社, 2001.

[6] ULLMAN J D. Principle of database and knowledge-base system, Vols. I and II[M]. Rockville, MD: Computer Science Press, 1988,1989.

编 辑 漆 蓉

(上接第533页)

参 考 文 献

[1] DENNING D E. Secure information flow in computer systems[D]. W. Lafayette, Ind.: Purdue Univ., 1975.

[2] DENNING D E. A lattice model of secure information flow[J]. COMM ACM, 1976, 19(5): 236-243.

[3] DENNING D E, DENNING P J. Certification of program for secure information flow[J]. COMM ACM, 1977, 20(7): 504-513.

[4] 肖军模, 刘 军, 周海刚. 网络信息安全[M]. 北京: 机械工业出版社, 2006

[5] 肖军模. 对军用安全模型的扩展[J]. 电子科技大学学报, 2005, 34(2): 186-189.

[6] 江义华. JAVA完美经典[M]. 北京: 中国铁道出版社, 2004.

编 辑 黄 莘