ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, Supplement, December 2009, pp.314−320

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

医疗数据发布中属性顺序敏感的隐私保护方法*

高爱强[1], 刁麓弘[2+]

[1](北京市电力公司,北京 100031)

[2](北京工业大学 应用数理学院,北京 100124)

Privacy Preservation for Attribute Order Sensitive Workload in Medical Data Publishing

GAO Ai-Qiang[1], DIAO Lu-Hong[2+]

[1](Beijing Electric Power Company, Beijing 100031, China)

[2](College of Applied Sciencs, Beijing University of Technology, Beijing 100124, China)

+ Corresponding author: diaoluhong@bjut.edu.cn

Gao AQ, Diao LH. Privacy preservation for attribute order sensitive workload in medical data publishing. *Journal of Software*, 2009,20(Suppl.):314−320. http://www.jos.org.cn/1000-9825/09036.htm

Abstract: Privacy becomes a more serious concern in applications involving microdata such as medical data publishing or medical data mining. Anonymization methods based on global recoding or local recoding or clustering provide privacy protection by guaranteeing that each released record will be indistinguishable to some other individual. However, such methods may not always achieve effective anonymization in terms of analysis workload using the anonymized data. The utility of attributes has not been well considered in the previous methods. This paper studies the problem of utility-based anonymization to concentrate on attributes order sensitive workload, where the order of the attributes is important to the analysis workload. Based on the multidimensional anonymization concept, a method is discussed for attributes order sensitive utility-based anonymization. The performance study using public data sets shows that the efficiency is not affected by the attributes order processing.

Key words: anonymization; privacy preservation; microdata; data analysis; attribute order sensitive

摘 要: 隐私保护已成为包含微数据应用诸如医疗数据发布共享或数据挖掘中的一个重要问题.基于全局重编码或局部重编码的匿名性方法,通过保证每一条数据记录都至少有某个数量的其他记录与其具有同样的特征来保护隐私性.如果考虑到对处理后的数据进行属性顺序敏感的数据分析任务,这类方法并不能很好地完成任务.研究基于数据可用性指标的匿名性方法,着重考虑数据分析任务中的属性顺序对于匿名性方法的影响.从多维数据匿名的概念出发,讨论用于该类情况下的数据匿名性方法.在公开数据集上的实验结果表明,该方法对于上述问题是有效的,并且效率并未受到影响.

关键词: 匿名性;隐私保护;微数据;数据分析;属性顺序敏感

1 数据共享中的隐私性保护问题

为了有效进行数据分析、知识发现等任务,组织需要发布其内部的数据供研究部门使用.在数据发布过程中,可能会牵扯到个人信息的共享.例如,医院发布病人的诊断记录供研究人员对各种疾病的特性、发生人群和年龄等进行研究.原始的数据,也叫作微数据,包含个人的标识信息(如姓名等),这些信息出于隐私保护的目的是

* Received 2009-05-03; Accepted 2009-09-30

不允许共享的.虽然如此,但是依然存在着这样的一组属性,能够与外部的某个数据库进行关联结合从而发现个人的标识信息.在表 1 所示的医院信息表中,并不显式地给出病人姓名.然而,如果某个攻击者能够获取如表 2 所示的投票注册列表,那么他就能够通过在两个表的属性集合{Age,Sex,Zipcode}上进行两个表的连接,以获得所有病人的身份信息,从而能够将病人与疾病联系到一起,泄漏病人关于其所患疾病的隐私信息.这样的一组属性,一般称作伪标识属性.

Table 1 An example of microdata

表 1 数据发布匿名性的一个例子(a)微数据(microdata)

Row#	Age	Sex	Zipcode	Disease
1 (Andy)	5	M	12 000	Gastric ulcer
2 (Bill)	9	M	14 000	Dyspepsia
3 (Ken)	6	M	18 000	Pneumonia
4 (Nash)	8	M	19 000	Bronchitis
5 (Joe)	12	M	22 000	Pneumonia
6 (Sam)	19	M	24 000	Pneumonia
7 (Linda)	21	F	58 000	Flu
8 (Jame)	26	F	36 000	Gastritis
9 (Sarah)	28	F	37 000	Pneumonia
10 (May)	56	F	33 000	Flu

Table 2 Voter registration list

表 2 数据发布匿名性的一个例子(b)投票注册表(voter registration list)

Name	Age	Sex	Zipcode
Andy	5	M	12 000
Bill	9	M	14 000
Ken	6	M	18 000
Nash	8	M	19 000
Mike	7	M	17 000
Joe	12	M	22 000
Sam	19	M	24 000
Linda	21	F	58 000
Jame	26	F	36 000
Sarah	28	F	37 000
May	56	F	33 000

基于匿名性原则的隐私保护是在向分析方发布数据时,通过数据隐藏、数据删除、数据泛化等策略保护数据所有者的个人身份和隐私信息,本报告主要围绕海量数据共享发布中的匿名性隐私保护问题进行研究.基于匿名性原则的隐私保护方法主要包括数据泛化和数据隐藏两大类,通过提升数据的粒度或者隐藏某些数据达到数据所有者信息匿名的目的.如较早提出的个体k匿名[1,2],通过把个体数据的伪标识属性进行粒度提升,使多个数据记录的伪标识属性在发布的数据集中一致,这样可以使具有相同伪标识属性的个体具有不可区分性,从而达到一定的匿名性保护.

匿名性隐私保护要保证充分的匿名性,即要保证一定的匿名性保护强度,从个体k匿名提出至今,这方面的研究得到了广泛的关注.匿名性是通过定义数据集要达到的保护强度和匿名性强度来定义,比如k-anonymity[1,2],l-diversity[3],t-closeness[4],(*alpha*,k)-anonymity[5],Anatomy[6],m-Invariance[7],(ε,m)-anonymity[8]等.

在进行数据匿名性处理之后,原来的数据精度和粒度都发生了改变,不可避免的会造成信息的丢失.为了保证处理后数据的数据质量以及数据分析和数据挖掘的质量,必须在数据处理的过程中尽可能的减少信息损失.在这方面的研究有discernability模型[9]等,在进行数据处理的过程中在保证k匿名性的隐私保护强度的前提下尽可能地减少相同元组的个数.

在本文中,我们针对在数据处理和共享后的数据分析任务对数据属性顺序敏感这样的问题进行研究,考虑数据可用性为指导下的匿名性数据处理方法.基于多维数据匿名处理的概念,讨论了一种方法对属性顺序敏感的数据分析任务场景下的匿名性数据共享问题进行处理.

本文第 2 节介绍一些数据发布和共享中的隐私保护概念.第 3 节讨论处理属性顺序敏感的数据分析任务场景下的匿名性隐私保护方法,并给出初步的实验结果.第 4 节对数据发布和共享中的隐私保护方面的相关工作

进行分析.第 5 节总结本文工作并讨论将来工作的方向.

2 相关概念

定义 1(伪标识属性(QID)). QID 属性集合 Q 是表 T 中具有如下特性的最小属性集合:这些属性能够与外部信息进行连接从而推测个体记录的标识属性及个体具有的敏感隐私属性.

设T为一个存储个人隐私信息的关系表,在关系T中的属性分成 4 类:(1) 标识属性A_i,唯一地标识一个个体,这样的属性必须在数据发布时从数据集中去掉;(2) 敏感属性A_s,比如表 1 中的疾病属性,是必须为个体保护的隐私信息;(3) QID属性,$A_{q1},\ldots,A_{qi}$,能够进行显式共享,但是可能会通过与其他外部数据表(见表 2)进行连接从而泄漏隐私信息(如表 1 中的{Age,Sex,Zipcode});(4) 其他非相关属性.

给定表T,一组属性$\{A_i,...,A_j\}$和元组$t\in T$,$t[A_i,...,A_j]$表示元组t在属性$A_i,...,A_j$上的取值,$T[A_i,...,A_j]$表示表T在属性$A_i,...,A_j$上的投影,维持在这组属性上投影的重复取值.$|T|$表示表T的势,即T中元组的个数.

在这里,我们要求敏感属性As为标量属性,其他的属性可以为数量或者标量.所有的属性都具有有限域.根据相关工作[3]所述,假定所有的标量属性A都有一个分类层次结构,表示属性A所有可能值的公知的层次关系.

为了进行隐私保护,防止上述连接的发生从而泄漏个人隐私信息,文献[1,2]等工作提出了 K 匿名性的概念,即通过一定的数据处理使发布后的数据分组中某个记录的 QID 属性至少与其他 k–1 条记录相同,从而使这 K 条记录达到一定的匿名保护目的.具有相同 QID 属性的数据分组是发布后数据的一个等价类(equivalence class),这样的一组等价类构成了处理后数据 T'的一个划分.

定义 2(等价类). 表T包含元组的多重集.表T相对于属性$A_1,\ldots,A_d$的一个等价类是表T在$(A_1\ldots A_d)$上具有相同取值的一组元组.在SQL语句中,这类似于在属性集$A_1\ldots A_d$上进行分组(GROUP BY)操作.

定义 3(k匿名). 表$T(A_1,...,A_n)$,QI为表T的一个QID属性,当且仅当$T[QI]$中的每一个值序列在$T[QI]$至少出现k次,称表T对于QI满足k匿名性.即,表T相对于属性$A_1\ldots A_d$的每一个等价类的大小至少为k.

K 匿名化即通过一定的处理使得表 T'具有 k 匿名性.

为了达到 K 匿名保护的目的,在相关研究工作中提出了许多方法,如泛化(generalization)、压制(suppression)等,如泛化的方法通过将 QID 属性的值采用其在分类树中更高级别和更抽象的值代替,使得只要 T 中元组在该属性上的取值包含在某个分类树结点下,这些元组在处理后的数据集 T'就处于同一个划分中.如表 3 就是对表 1 进行处理之后得到的一个 2-匿名数据表.

Table 3 2-anonymous table

表 3 数据发布匿名性的一个例子(c)2-匿名表(2-anonymous table)

Row#	Age	Sex	Zipcode	Disease
1	[1,10]	M	[10001,15000]	Gastric ulcer
2	[1,10]	M	[10001,15000]	Dyspepsia
3	[1,10]	M	[15001,20000]	Pneumonia
4	[1,10]	M	[15000,20000]	Bronchitis
5	[11,20]	M	[20001,25000]	Pneumonia
6	[11,20]	M	[20001,25000]	Pneumonia
7	[21,60]	F	[30000,60000]	Flu
8	[21,60]	F	[30000,60000]	Gastritis
9	[21,60]	F	[30000,60000]	Pneumonia
10	[21,60]	F	[30000,60000]	Flu

对于原始数据表 T,所要进行的处理就是 k-匿名化.

经过处理后的数据集 T'达到一定的隐私保护强度,保证某个个体处于 k 个同类别的个体集合之中,从而避免以较高的概率被识别.另一方面,如果处于同一个等价类中的 k 个元组在敏感属性上的取值相同,则其个体敏感信息被推理的概率依然是 100%.为了处理这种情况,文献[3]的作者提出了 l-diversity 的方法,即在一个分组中的敏感属性的取值至少有 l 个.沿着该方向进行研究的相关工作,如文献[4–8]都是对隐私保护策略进行研究.

对原始处理进行处理之后,虽然能够满足一定的隐私保护策略,但是由于在此过程中对数据进行了泛化或者某些具体值的隐藏,因此发布后的数据集 T'对于数据分析或者数据挖掘任务就有一定的失真,即损失一部分

数据可用性.隐私保护强度与数据可用性是一对矛盾的目标,达到较高的隐私保护强度只能得到较低的可用性,而为了较高的数据可用性其隐私保护强度就必然降低.发布原始数据集T对数据可用性的损失为0,但是其保护强度也最弱.因此数据可用性评价指标对于隐私保护策略和计算方法具有较大的影响.

目前较常用的数据可用性评价指标包括可区别性补偿、确定性补偿、KL差分等.在这里,设T为一个k匿名表,具有d个QID属性$A_1…A_d$,表T有n个划分或者叫n个等价类$P_1,…,P_n$.

定义 4(可区别性补偿). 表 T 的可区别性补偿:$DM(T)=\sum_{i=1}^{n}|P_i|^2$.

可区别性补偿测度为T中的每一个元组t,基于其所属的划分P_i分配了一个补偿值.

定义 5(确定性补偿). 表T的确定性补偿:$CM(T)=\sum_{t\in T}NCP(t)$,其中$NCP(t)$是为元组$t$分配的加权归一化确定性补偿.$NCP(t)=\sum_{i=1}^{d}\left(w_i\times\frac{|t.A_i|}{|T.A_i|}\right)$,$w_i$是QID属性$A_i$的权重,反映其在表匿名化过程中的重要性.如果$A_i$是数量属性,则$|t.A_i|$表示元组$t$在属性$A_i$上进行泛化后的区间,例如,如果$t.Age$=[20–30],则$|t.A_i|$=10.$|T.A_i|$是$T$中所有元组在属性$A_i$上的区间.如果$A_i$是标量属性,假设在属性$A_i$上存在分类树$H$,则$|t.A_i|$是分类树$H$中以$t.A_i$泛化后的值为根的子树种叶子节点的个数.$|T.A_i|$是整个树$H$中的叶子节点总数.

定义 6(KL-差分). $KL(T)=\sum_{t\in T}p_t^{(1)}\log\frac{p_t^{(1)}}{p_t^{(2)}}$,其中$p_t^{(1)}$和$p_t^{(2)}$分别表示元组$t$在原表和匿名表中的概率.

3 属性顺序敏感任务的隐私保护方法

在表 1~表 3 中的数据经常被用来讨论匿名化数据来进行隐私保护[10].数据发布后的数据分析任务包括多维分析和数据挖掘等,可能的数据挖掘任务包括关联规则分析、数据分类、聚类分析等,在文献[11]的研究工作中,其数据匿名化方法和数据可用性测度针对数据分类分析进行,在其进行匿名化的数据中包含了数据元组的类属性.在文献[6–8,10]的研究工作中,主要针对数据发布后的查询分析和各种聚集函数的分析,对于QID属性组中属性的前后顺序要求不明确.

如果经过匿名化处理之后的数据用来进行关联规则分析,则 QID 属性组中的属性顺序必须被考虑在内.假设现在有表 4 这样的一组数据,如果进行匿名化处理时按照先年龄后性别的处理顺序,则在关联规则分析中得到性别与某种疾病的关联规则的概率就会降低;同时,如果按照先性别后年龄的处理顺序,则在关联规则分析中得到年龄与某种疾病的关联规则的概率就会降低.

Table 4 Attribute order sensitive

表 4 属性顺序影响

Age	Sex	Value	处理后		
			Age	Sex	Value
[50~55]	F	A	[50~60]	F	A,C
[50~55]	M	B	[50~60]	M	B,D
[55~60]	F	C	[50~55]	*	A,B
[55~60]	M	D	[55~60]	*	C,D

如果数据分析任务对于属性顺序敏感,则进行数据匿名化的方法就必须考虑 QID 属性组中所有属性进行泛化处理的先后顺序.

下面给出基于文献[12]的 Mondrian 算法的面向 QID 属性顺序敏感的数据分析任务的数据匿名化方法,记为 AOS 算法,如图 1 所示.

在选择进行划分的维时,可以采取不同的启发式方法.在文献[12]中,作者根据值域最宽的方法进行选择.在这里,由于考虑到数据分析任务对于属性顺序敏感,采取预定义的属性优先顺序进行维的选择.

采用预定义属性优先顺序来进行待划分维的选择,不仅能够满足某些数据分析和数据挖掘任务对属性顺序敏感的要求,而且能够减少响应时间.根据文献[12]所采用的实验数据集以及文献[7,13]的参考代码实现,得到

了如图 2 和图 3 的一组实验结果,图 2 是随数据集大小变化时执行时间的变化,图 3 是执行时间随数据集中 QID 属性个数变化的情况.

```
Anonymize(partition)
    if (no allowable multidimensional cut for partition)
    return ϕ: partition→summary
    else
    dim←choose_dimension();
    fs←frequency_set(partition,dim)
    splitVal←find_median(fs)
    lhs←{t∈partition:t.dim≤splitVal}
    rhs←{{t∈partition:t.dim>splitVal}
    return Anonymize(rhs)∪Anonymize(lhs)
```

Fig.1 Algorithm AOS

图 1 AOS 算法

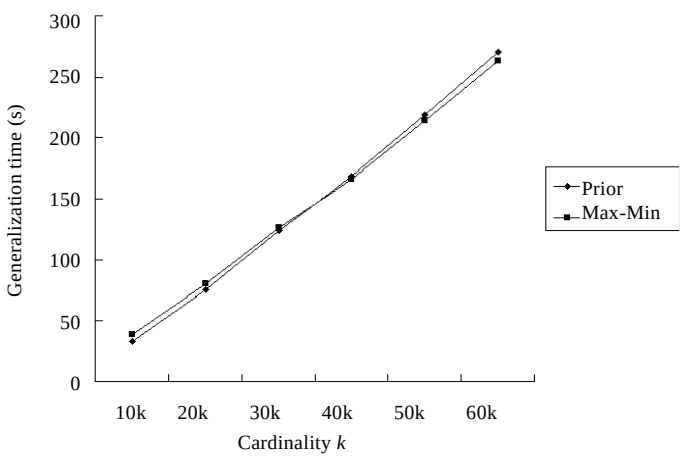

Fig.2 Computing time vs. the size of dataset

图 2 执行时间随数据集大小的变化

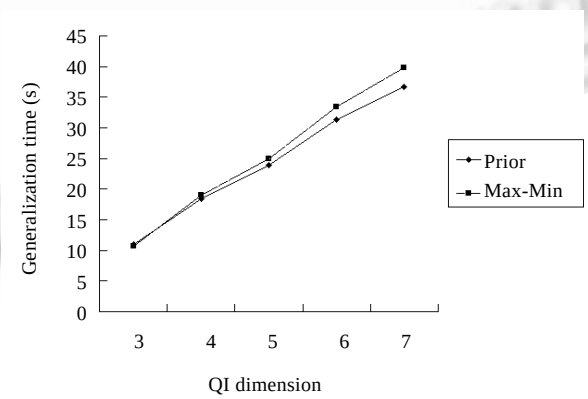

Fig.3 Computing time vs. the size of QID

图 3 执行时间随 QID 属性个数变化的情况

4 相关工作

为公共目的发布微数据进行科研的问题在统计学和计算机科学的研究领域中都已经有了较多的关注,文

献[14]给出了较为详尽的综述.在统计学研究领域中,主要关注识别和保护敏感的属性,主要有数据交换和数据隐藏两类方法.在计算机科学近几年的研究中,*k*-匿名[2]方法得到了很多的关注,*k*-匿名是通过在发布后数据的每一个分组中保持至少*k*个元组在一组属性上具有相同的取值,而这组属性恰好是能够通过与某些外部数据库进行连接得到隐私属性的一组属性.通过这种方法就保证了每一个元组至少有*k*个相同体,提高了隐私保护的强度.在文献[15]中*k*-匿名被证明为NP-hard问题.在*k*-匿名方法的研究中,主要是设计一些隐私保护的原则来提供保护性,文献[1,2]和文献[3]的方法是得到关注较多的两种方法.在文献[4]中提出的*t*-closeness方法要求每一个等价类内敏感属性(即受保护的属性,比如收入、疾病等)的取值的数据分布与整个数据集的数据分布相似.文献[16]讨论了基于聚类的两种方法,文献[10]探讨了个性化的隐私保护方法,其中每一个个体都有其特殊的隐私保护强度.

另一类工作则是在某个隐私保护原则下进行匿名化表的计算方法的研究,在计算匿名化之后的数据时要保证某些预定义的数据质量损失指标最小化(如果是可用性指标,则要最大化).如果对处理后的泛化表作一定的限制,可以对所有可能的泛化关系进行枚举并得到最优[17].在这种情况下,最优的结果可以通过一些启发式规则[9,18]对整个搜索空间进行裁剪.也有一些文献[11,12]对贪心算法进行探讨来得到次优解,不过速度却会更快一些.

如何在不影响隐私保护强度的前提下提高发布后数据的数据质量也是重点关注的研究内容之一.在文献[18]的工作中,作者提出的方法将微数据表投影到等价类属性和敏感属性的一个子集上.文献[6]提出的 anatomy 方法把等价类属性和敏感属性分别发布到两个表中,这样保证了匿名性,不必再对等价类属性进行泛化和匿名化处理.

5 结束语

本文讨论了基于数据可用性的隐私保护匿名方法,主要讨论数据分析任务如果对属性顺序敏感环境下的数据处理方法.基于多维数据匿名化概念,讨论了一种方法来进行基于数据可用性的数据发布共享和匿名性处理方法.简单的实验分析显示该方法在保证数据可用性的同时,性能并不受到影响.在下一步的工作中,需要进行更多的实验分析.同时,数据发布和匿名处理过程中的背景知识也是研究的重点.

References:

[1] Samarati P. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Engineering (TKDE), 2001,13(6):1010−1027.

[2] Sweeney L. *k*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness, and Knowlege-Based Systems, 2002,10(5):557−570.

[3] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond *k*-anonymity. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the Int'l Conf. on Data Engineering (ICDE). Atlanta: IEEE Computer Society, 2006. 24−35.

[4] Li NH, Li TC, Venkatasubramanian S. *t*-Closeness: Privacy beyond *k*-anonymity and l-diversity. In: Körpeoglu I, ed. Proc. of the Int'l Conf. on Data Engineering (ICDE). Istanbul: IEEE Computer Society, 2007. 106−115.

[5] Wong RCW, Li JY, Fu AWC, Wang K. (*alpha*,*k*)-Anonymity: An enhanced *k*-anonymity model for privacy preserving data publishing. In: Crissey M, ed. Proc. of the ACM Knowledge Discovery and Data Mining (SIGKDD). Philadelphia: ACM Press, 2006. 754−759.

[6] Xiao XK,Tao YF. Anatomy: Simple and effective privacy preservation. In: Dayal U, Whang KY, Lomet DB, Alonso G, Lohman GM, Kersten ML, Cha SK, Kim YK, eds. Proc. of the 32nd Int'l Conf. on Very Large Data Bases. Seoul: ACM Press, 2006. 139−150.

[7] Xiao XK, Tao YF. *m*-Invariance: Towards privacy preserving re-publication of dynamic datasets. In: Chan CY, Ooi BC, Zhou AY, eds. Proc. of the ACM Conf. on Management of Data (SIGMOD). Beijing: ACM Press, 2007. 689−700.

[8] Li JX, Tao YF, Xiao XK. Preservation of proximity privacy in publishing numerical sensitive data. In: Tsong J, Wang L, eds. Proc. of the ACM Conf. on Management of Data (SIGMOD). Vancouver: ACM Press, 2008. 473−486.

[9] Bayardo RJ, Agrawal R. Data privacy through optimal *k*-anonymization. In: Kitagawa H, Ishikawa Y, Morishima A, Takayama T, eds. Proc. of the Int'l Conf. on Data Engineering (ICDE). Tokyo: IEEE Computer Society, 2005. 217−228.

[10] Xiao XK, Tao YF. Personalized privacy preservation. In: Chaudhuri S, Hristidis V, Polyzotis N, eds. Proc. of the ACM Conf. on Management of Data (SIGMOD). Chicago: ACM Press, 2006. 229−240.

[11] Fung BCM, Wang K, Yu PS. Top-Down specialization for information and privacy preservation. In: Kitagawa H, Ishikawa Y, Morishima A, Takayama T, eds. Proc. of the Int'l Conf. on Data Engineering (ICDE). Tokyo: IEEE Computer Society, 2005. 205−216.

[12] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional *k*-anonymity. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the Int'l Conf. on Data Engineering (ICDE). Atlanta: IEEE Computer Society, 2006. 277−286.

[13] Xiao XK, Tao YF. Dynamic anonymization: Accurate statistical analysis with privacy preservation. In: Tsong J, Wang L, eds. Proc. of the ACM Conf. on Management of Data (SIGMOD). Vancouver: ACM Press, 2008. 107−120.

[14] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009,32(5):847−861 (in Chinese with English abstract).

[15] Meyerson A, Williams R. On the complexity of optimal *k*-anonymity. In: Deutsch A, ed. Proc. of the ACM Symp. on Principles of Database Systems (PODS). Paris: ACM Press, 2004. 223−228.

[16] Aggarwal G, Feder T, Kenthapadi K, Khuller S, Panigrahy R, Thomas D, Zhu A. Achieving anonymity via clustering. In: Vansummeren S, ed. Proc. of the ACM Symp. on Principles of Database Systems (PODS). Chicago: ACM Press, 2006. 153−162.

[17] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain *k*-anonymity. In: Özcan F, ed. Proc. of the ACM Conf. on Management of Data (SIGMOD). Baltimore: ACM Press, 2005. 49−60.

[18] Kifer D, Gehrke J. Injecting utility into anonymized datasets. In: Chaudhuri S, Hristidis V, Polyzotis N, eds. Proc. of the ACM Conf. on Management of Data (SIGMOD). Chicago: ACM Press, 2006. 217−228.

附中文参考文献:

[14] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847−861.

高爱强(1978－),男,山东海阳人,博士,高级工程师,主要研究领域为数据库系统理论,Web 服务,面向服务的架构.

刁麓弘(1978－),男,博士,讲师,主要研究领域为人工智能,模式识别,计算机视觉,计算机图形学.