基于属性相关性划分的多敏感属性隐私保护方法

谢 静,张健沛,杨 静,张 冰
(哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 近年来,基于 l-多样性的多维敏感属性的隐私保护研究日趋增多,然而大部分多敏感属性隐私保护方法都是基于有损分解的思想,破坏了数据间的关系,降低了数据效用.为此,提出了一种面向多敏感属性的隐私模型,首先给出一种 l-maximum 原则用以满足多敏感属性 l-多样性要求;其次,为了保护属性间的相关性,根据属性间的依赖度对属性进行划分;最后设计并实现了 MSA l-maximum(Multiple Sensitive Attributes l-maximum)算法.实验结果表明,提出的模型在保护隐私不泄露的同时,减少了元组的隐匿率,并且保护了数据间的关系.

关键词: 隐私保护; 多敏感属性; l-多样性; 属性相关性; 划分
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2014)09-1718-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.09.009

A Privacy Preserving Approach Based on Attributes Correlation Partition for Multiple Sensitive Attributes

XIE Jing, ZHANG Jian-pei, YANG Jing, ZHANG Bing
(*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: In recent years, l-diversity models are suitable not only for single sensitive attribute data tables, but also for multiple sensitive attributes data tables. However, most of the research is based on lossy join, it breaks the relationship between data. To address these problems, a model based on multiple sensitive attributes is proposed. The main idea of the model is that it proposes a l-maximum principle that can satisfy the multiple sensitive attributes l-diversity at first. Then, to protect the relationship between data, the model partitions attributes by the dependency degree between attributes. Finally, a multiple sensitive attributes l-maximum algorithm(MSA l-maximum)is proposed. The experiment results show that the proposed model can preserve the security of sensitive data, meanwhile it can also reduce the information hidden rate and keep a high data utility.

Key words: privacy preserving; multiple sensitive attributes; l-diversity; attributes correlation; partition

1 引言

信息化时代中,各个领域需要收集和分析的个人数据正以惊人的速度增长,这些原始形式的数据是公共资源分配,医学探索,趋势预测等研究的宝贵来源.但是,在进行数据研究的同时,往往会带来个人信息的泄露,为了保证个人数据被使用的同时不泄露机密信息,需要提出有效的隐私保护技术.对于个人数据,将某些可以唯一确定个体身份的属性,如姓名、身份证号等,称为身份标识符属性(identifier);然后,将通过组合可以确定出个体身份的属性称为准标识符属性(Quasi-Identifiers, QI),如生日、年龄、性别等;包含个人敏感信息的属性称为敏感属性(Sensitive Attributes, SA).

近几年,数据发布中的隐私保护技术受到越来越多的关注,相关研究成果也日趋增多[1~6].然而,目前的隐私保护模型大部分只能处理单维敏感属性,当攻击者在多维敏感属性的基础上进行推理披露时,数据发布者在发布多维敏感属性数据之前,应该考虑多维敏感属性带来的所有可能情况.因为多维敏感属性之间的关系、QI属性和 SA 属性之间的关系会带来更多攻击的可能性,因此直接将上述模型应用到多维敏感属性上不能保证信息的安全,为此研究者们相继提出了多维敏感属性的隐私保护方法[7~11].

本文在前人的工作基础上,研究数据发布中多维敏感属性的隐私保护问题.首先,提出了一种面向多维敏感属性的 l-maximum 原则;然后根据属性间的依赖度将属性进行划分;最后提出了一种 MSA l-maximum 算法,在满足隐私需求的同时采用属性聚类划分来保护相关性高的 SA 属性与 QI 属性之间的关系.

收稿日期:2013-07-29;修回日期:2014-01-24;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61370083, No.61073043, No.61073041);高等学校博士学科点专项科研基金(No.20112304110011, No.20122304110012);哈尔滨市科技创新人才研究专项资金(优秀学科带头人)(No.2011RFXXG015)

2　相关工作

现有的 k-匿名和 l-多样性以及它们的扩展模型大部分都是针对单维敏感属性进行研究的,这些模型不能直接适用于多维敏感属性的隐私保护问题.因此,针对多维敏感属性的问题,Yang 等人[7]首次对多维敏感属性的隐私保护问题进行研究,它采用有损连接的思想,提出了一种基于多维桶的分组技术,并且针对数据的重要性差异,提出了加权多维桶分组技术.然而,文献[7]中没有考虑相关敏感属性的性质,产生了一些不必要的信息损失.Yang 等人[10]利用拓扑空间中的覆盖思想,提出了一种针对多敏感属性的 l-多样性原则,要求等价类中的最小类覆盖 $\varphi_{\min}$ 满足 $|\varphi_{\min}|\geq l$,并且证明了该原则满足多敏感属性 l-多样性.文献[11]在前人工作的基础上,对多敏感属性的数据共享问题进行研究,提出一种基于属性分类的多敏感属性隐私保护模型.上述的方法都是基于有损连接的技术,虽然无需泛化,保护了数据的信息,但是有损连接要求 QI 属性和 SA 属性分开发布,破坏了 QI 属性和 SA 属性之间的关系,并且假设某一维的 SA 属性值泄露将会导致其他维的 SA 属性值随之泄露,不利于数据的隐私保护.

3　多维敏感属性 *l*-maximum 原则

假设用户要发布的数据表 $T=\{Q_1,\cdots,Q_i,\cdots,Q_n,S_1,\cdots,S_j,\cdots,S_m\}$,其中 $Q_i(1\leqslant i\leqslant n)$ 为 QI 属性,$SA_j(1\leqslant j\leqslant m)$ 为 SA 属性.对于 T 中的每个元组既包含 QI 属性又包含 SA 属性,也就是说,每条元组都包含敏感信息.

文献[10]引入拓扑空间中覆盖的思想来定义多敏感属性 l-多样性,提出了类覆盖和最小类覆盖的概念[10].文献[10]中要求等价类中的最小类覆盖 $\varphi_{\min}$ 满足 $|\varphi_{\min}|\geq l$,但是其已经证明寻求最小类覆盖集的问题是 NP-难问题,因此,最小类覆盖的计算过程比较复杂,特别是在等价类中元组个数较多的情况下.另外,最小类覆盖的约束只考虑了敏感值,没考虑敏感值出现的次数.

针对该问题,本文提出一种 l-maximum 原则用以多维敏感属性的隐私保护.

定义 1　多维敏感属性 *l*-maximum 原则　给定数据表 T,令 E 为 T 的一个等价类,$|E|$ 表示 E 中元组个数,$N=\{n_1,n_2,\cdots,n_k\}$,N 表示等价类中所有敏感值出现次数的降序排列.如果 N 中前 l 个元素之和不大于 $|E|$,即 $\sum_{i=1}^{l}n_i\leqslant|E|$,则称 E 满足 l-maximum 原则;如果表 T 中的每个等价类都满足 l-maximum 原则,则称表 T 满足 l-maximum 原则.

定理 1　给定数据表 T,T 中包含多维敏感属性,等价类 E 中的最小类覆盖为 $\varphi_{\min}$,如果 E 满足 l-maximum 原则,则 $|\varphi_{\min}|\geq l$.

证明　假设 E 中的最小类覆盖 $|\varphi_{\min}|<l$,令 $|\varphi_{\min}|=k$,$k<l$,因此,$\varphi_{\min}$ 中所包含的 k 个敏感值出现次数之和必大于或者等于 E 中元组个数,即 $\sum_{i=1}^{k}n_i\geq|E|$,n_i 表示 k 个敏感值中第 i 个敏感值在 E 中出现的次数,这与 l-maximum 原则相矛盾,故原假设不成立.因此,定理成立.

定理 2　给定数据表 T,T 中包含多维敏感属性,如果等价类 E 满足 l-maximum 原则,则 E 中至少有 l 个元组的敏感值互不相同.

证明　假设 E 中至多有 k 个元组不存在相同的敏感值,$k<l$,那么剩余的元组都存在与该 k 个元组相同的敏感值,可以知道 E 存在 k 个敏感值的最小类覆盖,即 $|\varphi_{\min}|=k<l$,由定理 1 的逆否命题可知,等价类 E 不满足 l-maximum 原则,所以原假设不成立,即等价类 E 中至少有 l 个元组不存在相同的敏感值.

4　基于聚类的属性划分方法

目前,大部分多维敏感属性的隐私保护方法都是采用有损连接的思想,有损连接技术没有改变原始数据,查询精度高,但是 QI 属性和 SA 属性之间的关系被破坏,降低了数据效用.针对该问题,本节采用聚类的思想将属性进行分割,依赖度高的属性放入一列,这样只是破坏了依赖度较低属性间的关系,保护了依赖度高的属性间的关系,提高了数据效用.依赖度根据属性间的平均互信息进行计算,下面给出相关定义.

设 A 和 B 是数据表 T 中的任意两维属性,属性 A 的值域为 $\{a_1,\cdots a_i,\cdots,a_m\}$,$1\leqslant i\leqslant m$,属性 B 的值域为 $\{b_1,\cdots,b_j,\cdots,b_n\}$,$1\leqslant j\leqslant n$.

定义 2　互信息　对于数据表 T 中的属性 A 和 B,属性 B 取值为 b_j 时,属性 A 取值为 a_i 的互信息定义为:

$$I(a_i;b_j)=\log\frac{p(a_i|b_j)}{p(a_i)}$$

其中 $p(a_i|b_j)=d_{ij}/D_j$,d_{ij} 表示 T 中 $t[A]=a_i$ 且 $t[B]=b_j$ 的元组个数,D_j 表示 T 中 $t[B]=b_j$ 的元组个数,$p(a_i)=d_i/|T|$,d_i 表示 T 中 $t[A]=a_i$ 的元组个数,$|T|$ 表示数据表 T 中元组个数.

定义 3　平均条件互信息　对于数据表 T 中的属性 A 和 B,属性 B 取值为 b_j 时,属性 A 的平均条件互信息定义为:

$$I(A;b_j)=\sum_{i=1}^{n}p(a_i|b_j)I(a_i;b_j)$$

定义 4　平均互信息　对于数据表 T 中的属性 A

和 B,属性 B 对于属性 A 的条件互信息,即依赖度为:

$$I(A;B) = \sum_{j=1}^{m} p(b_j) I(A;b_j)$$

定理 3 $I(A;B) = I(B;A)$,且 $I(A;B) \geqslant 0$.

证明

$$\begin{aligned} I(A;B) &= \sum_{j=1}^{m} p(b_j) \sum_{i=1}^{n} p(a_i \mid b_j) \log \frac{p(a_i \mid b_j)}{p(a_i)} \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} p(b_j) \frac{p(a_i;b_j)}{p(b_j)} \log \frac{p(a_i;b_j)}{p(a_i)p(b_j)} \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} p(a_i;b_j) \log \frac{p(a_i;b_j)}{p(a_i)p(b_j)} \end{aligned}$$

同理推导可知,

$$I(B;A) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(a_i;b_j) \log \frac{p(a_i;b_j)}{p(a_i)p(b_j)}$$

可以得出,$I(A;B) = I(B;A)$.

由平均互信息的公式可知,

$$-I(A;B) = \sum_{j=1}^{m} \sum_{i=1}^{n} p(a_i;b_j) \log \frac{p(a_i)p(b_j)}{p(a_i;b_j)}$$

由不等式 $\ln w \leqslant w-1$ 和关系式 $\log w = \ln w \log e$

$$\begin{aligned} -I(A;B) &\leqslant \sum_{j=1}^{m} \sum_{i=1}^{n} p(a_i;b_j) \left(\frac{p(a_i)p(b_j)}{p(a_i;b_j)} - 1\right) \log e \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} (p(a_i)p(b_j) - p(a_i;b_j)) \log e \leqslant 0 \end{aligned}$$

由推导可得,$I(A;B) \geqslant 0$,证毕.

在本文的聚类算法中,将数据表中的每个属性看作是聚类空间中的点,空间中两个属性 A,B 之间的距离定义为 $1/I(A;B)$,即属性 A 和 B 属性间的相关性越强则距离越近.在聚类过程中,每个簇中只包含一维敏感属性.划分之后的数据表中,相关性高的属性被划分在同一列中,保护了依赖度较高的数据间的关系,保持了更好的数据效用,更利于研究者对发布后的数据进行分析.而且,即使数据表中某一维的 SA 属性泄露,也不会导致其他维的 SA 属性泄露,能更好的保护隐私不泄露.此外,为了防止攻击者将不同列之间的属性值链接起来进行攻击,也可以在划分后,将各列中的属性值进行随机置换,这样即使攻击者通过额外的背景知识获得某个元组某一列的属性值,也不能推测出其他列的属性值.

5 多维敏感属性数据发布方法

5.1 信息损失度量方法

为了生成满足 l-maximum 原则要求的等价类,在数据发布过程中会产生一些隐匿元组,本文采用文献[7]中的隐匿率来衡量隐匿元组的比例,定义如下:

定义 5 隐匿率 给定数据表 T,n_s 为隐匿的元组个数,$|T|$ 为数据表中元组的总个数,隐匿率为隐匿的元组数占数据表 T 总元组数的比率,记为 SuppRatio = $n_s/|T|$.

分辨率度量[3,14](Discernability Metric,DM)是一种常见的信息损失度量方法.DM 的度量依赖于数据表中等价类的大小,具体定义为:

$$\mathrm{DM} = \sum_{i=1}^{c} |E_i|^2$$

其中,c 表示等价类的个数,$|E_i|$ 表示第 i 个等价类中元组个数.

5.2 多维敏感属性的 *l*-maximum 算法

本节提出一种面向多敏感属性的 l-maximum 算法(MSA l-maximum),算法分为分组阶段和处理剩余元组阶段,具体算法如下:

算法 1 MSA l-maximum **算法**

输入:属性划分后的数据表 T',参数 l;

输出:满足 l-maximum 原则的数据表 T^*.

```
//分组阶段
1. 计算 T' 中每个元组的敏感属性值频率之和 f(t_i);
2. Q = {按照 f(t_i) 将元组从大到小排列};
3. E = ∅;
4. While Q ≠ ∅
5.     选取 Q 中第一个元组 t_i 加入集合 E 中;
6.     Q = Q - {t_i};
7.     D = {Q 中与集合 E 敏感属性值不同的元组集合};
8.     While|E| < l
9.       if D ≠ ∅
10.          从 D 中选取第一个元组 t_j 加入到集合 E 中;
11.          更新 D;
12.          Q = Q - {t_j};
13.      else
14.          将集合 E 加入到集合 R 中;
15.          break;
16.      end
17.    end
18.    将集合 E 加入到集合 G 中;
19.    E = ∅;
20. end
//处理剩余元组阶段
21. for R 中的每一个元组 t
22.     if(存在分组 E,添加 t 后仍满足 l-maximum 原则)
23.       将 t 加入到分组 E 中;
24.     else
25.       隐匿元组 t;
26.     end
27. end
28. 将集合 G 以数据表 T* 形式发布;
```

6 实验结果及分析

6.1 数据集

本节中,通过实验分析 MSA l-maximum 的性能,并

将其与文献[7]提出的 MMDCF、文献[10]提出的 MSFM-PL-diversity 和文献[11]提出的 ACBG 进行比较.实验的数据集为 UCI 中的 Adult 数据集,选取数据集中的 10 个属性来进行实验,其中选取{Age, Final-weight , Country, Sex, Relationship}为 QI 属性,{Race, Marital-status, Work-class, Education, Occupation}为 SA 属性,SA 属性的组成如表 1 所示.

表 1　敏感属性个数及组成

|SA|	Sensitive Attributes
2	Race, Educaiton
3	Race, Educaiton, Marital-status
4	Race, Educaiton, Marital-status, Work-class
5	Race, Educaiton, Marital-status, Work-class, Occupation

6.2　隐匿率分析

图 1 ~ 3 分别给出了值 l、敏感属性维数 $|SA|$ 和数据集大小 $|T|$ 变化对四种算法隐匿率的影响.由图 1 可以看出,当 l 值增大时,所有算法的隐匿率都会增大,这是由于 l 值越大,生成等价类时包含的敏感值个数越多,对等价类中多样性的要求就越严格,导致不满足要求的元组增多,隐匿率会增大.当 l 值大于 4 时,隐匿率增长迅速,因为当 l 值增大时,l 值会接近甚至大于敏感属性的多样性,因此会产生大量不能满足要求的元组,隐匿率会迅速增长.由图 2 可以看出,当 $|SA|$ 增大时,算法的隐匿率都将增加大,这是由于 $|SA|$ 增大时,对等价类中多样性的要求就越严格,导致不满足要求的元组增多,隐匿率会增大.由图 3 可以看出当数据量 $|T|$ 增大时,四种算法的隐匿率都会降低,这是由于数据量增大时,生成等价类时可选择的元组增多,较容易生成满足多样性需求的等价类,所以隐匿的元组减少.

另外由图 1 ~ 3 可以看出,在同等条件下,MSA l-maximum 的隐匿率比其他三种算法的隐匿率低,这是由于 MSA l-maximum 算法分组阶段中对某一个元组只生成大小为 l 的分组,使得数据集可以产生更多的分组,这样在处理剩余元组阶段使得剩余元组有更多的分组可供选择,故隐匿率较低.对于 ACBG 算法虽然其在分组阶段生成大小为 l 的分组,但是在处理剩余元组阶段其对分组中的元组个数规定过于严格,导致元组将被隐匿,故隐匿率与 MSA l-maximum 相比较大.

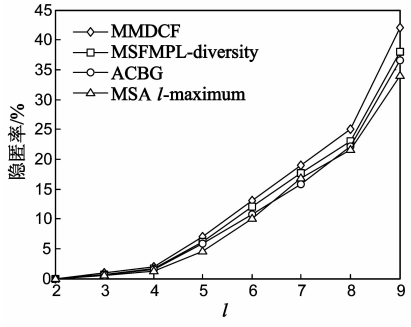


图1　不同l值下隐匿率的比较

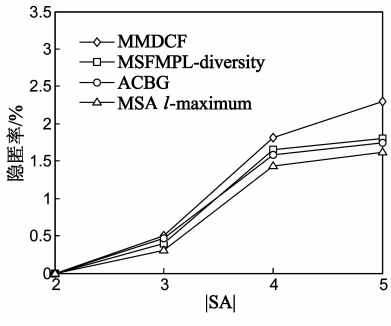


图2 不同|SA|值下隐匿率的比较

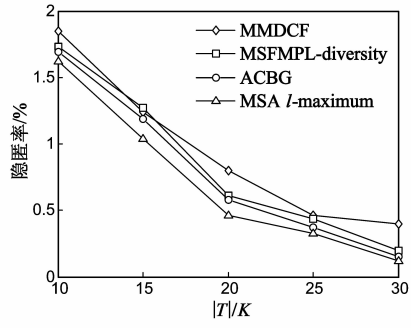


图3 不同|T|值下隐匿率的比较

6.3　DM 分析

图 4、图 5 分别给出了值 l、敏感属性维数 $|SA|$ 变化对四种算法 DM 值的影响.从图 4 中可以看出,当 l 值不大于 6 时,DM 值都将随着 l 值的增大而增大,这是因为 l 值增大导致等价类的大小将增大,所以 DM 值也会增加.当 $l=8$ 和 10 时,四种算法的 DM 值比 $l=6$ 时 DM 值小,从图 1 可以看出,l 值大于 7 时,算法的隐匿率都会大大增加,所以等价类中的元组值就会相应减少,DM 值会减小.从图 5 中可以看出,当 $|SA|$ 值增大时,DM 值都将增大,但是波动值非常小,这是由于四种算法在分组阶段都考虑了元组的整体选择度,减少了敏感属性维数的变化对分组阶段的影响,因此 DM 值得波动不大.

另外由图 4、图 5 可以看出,MSA l-maximum 的 DM 值比其他三种算法的 DM 值低,因为在 MSA l-maximum 算法分组阶段,对某一个元组只生成大小为 l 的分组,数据集可以产生更多的分组,每个分组中的元组数目要少于其他三种算法,因此 DM 值较低.

6.4　执行时间分析

图 6 ~ 8 分别给出了值 l、敏感属性维数 $|SA|$ 和数据集大小变化对四种算法执行时间的影响.由图 6 可以看出,l 值的变化对执行时间的影响不大,这是由于虽然 l 值变化了,但是数据集大小是固定的,整个数据的分布没有变化,因此时间波动不大.由图 7 可以看出,当 $|SA|$ 增大时,四种算法的执行时间都会增加,这是由于敏感属性维数的增加,数据集中敏感值的个数会增加,生成等价类时要处理的敏感值增多,增大时间开销.由图 8 可以看出,随着数据集 $|T|$ 的增加,四种算法执行时间都会增加,这是由于数据量的增加必然会造成算法的时间开销增大.

另外由图 6 ~ 8 可以看出,MSA l-maximum 的执行时间比其他三种算法的执行时间低.虽然 MSA l-maximum 算法聚类划分算法复杂度较高,但是划分消耗的时间可以忽略不计.此外,由于 MMDCF 和 MSFMPL-diversity 的分组阶段时间开销较大.而 ACBG 中每次选取元组时都要对元组进行扫描,寻找第一个未被屏蔽的元组,对于已经进行分组的元组也没有进行删除,造成了时间开销.然而,MSA l-maximum 算法分组完成后将相应元组删除,时间开销相对较小.

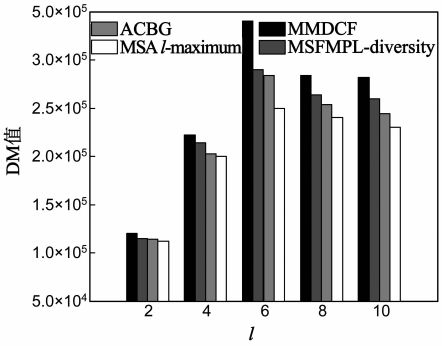


图4 不同l值下DM值的比较

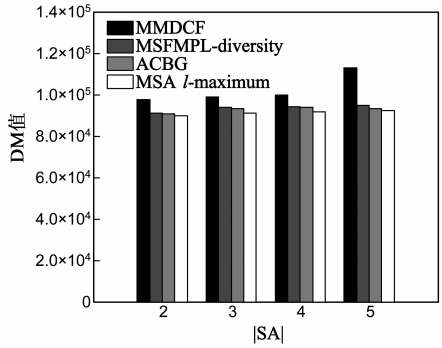


图5 不同|SA|值下DM值的比较

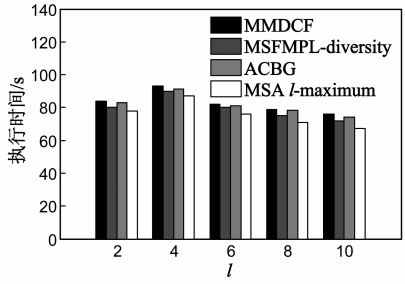


图6 不同l值下执行时间的比较

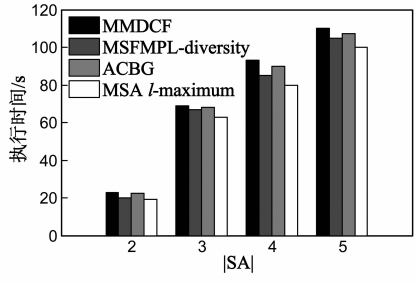


图7 不同|SA|值下执行时间的比较

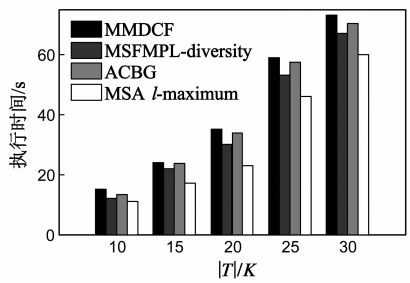


图8 不同|T|值下执行时间的比较

7 结束语

本文针对数据发布中多维敏感属性的隐私保护问题进行研究,提出了一种面向多敏感属性的隐私保护模型,给出了 l-maximum 原则,此外,针对现有方法破坏了数据间关系的局限,提出一种基于依赖度的属性划分方法,保护了相互依赖度高的属性间的关系,最后,提出一种 MSA l-maximum 算法用以实现该模型.实验结果表明,提出的模型能在有效地保护敏感信息不泄露的同时保持较高的数据效用.

参考文献

[1] Sweeney L. k-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 10(5): 557 – 570.

[2] Wong C R, Li J, Fu A, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[A]. 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining[C]. New York: ACM, 2006. 754 – 759.

[3] Machanavajjhala A, Gehrke J, Kifer D. l-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1 – 52.

[4] Li Ninghui, Li Tiancheng, Venkata S. t-closeness: privacy beyond k-anonymity and l-diversity[A]. Proc of the 23rd International Conference on Data Engineering[C]. Piscataway, NJ: IEEE, 2007. 106 – 115.

[5] Sha Chaofeng, Li Yi, Zhou Aoying. On t-closeness with KL-divergence and semantic privacy[A]. Proc of the 15th International Conference on Database Systems for Advanced Applications[C]. Berlin, German: Springer, 2010. 153 – 167.

[6] Rebollo-Monedero D, Forné J, Domingo-Ferrer J. From t-closeness-like privacy to postrandomization via information theory[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(11): 1623 – 1636.

[7] 杨晓春,王雅哲,王斌,等.数据发布中面向多敏感属性的隐私保护方法[J].计算机学报,2008,31(4):574 – 587.
Yang Xiaochun, Wang Yazhe, Wang Bin, et al. Privacypreserving approaches for multiple sensitive attributes in data publishing [J]. Chinese Journal of Computers, 2008, 31(4): 574 – 587. (in Chinese)

[8] Gal T S, Chen Zhiyuan, Gangopadhyay A. Aprivacy protection model for patient data with multiple sensitive attributes[J]. International Journal of Information Security and Privacy, 2008, 2

(3):28 – 44.

[9] Ye yang, Liu Yu, Lv dapeng, et al. Deconposition: privacy preservation for multiple sensitive attributes[A]. Proc of 14th Int Conf on Database Systems for Advanced Applications [C]. Berlin: Springer, 2009. 486 – 490.

[10] 杨静,王波.一种基于最小选择度优先的多敏感属性个性化 *l*-多样性算法[J].计算机研究与发展,2012,49(12):2603 – 2310.
Yang Jing, Wang Bo. Personalized *l*-diversity algotithm for multiple sensitive attributes based on minimum selected degree first[J]. Journal of Computer Research and Development, 2012, 49(12):2603 – 2610. (in Chinese)

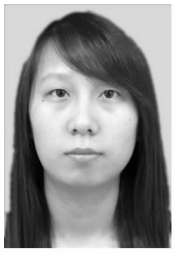
[11] 王茜,李艳军,刘泓.一种基于属性分类的多敏感属性隐私保护方法[J].计算机工程,2013,39(8):177 – 186.
Wang Qian, Li Yanjun, Liu Hong. A privacy preserving approach for multiple sensitive attributes based on attributes classification[J]. Computer Engineering, 2013, 39(8):177 – 186. (in Chinese)

[12] 韩建民,于娟,虞慧群,等.面向数值型敏感属性的分级 *l*-多样性模型[J].计算机研究与发展,2011,48(1):147 – 158.
Han Jianmin, Yu Juan, Yu Huiqun, et al. Amulti-level *l*-diversity model for numerical sensitive attributes[J]. Journal of Computer Research and Development, 2011, 48(1): 147 – 158. (in Chinese)

[13] Sun Xiaoxun, Li Min, Wang Hua. A family of enhanced(l, α)-diversity models for privacy preserving data publishing [J]. Future Generation Computer Systems: The Int Journal of Grid Computing: Theory, Methods and Applications, 2011, 27(3):348 – 356.

[14] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional *k*-anonymity[A]. Proc of the 22nd International Conference on Data Engineering[C]. New York: ACM, 2006.

作者简介



谢　静　女,1986 年生于湖北随州.哈尔滨工程大学计算机科学与技术学院博士研究生.主要研究方向为数据挖掘、隐私保护.
E-mail: xiejing@hrbeu.edu.cn

张健沛　男,1956 年生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院教授、博士生导师.主要研究方向为数据库与知识工程、数据挖掘、隐私保护、软件理论等.
E-mail: zhangjianpei@hrbeu.edu.cn