

支持多属性泛化的个性化(α, l, k)匿名模型

苏林萍,董子娴,李 为,吴克河,崔文超
(华北电力大学 控制与计算机工程学院,北京 102200)

摘 要:传统的个性化数据匿名模型一般可以分为两种机制:一种是面向个人的,一种是面向敏感值的。这两种方法一般都会因为追求敏感数据的个性化保护而过度泛化,造成大量的信息损失,使数据的可用性下降。为此,该文提出了一种个性化(α, l, k)匿名隐私保护模型。该模型有效结合了这两种传统的数据匿名机制,在最大程度地保证个性化匿名的需求下,根据敏感属性值敏感等级的不同,对各个等价组中的敏感属性值分别采取不同的匿名方式,优先泛化高敏感等级的属性值,使等价组中的每个敏感属性满足对出现频率 α 以及多样性 l 的约束条件,从而有效降低数据集中高敏感等级信息的泄露风险,并可以提高数据的可用性。实验结果表明,该模型能够在有限的运行时间内,相较于其他个性化匿名模型有更低的信息损失量和更好的隐私数据保护能力。

关键词: k -匿名;个性化隐私保护;泛化;敏感度评分;敏感等级

中图分类号:TP309

文献标识码:A

文章编号:1673-629X(2021)06-0088-06

doi:10.3969/j.issn.1673-629X.2021.06.016

A Personalized (α, l, k) Anonymous Model of Supporting Multi-attribute Generalization

SU Lin-ping, DONG Zi-xian, LI Wei, WU Ke-he, CUI Wen-chao

(School of Control and Computer Engineering, North China Electric Power University, Beijing 102200, China)

Abstract: The traditional personalized data anonymity model can be divided into two mechanisms: one is personal-oriented, and the other is sensitive value oriented. In general, these two methods tend to over-generalize due to the pursuit of personalized protection of sensitive data, resulting in a large number of information losses and declining availability of data. To this end, we propose a personalized (α, l, k) anonymous privacy protection model, which effectively is a combination of these two kinds of traditional data anonymous mechanism. Under the need of ensuring personalized anonymity in maximum extent, according to different sensitive attribute value level, the sensitive attribute values of each equivalent group respectively take different anonymously, priority generalization properties of high level of sensitivity, which makes each sensitive attribute in equivalent group meet the constraint conditions of occurrence frequency α and diversity l , so as to effectively reduce the leakage risk of high-sensitive information in the data set, and improve the availability of data. The experiment shows that this model has lower information loss and better privacy data protection capability than other personalized anonymous models in limited running time.

Key words: k -anonymity; personalized privacy protection; generalization; sensitivity score; sensitive level

0 引 言

在信息高度共享的今天,每天都有大量的数据被收录和发布。对信息的挖掘和分析,可以有效促进科学事业的发展,从而为人们营造更加便捷畅通的生活环境。但与此同时,不得不面临隐私数据的泄露问题。因此需要重点保护个体的隐私数据。

在数据匿名化的思想下,隐私数据保护模型的基本做法是:在满足信息安全发布的要求下,隐匿特定个体和敏感信息之间的联系,并提高数据的可用性^[1]。

但是传统的匿名方式并未考虑不同个体对敏感数据的个性化隐匿需求^[2]。不同的隐私信息其敏感等级必然不同,同一敏感信息对不同个体的敏感程度也会不同^[3]。

Xiao等在文献[4]中第一次提出了个性化匿名的思想,之后出现了很多增强型的改进算法。其中,文献[5-6]为数据集中的每一个元组都设置了不同的隐匿需求,尽管极大程度地满足了不同个体的个性化需求,但是工作量巨大,也造成了不必要的数据冗余。文献

[7]提出个性化(α, l)-多样化 k -匿名模型,归纳个性化隐匿方式包括两种机制,一种是面向个人的,一种是面向敏感值的。因为一般来说,仅仅面向个人的隐私保护模型容易由于个体的喜好造成信息的不必要损失和隐私泄露,而单纯面向敏感值的方式往往会欠缺特定个体的特定需求^[8]。但是这种方法因为层层的条件限制而造成了属性值的过度泛化。文献[9]则提出了个性化(p, α, k)-匿名模型,改善了匿名后数据损失较大和高敏信息泄露的缺陷,将敏感值划分不同的敏感级别,并且各等级应用不相同的匿名方式,但是其对个性化的需求体现不足^[10]。

针对上述文献中所提方法存在的不足,该文基于个性化(α, l)-多样化 k -匿名模型和个性化(p, α, k)-匿名模型,提出针对属性过度泛化的改进的个性化匿名模型:应用文献[11]中敏感度评分的概念给不同的敏感值定义不同的等级,再根据少量特定个体对自己敏感属性的评级为每条记录确定最终的敏感值等级,按照敏感属性泛化树将高敏值直接泛化到下一级,然后使此时等价组里中低级敏感值满足 l 和 α 约束。

1 匿名模型

数据匿名化,指对要发布数据集的各属性进行合理的脱敏操作,要求数据在对应运维、实施或者数据挖掘等场景下不影响使用的同时,不能反识别出对应的个体。

可将原始集中属性划分为以下四种^[12]:标识符属性(ID),是可以唯一标识到特定个体的属性,例如表1的“Name”。这部分属性在匿名处理中一般会被直接移除;准标识符属性(QI),是可通过和外部数据链接或背景知识的手段唯一确定出特定个体的属性^[13]。例如表1中的“Gender”、“Age”和“Zip code”属性;敏感属性(S),是指个体的敏感信息,指攻击者最想明确和关联的属性^[14],如表1中的属性“Disease”;非敏感属性(N),也就是其他属性,在脱敏时这部分属性不做处理。

表1 原始数据

Name	Gender	Age	Zip code	Disease
Emma	M	25	061259	Cancer
Lily	F	21	062243	HIV
Doris	F	25	062250	Flu
Martin	M	28	066500	Flu
Rose	F	23	062201	Cancer
Tom	M	56	102206	Cancer
Amy	M	59	102105	Cancer

定义1:等价组。给定数据集 $T = \{A_1, A_2, \dots,$

$A_n\}$, n 为 T 中属性的个数。则其中的准标识符属性集 $QI = \{q_1, q_2, \dots, q_i\}$ 里, i 为准标识符属性个数,值一致的记录则属于同一等价组。

定义2: k -匿名。给定数据集 T 和等价组 Q ,若对于 $\forall Q \subset T$, Q 中的记录条数都至少为 k ($k \geq 2$),则 T 满足 k -匿名。

由此可知,当数据集满足 k -匿名时,攻击者确定特定个体和元组数据之间关联关系的概率不超过 $1/k$,有效防止了链接攻击,不过因为并未破坏特定个体与敏感信息间的关系,所以还是会有背景知识攻击以及同质攻击的可能^[15]。例如,表2就是对原始表表1的2-匿名化实例,当知道Tom的性别、年龄和邮编信息时,因为表中记录6和7的疾病属性都是Cancer,所以由此可以确定Tom的所患疾病,很显然,这是Tom不想公开的属性信息。

表2 2-匿名表

ID	Gender	Age	Zip code	Disease
1	M	[21-25]	06****	Cancer
2	M	[21-25]	06****	Flu
3	F	[21-25]	0622**	Flu
4	F	[21-25]	0622**	HIV
5	F	[21-25]	0622**	Cancer
6	M	[56-60]	102***	Cancer
7	M	[56-60]	102***	Cancer

基于 k -匿名模型中存在的风险,需要破坏特定个体与敏感信息之间的关联关系,这就需要引入1-多样性模型。

定义3: l -多样性。给定数据集 T 和等价组 Q ,若对于 $\forall Q \subset T$,其敏感属性值的种类数都不小于 l ($l \geq 2$),则该数据集满足 l -多样性模型。

表3 2-多样性表

ID	Gender	Age	Zip code	Disease
1	M	[21-25]	06****	Cancer
2	M	[21-25]	06****	Flu
3	F	[21-25]	0622**	Flu
4	F	[21-25]	0622**	HIV
5	F	[21-25]	0622**	Cancer
6	M	[56-60]	102***	servious illness
7	M	[56-60]	102***	Cancer

表3即对表2中的敏感属性按2-多样性模型泛化后的示例,此时,对于每一个等价类,其敏感属性的种类个数都至少为2。但是该模型无法避免相似性攻击和偏斜攻击,以表3为例,当只能确定Amy为最后两条记录时,由于其疾病种类都是很严重的属性值,所以还是可以得知Amy得了绝症,这可能是Amy极度

不想公开的隐私信息。

定义 4: 个性化 (α, l) -多样化 k -匿名模型。给定数据集 $T = \{A_1, A_2, \dots, A_n\}$, 将敏感值按敏感度的不同划分不同的等级 Sid, 此时若有特定个体对自己记录的敏感值等级指定了等级 Ppl, 且 $Ppl > Sid$, 则按照 Ppl 的等级替换 Sid。匿名后数据集 T^* , 若 T^* 符合 k 匿名, 且各等价组里不同敏感值的个数不低于 l , 每个等价组里相同敏感等级的敏感值出现的频率不大于 α , 就可以称 T^* 符合个性化 (α, l) -多样化 k -匿名模型。

定义 5: 个性化 (p, α, k) -匿名模型。给定数据集 $T = \{A_1, A_2, \dots, A_n\}$, 将敏感值按敏感程度不同分为高中低三级, 将高等级敏感值直接泛化。匿名后数据集 T^* , 若此时 T^* 符合 k 匿名, 且此时各等价组里不同敏感值的个数不低于 p , 每个等价组中相同敏感值出现的频率不大于 α , 就可以称 T^* 符合个性化 (p, α, k) -匿名模型。

2 个性化 (α, l, k) 匿名模型相关概念

为了给特定个体提供有效的个性化服务, 同时要降低信息损失量来提高数据的可用性, 该文结合个性化 (α, l) -多样化 k -匿名和个性化 (p, α, k) -匿名两种模型, 提出了一种个性化 (α, l, k) 匿名模型。在个性化 (α, l) -多样化 k -匿名模型中, 虽然同时考虑了面向个人和面向敏感值两种个性化隐私机制, 但是通过实验也可以看出, 由于过度泛化造成了大量的信息损失, 极大地影响了数据的分析和挖掘。个性化 (p, α, k) -匿名模型则提出针对敏感度较高值泛化的思想, 将高等级敏感值直接泛化, 在让其余中低敏感等级的值满足 p, α 约束时, 也是优先泛化较高敏感级的属性, 由此不仅有效降低了数据集的敏感度, 还降低了信息损失量, 但是个性化的思想体现不足。由此, 该文提出一种个性化 (α, l, k) 匿名模型。

2.1 敏感属性的敏感度划分与频率约束 α

下面将引入文献[11]里敏感度评分的思想, 以其结果作为敏感值的敏感等级, 这样较个性化 (α, l) -多样化 k -匿名模型更加体现个性化需求。

定义 6: 敏感度评分。统计个体对每个敏感值敏感程度的评分结果, 并以统计结果的有效区间划分等级, 作为敏感值敏感等级的预设参数。用这种方式设定的参数能更加满足大众用户对敏感度的要求。

例如散点图 1 所示, 该图为对敏感属性疾病的调查结果, 横坐标依次代表 Flu、Indigestion、Heart disease、Asthma、Phthisis、Hepatitis、HIV 和 Cancer 这八种疾病, 将评分满分设置为 80 分, 每个个体根据自己对疾病的重视程度对各个疾病进行打分, 得分越高表

示重视程度越高。去掉离群点, 可以看到数据依次集中在区间 $[0, 20)$ 、 $[0, 20)$ 、 $[20, 40)$ 、 $[20, 40)$ 、 $[40, 60)$ 、 $[40, 60)$ 、 $[60, 80)$ 、 $[60, 80)$ 中, 所以疾病的敏感属性可划分为 4 个等级, 分别为 1、2、3 和 4, 并将这个值作为敏感属性的预设参数 C 。划分结果如表 4 所示。

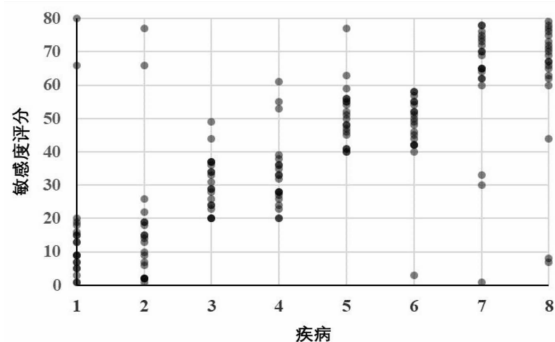


图 1 敏感度评分结果

表 4 敏感属性预设参数

Id	Sensitive value	C
1	Flu, Indigestion	1
2	Heart disease, Asthma	2
3	Phthisis, Hepatitis	3
4	HIV, Cancer	4

定义 7: 频率约束 α 。给定数据集 T 、等价组 Q , 指定的敏感属性 S 中各个敏感值的出现频率 α ($0 \leq \alpha \leq 1$)。若在任意等价组 Q 中, 任意属性值都满足 $|Q(S)|/|Q| \leq \alpha$, 那么数据集 T 满足出现频率约束 α 。其中, $|Q(S)|$ 指等价组 Q 中敏感属性为 S 的记录个数, $|Q|$ 是等价组 Q 的大小。

2.2 敏感属性泛化树

构建敏感属性泛化树, 如图 2 所示。各个原始敏感值作为泛化树的叶节点, 树的高度至少是敏感属性的总等级数, 要求被泛化的每个父节点均满足各敏感值的行业规范。

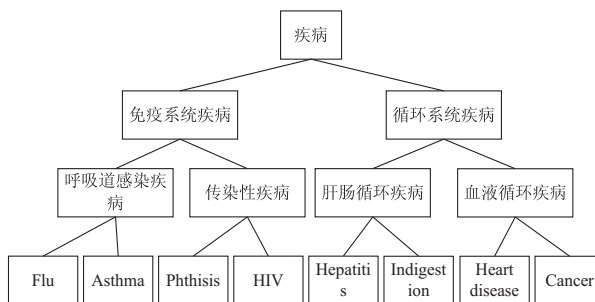


图 2 敏感属性泛化树

2.3 个性化隐私保护规则

在通过应用敏感度评分的方法, 达到了面向敏感值的个性化需求基础上, 本模型还支持面向个人的个性化需求, 允许用户给自己记录的敏感值认定敏感等

级。需要注意的是,自定义的敏感等级值 p 不得超过泛化树的高度 H 。仅当 $p > C$ 时,需用 p 值替换 C 的值。

如表5所示,在ID为3的记录中 $p = 2 \geq C = 1$,所以会对该记录执行进一步脱敏操作,用对应等级父节点“呼吸道感染”属性值代替“Flu”。其中,用户指定的敏感等级 p 列中,“-”代表用户未指定等级。

表5 隐私保护级别

ID	Gender	Age	Zip code	Disease	C	p
1	M	[21-25]	06****	Cancer	4	3
2	M	[21-25]	06****	Flu	1	-
3	F	[21-25]	0622**	Flu	1	2
4	F	[21-25]	0622**	HIV	4	2
5	F	[21-25]	0622**	Cancer	4	3
6	M	[56-60]	102***	Cancer	4	4
7	M	[56-60]	102***	Cancer	4	-

2.4 信息损失度量

不管是准标识符属性还是隐私属性的泛化操作都会带来信息的损失^[16]。信息的损失反映了信息的可用性,但在一定程度上也反映了敏感信息的保护程度。

定义8:信息损失量。给定数据集 T ,规定 T 里属性 A 的阈值是 $\text{size}(A)$,那么 A 被泛化成 A^* 的信息损失量为:

$$IL_{(A)} = \frac{|A^*|}{|\text{size}(A)|} \quad (1)$$

其中, $|A^*|$ 为属性 A 被泛化后的值。 $|\text{size}(A)|$ 是 A 的可能取值,若该属性是连续性数据,取区间长度,若是分类型数据,取值域的基数。

所以 T 中记录 t_i ($1 \leq i \leq n$) 的信息损失量如下,其中 W_i 是 A 的信息损失量权重:

$$IL_{(t_i)} = \sum_{i=1}^n W_i \cdot IL_{(A)} \quad (2)$$

则 T 中所有属性的信息损失量为:

$$IL_{(T)} = \sum_{i \in T} \sum_{i=1}^n W_i \cdot IL_{(A)} \quad (3)$$

3 算法设计

该文提出了个性化(α, l, k)匿名算法,本算法结合了个性化匿名的两种机制,在极大程度满足个性化的前提下,有效降低了数据损失量。

算法步骤如下:

(1)引用文献[17]中提出的多属性泛化的方法得到符合 k 匿名的数据集;

(2)比较自定义的敏感等级 p 和敏感属性的预设参数 C ,若 $p > C$,则修改对应记录的等级,用 F 代表敏感属性的最终级别,再将敏感值泛化到相应的级别;

(3)将每个等价组中的记录按敏感值的敏感等级由高到低进行排序,将 F 最高的属性信息直接泛化到下一级;

(4)统计各等价类里不一致的敏感值个数,若小于 l 则将 F 相对较高的值进行泛化,并使其满足出现频率 α 的约束,直到满足个数值大于等于 l ;

(5)计算各等价类中各敏感值出现的频率,若大于 α 则将敏感度相对较高的属性进行泛化,直到满足频率值小于等于 α 。

如表6所示,是对表2中隐私属性 Disease 的进一步泛化,使其满足个性化(0.5,2,2)匿名模型的要求。最后一列是敏感属性的最终敏感等级,在满足个性化隐私保护规则的同时,满足对 α 、 l 和 k 值的要求。其中共三个等价组,分别为记录1~2、3~4和5~7,各等价组均包括两条及以上记录,等价组里不同种类的敏感值最少为两种,且符合出现频率 α 为0.5。

表6 个性化(0.5,2,2)匿名模型

ID	Gender	Age	Zip code	Disease	C	p	F
1	M	[21-25]	06****	血液循环疾病	4	3	4
2	M	[21-25]	06****	Flu	1	-	1
3	F	[21-25]	0622**	血液循环疾病	4	-	4
4	F	[21-25]	0622**	传染性疾病	4	-	4
5	F	[21-25]	0622**	呼吸道感染疾病	1	2	2
6	M	[56-60]	102***	循环系统疾病	4	-	4
7	M	[56-60]	102***	血液循环疾病	4	-	4

生成个性化(α, l, k)匿名算法:(以上文中疾病这一敏感属性为例)

输入:数据集 T ,参数 α 、 l 、 k ,敏感等级 C 与 p

输出:满足发布条件的数据集 T^*

(1)对 T 中的所有属性构建泛化树;

(2)计算 T 中记录的总条数 count, if (count = =

0), 则执行(7); else, 则继续执行(3);

(3) 引用文献[11]中提出的多属性泛化方法得到符合 k 匿名的数据集 T^1 ;

(4) 用 C 列值给 F 列赋初值, 对于每一条记录, if ($C < p$), 则修改 F 值为 p 值, 并按照泛化树将该敏感值泛化到相应等级, 获得 T^2 ;

(5) 依次在 T^2 中取出 k 条记录, 并将其存放到 t_i 中;

(6) i 从 $1 \sim n$ 遍历 $t = \{t_1, t_2, \dots, t_n\}$:

①若 $t_i \cdot \text{length}()$ 不为 0, 则 j 从 $1 \sim m$ 遍历 t_j , 将 t_j 中的记录按敏感值的 F 由高到低进行排序, 将 F 为最高级的敏感值泛化一级并替换原值;

②若 $t_i \cdot \text{length}()$ 不为 0, 则 k 从 $1 \sim m$ 遍历 t_k , 若 t_k 中不同敏感值个数小于 l , 则泛化较高 F 的敏感值,

且保证该敏感值满足 α 约束, 直到 t_k 中的敏感值均满足 l 和 α 约束;

③合并 t 作为 T^* 。

(7) 返回 T^* 。

4 实验结果与分析

本实验采用 UCI 机器学习仓库中 Adult 数据集, 此数据集被广泛应用于脱敏领域的研究实验。其中有 48 842 条记录, 可筛选出 30 162 条有效数据作为原始数据集 T 。选取 T 中的 6 个属性为 QI, 并添加一列 Disease 为 S 列, 属性的基本情况如表 7 所示。其中敏感属性列的敏感等级 C 值依旧应用上文的用户评分结果, 随机选取 2/5 的数据记录添加用户自定义的疾病敏感属性值, 并为表 7 中所有属性构建泛化树。

表 7 各属性基本情况

ID	名称	数据类型	属性类型	取值个数	泛化树高度
1	年龄	连续型	QI	74	6
2	性别	分类型	QI	2	2
3	种族	分类型	QI	5	3
4	国籍	分类型	QI	41	3
5	受教育程度	分类型	QI	16	4
6	工作类型	分类型	QI	8	3
7	疾病	分类型	S	8	4

实验环境: 硬件环境为 Intel Core i7-6700 3.40 GHz CPU, 8 GB RAM; 操作系统为 Windows 10; 编程语言为 Java。为了验证分析该算法的实用性, 将个性化 (α, l)-多样化 k -匿名模型和个性化 (p, α, k)-匿名模型在运行时间和信息损失量上作比较。固定 l 和 α 的值分别为 4 和 0.7, 每组实验反复运行 10 次, 剔除离群数据, 并取剩余值的平均数作为最后的取值。

4.1 运行时间分析比较

在相同实验条件下, 比较三种算法在 k 值大小的变化下, 运行时间的变化。由图 3 可知, 随 k 值变大, 三种模型的运行时间都会减少, 是由于等价组数量变少了, 要处理的次数也就变少了。由于该文所提算法

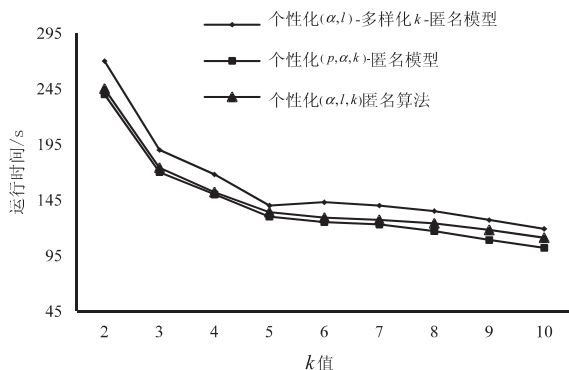


图 3 运行时间与 k 值的关系

较个性化 (p, α, k)-匿名算法增加了很多个性化的处理, 所以运行时间上会稍多。随着 k 值的增大, 对于多样性的要求越容易满足, 所以折线的斜率会小一些。

4.2 信息损失量分析比较

采用上文所给信息损失量公式(3), 在相同实验条件下, 比较三种算法在 k 值大小的变化下, 信息损失量的变化。如图 4 所示, 随 k 值变大, 三种模型的损失量也在变大, 是由于等价组里记录数变多造成的。该文所提算法不仅应用了个性化 (p, α, k)-匿名算法中对不同敏感等级的敏感值采取不同匿名方式的思想, 还引入了文献[11]中的多属性泛化方法来使数据集满足 k 匿名, 该方法针对属性过度泛化进行了深入研究, 因此所提模型信息损失量要低。

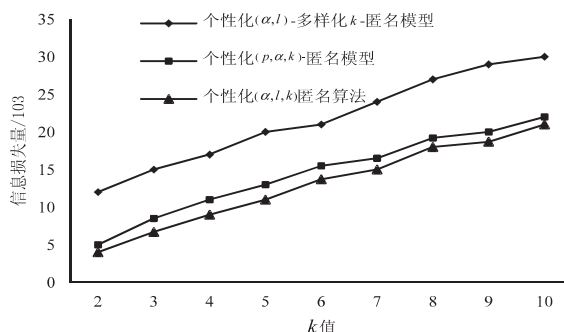


图 4 信息损失量与 k 值的关系

5 结束语

提出一种个性化 (α, l, k) 匿名模型。该模型在应用多属性泛化算法得到符合 k -匿名模型的基础上,将敏感属性在面向个人和面向敏感值这两方面进行匿名操作,且针对不同敏感等级的敏感值执行不一样的操作,同时使其满足各等价组里敏感值的多样性和频率。实验表明,该模型在有限的运行时间内,达到了较好的个性化隐私保护效果。但是该模型是面向单敏感值的匿名操作,在实际生活中会经常面临多敏感值的情况,所以提出可以应用于任意敏感属性个数的个性化匿名模型是后续研究工作的主要内容。

参考文献:

- [1] 魏大林. 支持隐私保护的数据发布技术研究[D]. 北京:北京交通大学,2015.
- [2] MIJUMBI R, SERRAT J, GORRICHIO J L, et al. Network function virtualization: state-of-the-art and research challenges[J]. IEEE Communications Surveys & Tutorials, 2016, 18(1): 236–262.
- [3] 康海燕, 杨孔雨, 陈建明. 基于K-匿名的个性化隐私保护方法研究[J]. 山东大学学报: 理学版, 2014, 49(9): 142–149.
- [4] TAO Y, XIAO X. Personalized privacy preservation[M]// Personalized privacy preservation. Berlin: Springer, 2006: 229–240.
- [5] LIU Xiangwen, XIE Qingqing, WANG Liangmin. A personalized extended (α, k) -anonymity model[C]//2015 third international conference on advanced cloud and big data (CBD). Yangzhou, China: IEEE, 2015.
- [6] YE Xiaojun, ZHANG Yawei, LIU Ming. A personalized (α, k) -anonymity model[C]//International conference on web-age information management. Zhangjiajie, Hunan: IEEE, 2008.
- [7] 曹敏姿, 张琳琳, 毕雪华, 等. 个性化 (α, l) -多样性 k -匿名隐私保护模型[J]. 计算机科学, 2018, 45(11): 180–186.
- [8] 张旭, 姚龙, 宋增源. 基于位置服务中关联攻击的隐私保护研究[J]. 电脑知识与技术, 2019, 15(4): 58–59.
- [9] 蒲东, 方睿. 个性化 (p, α, k) -匿名隐私保护算法[J]. 计算机应用与软件, 2020, 37(2): 301–307.
- [10] 陈菲. 数字图书馆信息服务平台的匿名发布研究[D]. 兰州: 西北师范大学, 2017.
- [11] 贾俊杰, 闫国蕾. 一种个性化 (p, k) 匿名隐私保护算法[J]. 计算机工程, 2018, 44(1): 176–181.
- [12] 张志祥, 金华, 朱玉全, 等. 基于有损连接的个性化隐私保护[J]. 计算机工程与设计, 2011, 32(9): 2938–2942.
- [13] 王伟, 郭献彬. 一种增强的匿名化隐私保护模型[J]. 信息通信, 2016(1): 1–4.
- [14] 彭福荣, 张怀锋. 统计基层数据中脱敏技术的应用[J]. 信息技术与信息化, 2018(11): 19–21.
- [15] 王波, 杨静. 一种基于逆聚类的个性化隐私匿名方法[J]. 电子学报, 2012, 40(5): 883–890.
- [16] 刘晓冰. 匿名化隐私保护算法的研究与实现[D]. 成都: 电子科技大学, 2011.
- [17] 宋明秋, 王琳, 姜宝彦, 等. 多属性泛化的K-匿名算法[J]. 电子科技大学学报, 2017, 46(6): 896–901.