

数据发布中面向多敏感属性的隐私保护方法

杨晓春 王雅哲 王 斌 于 戈

(东北大学信息科学与工程学院 沈阳 110004)

摘 要 现有的隐私数据发布技术通常关注单敏感属性数据,直接应用于多敏感属性数据会导致大量隐私信息的泄漏.文中首次对多敏感属性数据发布问题进行详细研究,继承了基于有损连接对隐私数据进行保护的思想,提出了针对多敏感属性隐私数据发布的多维桶分组技术——MSB(Multi-Sensitive Bucketization).为了避免高复杂性的穷举方法,首先提出 3 种不同的线性时间的贪心算法:最大桶优先算法(MBF)、最大单维容量优先算法(MSDCF)和最大多维容量优先算法(MMDCF).另外,针对实际应用中发布数据的重要性差异,提出加权多维桶分组技术.实际数据集上的大量实验结果表明,所提出的前 3 种算法的附加信息损失度为 0.04,而隐匿率都低于 0.06. 加权多维桶分组技术对数据拥有者定义的重要信息的可发布性达到 70%以上.

关键词 数据发布;数据隐私;多敏感属性;有损连接; l -多样性

中图法分类号 TP311

Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing

YANG Xiao-Chun WANG Ya-Zhe WANG Bin YU Ge

(School of Information Science and Engineering, Northeastern University, Shenyang 110004)

Abstract Current privacy preserving data publishing techniques concentrate on tables with only one sensitive attribute. However, most of the real-world applications contain multiple sensitive attributes. Directly applying the existing single-sensitive-attribute privacy preserving techniques often causes unexpected private information disclosure. This paper firstly discusses the problem of secure publishing data when sensitive data contains multi attributes, and then propose a multi-dimensional bucket grouping approach on the idea of lossy join, called Multi-Sensitive Bucketization (MSB). In order to avoid exhausting search, three specific line-time greedy based MSB algorithms are proposed, which are maximal-bucket first algorithm (MBF), maximal single-dimension-capacity first algorithm (MSDCF), and maximal multi-dimension-capacity first algorithm (MMDCF). In addition, according to the differences among published data, a weighted MSB approach is further proposed. Experimental results on the real-world datasets show that the addition information loss of the proposed MSB methods were not more than 0.04 and the suppression ratios were less than 0.06. The weighted MSB approach can guarantee more than 70% publishing ratio.

Keywords data publishing; data privacy; multi-sensitive attributes; lossy join; l -diversity

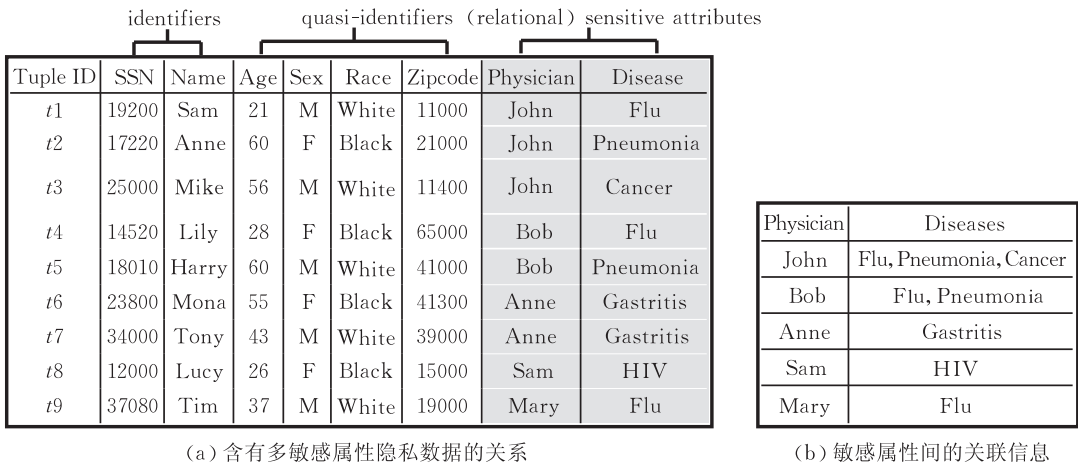
收稿日期:2007-11-26. 本课题得到新世纪优秀人才支持计划(NCET-06-0290)、国家自然科学基金(60503036)和霍英东教育基金会青年教师优选资助课题(104027)资助. 杨晓春,女,1973年生,博士,副教授,主要研究方向为数据库理论与技术. E-mail: yangxc@mail.neu.edu.cn. 王雅哲,女,1984年生,硕士研究生,主要研究方向为数据隐私保护. 王 斌,男,1972年生,博士研究生,主要研究方向为 P2P、Web 查询处理. 于 戈,男,1962年生,教授,博士生导师,主要研究领域为分布式数据库、Web 服务、数据流.

1 引言

信息化的现代社会中,高度的数据收集和共享机制为各种组织机构间的合作和研究工作提供了巨大的便利条件,同时也增加了个人隐私信息泄露的风险.虽然独力的数据发布单位分别会采取措施隐藏发布数据中的个人身份标识或者某些隐私数据,但是值得注意的是通过在多个公开的数据源间进行连接操作往往会导致意想不到的隐私信息泄漏问题.文献[1]中的研究表明,通过 Zipcode, Sex, Data of Birth 等属性对选民登记表和隐匿了个体标识的医疗信息表进行连接操作,超过 87% 的美国公民的

身份都可以被唯一标识.因此,防止这种链接攻击(linking attack)^[1],保护个人隐私信息的安全,是敏感数据发布与共享应用中的重要问题之一.

本文侧重关系型的隐私数据.例如,图 1(a)为一个典型的包含个人医疗信息的共享医疗记录表.用于发布或共享的关系中主要涉及四类属性:(1)个体身份的标识属性(identifier),如姓名、社会保险号等,通常在数据发布时被隐匿;(2)与外部数据源进行连接可标识个体身份的属性,如 Age, Sex, Race, Zipcode 等,叫做准标识符 QI(Quasi-Identifier);(3)包含个体隐私信息的敏感属性,如 Disease 等,需要被保护;(4)其他属性.



(a) 含有多敏感属性隐私数据的关系 (b) 敏感属性间的关联信息

图 1 多敏感属性数据集和敏感属性间的关联信息

现有的敏感数据发布方法^[1-18]主要针对单一敏感属性的数据.然而,在很多现实应用中,发布的数据中往往涉及到多个敏感属性.特别是有一些属性虽然本身并不直接包含个体的隐私信息,但是却与隐私信息之间有着明显的特定联系.发布这类属性信息相当于间接发布敏感信息,因此这类属性也要作为敏感属性被保护,叫做相关敏感属性(relational sensitive attribute)(例如图 1(a)中的 Physician 属性).通常每个医生只主治某几类疾病,Physician 与 Disease 之间的这种关联信息(如图 1(b)所示)很容易通过其它渠道获取,因此 Physician 与 Disease 构成一对相关敏感属性.通常情况下,现有的单敏感属性隐私数据发布方法并不适用于多敏感属性,特别是相关多敏感属性的隐私数据发布的问题.直接将现有的方法应用于包含多个敏感属性的数据会导致隐私信息的泄漏.

本文在前人工作的基础上首次对多敏感属性,特别是相关多敏感属性隐私数据发布问题进行了

详细研究.提出了一种基于多维桶的分组技术——MSB(Multi-Sensitive Bucketization).其核心思想是将多个敏感属性构成的复合敏感属性作为一个高维向量,并引进一种多维桶结构,将关系表中的记录按照复合敏感属性向量映射到不同的桶中,在这些桶上按照某种策略进行分组,保证每个分组内的记录在每一维敏感属性上取值都满足隐私信息保护要求,保证多敏感属性情况下隐私数据发布的安全性.本文的主要贡献如下:

(1)提出了基于多维桶分组技术的隐私数据发布方法.该方法适用于含有任意多个敏感属性的关系型数据,能很好地保证在多敏感属性数据的安全发布.

(2)给出了 3 种基于 MSB 的贪心算法,包括最大桶优先算法、最大单维容量优先算法和最大多维容量优先算法.在保障信息安全发布的前提下,降低了连接的有损程度.

(3)讨论了实际应用中发布数据存在的差异.

对于发布数据中信息的不同要求,数据拥有者可以指定数据的不同的发布程度.在提出的3种算法的基础上,提出了加权多维桶分组技术,提高发布的数据对于具体应用的实用性.

(4)采用实际数据集进行大量实验,对所提出的方法进行了验证与分析,测试了数据的发布质量和执行效率,说明所提出的方法能很好地保护多敏感属性隐私数据的安全,发布具有高质量的数据.

2 相关工作

敏感数据发布与共享环境中的个体隐私信息的安全性问题一直是数据隐私研究的热点. Sweeney 和 Samarati 提出 k -匿名模型保护隐私数据不受链接攻击^[1,2]. k -匿名是指关系中的每条记录都至少和其它的 $k-1$ 条记录具有相同的准标识符属性取值. 概括和隐匿是实现数据 k -匿名化的主要方法^[1-2]. 其重点在于保证数据安全性的前提下,使通过概括和隐匿损失的信息最少或者使数据的可用性最高. 为达到这一目的,文献[5-6]分别提出了最坏情况下指数时间复杂度的最优化 k -匿名算法. Meyerson 等在文献[7]中证明了通过最少的概括实现数据 k -匿名化($k>2$)的问题是 NP 难的. 因此文献[8-11]采用了近似算法实现 k -匿名. 文献[9]给出了近似比为 $O(\log k)$ 的近似算法. 另外,文献[6]根据不同的概括策略对基于概括和隐匿的 k -匿名化算法进行了分类. 文献[12]提出了一种自顶向下逐步特化启发式方法,在进行 k -匿名化的同时保留数据中大量的分类信息. 文献[15-16]分别讨论了多约束 k -匿名化问题和数据更新时 k -匿名的增量维护问题. 文献[17]提出了一种利用经典的空间索引技术来实现 k -匿名的技术,具有很好的扩展性且能很好地支持增量的数据发布. 目前的 k -匿名算法都倾向于保证发布数据中最小的信息损失,但是恶意用户很可能利用这一知识推断出隐私信息,文献[18]讨论了这一问题,并提出了 m -confidentiality 的概念.

文献[4]提出在某些情况下 k -匿名并不能保证隐私信息的安全. 例如,具有相同准标识符属性取值的所有或大部分记录都具有相同的敏感属性值,恶意的用户还是能够轻易地以高概率推断出隐私信息. 因此文献[4]提出了 l -多样性的概念,对匿名化的数据记录中的出现频率最高的敏感属性值的个数做出了约束,要求它不大于 $1/l$.

当数据表中准标识符属性的个数较多时,通过

概括和隐匿的方法对数据进行 k -匿名化会损失大量信息^[13]. 因此文献[3]提出了一种不基于概括和隐匿的新颖的方法——Anatomy. 它通过将原始关系的准标识符属性和敏感属性以两个不同的关系发布,利用它们之间的有损连接来保护隐私数据的安全,并且给出了基本的 Anatomy 算法保证发布的数据满足 l -多样性的要求. 同样,文献[14]也提出了另一种基于有损连接的隐私数据发布算法.

以上提出的敏感数据发布方法都主要针对单一敏感属性数据的情况,对于具有多个敏感属性,特别是多个相关敏感属性的数据集,直接应用以上方法并不能保证隐私数据的安全. 本文假设安全数据发布需满足 l -多样性约束,基于此首次详细讨论了多敏感属性隐私数据发布问题,提出了基于有损连接技术的支持多敏感属性的隐私数据发布多维桶分组技术.

3 多敏感属性隐私数据发布问题

下面对多敏感属性隐私数据发布问题进行形式化定义.

设用户要发布关系 $T\{A_1, A_2, \dots, A_p, S_1, S_2, \dots, S_d\}$. 其中 $A_i (1 \leq i \leq p)$ 为准标识符属性, $S_j (1 \leq j \leq d)$ 为敏感属性. 设 T 中含有 n 条记录,即 $|T|=n$,其中每条记录记为 $t_i (1 \leq i \leq n)$. 另 $t[X]$ 表示记录 t 的 X 属性值.

定义 1(复合敏感属性). 关系 T 的所有敏感属性的整体构成一个复合敏感属性(composite sensitive attribute),记为 S . 其中第 i 个敏感属性作为复合敏感属性的第 i 维,记做 $S_i (1 \leq i \leq d)$, $D(S_i)$ 为 S_i 的值域, $|S_i|$ 表示 $D(S_i)$ 的基数.

定义 2(复合敏感属性向量). 关系 T 中记录 t 的所有敏感属性取值构成向量形式 $\langle t[S_1], t[S_2], \dots, t[S_d] \rangle$, 叫做复合敏感属性向量.

定义 3(分组^[3]). 一个分组是 T 中记录的子集. T 中每条记录属于且仅属于一个分组. 关系 T 的分组记为 $GT\{G_1, G_2, \dots, G_m\}$, $\bigcup_{j=1}^m G_j = T$ 并且 $QI_i \cap QI_j = \emptyset (1 \leq i \neq j \leq m)$.

定义 4(单敏感属性 l -多样性^[4]). 对于一组单敏感属性记录 G , 设 v 为 G 中记录的敏感属性取值中最大频繁取值, $c(v)$ 为其出现的次数, 如果 $\frac{c(v)}{|G|} \leq \frac{1}{l}$ ($|G|$ 为 G 中记录的条数), G 就满足 l -多样性

性质.

定理 1. 一个分组中, 如果复合敏感属性的每一维敏感属性都满足 l -多样性性质, 则复合敏感属性必满足 l -多样性性质.

证明. 用反证法证明. 假设分组 $G(t_1, t_2, \dots, t_m)$ 中, 复合敏感属性的每一维敏感属性都满足 l -多样性性质, 但是复合敏感属性不满足 l -多样性性质. 设 G 中记录的所有复合敏感属性向量中出现最频繁的向量为 $v_0 \langle t_0[S_1], t_0[S_2], \dots, t_0[S_q] \rangle$, 其出现的次数为 $c(v_0)$. 对于每一个单独的敏感属性 S_j , 其值 $t_0[S_j]$ 出现的次数 $c(t_0[S_j]) \geq c(v_0)$. 复合敏感属性不满足 l -多样性性质意味着 $\frac{c(v_0)}{|G|} \leq \frac{1}{l}$, 那么, 对于其中每一维敏感属性 $\frac{c(t_0[S_j])}{|G|} \geq \frac{c(v_0)}{|G|} \geq \frac{1}{l}$, 与条件中每一维敏感属性都满足 l -多样性性质矛盾. 因此定理得证. 证毕.

由以上定理得复合敏感属性 l -多样性的定义.

定义 5(复合敏感属性 l -多样性). 设一个分组 G , 如果其中的所有数据记录的复合敏感属性的每一维敏感属性上的取值分别满足 l -多样性性质, 则 G 对于复合敏感属性满足 l -多样性性质.

定义 6(复合敏感属性 l -多样性分组). 对于关系 T 上的分组 $GT\{G_1, G_2, \dots, G_m\}$, 如果其中每个分组 $G_i (1 \leq i \leq m)$ 都满足复合敏感属性 l -多样性性质, 则 GT 为 T 上的复合敏感属性 l -多样性分组.

定理 2. 关系 T 上的复合敏感属性 l -多样性分组 GT 对于发布的数据来说是安全的.

证明. 由定义 5 可知, 满足复合敏感属性 l -多样性性质的每个分组中的记录的每一维敏感属性的取值都满足 l -多样性性质. 将这个分组内数据的准标识符属性和复合敏感属性分别发布为 QIT 表和 ST 表, 通过连接操作, 恶意用户对于分组中记录的每一维敏感属性都无法以大于 $1/l$ 的概率推断出其真实值. 因此, 对这个分组来说发布的数据是安全的. 由定义 6, 关系 T 上的复合敏感属性 l -多样性分组 GT 中每个分组 G_i 都满足复合敏感属性 l -多样性性质, 因此按照这个分组原则发布 QIT 表和 ST 表是安全的, 也可以说 T 上的复合敏感属性 l -多样性性质的分组 GT 是安全的. 定理得证. 证毕.

由以上定理可知, 对于包含多个敏感属性的关系 T , 如果其上的分组 GT 为复合敏感属性 l -多样性分组, 则按照这种分组原则发布的数据是安全的. 要使 T 上的分组满足复合敏感属性 l -多样性性质,

显然 l 不能大于 T 中每个敏感属性的不同取值个数, 即 $l \leq \min(|S_i|), 1 \leq i \leq d$ (d 为 T 中敏感属性的个数). 下面的定理给出了关系 T 可被划分成复合敏感属性 l -多样性分组的条件.

定理 3. 关系 T 可被划分成复合敏感属性 l -多样性分组的充要条件是对于每一维敏感属性 $S_i (1 \leq i \leq d)$, 其最大频繁属性值 v_i 出现的次数 $c(v_i)$ 不大于 n/l .

证明. 充分性. 如果对于关系 T 的每一维敏感属性 S_i , 其最大频繁属性值 v_i 出现的次数 $c(v_i) \leq \frac{n}{l}$, 则一定存在 T 上的复合敏感属性 l -多样性分组. 如果设不存在 T 上的复合敏感属性 l -多样性分组, 则在最坏的情况下, 将 T 整体作为一个分组时也不满足复合敏感属性 l -多样性性质. 也就是说必然存在着至少一维敏感属性, 其最大频繁属性值 v_i 出现的次数 $c(v_i) > \frac{n}{l}$.

必要性. 如果关系 T 可被划分成复合敏感属性 l -多样性分组则对于每一维敏感属性 S_i , 其最大频繁属性值 v_i 出现的次数 $c(v_i) \leq \frac{n}{l}$. 设 T 上的复合敏感属性 l -多样性分组为 G_1, G_2, \dots, G_m . 另 $x_{ki}(v_j)$ 为第 $k (1 \leq k \leq m)$ 个分组 G_k 内第 $i (1 \leq i \leq d)$ 维敏感属性的值 v_j 出现的次数, 对于敏感属性 S_i 的最大频繁取值 v_0 有 $\sum_{1 \leq k \leq m} x_{ki} = c(v_0)$. $|G_k|$ 为分组 G_k 内记录的条数, $\sum_{1 \leq k \leq m} |G_k| = n$. 由于 G_1, G_2, \dots, G_m 都满足复合敏感属性 l -多样性性质, 对于第 i 维敏感属性最大频繁取值 v_i 都有 $\frac{c(v_0)}{|G|} \leq \frac{1}{l}$, 有 $\frac{c(t_0[S_j])}{|G|} \geq \frac{c(v_0)}{|G|} \geq \frac{1}{l}$, 即 $c(v_i) \leq n/l$. 证毕.

根据以上定义和定理, 多敏感属性隐私数据发布问题的关键就是找到包含多个敏感属性的关系 T 上的复合敏感属性 l -多样性分组 GT . 对于不满足定理 3 中复合敏感属性 l -多样性分组条件的关系 T , 可以适当隐匿其中的某些数据记录, 通过少量的数据损失来保证发布数据的安全性.

4 多敏感属性数据发布方法

4.1 基于多维桶的安全数据发布的基本策略

本文提出基于多维桶分组技术——MSB 方法解决多敏感属性隐私数据发布的问题. 其目标是找

到多敏感属性的关系 T 上的分组方案,使每个分组都满足复合敏感属性 l -多样性. 因此, MSB 方法首先引进多维桶结构重新组织 T 中的记录. 多维桶构造方法如下: 复合敏感属性的每一维对应桶的一维. 将 T 中的数据记录按照其复合敏感属性向量每一维的值分别映射到相应的桶中. 设关系中复合敏感属性维数为 d , 在其上建立的 d 维桶记为 $Bucket(S_1, S_2, \dots, S_d)$. 其中每个桶记为 $buk\langle s^1, s^2, \dots, s^d \rangle, s^j \in D(S_j) (1 \leq j \leq d)$. 每个桶大小为 $size(buk\langle s^1, s^2, \dots, s^d \rangle)$, 表示包含的记录个数. 例如, 图 2 为图 1(a) 关系中记录的复合敏感属性向量和其构造的 d 维桶 ($d=2$). 然后, MSB 方法在构造的 d 维桶上按照某种策略提取记录构成分组, 使分组中的记录尽可能来自在每一维上的值都互不相同的桶.

	Flu	Pneumonia	Gastritis	HIV	Cancer
John	{t1}	{t2}			{t3}
Bob	{t4}	{t5}			
Anne			{t6, t7}		
Sam				{t8}	
Mary	{t9}				

图 2 复合敏感属性数据集及其对应的 d 维桶 ($d=2$)

MSB 方法对于发布数据, 应用组内有损连接技术, 保护隐私数据, 因此在不破坏 l -多样性的前提下, 分组越小有损连接造成的信息损失越少. 在理想情况下, 最小的分组大小为 l . 对于大小为 l 的分组, 要使其满足复合敏感属性 l -多样性, 要求每个分组中的所有记录在每一维敏感属性上的取值都互不相同. 但是, 满足构成复合敏感属性 l -多样性分组的条件 (见定理 3) 的实际数据集 T 未必总是可以达到使得每个复合敏感属性 l -多样性分组大小都恰好是 l 的理想情况. 在保证敏感数据的安全性的前提下, 分组的大小可能超过 l , 这样就会造成附加的有损连接信息损失.

定义 7(附加信息损失度). 对于关系 T 上满足复合敏感属性 l -多样性的分组 $GT\{G_1, G_2, \dots, G_m\}, |G_i| \geq l (1 \leq i \leq m)$, 附加信息损失度 (additional information loss) 为 $\sum_{1 \leq i \leq m} (|G_i| - l) / ml$.

采用了一种固定分组大小为 l 的贪心式分组方案, 即以 l 为基本分组大小, 采用贪心策略在多维桶上按一定的顺序选择 l 个每维取值都互不相同的桶提取记录构成分组, 重复进行, 得到尽量多的 l 大小的分组. 然后对于剩余的记录在不破坏复合敏感属

性 l -多样性的前提下将其添加到某一分组中. 最后, 将不包含在任何分组中的记录从发布的数据中隐匿. 因此需要用隐匿率 (suppression ratio) 来衡量隐匿的记录数占关系 T 中记录总数的比例. 定义隐匿率为

$$SuppRatio = n_s / |T| \tag{1}$$

其中, n_s 为隐匿的记录数. 显然, 隐匿率越小损失的记录数越少, 理想情况下的隐匿率为 0. 将隐匿率与附加信息损失度一起作为算法发布数据质量的衡量标准.

因此, Multi-Sensitive Bucketization 方法在多维桶上的分组过程可分为两个阶段: (1) 分组阶段 (grouping phase). 该阶段按照某种贪心策略选择每一维取值都互不相同的 l 个桶, 在每个桶中各提取 1 条记录构成一个分组, 循环进行直到无法构成满足要求的分组; (2) 处理剩余记录阶段 (residual processing phase). 对于分组后多维桶中剩余的记录, 在不破坏复合敏感属性 l -多样性的前提下尽可能将其添加到现有的分组中. 最后, 将不属于任何分组的记录从发布的数据中隐匿. 经过上述步骤处理后, 得到关系 T 上的分组. 将每个分组的准标识符发布为 QIT 表, 将敏感属性发布为 ST 表就完成了隐私数据的发布过程.

根据不同的多维桶上的分组策略, 本文给出 3 种线性时间的贪心算法对多敏感属性隐私数据发布问题进行求解: 最大桶优先 MBF 算法、最大单维容量优先 MSDCF 算法和最大多维容量优先 MMDCF 算法. 随着贪心策略的改进, 3 种算法发布数据的质量逐步提高, 但运行时间逐步增加. 将通过大量实验对实际情况中算法的发布数据的质量与执行时间进行验证.

4.2 最大桶优先算法

首先给出一种最简单的最大桶优先算法, 记为 MBF (Maximal-Bucket First). 在多维桶上进行分组时, 要构成尽可能多的 l 大小的分组. 基本思想是优先选择最大 (即含有记录数最多) 的非空桶提取记录构成分组. 每选择一个桶提取一条记录就在每一维上将与此桶在该维取值相同的所有桶全部屏蔽, 之后不选择已屏蔽的桶中的记录. 重复进行, 选出来自最大的且每一维取值都互不相同的 l 个桶中的 l 条记录构成一个分组, 这样就能保证同一个分组中的 l 条记录没有任何两条在某一维敏感属性上取得相同的值. 取消所有桶上的屏蔽标志, 并重复进行这一分组过程, 直到无法构成一个完整的分组为止. 对

于每条剩余的记录, 搜索所有已构成的分组, 检查是否可以将其添加到该分组中而不破坏分组的复合敏感属性 l -多样性性质. 最后将无法添加到任何分组的记录在发布的数据中隐匿. MBF 算法具体步骤如算法 1 所示. 由算法过程可知算法的时间复杂度为 $O(n)$.

算法 1. 最大桶优先算法(MBF).

输入: 关系 $T\{A_1, A_2, \dots, A_p, S_1, S_2, \dots, S_d\}$, 多样性参数 l

输出: 准标识符属性表 QIT, 敏感属性表 ST

//分组阶段.

1. 关系 T 上的分组 $G_s = \emptyset$, 并根据 T 的复合敏感属性建立 d 维桶 $\text{Bucket}(S_1, S_2, \dots, S_d)$;
 2. while 可以提取记录构成分组
 3. 对所有桶设未屏蔽标记, 分组 $G = \emptyset$;
 4. for $i = 1:l$
 5. if 存在未被屏蔽的非空桶
 6. 在未屏蔽桶中选择容量最大的桶 buk , 提取一条记录 t 添加到 G ;
 7. 在桶中删除 t , 且桶 $\text{size}(buk) = \text{size}(buk) - 1$;
 8. 屏蔽所有与 t 在某个维上有相同取值的桶;
 9. else 结束分组过程;
 10. end if
 11. end for
 12. 将构成的分组 G 添加到 G_s ;
 13. end while
- //处理剩余记录阶段.
14. for each 剩余记录 rt
 15. 如果存在分组 G , 添加 rt 后仍满足复合敏感属性 l -多样性性质, 添加 rt 到 G ;
 16. end for
 17. 隐匿所有无法添加到分组的剩余记录;
 18. 将所有分组 G_s 以 QIT, ST 形式输出.

以图 2 中的 2 维桶为例对算法执行过程进行说明. 按照 $l=3$ 为参数分组, MBF 算法发布的数据如图 3 所示. MBF 算法首先选择一个最大的桶 $\langle \text{Anne}, \text{Gastritis} \rangle$ (桶大小相同则任取其一), 提取记录 $t6$, 桶 $\langle \text{Anne}, \text{Gastritis} \rangle$ 的大小变为 1, 并屏蔽属性值 Anne 对应行和 Heart Disease 对应列上的所有非空桶. 在剩余桶中继续选择桶 $\langle \text{Bob}, \text{Pneumonia} \rangle$, 提取 $t5$ 和桶 $\langle \text{Sam}, \text{HIV} \rangle$, 提取 $t8$. $\{t5, t6, t8\}$ 构成一个分组. 算法循环进行, 继续下一分组. 由于所有非空桶的当前容量都为 1, 算法任意选择. 假设算法提取 $t1, t7$ 后, 所有非空桶都被屏蔽, 分组无法继续进行. 最后得到的一种分组结果为 $\{t5, t6, t8\}$. 剩余记录 $\{t1, t2, t3, t4, t7, t9\}$, 可以将 $t1$ 添加到分组

$\{t5, t6, t8\}$ 而不违反复合敏感属性 l -多样性. 因此最后的一种分组结果为 $\{t1, t5, t6, t8\}$, 隐匿的记录为 $\{t2, t3, t4, t7, t9\}$. 隐匿率为 $5/9$, 附加信息损失度约为 0.3.

由上文例子中的分组过程可以看出, 虽然最大桶优先(MBF)算法的执行过程十分简单, 但是算法的附加信息损失度和隐匿率都比较大, 分组效果并不理想. 4.3 节将介绍两种分组效果较好的最大维容量优先的算法.

QIT			ST	
Tuple ID	QIs	Group ID	Group ID	Sensitive Attributes
$t1$...	G_1	G_1	$\langle \text{Bob}, \text{Flu} \rangle$
$t5$...	G_1		$\langle \text{John}, \text{Pneumonia} \rangle$
$t6$...	G_1		$\langle \text{Sam}, \text{HIV} \rangle$
$t8$...	G_1		$\langle \text{Anne}, \text{Gastritis} \rangle$

图 3 最大桶优先算法发布的数据结果

4.3 最大维容量优先算法

最大维容量优先算法与最大桶优先算法的主要区别在于分组时的优先选择策略不同. 最大维容量优先算法在分组时并不仅考虑每个桶的大小, 而且综合考虑同一维上取值相同的所有桶. 在多维桶上对每一维上的每个不同取值计算容量 Capa . 多维桶上 S_j 维的某个取值 $s_0^j \in D(S_j)$ 的容量为所有在这一维上取值为 s_0^j 的所有桶大小的和, $\text{Capa}(s_0^j) = \sum_{s^j = s_0^j} \text{size}(buk \langle s_0^1, s_0^2, \dots, s_0^d \rangle)$. 根据维容量, 对每个非空桶计算一个选择度, 用来在分组时确定对各个桶的优先选择程度. 桶 $buk \langle s_0^1, s_0^2, \dots, s_0^d \rangle$ 的选择度记为 $\text{Select}(buk \langle s_0^1, s_0^2, \dots, s_0^d \rangle)$. 算法优先选择选择度最大的桶提取记录构成分组. 最大维容量优先算法的基本步骤如算法 2 所示. 算法的时间复杂度仍然 $O(n)$. 根据选择度计算方法的不同给出了最大单维容量优先的算法 MSDCF (Maximal Single-Dimension-Capacity First) 和最大多维容量优先算法 MMDCF (Maximal Multi-Dimension-Capacity First).

算法 2. 最大维容量优先算法.

输入: 关系 $T\{A_1, A_2, \dots, A_p, S_1, S_2, \dots, S_d\}$, 多样性参数 l

输出: 准标识符属性表 QIT, 敏感属性表 ST

//分组阶段.

1. 关系 T 上的分组 $G_s = \emptyset$, 并根据 T 的复合敏感属性建立 d 维桶 $\text{Bucket}(S_1, S_2, \dots, S_d)$;
2. 计算每一维上所有不同取值对应的容量;
3. while 可以提取记录构成分组

4. 对所有桶设未屏蔽标记,分组 $G=\varnothing$;

5. 计算非空桶的选择度;

6. for $i=1:l$

7. if 存在未被屏蔽的非空桶

8. 在未屏蔽桶中选择选择度最大的桶 buk ,提取一条记录 t 添加到 G ;

9. 在桶中删除 t ,且桶 $size(buk)=size(buk)-1$;

10. 重新计算桶 buk 每一维的取值的容量;

11. 屏蔽所有与 t 在某些维上有相同取值的桶;

12. else 结束分组过程;

13. end if

14. end for

15. 将构成的分组 G 添加到 G_s ;

16. end while

... .. //(处理剩余记录的过程同算法 1)

4.3.1 最大单维容量优先算法

最大单维容量优先算法 MSDCF 将每个桶对应的每维上的容量的最大值与该桶的大小的和作为选择度,因此,桶 $buk\langle s_0^1,s_0^2,\cdots,s_0^d\rangle$ 的选择度如式(2)

所示.

$$Select(buk\langle s_0^1,s_0^2,\cdots,s_0^d\rangle)=$$
$$Max_{1\leq j\leq d}Capa(s_0^j)+size(buk\langle s_0^1,s_0^2,\cdots,s_0^d\rangle)\quad (2)$$

图 4 仍以 2 维桶为例,给出各个维对应不同取值的 $Capa$ 值和 $l=3$ 时 MSDCF 算法发布的数据. 算法按照单维容量最大优先原则在多维桶上分组. 由计算 $Select(buk\langle Anne,Gastritis\rangle)=4$, 桶 $buk\langle Anne,Gastritis\rangle$ 为选择度最大的桶之一. 在桶 $\langle Anne,Gastritis\rangle$ 中提取记录 $t6$,并屏蔽对应的每一维上所有与其取值相同的非空桶. 在剩余桶中继续根据选择度进行选择,并分别在桶 $\langle Bob,Flu\rangle$ 和桶 $\langle John,Cancer\rangle$ 中提取记录 $t3,t4$,构成分组 $\{t3,t4,t6\}$. 算法循环进行,最后可得到分组 $\{t3,t4,t6\}$ 和 $\{t5,t8,t9\}$,剩余记录 $\{t1,t2,t7\}$. 经过处理剩余记录过程,可将 $t7$ 添加到第 2 个分组,因此最后分组结果为 $\{t3,t4,t6\}$ 和 $\{t5,t7,t8,t9\}$,剩余记录 $\{t1,t2\}$. 隐匿率为 $2/9$,附加信息损失度约为 0.17 .

						QIT			TS		
Flu Pneumonia Gastritis HIV Cancer						Tuple ID	QIs	Group ID	Group ID	Sensitive Attribute	
John	$\{t1\}$	$\{t2\}$			$\{t3\}$	3	$t3$...	G_1	G_1	$\langle Anne, Gastritis \rangle$
Bob	$\{t4\}$	$\{t5\}$				2	$t4$...	G_1		$\langle Bob, Flu \rangle$
Anne			$\{t6, t7\}$			2	$t5$...	G_2		$\langle John, Cancer \rangle$
Sam				$\{t8\}$		1	$t6$...	G_1	G_2	$\langle Mary, HIV \rangle$
Mary	$\{t9\}$					1	$t7$...	G_2		$\langle Bob, Pneumonia \rangle$
							$t8$...	G_2		$\langle Sam, HIV \rangle$
							$t9$...	G_2		$\langle Anne, Gastritis \rangle$
	3	2	2	1	1						

图 4 d 维桶($d=2$)每一维上的容量和 MSDCF 算法发布的数据

由上面例子中的算法执行过程可以看出,最大单维容量优先的 MSDCF 算法综合考虑了每个桶单维上的值对应的最大容量,其附加信息损失度和隐匿率都小于最大桶优先的 MBF 算法,发布数据的质量较好. 但是,算法需要重复计算维容量,执行过程要比简单的 MBF 算法复杂.

4.3.2 最大多维容量优先算法

最大多维容量优先的 MSB 算法,记为 MMDCF. 与最大单维容量优先算法不同,在计算选择度时, MMDCF 算法不仅考虑桶的单个维上取值对应的容量,而且全面考虑桶的每一维上的所有取值对应的容量的和. 因此,桶 $buk\langle s_0^1,s_0^2,\cdots,s_0^d\rangle$ 的选择度如式(3)所示.

$$Select(buk\langle s_0^1,s_0^2,\cdots,s_0^d\rangle)=$$
$$\sum_{1\leq j\leq d}Capa(s_0^j)+size(buk\langle s_0^1,s_0^2,\cdots,s_0^d\rangle)\quad (3)$$

QIT			TS	
Tuple ID	QIs	Group ID	Group ID	Sensitive Attribute
t1	...	G ₁	G ₁	⟨Anne, Gastritis⟩
t2	...	G ₂		⟨Bob, Pneumonia⟩
t3	...	G ₃		⟨John, Flu⟩
t4	...	G ₂	G ₂	⟨Bob, Flu⟩
t5	...	G ₁		⟨John, Pneumonia⟩
t6	...	G ₁		⟨Anne, Gastritis⟩
t7	...	G ₂	G ₃	⟨Mary, HIV⟩
t8	...	G ₃		⟨Sam, HIV⟩
t9	...	G ₃		⟨John, Cancer⟩

图 5 MMDCF 算法发布的数据

算法根据式(3)定义的选择度进行分组. 仍然以图 4 中的 2 维桶为例,按照最大多维容量优先原则,以 $l=3$ 分组,发布的数据结果如图 5 所示. 算法首先计算每个桶的选择度,得到最大的选择度

$Select(buk\langle John, Flu \rangle) = 7$. 在桶 $\langle John, Flu \rangle$ 中提取 $t1$, 并屏蔽该桶对应的每一维上所有值相同的非空桶. 在剩余桶中继续计算选择度进行选择, 分别在桶 $\langle Bob, Pneumonia \rangle$ 和桶 $\langle Anne, Gastritis \rangle$ 中提取记录 $t5, t6$, 构成分组 $\{t1, t5, t6\}$. 算法循环进行, 最后可得到分组 $\{t1, t5, t6\}$, $\{t2, t4, t7\}$ 和 $\{t3, t8, t9\}$. 因为没有剩余记录, 其隐匿率为 0, 附加信息损失度为 0.

从上文例子中可以看出, 由于 MMDCF 算法的选择策略综合考虑在各个维上有相同取值的所有桶大小之和, 优先选择多维容量之和最大的桶, 得到了最优的分组结果, 发布数据的质量最好.

4.4 加权多维桶分组技术

由于 MSB 方法采用了一种固定分组大小为 l 的贪心式分组方案, 在分组过程中为了保证隐私信息的安全性, 会隐匿部分元组. 但是, 在实际应用中某些元组中包含着更重要的信息, 需要尽量保留在发布的数据中, 从而提高发布数据的可用性. 例如, 通常对于疾病分析研究来说, 医疗记录表中 HIV 患者的信息通常要比 Flu 患者的信息更加重要, 也就是说对于图 1(a) 的医疗信息表, 记录 $t8$ 包含的信息要比 $t1$ 更加重要, 需要优先考虑保留在发布的数据中. 当然, 这种信息的重要性也可以根据发布数据的目的与数据的使用者的要求的不同而不同. 为了解决这一问题, 提出了一种加权多维桶分组技术——WMSB(Weighted Multi-Sensitive Bucketization).

WMSB 的基本思想是根据用户的需求和每个多维桶各个维的取值为每个多维桶指定一个权值, 在分组时综合考虑权值来计算每个桶的加权选择度, 保证在发布的数据中保留更多的重要信息. 权值的指定可由用户对于数据的需求自行设定. 用户可以对每个敏感属性的不同取值指定一个权值. 即对于敏感属性 S_i 的一个取值 s^j , 用户指定权值为 $w_i^{s^j}$, $w_i^{s^j} \in [0, 1]$. 那么对于桶 $buk\langle s^1, s^2, \dots, s^d \rangle$, 其权值记为

$$weight\langle s^1, s^2, \dots, s^d \rangle = \frac{1}{d} \sum_{i=1}^d w_i^{s^i} \tag{4}$$

对于最大桶优先算法, 如果考虑权值, 在分组时不是优先选择最大的非空桶提取记录构成分组, 而是要计算加权选择度 $WSelect(buk\langle s^1, s^2, \dots, s^d \rangle) = size(buk\langle s_0^1, s_0^2, \dots, s_0^d \rangle) \times weight\langle s^1, s^2, \dots, s^d \rangle$, 然后选择 $WSelect$ 值最大的桶优先构成分组. 对于最大维容量优先算法也同样进行这样的改进, 计算加权选择度. 桶 $buk\langle s^1, s^2, \dots, s^d \rangle$ 的加权选择度如式(5)所示.

$$\begin{aligned} WSelect(buk\langle s^1, s^2, \dots, s^d \rangle) = \\ Select(buk\langle s_0^1, s_0^2, \dots, s_0^d \rangle) \times weight\langle s^1, s^2, \dots, s^d \rangle \end{aligned} \tag{5}$$

为了衡量在计算权值后在发布的数据中保留的用户自定义的重要信息的量, 给出了用户自定义重要信息的可发布性(Publishability)进行度量, 如式(6)所示.

$$Publishability = \frac{\sum size'(buk\langle s^1, \dots, s^d \rangle) \times weight(buk\langle s^1, \dots, s^d \rangle)}{\sum size(buk\langle s^1, \dots, s^d \rangle) \times weight(buk\langle s^1, \dots, s^d \rangle)} \tag{6}$$

其中, $size'(buk\langle s^1, s^2, \dots, s^d \rangle)$ 表示将发布数据中保留的元组重新映射到多维桶之后每个桶的大小. 数据拥有者自定义重要信息可发布性, 即为发布数据中包含的用户定义的重要信息量与原始数据中包含的数据拥有者定义的重要信息量的比值. 因此, 用户自定义重要信息可发布性量度越大说明发布的数据中保留的重要信息越多, 发布的数据对于具体用户的可用性就越高.

	Flu	Pneumonia	Gastritis	HIV	Cancer	
John	$\{t_1\}0.1$	$\{t_2\}0.8$			$\{t_3\}1$	3
Bob	$\{t_4\}0.8$	$\{t_5\}0.4$				2
Anne			$\{t_6, t_7\}0.9$			2
Sam				$\{t_8\}0.4$		1
Mary	$\{t_9\}0.6$					1
	3	2	2	1	1	

图 6 加权值的 d 维桶($d=2$)

以最大单维容量优先算法 MSDCF 为例进行简单说明. 图 6 给出每个桶的权值. 按照式(5)计算的加权选择度进行分组, 结果为 $\{t3, t4, t6\}$, $\{t2, t7, t8, t9\}$, 可以看出在附加信息损失度和隐匿率基本不变的情况下, 发布的数据结果中包含了数据提供者认为更重要记录 $t2$ 代替基本的非加权方法发布的 $t5$. 加权与非加权方法的自定义重要信息的可发布性分别为 0.915 和 0.847. 可见加权多维桶分组方法比基本的非加权方法发布数据对于用户来说具有更高的可用性.

5 实验结果及分析

采用实际数据集对本文提出的算法进行大量实验测试, 给出以下实验结果, 并进行对比分析. 采用

UCI machine learning repository 的人口统计数据
集进行实验测试. 实际数据集来自 <http://kdd.ics.uci.edu>. 对原数据集过滤掉不完整记录, 并进行数
据格式转换后, 随机提取 10K (1K=1000) 记录, 并
选择 5 个属性作为敏感属性 (见表 1). 实验的硬件
环境为 Intel Pentium4 2.4GHz CPU, 512MB 内
存; 操作系统平台为 Microsoft Windows XP Pro-
fessional; 编程环境为 Microsoft Visual C++ 6.0
编译器.

表 1 实验数据集信息

敏感属性	Occupation	Salary-class	Work-class	Education	Race
基数	50	50	10	17	9

表 2 实验中采用的复合敏感属性

复合敏感 属性维数 (敏感属性 个数)	复合敏感属性
CS =2	⟨Occupation, Salary-class⟩
CS =3	⟨Occupation, Salary-class, Work-class⟩
CS =4	⟨Occupation, Salary-class, Work-class, Education⟩
CS =5	⟨Occupation, Salary-class, Work-class, Education, Race⟩

算法发布数据的安全性由复合敏感属性 l -多样
性性质保证, 因此发布的数据一定是安全的. 通过附
加信息损失度和隐匿率来衡量基本的多维桶分组算
法发布数据的信息损失, 从而评价算法的性能. 附加
信息损失度越小说明发布的数据中由于有损连接造
成的信息损失越小. 隐匿率越小则说明由隐匿数据
记录造成的信息损失越小. 附加信息损失度和隐匿
率都为 0 时为最优的结果. 因此算法发布的数据的
附加信息损失度和隐匿率越小就越接近最优结果,
算法发布数据的质量就越好. 对于改进的加权多敏
感属性隐私数据发布算法, 通过数据所有者定义的
重要信息可发布性来衡量算法发布数据的可用性.

实验主要从以下几个方面对算法的各个性能指
标进行比较分析: (1) 变化的数据集大小 (数据量取
1K~10K); (2) 变化的多样性参数取值 (l 取值
为 2~9); (3) 变化的敏感属性个数. 通过选择不同
的参数 (d 的取值为 2~5). 测试不同算法发布数据
的附加信息损失度、隐匿率和用户自定义重要信息
可发布性, 并对算法运行时间等进行分析. 表 2 给出
了实验采用的不同敏感属性集描述.

5.1 基本多敏感属性隐私数据发布算法性能分析

首先通过大量实验测试前 3 种不同的多敏感属
性隐私数据发布算法 MBF, MSDCF 和 MDDCF 的
附加信息损失度和隐匿率, 分析算法发布数据的信
息损失, 从而评价算法的性能.

测试不同数据集中的数据量对附加信息损失度
和隐匿率的影响. 图 7 给出了 l 取值为 3, 敏感属性
个数 $d=3$ 时的实验结果. 由图 7 可以看出, 多敏感
属性隐私数据发布算法 MBF, MSDCF 和 MDDCF
对于不同大小的数据集, 附加信息损失度都不超过
0.15, 而隐匿率也不大于 6.5%. 而且当数据量超过
5K 时附加信息损失度都在 0.03 以下. 这说明算法
MBF, MSDCF 和 MDDCF 发布数据的信息损失都
比较小. 从图中可以观察到, 附加信息损失度与隐
匿率都随着数据量的增大而减小. 当数据量增大到
8K 时, 算法 MBF, MSDCF 和 MDDCF 的隐匿率都
为 0, 也就是没有数据被隐匿. 这是由于随着数据量
的增大, 数据记录的各个敏感属性取值的多样化程
度越来越好, 分桶更加均匀, 使得分组的效果逐渐变
好, 需要隐匿的记录个数也逐渐减少. 另外, 从图 7
中还可以看出, 对于不同的数据集, 无论是附加信息
损失度还是隐匿率, 最大单维容量优先 MSDCF 算
法都要小于最大桶优先 MBF 算法, 而最大多维容
量优先 MDDCF 算法小于最大单维容量优先 MSD-
CF 算法. 特别是, 图 7(a) 显示 MDDCF 算法的附加
信息损失度一直都很接近于 0, 且图 7(b) 显示 MM-
DCF 算法的隐匿率也都很小 (不超过 1.5%), 说明综
合考虑每一维容量进行优先选择的最大多维容量优
先算法在只损失很少的数据记录的情况下, 发布数
据的分组大小近于最优, 算法具有很好的性能. 因
此, 由实验结果可以看出 3 种算法都能保证发布较
高质量的数据, 且对于发布数据的质量来说, MBF,
MSDCF 和 MDDCF 算法性能逐步提高.

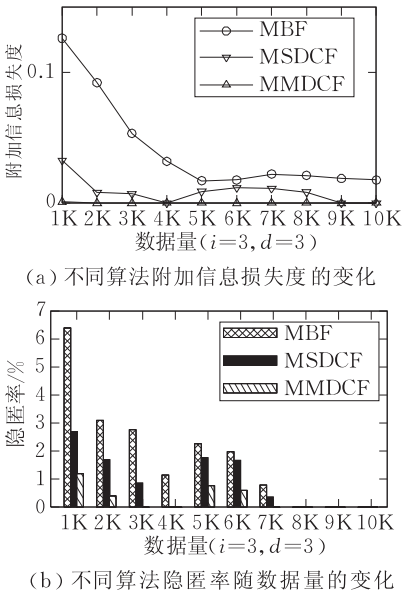


图 7 不同算法的附加信息损失度和隐匿率随数据量的变化

实验进而分析了多样性参数取值的变化对不同算法的附加信息损失度和隐匿率的影响。实验选择数据量为 5K,复合敏感属性维数为 3 的数据集进行测试。结果如图 8 所示。附加信息损失度和隐匿率几乎都随着 l 值增大而增大,当 l 取值不超过 4 时,算法 MBF,MSDCF 和 MDDCF 的附加信息损失度不大于 0.02,隐匿率小于 5%,算法具有很好的性能。但是,当 l 取值大于 5 时算法的隐匿率和 MBF 算法的附加信息损失度都迅速增加。这是由于实验中数据集涉及到的 3 个敏感属性中,不同取值的个数最少的是 Work-class 属性,为 10。 l 的取值越接近于

这个值,在这一维上保证分组的 l -多样性越困难,这样就使得整体的分组效果显著变坏。但是,从图 8 中还可以观察到,即使 l 取值不断增大,接近于最小的不同属性值个数,最大单维容量优先的 MSDCF 算法和最大多维容量优先的 MMDCF 算法的附加信息损失度始终保持在一个很小的水平,接近于最优。并且它们的隐匿率也都小于 MBF 算法。而且,MMDCF 算法在两个衡量标准上都略优于 MSDCF 算法。说明即使在 l 取值不好的情况下,综合考虑的信息越多的算法发布数据的信息损失越小。

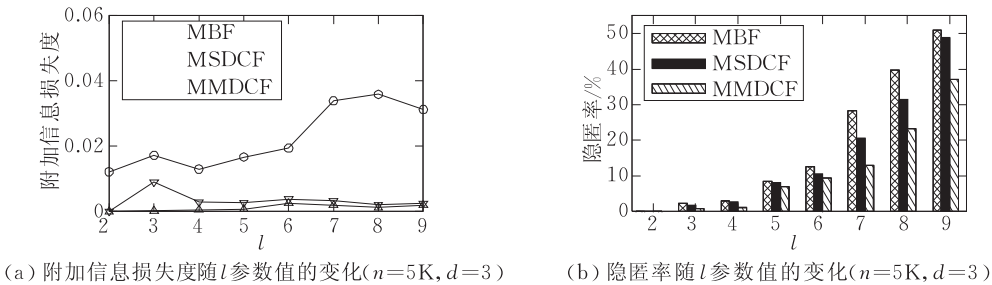


图 8 不同算法的附加信息损失度和隐匿率随 l 参数取值的变化

实验测试了具有不同敏感属性个数的数据集对不同算法附加信息损失度和隐匿率的影响。图 9 给出了 5K 大小的数据集,多样性参数 l 取值为 3 时的实验结果。从图中可以看出,对于不同敏感属性个数的数据集,算法 MBF,MSDCF 和 MDDCF 的附加信息损失度都小于 0.03,且隐匿率都不超过 3%。尤其是当敏感属性个数为 2 时,算法 MBF,MSDCF 和 MDDCF 的附加信息损失度和隐匿率都为 0,得到最优的分组结果。这说明对于具有不同敏感属性个数的数据集,算法 MBF,MSDCF 和 MDDCF 都具有较好的性能。另外,从图 9 中还可以观察到,随着

复合敏感属性维数的增加,MBF 算法的附加信息损失度和 3 种算法的隐匿率都有增大的趋势。这是由于敏感属性的个数越多,即复合敏感属性的维数越高,得到在每一维上都满足 l -多样性的分组越困难。然而,对于最大单维容量优先的 MSDCF 算法和最大多维容量优先的 MMDCF 算法,复合敏感属性维数的增大对其分组效果的影响并不显著。这是由于这两种算法都不只单独考虑单个桶的容量来决定分组的策略,而是在不同程度上综合考虑了其它维上整体的容量进行分组选择,这样就降低了由于维数增加而导致的对分组效果的影响。

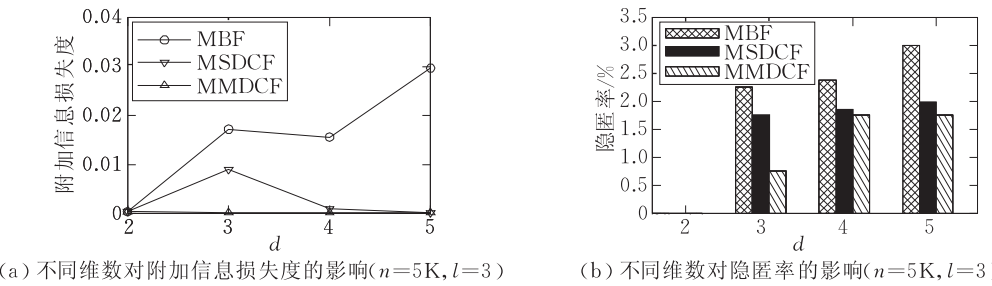


图 9 不同复合敏感属性维数对不同算法附加信息损失度和隐匿率的影响

5.2 基本多敏感属性隐私数据发布算法执行时间分析

图 10 给出了不同数据集和不同参数下算法 MBF,MSDCF 和 MDDCF 的执行时间。图 10(a)为不同的数据集大小对于算法运行时间的影响。从

图 10 中可以看出,算法执行时间随着数据量的增大呈近似线性的增长。另外,对于相同的数据集,算法 MBF,MSDCF 和 MDDCF 的执行时间逐渐增加。这是由于为了得到较好的分组结果,算法的选择策略的计算量逐渐增加。图 10(b)为 l 取值不同时 3 种算

法的运行时间. 由于数据量相同, 且 l 取值并不影响多维桶的结构, 因此 l 取值对算法的执行时间影响不大. 图 10(c) 给出了复合敏感属性维数不同时算法的执行时间. 从图中可以看出, 算法的执行时间都随着复合敏感属性维数的增大而增大. 这是由于复合

敏感属性的维数越多, 多维桶的个数越多, 算法对每个桶计算选择度的执行时间就越长. 另外, 图 10(c) 显示对于相同数据集 MBF 算法、MSDCF 算法与 MDDCF 算法的执行时间逐步增加. 这与图 10(a) 中的结论一致.

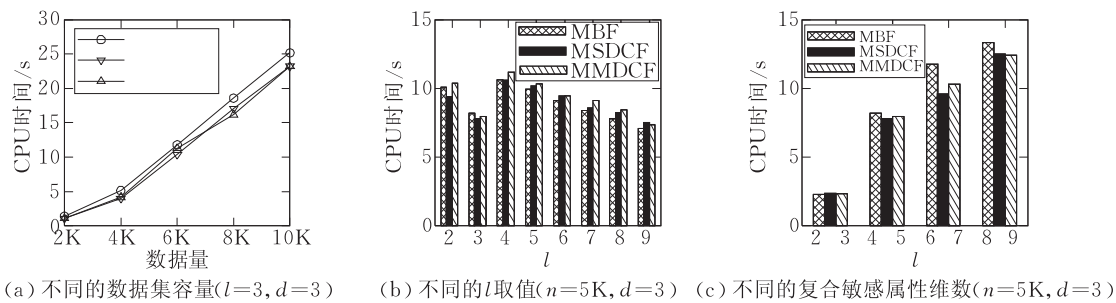


图 10 不同算法运行的 CPU 时间

5.3 加权多敏感属性隐私数据发布算法性能分析

我们针对加权多敏感属性数据发布方法的效果进行了实验分析. 实验通过随机数生成器对每个桶生成模拟权值, 模拟数据拥有者自定义的数据的重要性. 测试改进加权多敏感属性隐私数据发布算法对于用户自定义重要数据信息的保留程度. 在所提出的 3 种算法 MBF, MSDCF 和 MDDCF 的基础上, 引入加权选择度, 得到相应的 3 种加权算法, 分别记为 WMBF, WMSDCF 和 WMDDCF.

图 11 为加权多敏感属性隐私数据发布算法和基本的非加权多敏感属性隐私数据发布算法的附加信息损失度和隐匿率的比较. 从图 11 中可以看出,

加权的多维桶分组算法的附加信息损失度和隐匿率的值都接近基本的非加权算法. 可见, 单纯从分组的效果角度考虑, 加权算法同样具有很好的性能. 但是, 由于加权算法在计算选择度并进行分组时要考虑到桶中信息对于数据拥有者的重要性, 即考虑到数据拥有者定义的权值, 而不是单纯的考虑到趋近最优的分组策略. 在某些情况下, 加权算法的附加信息损失度和隐匿率略大于基本的非加权算法. 但是, 并不说明加权算法发布数据的可用性低于非加权算法, 下面给出不同算法对于重要信息可发布性的测试结果, 并加以分析.

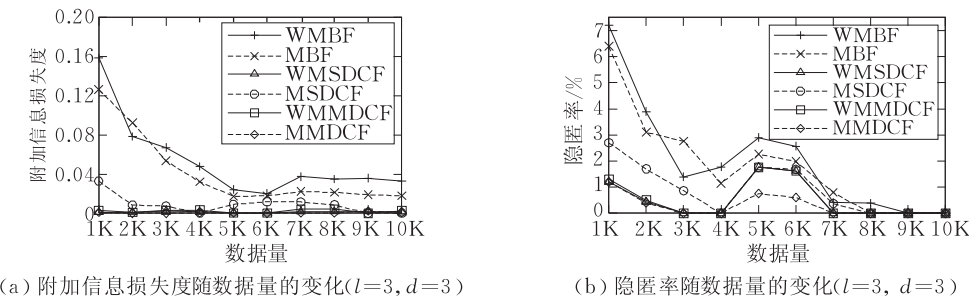


图 11 加权算法和非加权算法的附加信息损失度和隐匿率比较

图 12 为 3 种基本多敏感属性隐私数据发布算法 MBF, MSDCF 和 MDDCF 与其对应的加权算法对于发布数据的可发布性的比较. 通过选择不同的数据量 (见图 12(a)~(c))、不同的 l 取值 (图 12(d)~(f))、不同的多敏感属性个数 (图 12(g)~(i)) 等参数进行测试, 每种加权算法的重要信息可发布性的值都不小于 0.6, 进一步说明本文提出的改进加

权算法对于不同需求的用户, 发布的数据具有较高的可用性. 且从图 12 中还可看出, 加权算法的重要信息可发布性都要高于基本的非加权算法. 这说明加权算法发布的数据能更好地保留数据拥有者自定义的重要信息. 从而, 使得发布的数据对于实际应用具有更高的实用性.

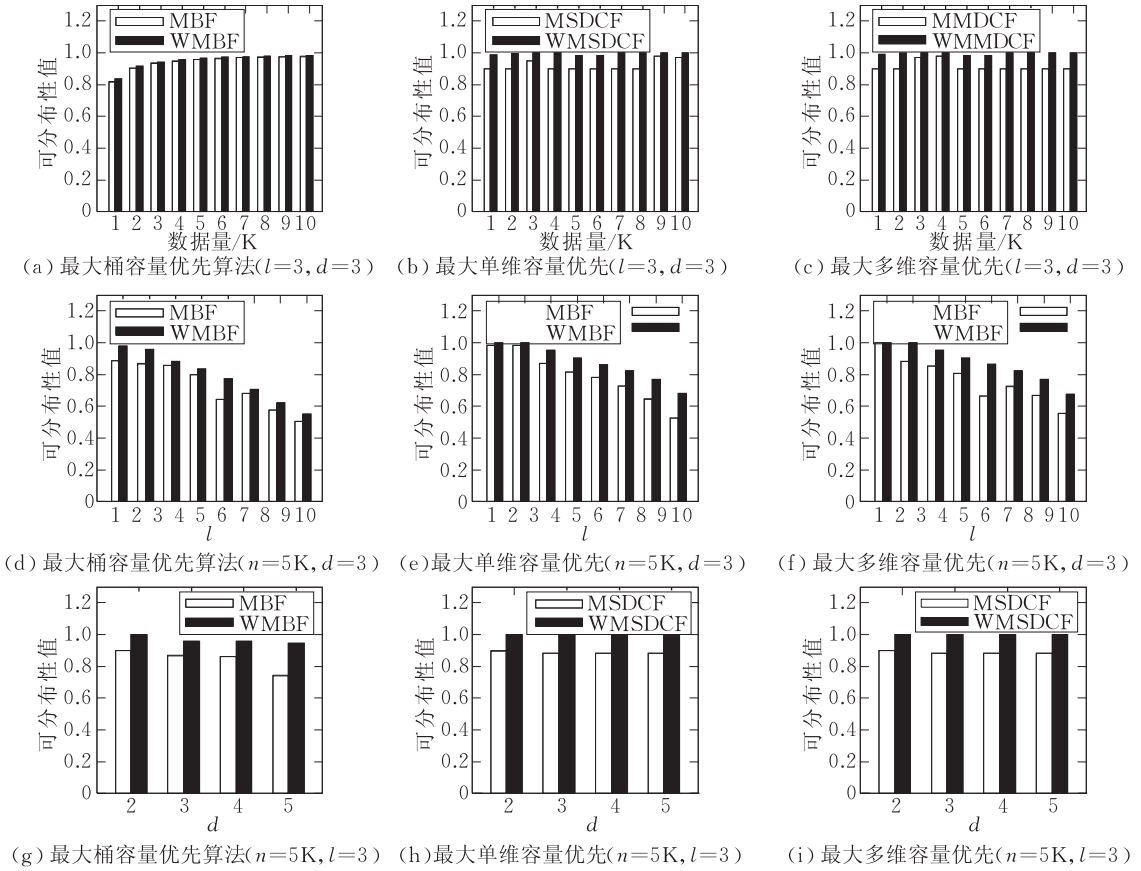


图 12 加权算法和非加权算法用户自定义重要信息可发布性比较

6 结论及未来工作

隐私信息的安全性是数据发布与共享环境中面临的重要问题. 现有的隐私数据发布技术通常只针对具有单一敏感属性的数据, 对于现实中大量存在的多敏感属性数据却无法保证其中隐私信息的安全. 本文详细研究了多敏感属性隐私数据发布问题, 提出了一种基于有损连接的多维桶分组技术. 该技术适用于具有任意多个敏感属性的隐私数据的安全发布问题. 提出了基于贪心策略的分组技术, 并给出了 3 种采用不同的贪心策略的具体算法, 包括最大桶优先 MBF 算法、最大单维容量优先 MSDCF 算法和最大多维容量优先 MMDCF 算法. 另外, 本文还讨论了在实际应用中可能面临的数据对于不同的数据拥有者的重要程度不同的问题, 给出了基于 3 种算法 MBF, MSDCF 和 MDDCF 的加权多维桶分组技术. 在实际数据集上进行大量的实验, 结果表明在保证多敏感属性数据中隐私信息安全性的前提下, 3 种算法 MBF, MSDCF 和 MDDCF 的附加信息损失度和隐匿率都比较小, 发布的数据质量较高. 在此

基础上派生的加权多维桶分组技术能更好地满足不同数据拥有者对于发布数据的需求.

在今后的工作中将继续研究隐私数据发布和共享问题中数据的各种属性之间的关联信息对于发布数据结果的影响, 并研究关系数据库中属性间存在的数据依赖对于隐私数据发布问题的影响.

参 考 文 献

[1] Sweeney L. *K*-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 2002, 10(5): 557-570

[2] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information//*Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Seattle, Washington, 1998: 188

[3] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation//*Proceedings of the 32nd International Conference on Very Large Data Bases*. Seoul, Korea, 2006:139-150

[4] Machanavajhala A, Gehrke J, and Kefer D. *l*-diversity: Privacy beyond *k*-anonymity//*Proceedings of the 22nd International Conference on Data Engineering*. Atlanta, Georgia,

2006; 24

- [5] Bayardo R, Agrawal R. Data privacy through optimal k -anonymization//Proceedings of the 21st International Conference on Data Engineering. Tokyo, Japan, 2005; 217-228
- [6] LeFevre K, DeWitt D, Ramakrishnan R. Incognito: Efficient full-domain k -anonymity//Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, 2005; 49-60
- [7] Meyerson A, Williams R. On the complexity of optimal k -anonymity//Proceedings of the 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Paris, France 2004; 223-228
- [8] LeFevre K, DeWitt D, Ramakrishnan R. Mondrain multidimensional k -anonymity//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, Georgia, 2006; 25
- [9] Park H, Shim K. Approximate algorithms for k -anonymity//Proceedings of the ACM SIGMOD International Conference on Management of Data. Beijing, China, 2007; 67-78
- [10] Sweeney L. Achieving k -anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002, 10(5); 571-588
- [11] Winkler W. Using simulated annealing for k -anonymity. Washington D. C: Census Bureau, Statistical Research Division, Technical Report Statistics; 2002-7, 2002
- [12] Fung B, Wang K, Yu P. Top-down specialization for information and privacy preservation//Proceedings of the 21st International Conference on Data Engineering. Tokyo, Japan, 2005; 205-216
- [13] Aggarwal C. On k -anonymity and the curse of dimensionality//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway, 2005; 901-909
- [14] Wong R, Liu Y, Yin J, Huang Z, Fu A, Pei J. (α , k)-Anonymity based privacy preservation by lossy join//Proceedings of the Advances in Data and Web Management, Joint 9th Asia-Pacific Web Conference, APWeb 2007, and 8th International Conference, on Web-Age Information Management, WAIM 2007, Huangshan, Anhui, China, 2007; 733-744
- [15] Yang X C, Liu X Y, Wang B, Yu G. K -anonymization approaches for supporting multiple constraints. Journal of Software, 2006, 17(5): 1222-1231
(杨晓春, 刘向宇, 王斌, 于戈. 支持多约束的 K -匿名化方法. 软件学报, 2006, 17(5): 1222-1231)
- [16] Pei J, Xu J, Wang Z. B, Wang W, Wang K. Maintaining k -anonymity against incremental updates//Proceedings of the 19th International Conference on Scientific and Statistical Database Management. Banff, Canada, 2007; 5
- [17] Lwuchukwu T, Naughton J. K -anonymization as spatial indexing: Toward scalable and incremental anonymization//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007; 746-757
- [18] Wong R, Fu A, Wang D, Pei J. Minimality attack in privacy preserving data publishing//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007; 543-554



YANG Xiao-Chun, born in 1973, Ph. D., associate professor. Her current research interests include database theory, data privacy, and data quality.

WANG Ya-Zhe, born in 1984, M. S. candidate. Her current research interests include database and data privacy.

WANG Bin, born in 1972, Ph. D. candidate. His current research interests include query processing, data quality.

YU Ge, born in 1962, professor, Ph. D. supervisor. His current research interests include distributed database system, Web services, and data stream.

Background

This research is supported by Program for New Century Excellent Talents in University under grant No. NCET-06-0290, the National Natural Science Foundation of China under grant No. 60503036, and Fok Ying Tung Education Foundation under grant No. 104027.

This paper focuses on the field of privacy preserving data publishing. The research group has done much research work in designing high efficient and practical privacy preserving data publishing algorithms and other techniques in data privacy, such as privacy preserving data outsourcing, loca-

tion privacy, and so on.

Privacy preserving data publishing problem is an important branch of data privacy. The problem of linking-attack, one of the main cause of revealing private information, was firstly formalized by L. Sweeney and P. Samaranti. And they presented the k -anonymity model to prevent such linking-attack. Generalization and suppression are the general way to achieve k -anonymity. Then many work have been done to design efficient, scalable, and flexible k -anonymity algorithms, which balance the trade-off between privacy and data usability.

ty as well as possible. The classical k -anonymity algorithms include Bayardo-Agrawal's k -Optimize algorithm, Incognito algorithm, Mondrain algorithm, and so on. There are also lots of researches focus on the ineffectiveness of the k -anonymity model under some circumstances. They present several models as complementary to the traditional k -anonymity model, such as l -diversity, t -closeness, m -confidentiality, and so on. Recently, there presented several lossy-join based methods. The most important advantage of those methods is that it can assure the accuracy of the published data than those generalization and suppression based techniques. How-

ever, all the current privacy preserving data publishing techniques concentrate on table with only one sensitive attribute. As far as we are concerned, most of the real-world applications contain multiple sensitive attributes. The existing single-sensitive-attribute privacy preserving techniques cannot guarantee the security of the sensitive information with multiple sensitive attributes. In this work, the authors firstly illustrate the multi-sensitive-attribute private data publishing problem and present several efficient algorithms to solve the problem.

全国第 15 届计算机辅助设计与图形学(CAD/CG'2008)学术会议征文通知

2008 年 7 月 22 日~24 日 中国·大连

由中国计算机学会主办、辽宁师范大学承办的全国第 15 届计算机辅助设计与图形学学术会议(CAD/CG'2008)将于 2008 年 7 月 22 日在中国大连举行. 本次学术会议是中国计算机学会恢复学术活动 30 周年的纪念, 也是计算机辅助设计与图形学专业委员会学术年会 30 周年纪念.

本次会议内容包括中国计算机学会计算机辅助设计与图形学专业委员会恢复学术活动 30 周年纪念座谈、大会学术报告、计算机辅助设计与图形学热点问题专题研讨、最新成果和应用系统演示, 并将邀请国内外学术界和产业界的著名专家学者到会作特邀报告.

会议录用的部分优秀论文将推荐至《计算机学报》、《计算机研究与发展》、《计算机辅助设计与图形学学报》、《工程图学学报》、《软件学报》(增刊)、《中国图象图形学报》、《系统仿真学报》发表. 大会录用论文将由正式出版社出版. 热诚欢迎一切从事计算机辅助设计与图形学研究、应用及软件开发的专家、学者和专业技术人员踊跃投稿.

- 截稿日期: 2008 年 4 月 20 日
- 投稿邮箱: cadcg2008@lnnu.edu.cn 和 cadcg20008@gmail.com
- 会议网址: www.cadcg2008.lnnu.edu.cn
- 电话/传真: 0411-82158874
- 联系人: 孙晓鹏 博士