ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.4, April 2010, pp.771−781
doi: 10.3724/SP.J.1001.2010.03466

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

隐私保护数据发布中身份保持的匿名方法*

童云海, 陶有东+, 唐世渭, 杨冬青

(机器感知与智能教育部重点实验室(北京大学),北京 100871)

Identity-Reserved Anonymity in Privacy Preserving Data Publishing

TONG Yun-Hai, TAO You-Dong+, TANG Shi-Wei, YANG Dong-Qing

(Key Laboratory of Machine Perception of Ministry of Education (Peking University), Beijing 100871, China)

+ Corresponding author: E-mail: taoyd@pku.edu.cn

Tong YH, Tao YD, Tang SW, Yang DQ. Identity-Reserved anonymity in privacy preserving data publishing. *Journal of Software*, 2010,21(4):771−781. http://www.jos.org.cn/1000-9825/3466.htm

Abstract: In the research of privacy preserving data publishing, the present method always removes the individual identification attributes and then anonymizes the quasi-identifier attributes. This paper analyzes the situation of multiple records one individual and proposes the principle of identity-reserved anonymity. This method reserves more information while maintaining the individual privacy. The generalization and loss-join approaches are developed to meet this requirement. The algorithms are evaluated in an experimental scenario, reserving more information and demonstrating practical applicability of the approaches.

Key words: privacy preservation; data publishing; anonymity; identity-reserved; lossy join; generalization

摘 要: 在隐私保护的数据发布研究中,目前的方法通常都是先删除身份标识属性,然后对准标识属性进行匿名处理.分析了单一个体对应多个记录的情况,提出了一种保持身份标识属性的匿名方法,它在保持隐私的同时进一步提高了信息有效性.采用概化和有损连接两种实现方式.实验结果表明,该方法提高了信息有效性,具有很好的实用性.

关键词: 隐私保护;数据发布;匿名;身份保持;有损连接;概化

中图法分类号: TP309 **文献标识码**: A

当今社会,个人的信息记录被不同的政府部门或者机构广泛地收集和分析.为了有利于数据分析,有些机构会发布这些涉及到个人数据的信息.这些数据在发布时,一方面要保护个人隐私不被泄露,另一方面又要具备足够的信息供分析使用.近年来,数据发布中的隐私保护方法从信息技术方面得到了进一步广泛的研究,并对隐私保护的程度和泄露都作了一定的定量研究.数据发布中的隐私保护试图在保护个体隐私的同时保存更多的信息有效性.*K*-匿名[1−3]是其中的核心思想.

待发布的数据表通常包含 3 类属性:(1) 个体标识属性(individually identifying attribute,简称 ID),包括可以显式表明个体身份的属性,比如姓名、身份证号码和手机号码.(2) 准标识属性(quasi-identifier attribute,简称 QI),可以用于链接攻击的属性,并可用于表明数据保护的程度,比如性别、年龄和邮政编码.(3) 敏感属性(sensitive

* Supported by the National Natural Science Foundation of China under Grant No.60403041 (国家自然科学基金)
Received 2008-01-09; Revised 2008-06-11; Accepted 2008-08-28

attribute,简称 ST),描述个体隐私的细节信息,比如疾病和收入.例如,表 1 中的性别和邮政编码构成了准标识属性组,姓名是个体标识属性,疾病是敏感属性.如果发布数据表仅仅简单删除了个体标识属性,隐私信息仍然有可能被推理获得,因为某些准标识属性组的取值是唯一的.如果攻击者具有一个外部数据库,包含了个体的准标识属性取值,他可以通过连接两个表来发现敏感属性取值.我们由宁波城区的人口数据发现,78%的人可以被出生日期和 6 位邮政编码唯一确定.

具有相同准标识属性取值的元组组成了一个准标识属性分组(简称 QI 分组).为了达到匿名效果,*k*-匿名模型要求每个 QI 分组中至少包含 *k* 个元组,可以通过概化元组的准标识属性取值来满足 *k*-匿名要求.概化就是把较为准确的取值替换为较为概要的取值,比如把出生日期“1980 年 5 月 20 日”替换为“1980 年 5 月”,或者进一步替换为“1980 年”.例如,对表 1,如果我们要求 2-匿名,那么头 2 行就可以构成一个 QI 分组.元组 3 和元组 4 通过概化性别属性组成了一个 QI 分组,元组 5 和元组 6 通过概化邮编属性构成了一个 QI 分组.这样得到了匿名结果表 2,每个 QI 分组中都有两个元组,从而满足了 2-匿名(事实上,如果按照 *l*-多样化模型,表 2 同样满足了 2-多样化[4]).

Table 1 A patient table in which someone appears more than once

表 1 存在单一个体对应多个记录的病人情况表

No.	Name	Gender	Postcode	Disease
1	Mike	M	10085	Hypertension
2	Mike	M	10085	Hyperlipemia
3	Emily	F	10075	Diabetes
4	Tim	M	10075	Heart
5	Jane	F	10086	Cancer
6	Ella	F	10087	Flu

Table 2 Published table of Table 1 (satisfying 2-anonymity and 2-diversity) after common generalization

表 2 对表 1 概化匿名处理后的发布表(满足 2-匿名和 2-多样化)

Group ID	Sex	Postcode	Disease
1	M	10085	Hypertension
1	M	10085	Hyperlipemia
2	*	10075	Diabetes
2	*	10075	Heart
3	F	1008*	Cancer
3	F	1008*	Flu

现有的匿名模型,比如 *k*-匿名和 *l*-多样化,通常先删除个体身份属性,然后概化准标识属性来满足匿名要求.这些匿名模型通常假定单一个体在数据表中最多只出现一次.但在实际环境中,这个前提可能是难以满足的.比如在某个医院发布的病人情况表中,某些病人由于同时患有不同疾病,因而出现了多次.在表 1 中,Mike 同时患有高血压和高血脂,因此出现了 2 次.在表 2(表 1 的匿名结果表)中,如果一个攻击者发现 Mike 出现在 QI 分组 1 中,则其会推理 Mike 患有高血压或者高血脂.实际上,此时,这两种推测都是正确的.因此这时的隐私泄漏概率是 100%.这就是一种发生在单一个体对应多条记录情况下的隐私保护程度的降低现象.而且,身份信息的删除丢失了同一个人多个敏感属性之间可能的关联信息,而这种信息在很多研究中是非常重要的,比如医疗中并发症的研究.

为了避免这种隐私泄漏,在本文中我们分析了单一个体对应多个记录的情况,提出了身份保持的匿名模型.保留身份信息可以提高信息有效性.比如在表 2 中,保留身份信息可以帮助研究人员发现不同疾病的并发情况.我们采用概化和有损连接两种方式实现身份保持的匿名要求,并在实验环境中进行了检验.本文工作的贡献在于以下两点:

(1) 分析了单一个体对应多个记录的情况,提出了一种身份保持的匿名模型.针对具体的匿名要求,提出了 3 种身份保持的匿名要求.

(2) 给出了概化和有损连接两种实现方式,有效地满足了身份保持的匿名要求.在实验中检验了原有匿名模型在单一个体对应多个记录情况下的脆弱性,然后比较了原有匿名模型和身份保持的匿名模型的信息损失,最后比较了身份保持的匿名模型在两种实现方式下的有效性.

本文第 1 节回顾相关工作.第 2 节提出身份保持的匿名模型.第 3 节给出概化和有损连接两种实现方法.第 4 节给出实验结果.第 5 节给出结论.

1 相关工作

2001 年以来,隐私保护的数据发布得到了广泛的重视和研究.Samarati 和 Sweeney 提出了 k-匿名模型[1−3].它要求发布表中的每个元组都至少与其他(k–1)个元组在准标识属性上完全相同.Machanavajjhala 等人进一步提出了 l-多样化模型[4].它要求每个 QI 分组中至少包含 l 个多样化的敏感属性取值.这一模型扩展了 k-匿名模型,对敏感属性的多样化提出了要求,并对敏感属性的多样化给出了多种不同的解释.Wong 等人提出了(σ,k)-匿名模型[5].它在 k-匿名模型的基础上,要求每个 QI 分组中每个敏感属性的取值频率不能超过给定的值σ.Li 等人提出了 t-接近模型,它要求 QI 分组中每个敏感属性取值的分布较为接近[6].这些模型都是首先删除身份标识属性,然后对准标识属性进行概化.Xiao 和 Tao 提出了一种个性化的匿名模型[7].在分析隐私泄漏的概率时,他们区分了两种情况:一种称为主键场景,其中单一个体对应的元组不能超过一个;另一种称为非主键场景,其中单一个体可以对应任意多个元组.这种非主键场景类似于本文讨论的情况.

匿名的概化实现算法上有全局重编码和局部重编码两类.在全局重编码[1,2,8,9]中,每个准标识属性的取值要求概化到概念层次树的同一层次.但是全局重编码通常会过度概化,从而损失了更多的信息.局部重编码[2,10]放弃了这一要求,准标识属性的取值可以概化到不同层次.比如表 2 就是表 1 的一个局部重编码结果,有些元组的性别属性概化为(*),而有些元组的性别属性则保持不变.

最近的研究提出可采用有损连接方法来达到匿名要求[11,12].这种方法发布两个表,两个表之间通过一个非主键的分组标识属性加以连接.它利用有损连接的方式来保护隐私.

2 身份保持的匿名模型

个体标识属性一般由信息发布者指定.通常的匿名模型首先删除身份属性.而身份保持的匿名首先对个体身份属性进行重编码,然后在发布表中保留编码后的个体属性.可以简单地用数字来替换原来有意义的身份属性取值.保留这样的个体信息可以提高信息有效性.比如在表 2 中,保留个体信息可以帮助研究人员发现不同疾病的并发情况.

在 k-匿名模型中,每个 QI 分组包含至少 k 个元组.类似地,在身份保持的 k-匿名中,每个 QI 分组至少包含 k 个不同的个体.

定义 1(身份保持的 k-匿名需求). 在数据发布集中,任何一个准标识属性分组中至少包含 k 个不同的个体.

事实上,这个定义与 Samarati 在文献[1]中给出的要求相一致.但在文献[1]中,它默认每个个体最多只对应一个元组,因此它首先删除了身份属性.由于对身份属性进行了重编码,我们在个体的基础上重新定义了该需求.

在 k-匿名中,发布数据表的模式是 $T(QI,ST)$.而在本文中,发布数据表的模式是 $T(ID,QI,ST)$.ID 是重编码后的身份属性.设发布的数据表为 T,$A=\{a_1,a_2,\ldots,a_b\}$是 T 中的个体集合,$S=\{s_1,s_2,\ldots,s_t\}$是敏感属性取值集合.对于每个 a_i,设 $S(a_i)$是个体 a_i 对应的所有元组中的敏感属性取值集合.对于每个 s_j,设 $A(s_j)$是敏感属性取值为 s_j 的所有元组对应的个体集合.在 QI 分组 Q 中,令 $m=\left|\bigcup_{a_i\in Q.ID} S_{a_i}\right|$,$n=\left|\bigcup_{s_j\in Q.ST} A_{s_j}\right|$.

根据不同的隐私保护需求,我们提出 3 种不同的身份保持的匿名需求.

定义 2(身份保持的 k-匿名). 考虑数据表 D,包含身份属性、准标识属性和敏感属性.发布数据表 D'从 D 转化而来.D'称为满足身份保持的 k-匿名,如果 D'中任何一个 QI 分组中至少包含 k 个不同的个体.即在任何 QI 分组 Q 中,$n\geq k$.

定义 3(身份保持的(*k*,*l*)-匿名). 考虑数据表 *D*,包含身份属性、准标识属性和敏感属性.发布数据表 *D′*从 *D* 转化而来.*D′*称为满足身份保持的(*k*,*l*)-匿名,如果 *D′*中任何一个 QI 分组中至少包含 *k* 个不同的个体和 *l* 个不同的敏感属性取值,即 $m \geq l$ 并且 $n \geq k$.

定义 4(身份保持的(α,β)-匿名). 考虑数据表 *D*,包含身份属性、准标识属性和敏感属性.发布数据表 *D′*从 *D* 转化而来.*D′*称为满足身份保持的(α,β)-匿名,如果 *D′*中任何一个 QI 分组中任何个体的元组比例不超过α,任何敏感属性取值比例不超过β,其中,$0<\alpha,\beta<1$.

这 3 个原则分别类似于 *k*-匿名、*l*-多样化和(α,*k*)-匿名.(α,*k*)-匿名考虑了敏感属性取值的频率,它要求:在 QI 分组中,每个敏感属性取值概率不能超过α,并且 QI 分组中至少有 *k* 个元组.引入个体属性之后,身份保持的(α,β)-匿名要求 QI 分组中每个个体的概率不能超过α,并且每个敏感属性取值的概率不能超过β.由于我们要求了个体的取值概率,因此,在(α,β)-匿名中可以取消参数 *k*.

身份保持的匿名方法考虑了单一个体对应多个记录的情况,我们定义单一个体的平均元组数来评价个体的重复程度,记为 *rpi*=(the size of dataset)/(the number of individuals).如果 *rpi*=1,那么每个个体出现 1 次,可应用普通的匿名方法来处理;如果 *rpi*>1,那么应用本文的方法更为合适.本文的方法保持了同一个体不同敏感属性取值之间的关系,而采用普通的匿名方法丢弃了这类信息.

3 匿名方法

概化和有损分解是达到匿名的两种方法[1,12].在本节中,我们分别用这两种方法实现身份保持的匿名.

Table 3 A patient table for publish in which someone appears more than once

表 3 存在单一个体对应多个记录的病人情况表

Tuple-No	ID	Zip	Disease
1	1	10085	Hypertension
2	1	10085	Hyperlipemia
3	2	10086	Diabetes
4	3	10087	Heart
5	4	10075	Hypertension
6	4	10075	Diabetes
7	5	10076	Heart
8	6	10077	Flu
9	7	10050	Heart

3.1 概化方法

概化方法对准标识属性进行概化,得到一张关系表,其中准标识属性取值相同的元组形成一个 QI 分组.为了减小匿名表的概化层次,可以采用允许适量的元组抑制(suppression)的方法,但是元组抑制同样减少了发布表的信息有效性.

概化方法分为全局重编码和局部重编码两类.全局重编码要求同一个准标识属性上的取值都概化到概化层次树的相同层次.文献[9]中提出了全局重编码算法,它同样可以应用到身份保持的概化方法中.这种算法较为简单,它每次选择一个准标识属性进行概化,直到满足匿名需求.全局重编码通常会过度概化.

局部重编码不要求同一属性的取值概化到相同层次.Wong 等人提出了一种自顶向下的局部重编码算法[5].这种方法首先将所有元组都概化到一个全概化分组中,然后在维持匿名原则的基础上,对元组进行特化(specialization).这样循环下去,直到特化会破坏匿名需求为止.本文提出了一种自底向上的局部重编码算法.我们首先检查所有元组,将满足匿名原则的元组添加上分组标号.然后选择一个准标识属性进行概化,检查还未分组的元组,将满足匿名要求的元组添加上组标号.这一过程重复下去,直到所有元组都被添加了组标号或者剩余元组的数目达不到匿名要求(比如当 *k*=7 时,剩余 5 个元组)为止.这些剩余的元组,我们称其为“孤儿元组”.对孤儿元组,我们采用迁移和合并的方法进行处理.如果其他分组中有多余的元组,可以移出多余的元组加入孤儿元组使其符合匿名要求.如果其他分组中没有多余元组,则可简单地将孤儿元组合并到临近的 QI 分组中.关于这一算法的描述见算法 1.

算法 1. 自底向上的局部重编码概化算法.

输入:待发布表 *T*;

输出:发布表 *PT*.

1. *PT* 为对 *T* 进行身份属性重编码的结果表

2. 检查 *PT*,将满足身份保持的匿名要求的元组添加分组标号

3. While (*PT* 中没有分组标号的元组数>0) and (准标识属性组没有概化到最高层次) do

3.1 选择一个准标识属性;

3.2 对剩余元组的选定属性进行概化;

3.3 对满足匿名要求的元组添加分组标号

4. If (*PT* 中没有分组标号的元组数>0) then

4.1 从可以移出元组的 QI 分组中移出元组,加入剩余元组中;

4.2 对满足匿名要求的元组添加分组标号

5. If (*PT* 中没有分组标号的元组数>0) then 合并剩余元组到临近 QI 分组中

6. Return *PT*

下面我们以表 3 为例简单说明一下算法 1 的过程.设 QI 属性仅包含邮编,该表已经对身份属性进行了重编码.个体 1 和个体 2 分别出现 2 次,其他 5 个个体分别出现 1 次.首先检查表 3,发现没有元组可以直接进入分组.于是邮编概化一次,这样元组 1 到元组 4 被分为一个分组,标为组 1;元组 5 到元组 8 被分为一个分组,标为组 2.元组 9 被剩下,成为孤儿元组(见表 4).然后检查是否有分组中包含多余元组可以被移出.如果要求身份保持的 *k*-匿名并设定 *k*=2,可以移动一个元组(比如元组 3)加入孤儿元组,形成分组 3(见表 5).如果要求身份保持的(α,β)-匿名并设α=0.5 且β=0.5,则找不到多余元组.于是将孤儿元组合并到分组 1 中.

Table 4 Generalized once result of Table 3

表 4 对表 3 一次概化处理的结果

Goup-ID	ID	Zip	Disease
1	1	1008*	Hypertension
1	1	1008*	Hyperlipemia
1	2	1008*	Diabetes
1	3	1008*	Heart
2	4	1007*	Hypertension
2	4	1007*	Diabetes
2	5	1007*	Heart
2	6	1007*	Flu
—	7	1005*	Heart

Table 5 Published result of Table 3 satisfying identity-reserved 2-anonymity

表 5 对表 3 满足身份保持的 2-匿名的一个发布表

Goup-ID	ID	Zip	Disease
1	1	1008*	Hypertension
1	1	1008*	Hyperlipemia
1	3	1008*	Heart
2	4	1007*	Hypertension
2	4	1007*	Diabetes
2	5	1007*	Heart
2	6	1007*	Flu
3	7	100**	Heart
3	2	100**	Diabetes

算法的空间复杂性是 $O(n)$,因为它是在原有数据的基础上进行处理的.算法的时间复杂性是多项式的,时间主要花费在步骤 3 的循环上.

3.2 有损连接方法

有损连接方法发布两张关系表,一张是准标识属性表,另一张是敏感属性表.两个表都包含一个共同的组标识属性.组标识属性记录了每个元组所在 QI 分组的分组号.通过分组号可以将两表关联起来,得到需要的发布信息.由于分组号并不是表的主键,因此两表之间的关联形成一种有损连接.具有相同分组号的准标识属性和敏感属性可以配对组成一个元组.

在概化方法中,每个 QI 分组中的准标识属性取值是单一的.而在有损连接方法中,每个 QI 分组中包含了准标识属性的原始取值.

在有损连接方法中,发布内容包括两张表:准标识属性表(QI table,简称 QIT)和敏感属性表(ST table,简称 STT).准标识属性表中包含了准标识属性的原始取值和分组标号.分组标号属性记录了该元组所在的 QI 分组的分组号.但在单一个体对应多个记录的情况下,QIT 中就会存在完全相同的多个元组,我们在 QIT 中删除这些完全相同的重复元组.否则,攻击者可以轻易地利用这种重复现象,只要与敏感属性表中这些重复 ID 相对应,就可以发现个体的隐私信息.准标识属性表的模式表示为

Original QI attributes	Group No

敏感属性表包含 3 个属性:身份属性 ID、敏感属性 ST 和分组标号.ID 是重编码后的身份属性.如果存在单一个体对应多个记录的情况,一个 ID 取值可以在一个 QI 分组中出现多次.ST 保存了敏感属性取值.分组标号记录了该元组所在的 QI 分组的分组号.

敏感属性表的模式可以表示为

ID	Sensitive attribute	Group No

这时,QIT 的元组个数可能小于 STT 的元组个数,这是因为 QIT 删除了同一个体的重复元组,而 STT 则保留了同一个体的多种敏感属性取值.比如在表 6 中有 7 个元组,而在表 7 中有 9 个元组.通过这种方式我们可以保存更多的信息有效性.我们不仅保存了准标识属性和敏感属性的原始取值,而且保留了个体和敏感属性之间的对应关系.

Table 6 Published QIT of Table 3 satisfying identity-reserved 2-anonymity, after loss-join processing

表 6 采用有损连接,对表 3 进行身份保持的 2-匿名处理后的 QIT 表

Group-No	Zip
1	10085
1	10086
2	10087
2	10075
2	10076
3	10077
3	10050

Table 7 Published STT of Table 3 satisfying identity-reserved 2-anonymity after loss-join processing

表 7 采用有损连接,对表 3 进行身份保持的 2-匿名处理后的 STT 表

Group-No	ID	Disease
1	1	Hypertension
1	1	Hyperlipemia
1	3	Heart
2	4	Hypertension
2	4	Diabetes
2	5	Heart
2	6	Flu
3	7	Heart
3	2	Diabetes

例如,表 3 是需要发布的病人表.它已经对身份属性进行了重编码,然后按照概化方法进行分组,最后把表划分为两个发布表.一个是带有组标号属性的准标识属性表(QIT),见表 6;另一个是带有组标号属性的敏感属性表(STT),它包括了重编码后的身份属性,见表 7.由于个体身份属性和敏感属性都保留在 STT 中,研究者可以直接发现这两者之间的关系.比如,从表 7 中可以发现前两个记录中不同的病症,高血压和高血脂,对应于同一个人.这一事实可以提高数据的有效性,有利于研究者的进一步研究.采用有损连接,每个准标识属性取值都链接到一个分组标号上,通过这个分组标号可以对应于多个个体身份和敏感属性取值.从表 6 可以看到,第 1 行记录(*postcode*=10085)可以链接到两个不同的个体(ID=1 or 3)和 3 种不同的病症(hypertension,hyperlipemia 和 heart).这样,发布表就满足了身份保持的 2-匿名要求.

有损连接的算法可以基于概化算法.它在概化算法的基础上进行表分解.具体地,包括以下几个步骤:

(1) 对身份标识属性重编码,将原表 T 转化为 T'.

(2) 采用概化算法,把 T'转化为 T^*,使得 T^*满足身份保持的匿名模型.这样,T'中的每个元组按照 T^*的分组结果都对应了一个分组号.

(3) 将 T'的准标识属性投影到 QIT 中,同时把每个元组对应的分组号填入到分组标号属性中,然后删除重复的记录.

(4) 将 T'的 ID 和敏感属性取值投影到 STT 中,同时把每个元组对应的分组号填入到分组号属性中.

步骤 2 中采用了概化算法来实现身份保持的匿名模型,这一算法在第 3.1 节进行了描述.有损连接算法的空间复杂性是 $O(n)$,因为它是在原有数据的基础上进行处理的.算法的时间复杂性是多项式的,时间主要花费在步骤 2 的概化算法上,而步骤 3 和步骤 4 的复杂性都是 $O(n)$.

4 实验结果

在本节中,我们在实验环境中评价身份保持的匿名模型.首先,我们检测在单一个体对应多个记录的情况下,原有匿名模型的脆弱性.然后,我们比较了原有匿名模型和身份保持的匿名模型的信息损失差异,最后比较概化方法和有损连接方法在信息有效性上的差异.系统平台是 Windows XP 和 Microsoft SQL Server 2000.

Table 8 Description of Adult dataset

表 8 Adult 数据集的描述

	Attribute name	Value number	Generalization type	Height of hierarchy
1	Age	74	Range-5, 10, 20	4
2	Sex	2	Suppression	1
3	Race	5	Suppression	1
4	Marital status	7	Taxonomy tree	2
5	Education	16	Taxonomy tree	3
6	Country	41	Taxonomy tree	2
7	Work class	7	Taxonomy tree	2
8	Salary class	2	Suppression	1
9	Occupation	14	Sensitive attribute	–
10	Id-number	40 000	–	–

实验数据来自于 Adult database of UCI Machine Learning Repository[13].该数据库包含 45 222 个美国普查人口数据.我们删除那些存在缺失值的元组.由于我们要检验单一个体对应多个元组的情况,在数据库中添加了表示重编码后的身份属性字段 Id-number.表 8 是对数据集的描述,包括属性名称、属性取值个数、取值类型、概化方式和概化层次.

我们按照一定比例添加单一个体的多个记录.首先选出 40 000 个元组,在每个元组的 Id-number 属性中填入不同的身份标号.然后把这些元组按照一定比例划分到 3 个不相交的子集中,假设分别称为 A,B 和 C.对于子集 A 中的每个元组,我们直接将其加入待发布数据集中.对于子集 B 中的每个个体,我们添加一个新的元组,它复制原有元组的身份标识取值和准标识取值,然后赋予一个不同于原有取值的敏感属性取值.这样,每个 B 中的个体即对应于待发布数据集中的两个元组.对于子集 C 中的每个元组,我们添加两个新的元组,它复制原有元组的身份标识取值和准标识取值,然后赋予两个不同的敏感属性取值.这样,每个 C 中的个体对应于待发布数据集中的 3 个元组.这样我们得到的待发布数据集包含 $|A|+2|B|+3|C|$ 个元组.根据 rpi 的定义,待发布数据集的 $rpi=(|A|+2|B|+3|C|)/(|A|+|B|+|C|)=Ratio_A+2Ratio_B+3Ratio_C$.我们根据表 9 的比例产生 4 个不同 rpi 的数据集进行测试,rpi 分别等于 1.2,1.4,1.6 和 1.8.

首先我们检测在单一个体对应多个记录的情况下,原有匿名模型隐私的脆弱性.忽略身份属性,采用类似算法 1 的 k-匿名局部重编码概化算法得到匿名结果表 PT.这时,每个 QI 分组中至少有 k 个元组,它并未考虑单一个体对应多条记录的情况.而按照身份保持的 k-匿名,要求每个 QI 分组中至少有 k 个不同的个体.我们定义 PT 中包含至少 k 个元组但却只有最多$(k-1)$个体的 QI 分组为脆弱分组,即该分组满足了 k-匿名,但不满足身份保持的 k-匿名.那么脆弱分组比例就是脆弱分组数/QI 分组总数.当 k 增加或者 rpi 减小时,QI 分组中的元组总数会随之增加,不同个体数也会增加,那么脆弱分组比例就会相应减小.图 1 显示了脆弱分组比例与 k 值和 rpi 的这种

关系.

特别地,*PT* 中一些分组中只包含了唯一一个个体.正如第 1 节中所指出的,这种情况下隐私泄漏的概率更大.我们把这样的 QI 分组称为单值分组.只有在 *k* 值不大于单一个体对应的最多元组数时,才可能存在单值分组.与图 1 类似,图 2 显示,单值分组比例随着 *k* 值的增加而增加或者随着 *rpi* 值的减小而减小.

Table 9 4 datasets with different *rpi*s after processing Adult dataset

表 9 对 Adult 数据集进行处理得到 4 个不同 *rpi* 的数据集

A ratio	*B* ratio	*C* ratio	*rpi*	Tuple number
0.85	0.1	0.05	1.2	48 000
0.70	0.2	0.10	1.4	56 000
0.55	0.3	0.15	1.6	64 000
0.60	0.4	0.20	1.8	72 000

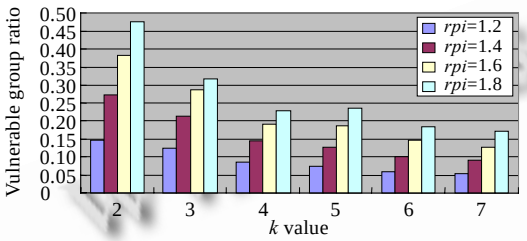


Fig.1 Vulnerable group ratio with *rpi* and *k*

图1 脆弱分组比例与 *rpi* 和 *k* 的关系

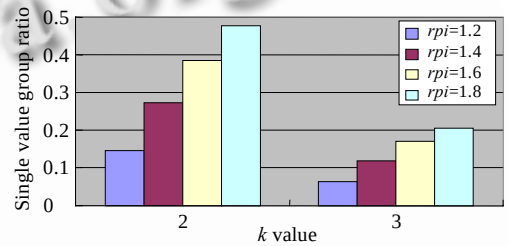


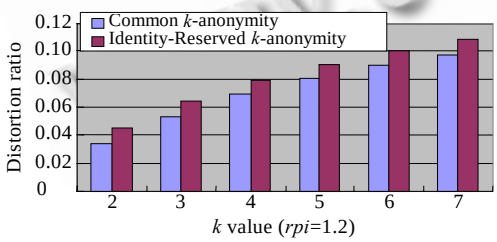
Fig.2 Single value group ratio with *rpi* and *k*

图 2 单值分组比例与 *rpi* 和 *k* 的关系

其次,我们比较普通 *k*-匿名和身份保持的 *k*-匿名之间的信息损失差别.我们用概化度(distortion ratio)作为描述信息损失的指标.在文献[5]中,一个元组的概化度定义为各个准标识属性概化结果的所在层次高度之和与全概化结果的概化高度之和的比值.设 $height_i$ 是第 i 个元组的各个准标识属性的概化高度之和,*Height* 是元组全概化的概化高度之和.那么数据集的概化度定义为

$$distortion\ ratio = \frac{\sum_{i=1}^{TupleCount} height_i}{TupleCount \times Height}.$$

图 3 显示了普通 *k*-匿名和身份保持的 *k*-匿名之间的概化度差别.我们注意到,普通 *k*-匿名取得相对稍低的概化度,但是差别并不大,并且概化度的差别随着 *rpi* 的增大而增大.这是由于,*rpi* 增大,那么同一个体的重复度增加,因而身份保持的匿名结果中 QI 分组需要的元组数增大,从而导致概化度的增大.



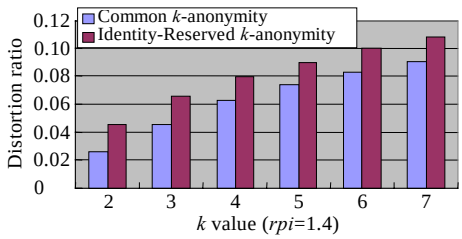


Fig.3 Distortion ratio between common *k*-anonymity and identity-reserved *k*-anonymity when *rpi*=1.2 and 1.4

图 3 *k*-匿名与身份保持的 *k*-匿名的概化度比较(*rpi*=1.2,1.4)

最后,我们比较身份保持的匿名中概化与有损分解两种实现方法在信息有效性上的差别.这里的概化方法采用第 3.1 节中提到的算法,有损分解方法如第 3.2 节中所描述.我们用相对查询正确率来评价信息有效性.首先在原始数据集上提交查询,得到一个正确结果(称为 *act*);然后在发布的匿名表上提交查询,得到近似结果(称为 *apr*).相对查询正确率(correctness)定义为

$$correctness = 2^{-\left|\frac{act-apr}{act}\right|}.$$

这样,相对查询正确率的取值总在 0 和 1 之间.如果 *apr* 等于 *act*,那么相对查询正确率为 1.查询语句定义为如下格式:

Select count(*)

From data-table

Where pred(A_1) and … and pred(A_{qd}) and pred(A_s)

其中,data-table 是原始表或者概化方法的发布表.Qd 表示准标识属性个数,A_s 表示敏感属性.每个谓词 *pred*(*A*)格式为($A=x_1$ or $A=x_2$ or… or $A=x_b$).由于有损连接方法发布两张表,因此我们计算有损连接方法的近似结果 *apr* 分为两个步骤.

(1) 找到包含满足准标识属性谓词的元组的所有分组序号,这可以通过执行如下语句来实现:

Select groupNo

From QIT

Where pred(A_1) and … and pred(A_{qd})

(2) 对于(1)的结果中的每个分组 g_i,计算 3 个数:(a) QIT 表的分组 g_i 中满足准标识属性谓词的所有元组个数,记为 qi_hit_i.(b) QIT 表的分组 g_i 中所有元组的个数,记为 $groupSize_i$.(c) 在 STT 表的分组 g_i 中,满足敏感属性谓词的元组个数,记为 st_hit_i.

那么近似结果 *apr* 可以计算为

$$apr = \sum_{g_i} st\_hit_i \times \frac{qi\_hit_i}{groupSize_i}.$$

在实验中,我们从 3 个方面来比较概化方法和有损连接方法的信息有效性:参数取值、准标识属性数、查询涉及维度数.每次执行 1 000 条查询,取其平均值作为相对查询正确率.

图 4 显示了 *k* 值对于匿名结果的影响.当 *k* 增加时,两种方法的相对正确率都随之降低,这是由于 QI 分组中的元组数不断增加,匿名程度不断增加.但是有损连接方法的相对查询正确率更高.图 5 显示了 *l* 值对于匿名结果的影响.当 *l* 增加时,两者的相对查询正确率也都随之降低,这是由于 QI 分组中的元组数不断增加,匿名程度不断增加.因为有损连接方法在 QIT 和 STT 中保留了原始的取值,它能较为准确地找到符合要求的那些元组和 QI 分组,因此其相对查询准确率相对较高,而且变化不大.

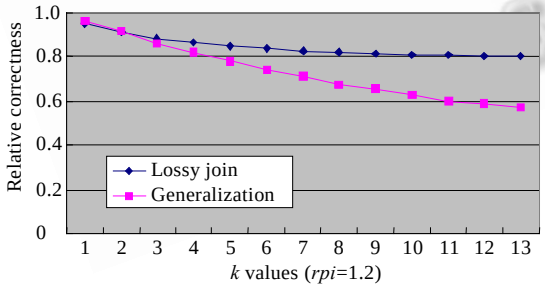


Fig.4 Relation between *k* and relative correctness in identity-reserved *k*-anonymity

图 4 身份保持的 *k*-匿名,*k* 值对两种方法的影响

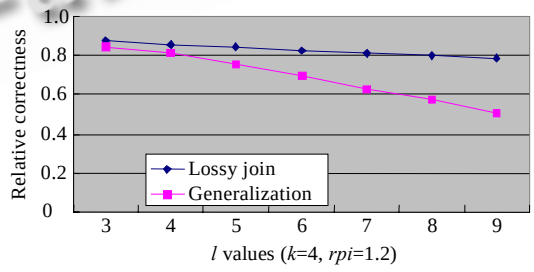


Fig.5 Relation between *l* and relative correctness in identity-reserved (*k*,*l*)-anonymity

图 5 身份保持的(*k*,*l*)-匿名,*l* 值对两种方法的影响

图 6 显示了身份保持的(α,β)-匿名中β值对于结果的影响(当α=0.5),图 7 显示了α值对于结果的影响(当β=0.5).当α(或β)增加时,相对查询正确率随之增加,这是因为,对于敏感属性(或身份属性)的限制逐渐减小.当取值较小时,β值对于结果的影响要显著于α值的影响.因为在通常情况下,ID 属性包含更多的不同取值,而敏感属性取值个数较少,β的变化会显著地影响结果.当取值较大时,两者对匿名结果的影响都很小,相对查询正确率都

较高.因为有损连接在 QIT 和 STT 中保留了原始的取值,它表现出了更高的正确率.

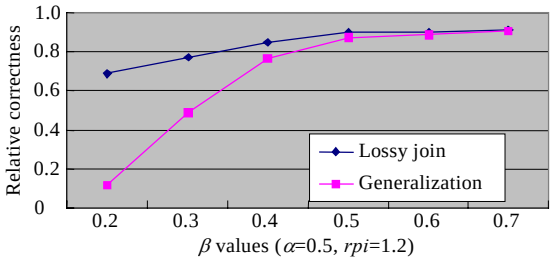


Fig.6 Relation between β and relative correctness in identity-reserved (α,β)-anonymity

图 6 身份保持的(α,β)-匿名中,β值对两种方法的影响

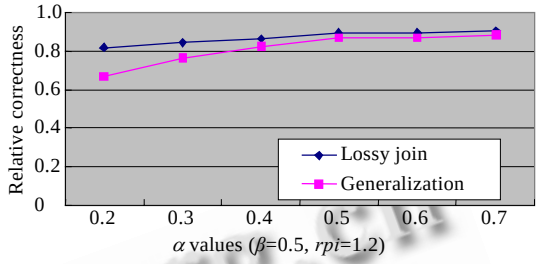


Fig.7 Relation between α and relative correctness in identity-reserved (α,β)-anonymity

图 7 身份保持的(α,β)-匿名中,α值对两种方法的影响

图 8 显示了准标识属性数对于结果的影响.当准标识属性数增加时,需要概化的 QI 属性增加,概化程度增加,因而两种方法的相对查询正确率随之下降.因为有损连接方法在 QIT 和 STT 中保留了原始的取值,它表现出了更高的正确率.图 9 显示了查询涉及的维度数对于结果的影响.当查询维度增加时,查询范围逐渐减小,因为涉及的元组数和 QI 分组数减小,因此这时概化方法的相对查询准确率有所提高.而因为有损分解方法在 QIT 和 STT 中保留了原始的取值,在不同的查询维度情况下,它都能较为准确地找到符合要求的那些元组和 QI 分组,因此其相对查询准确率相对较高,而且变化不大.

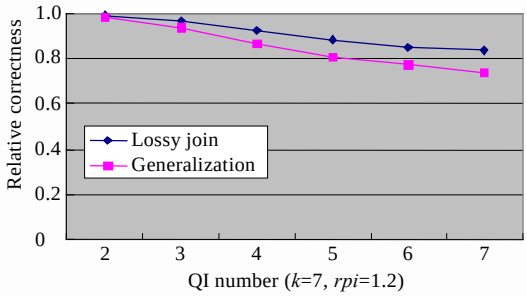


Fig.8 Relation between QI and relative correctness in identity-reserved k-anonymity

图 8 身份保持的 k-匿名中,QI 数对两种实现方法的影响

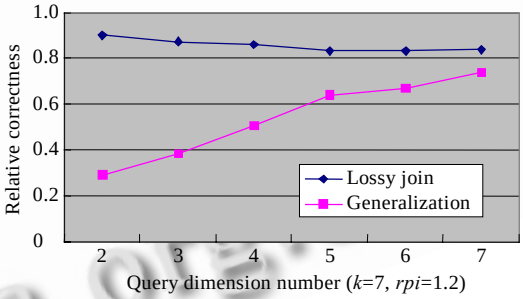


Fig.9 Relation between query dimension number and relative correctness in identity-reserved k-anonymity

图 9 身份保持的 k-匿名中,查询维度数对两种实现方法的影响

5 结 论

在本文中,我们分析了隐私保护的数据发布中单一个体对应多个记录的情况,发现原有的匿名模型在该情况下存在一定的隐私泄漏.因此我们提出了身份保持的匿名模型.它保存了身份重编码后的信息,提高了信息有效性.根据不同的隐私需求,我们提出了 3 种具体的身份保持的匿名模型.概化方法和有损连接方法是达到身份保持的匿名模型的两条途径.我们在实验环境中从多个方面比较了原有匿名模型和身份保持的匿名模型,并检验了本文方法的有效性.

References:

[1] Samarati P. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Engineering, 2001, 13(6):1010−1027. [doi: 10.1109/69.971193]

[2] Sweeney L. Achieving *K*-anonymity privacy protection using generalization and suppression. Int'l Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2002,10(5):571−588. [doi: 10.1142/S021848850200165X]

[3] Sweeney L. *K*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[4] Machanavajjhala A, Gehrke J, Kifer D. *l*-Diversity: Privacy beyond *K*-anonymity. In: Liu L, Reuter A, Whang KY, Zhang J, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Atlanta: IEEE Computer Society, 2006. 24−35.

[5] Wong RC, Li J, Fu AW, Wang K. (α,k)-Anonymity: An enhanced *K*-anonymity model for privacy-preserving data publishing. In: Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D, eds. Proc. of the 12th Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2006. 754−759.

[6] Li N, Li T, Venkatasubramanian S. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Dogac A, Ozsu T, Sellis T, eds. Proc. of the 23rd Int'l Conf. on Data Engineering. Istanbul: IEEE Computer Society, 2007. 106−115.

[7] Xiao X, Tao Y. Personalized privacy protecting. In: Chaudhuri S, Hristidis V, Polyzotis N, eds. Proc. of the Int'l Conf. on Management of Data. Chicago: ACM Press, 2006. 229−240.

[8] Fung BCM, Wang K, Yu PS. Top-Down specialization for information and privacy preservation. In: Aberer K, Franklin M, Nishio S, eds. Proc. of the 21st Int'l Conf. on Data Engineering. Tokyo: IEEE Computer Society, 2005. 205−216.

[9] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain *K*-anonymity. In: Özcan F, ed. Proc. of the Int'l Conf. on Management of Data. Maryland: ACM Press, 2005. 49−60

[10] Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A. Anonymizing tables. In: Eiter T, Libkin L, eds. Proc. of the 10th Int'l Conf. on Database Theory. Edinburgh: Springer-Verlag, 2005. 246−258.

[11] Wang K, Fung BCM. Anonymizing sequential releases. In: Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D, eds. Proc. of the 12th Int'l Conf. on Knowledge Discovery and Data Mining. Philadelphia: ACM Press, 2006. 414−423.

[12] Xiao X, Tao Y. 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: VLDB Endowment, 2006. 139−150.

[13] UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Adult

童云海(1971－),男,浙江鄞县人,博士,副教授,主要研究领域为数据仓库和数据挖掘,数据隐私保护,多媒体智能信息处理.

唐世渭(1939－),男,教授,博士生导师,CCF高级会员,主要研究领域为数据仓库和数据挖掘,海量信息处理,数字图书馆.

陶有东(1977－),男,博士,主要研究领域为数据仓库和数据挖掘,数据隐私保护,多媒体数据处理.

杨冬青(1945－),女,教授,博士生导师,CCF高级会员,主要研究领域为数据模型,数据库系统,Web 环境下的信息集成与共享.