多属性泛化的*K*-匿名算法

宋明秋，王 琳，姜宝彦，邓贵仕

(大连理工大学系统工程研究所 大连 辽宁 116024)

【摘要】针对现有的*K*-匿名模型中存在泛化属性选取不唯一和数据过度泛化的问题，提出多属性泛化的*K*-匿名算法。在*K*-匿名模型实现的过程中，引入属性近似度概念，定量刻画准标识符属性的离散程度，进而确定泛化的准标识符属性；同时采用广度优先泛化的方法，避免数据被过度泛化，最终实现数据表的*K*-匿名要求。实验结果表明，多属性泛化的*K*-匿名模型可以提高泛化后数据精度，其处理效率和Datafly算法相当。该算法有效地解决了取值最多准标识符属性存在多个时的泛化属性选取问题，并且防止属性被过度泛化，提高数据的可用性。

关 键 词 泛化；*K*-匿名；隐私保护；关系型数据

中图分类号 TP301.6 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.06.018

K-Anonymity Algorithm Based on Multi Attribute Generalization

SONG Ming-qiu, WANG Lin, JIANG Bao-yan, and DENG Gui-shi

(Institute of Systems Engineering, Dalian University of Technology Dalian Liaoning 116024)

Abstract Aiming at the major issues for data over-generalization and no unique attributes of *K*-anonymity model, a modified *K*-anonymity algorithm based on multiple attributes generalization is proposed in this paper. The conception of attribute approximation degree is introduced which describes the discrete degree of quasi-identifiers, and determines the candidate quasi-identifier attribute to be generalized. In the meantime, breadth-first generalization is exploited to avoid over-generalization and meets the *K*-anonymity requirements ultimately. The experimental results show that the new *K*-anonymity algorithm based on multiple attribute generalization can improve data precision and its efficiency is equal to Datafly algorithm. The proposed algorithm can effectively solve the issue of generalization attribute selecting when quasi-identifiers are not unique, the over-generalization of quasi-identifiers attributes can be avoided, and the usability of data can be improved.

Key words generalization; *K*-anonymity; privacy protecting; relational data

随着网络技术的高速发展，大量的个人信息被政府部门、科研机构等有关组织存储、发布，导致隐私信息被曝光，先进的数据挖掘算法在提高信息有效性的同时，也导致了隐私泄露的问题[1]。如何在数据共享的同时，实现有效合理的隐私保护方法[2]就显得尤为重要。

早在20世纪80年代初，文献[3]首次提出了匿名化的概念，并指出这种技术手段可应用于隐私信息的保护。文献[4]提出*K*-匿名模型的数据匿名化隐私保护方法，通过泛化和抑制[5]、分解和排列[6]以及微聚集和凝聚[7]等方式对原始数据进行匿名化处理，有效地解决了链接攻击问题。文献[8]在*K*-匿名模型的基础上提出了一种新的隐私保护模型即l-多样性模型，对数据表中的敏感属性进行相关约束，提升发布的数据表对于同质攻击或背景知识攻击等的防范。此后，(*a*, *K*)-匿名模型[9]使用*a*阈值对敏感属性进行约束；针对l-多样性模型在一些特殊情况下不适用的问题提出了*t*-closeness模型[10]，要求敏感属性接近全局分布；而(*K*, *e*)-匿名模型[11]和(*K*, l)-匿名模型[12]等模型为针对敏感属性为数值型数据的近似攻击提供了解决方案[13]。文献[14]用信息熵模型刻画属性的隐私程度，进而为信息泄露风险量化提供支撑。此外，*K*匿名模型在多领域的应用也成为现阶段的研究热点，文献[15]将*K*-匿名技术应用到社会网络图的隐私保护，应用位置服务数据[16]和快递信息[17]等隐私保护也采用*K*-匿名技术。

现有的*K*-匿名模型研究主要集中于高效近似算法的设计和多领域的应用，在模型算法实现中，没

收稿日期：2017－03－31；修回日期：2017－06－12
基金项目：国家自然科学基金面上项目(71171028)；国家科技支撑计划(2013BAH01B03)
作者简介：宋明秋(1967－)，女，博士，副教授，主要从事信息安全、隐私保护方面的研究.

有考虑取值最多的准标识符属性不唯一的情况，以及选取的准标识符属性一直被泛化，从而导致数据精度过低的问题。针对这一问题，本文讨论泛化属性选取方法，每次泛化操作之前通过准标识符属性取值的种类及其近似度选定泛化属性，有效避免单一属性的过度泛化，提高泛化后数据集的可用性。

1 *K*-匿名模型相关定义

1.1 数据匿名

定义 1 准标识符属性。给定数据集Ω，数据表$\mathrm{PT}(A_1,A_2,\cdots,A_n)$，且存在映射关系$f_c:\Omega\to\mathrm{PT}$以及$f_g:\mathrm{PT}\to\Omega'$，其中$\Omega\subseteq\Omega'$。准标识符属性是通过与外表关联可以唯一标识某一记录的属性[17]。

定义 2 *K*-匿名。给定数据表$\mathrm{PT}(A_1,A_2,\cdots,A_n)$，$A'$是与PT相关联的准标识符属性，当且仅当$\mathrm{PT}[A']$中的每个值序列至少在$\mathrm{PT}[A']$中出现*K*次，则称该数据集PT满足*K*-匿名[18]，如表1所示。

表1 *K*-匿名模型实例

序号	肤色	邮编	出生日期	性别
T1	白	1160**	1991	女
T2	白	1160**	1991	女
T3	白	1160**	1991	女
T4	黄	1161**	1989	男
T5	黄	1161**	1989	男
T6	黄	1161**	1989	男
T7	黑	1162**	1990	女
T8	黑	1162**	1990	女
T9	黑	1162**	1990	女
T10	黑	1162**	1990	女

定义 3 泛化。在数据匿名处理过程中，对数据采用模糊的、抽象的值替代原始数据的方式，称为泛化[19]，即用高层次的节点代替低层次的节点，泛化处理层次如图1所示。

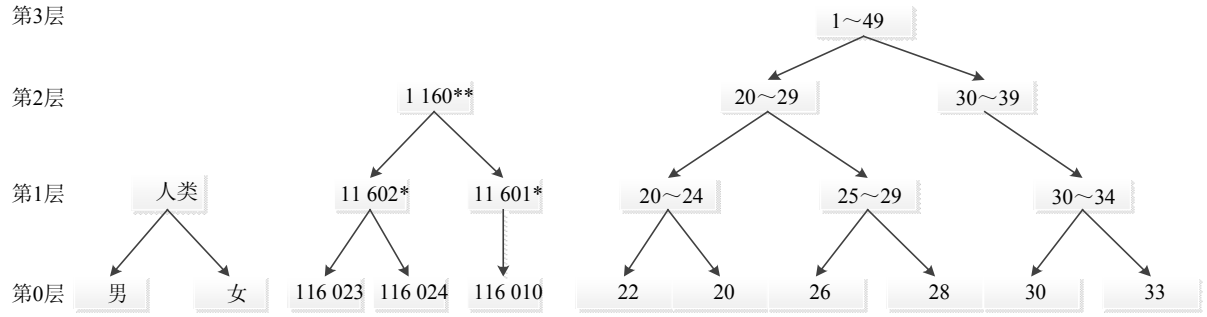

图1 准标识符属性泛化层次图

定义 4 等价组。数据表中准标识符属性取值完全相同的记录组合在一起称为一个等价组[20]。

1.2 数据精度

定义 5 数据精度。数据精度是衡量经过泛化后的数据与真实数据的逼近程度，即泛化后数据损失的度量。泛化后的数据与真实数据越接近，信息的损失量越少，泛化后的数据可用性越高。因此，数据精度是评价*K*-匿名模型实现算法的一个重要指标。

这里采用的数据精度度量标准是基于泛化层级的数据表精度度量标准Precision(PT)[21]，定义如下：

$$\mathrm{Precision(PT)}=1-\frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\frac{i}{\left|A_{ij}\right|}}{\left|\mathrm{PT}\right|\left|N_A\right|} \tag{1}$$

式中，$\left|A_{ij}\right|$是该属性可泛化的最高层次数；i是*K*-匿名后数据表PT中第j条记录的第i个属性被泛化的次数；$\left|\mathrm{PT}\right|$是给定数据表PT中记录的数量；$\left|N_A\right|$是给定数据表PT中所包含的准标识符属性数目；总的来看，Precision(PT)值越小，数据精度越小，数据的可用性越差。

1.3 Datafly算法分析

给定关系型数据表PT，准标识符属性集$\{A_1,A_2,\cdots,A_j\}$，每个准标识符A_j属性值的种类数为count(A_j)，属性A_i是count(A_j)最大的准标识符属性，按照泛化处理层次图对准标识符属性A_i进行泛化，直至数据表PT满足*K*-匿名要求。

在Datafly算法实现过程中，选取count(A_j)最大的某一准标识符属性A_i，且一直对A_i进行泛化，直至满足*K*-匿名要求，这就容易导致A_i被过度泛化，进而导致数据表PT的数据精度下降。除此之外，A_i可能同时存在多个，Datafly算法采用随机的方式选取泛化属性，若选取离散程度较大的A_i进行泛化，会减慢实现*K*-匿名要求的速度。针对上述问题，本文提出一种改进的Datafly算法，即多属性泛化的*K*-匿名算法，以期解决存在多个A_i时泛化属性的选取和数据被过度泛化的问题。

2 多属性泛化的*K*-匿名算法

2.1 基本思想

在多属性泛化的*K*-匿名算法中，需要匿名化处理的准标识符属性是由数据表中的准标识符属性值的种类决定。选取取值种类最多的属性作为优先泛化的属性，按其预先给定的泛化层次进行泛化。

首先，针对属性过度泛化问题。多属性泛化的*K*-匿名算法在每次泛化和*K*-匿名检验后都重新选取需要泛化的准标识符属性。这样降低了给定的数据表被过度泛化的可能性，加快关系型数据表满足*K*-匿名的要求，提高泛化后数据的可用性。

其次，针对属性选取不唯一问题，Datafly算法在泛化的准标识符属性选取这一环节中，都没有考虑取值最多的准标识符属性同时存在多个的情况。因此，引入属性近似度这一概念，依据准标识符属性近似度的值，选取近似度最大的准标识符属性优先进行泛化。

定义 6 属性近似度。准标识符属性的近似度即准标识符属性的域值之间的离散程度。准标识符属性的近似度越高，其属性值分布越不均匀，对其进行泛化不仅可以降低背景知识攻击等的威胁，还可以加快数据表满足*K*-匿名模型。

2.2 泛化属性选取

在多属性泛化的*K*-匿名算法中，只对关系型数据表中的准标识符属性进行泛化处理。实际的关系型数据表中通常存在多个准标识符属性，需要进行如下分析来选取优先泛化的属性：

1) 取值最多的准标识符属性只存在一个。

当取值最多的准标识符属性只存在一个的时候，多属性泛化的*K*-匿名算法选取这一准标识符属性进行泛化处理。

2) 属性值种类最多的准标识符属性不唯一。

当属性值种类最多的准标识符属性存在多个时，多属性泛化的*K*-匿名算法选取近似度值高的准标识符属性作为泛化属性，优先进行泛化。近似度高的准标识符属性的取值离散程度大，分布不均匀，使得某些等价组内记录条数过少，无法满足*K*-匿名要求。对近似度高的准标识符属性进行泛化，可以增加等价组内记录的条数，减少包含记录条数过少的等价组，进而加速实现*K*-匿名模型。

根据前面对准标识符属性近似度的定义，标准差反映一组数据的离散程度，故用标准差来描述准标识符属性的近似度，计算步骤和公式如下：

1) 统计数据表中准标识符属性各取值在数据表中的频数 f_i 及属性域值的总数量 n。

2) 该准标识符属性各取值的出现概率 p_i 和该属性取值的平均概率 $\overline{p_i}$ 为：

$$p_i = \frac{f_i}{n} \tag{2}$$

$$\overline{p_i} = \frac{1}{n}\sum_{1}^{n} p_i \tag{3}$$

3) 求出该准标识符属性的方差为：

$$D(x) = \frac{1}{n}\sum_{i=1}^{n}(p_i - \overline{p_i})^2 \tag{4}$$

4) 对 $D(x)$ 开方，得标准差 $\sigma(x)$ 即反映属性的近似度为：

$$\tau \Leftrightarrow \sigma(x) = \sqrt{D(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \overline{p_i})^2} \tag{5}$$

表2 医疗信息表

年龄	性别	邮编	身体状况
22	男	110024	胃溃疡
22	女	110031	感冒
33	男	110024	健康
33	女	110032	胃炎
28	男	110024	发烧
28	女	110033	感冒
30	男	110024	咽炎
30	女	110034	心脏病
26	男	110024	胃溃疡
26	女	110034	咽炎

例如，表2是医疗信息发布表，它包含3个准标识符属性{年龄，性别，邮编}和1个敏感属性{身体状况}。准标识符属性“年龄”有5个取值，依次为{22, 26, 28, 30, 33}：同样，准标识符属性“性别”、“邮编”的取值种类数依次为2、5。对该数据表的准标识符属性进行泛化处理，年龄和邮编这两个准标识符属性种类数取值最多且均为5。若采用多属性泛化的*K*-匿名算法，计算这两个属性的近似度 τ，得 $\tau_{\text{年龄}} = 0$，$\tau_{\text{邮编}} = 0.2$。由于 $\tau_{\text{邮编}} > \tau_{\text{年龄}}$，选取准标识符属性“邮编”先进行泛化。若采用Datafly算法将会在“邮编”和“年龄”这两个准标识符属性中随机选取一个进行泛化。假设要求表2泛化后满足*K*=2的要求，则经Datafly算法泛化后数据精度是0.583，而经多属性泛化的*K*-匿名算法泛化后数据精度是0.667，可见多属性泛化的*K*-匿名算法泛化后的数据可用性高于Datafly算法。

2.3 多属性泛化的*K*-匿名算法

在算法运行前，需要输入数据：泛化层次*K*值和

数据表中的准标识符属性及其泛化层次。

根据初始的设定对所输入的数据表进行*K*-匿名检验。如果数据表满足*K*-匿名，那么系统会自动将所输入的数据输出。如果数据表不满足*K*-匿名，多属性泛化的*K*-匿名算法进入准标识符属性分析选取阶段，即计算各准标识符属性的取值种类数。如果存在多个属性值种类最多的准标识符属性，那么计算种类最多的准标识符属性的近似度，选取其近似度最大的属性作为优先泛化属性。对该属性进行一次泛化。泛化完毕后，再次对处理后的数据表进行检验，验证数据表是否满足*K*-匿名要求。如果检验结果为“是”，那么系统将处理后的数据输出；如果检验结果为“否”，那么表格将再次进入泛化属性选取和*K*-匿名检验的循环，直到其符合*K*-匿名要求为止，步骤如下所示。

输入：关系型数据表PT，准标识符属性名称，给定*K*值，准标识符属性的泛化层次

输出：匿名处理后的数据表PT*

步骤：

m=0;

if(关系型数据表满足*K*-匿名)

输出匿名处理后的数据表;

else

计算每一个准标识符属性值的种类数

and找到属性取值种类数最多的准标识符属性;

if(属性值种类最多的准标识符属性为1)

选取该准标识符属性*A*；

else计算属性值种类最多的准标识符属性近似度

and选择近似度最高的属性*A*；

end if;

将该属性*A*按其泛化层次图从*m*层泛化至*m*+1层，得到数据表;

return 关系型数据表;

end if。

3 实验结果与分析

实验目的是对多属性泛化的*K*-匿名算法性能进行评价，评价指标为算法运行时间和泛化后数据精度，并将本实验的结果与经典的Datafly算法运行结果进行对比，客观地评价多属性泛化的*K*-匿名算法的性能。

3.1 实验数据集

本文实验选取了UCI Machine Learning Repository Adult数据集中的Adult.test文本文件作为实验的数据样本集[22]。采用文献[23]中的数据预处理方法对原数据集进行预处理，得到实验数据集中的16 008条数据作为最终的实验数据。实验数据囊括8个属性作为准标识符属性，1个属性作为敏感属性，实验数据的情况如表3所示，其中，QID为准标识符属性，SA为敏感属性。

表3　数据结构示意表

编号	名称	类型	属性类型	不同值个数	可泛化层级
1	年龄	数值型	QID	74	4
2	受教育时间	数值型	QID	16	4
3	婚姻状况	文本型	QID	7	2
4	种族	文本型	QID	5	2
5	性别	文本型	QID	2	2
6	每周工作时长	数值型	QID	99	2
7	国籍	文本型	QID	41	4
8	年收入	数值型	SA		

3.2 实验环境

硬件环境：Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz；8.00 GB内存；操作系统：Windows10；编程语言：C#。

3.3 实验分析

在算法运行时间这一维度上，多属性泛化的*K*-匿名算法除了受软硬件环境等一些客观因素的影响外，还受*K*取值和样本数据大小的影响。因此，在有关算法运行时间的测算中，着重从*K*的取值和样本数据大小这两个方面对该算法的运行时间进行测量。

1) 算法运行时间随*K*值变化情况

在样本数据量一定，算法运行时间随*K*值变化趋势的测量实验中，样本数据选定为处理后的数据集Adult.test中的所有数据(16 008条)。由于*K*值表示等价组内完全相同的记录数，且避免数据表的链接攻击，故*K*值为大于等于2的正整数。

上述实验结果中，当*K*大于10时，*K*值的增大对算法运行时间影响不大。因此*K*值均从2～200中递增选取，其中最小值为2，最大值为200，并将实验的执行结果以折线图的形式进行输出，如图2所示。

实验结果表明：当数据量一定时，*K*-匿名模型实现算法的运行时间会随着*K*值的增大而增加。当*K*值较小时，多属性泛化的*K*-匿名算法与Datafly算法的运行时间基本相同；但随着*K*值增大，多属性泛化的*K*-匿名算法的运行时间略微高于Datafly算法的运行时间。

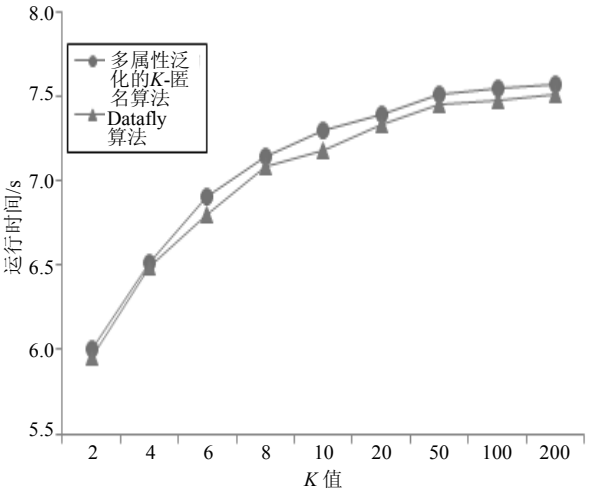

图2 算法运行时间随*K*值变化情况

2) 算法运行时间随数据量变化情况

在保持*K*值不变，算法运行时间随样本数据量变化趋势的测算试验中，设定*K*值为2，样本数据是从处理后的数据集Adult.test(16 008条数据)中依次选取10、20、50、100、500、1 000、10 000条数据作为实验测算样本，结果如图3所示。

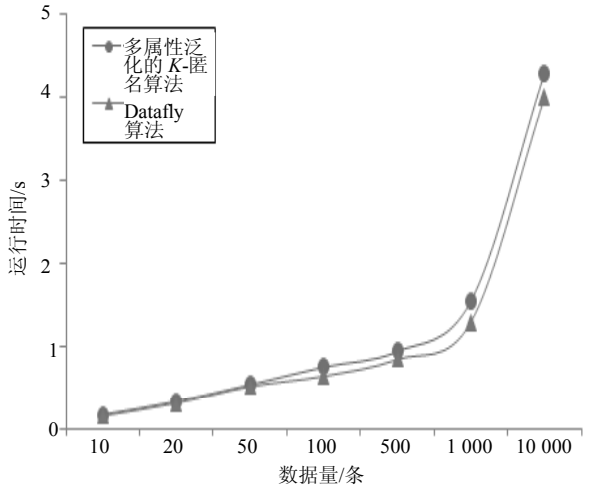

图3 算法运行时间随数据量的变化情况

实验结果表明：当*K*值一定时，*K*-匿名模型实现算法的运行时间会随着数据量的增大而增加。当数据量较小时，多属性泛化的*K*-匿名算法的运行时间与Datafly算法的运行时间基本相同。随着数据量的不断增大，多属性泛化的*K*-匿名算法的运行时间要略高于Datafly算法。原因是多属性泛化的*K*-匿名算法优先选取近似度大的准标识符属性进行泛化，增加等价组内记录条数，加快实现*K*-匿名要求。但在每次泛化前，均需重新计算确定取值最多的准标识符属性，选取近似度高的准标识符属性进行泛化，故多属性泛化的*K*-匿名算法和Datafly算法的总体运行时间相仿。

3) 数据精度测算结果及分析

实验选取Precision测算公式对泛化后数据进行精度测量。在*K*-匿名算法中，各准标识符属性的泛化程度是影响泛化后数据精度的最主要因素，而各准标识符属性的泛化程度由实验开始前所选取的*K*值决定。样本数据是经处理后的Adult.test中的所有数据(16 008条)，*K*值是从2～200中递增选取，其中最小值为2，最大值为200，并和经典Datafly算法进行对比，如图4所示。

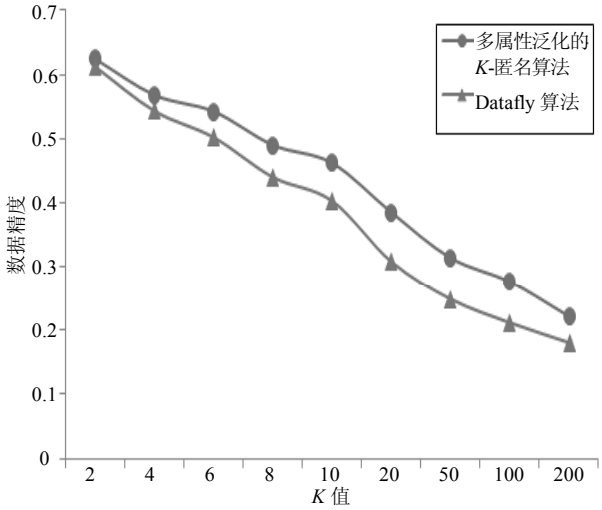

图4 数据精度随*K*值变化情况

实验结果表明：当数据量一定时，经*K*-匿名模型实现算法处理后的数据精度会随着*K*值的增大而减小。在*K*值为2时，两种算法处理后的数据精度几乎相同。不过，随着*K*值的不断增大，多属性泛化的*K*-匿名算法处理后的数据精度要明显高于Datafly算法处理后的数据精度。在多属性泛化的*K*-匿名算法中，每次泛化前均需重新选择被泛化的属性，有效解决Datafly算法中某一取值最多的准标识符属性达到最高泛化等级时，数据表仍旧不能满足*K*-匿名要求而导致属性被过度泛化的问题，故经过多属性泛化的*K*-匿名算法泛化后的数据精度高于经典Datafly算法泛化后的数据精度。

4 结 束 语

匿名算法的效率和处理后数据的可用性是衡量*K*-匿名算法的两个重要指标。针对经典Datafly算法存在泛化属性选取过于单一的问题，提出了多属性泛化的*K*-匿名算法。在该算法中，由准标识符属性值的种类数量确定需要优先泛化的准标识符属性；并针对泛化过程中可能出现取值最多的准标识符属

性同时存在多个的情况，引入属性近似度的概念，选取属性近似度最大的准标识符属性优先泛化，有效地控制属性过度泛化的问题，提高泛化后数据的可用性。通过与经典Datafly算法进行实验对比，多属性泛化的*K*-匿名算法泛化后数据精度更高，运算时间和Datafly算法相当，具有更好的实际应用价值。

参 考 文 献

[1] LIN Chi, SONG Zi-hao, SONG Hou-bing, et al. Differential privacy preserving in big data analytics for connected health[J]. Journal of Medical Systems, 2016, 40(4): 1-9.

[2] CHEN De-yan, ZHAO Hong. Data security and privacy protection issues in cloud computing[C]//2012 International Conference on Computer Science and Electronics Engineering. Hangzhou, China: IEEE, 2012, 1: 647-651.

[3] COX L H. Suppression methodology and statistical disclosure control[J]. Journal of the American Statistical Association, 1980, 75(370): 377-385.

[4] SWEENEY L. *K*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[5] SWEENEY L. Achieving *K*-anonymity privacy protection using generalization and suppression [J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002,10(5): 571-588.

[6] SUN Xiao-xun, WANG Hua, LI Jiu-yong, et al. Publishing anonymous survey rating data[J]. Data Mining and Knowledge Discovery, 2011, 23(3): 379-406.

[7] SORIACOMAS J, DOMINGOFERRER J, SANCHEZ D and MARTINEZ S. Enhancing data utility in differential privacy via microaggregation-based *K*-anonymity[J]. The VLDB Journal, 2014, 23(5): 771-794.

[8] MACHANAVAJJHALA A, KIFER D, GEHRKE J. L -diversity: Privacy beyond *K*-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2006, 1(1): 24.

[9] CHEN Rui, FUNG B C M, MOHAMMED N, et al. Privacy-preserving trajectory data publishing by local suppression[J]. Information Sciences, 2011, 231(1): 83-97.

[10] SORIACOMAS J, DOMINGOFERRER J, SANCHEZ D, et al. T-Closeness through microaggregation: Strict privacy with enhanced utility preservation[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(11): 3098-3110.

[11] 夏赞珠, 韩建民, 于娟, 等. 用于实现 (*k*,*e*)-匿名模型的 MDAV 算法[J]. 计算机工程, 2010, 36(15): 159-161.
XIA Zan-zhu, HAN Jian-ming,YU Juan, et al. MDAV Algorithm for implementing (*k*,*e*)-Anonymity model[J]. Computer Engineering, 2010, 36(15): 159-161

[12] 杨高明, 李敬兆, 杨静, 等. (*k*,*l*)-多样性数据发布研究[J]. 计算机科学, 2013, 40(8): 140-145.
YANG Gao-ming, LI Jing-zhao, YANG Jing, et al. Achieving(*k*,*l*)-diversity in privacy preserving data publishing[J]. Computer Science, 2013, 40(8): 140-145.

[13] LIU Qinghai, SHEN Hong, SANG Ying-peng. Privacy-preserving data publishing for multiple numerical sensitive attributes[J]. Tsinghua Science and Technology, 2015, 20(3): 246-254.

[14] 彭长根, 丁红发, 朱义杰, 等. 隐私保护的信息熵模型及其度量方法[J]. 软件学报, 2016, 27(8): 1891-1903.
PENG Chang-gen, DING Hong-fa, ZHU Yi-jie, et al. Information entropy models and privacy metrics methods for privacy protection[J]. Journal of Software, 2016, 27(8): 1891-1903.

[15] 刘向宇, 李佳佳, 安云哲, 等. 一种保持结点可达性的高效社会网络图匿名算法[J]. 软件学报, 2016, 32(8): 1904-1921.
LIU Xiang-yu, LI Jia-jia, AN Yun-zhe, et al. On reachability preserving graph anonymization in social networks[J]. Journal of Software, 2016, 32(8): 1904-1921.

[16] LI Xiu-hua, MIAO Mei-xia, LIU Hai, et al. An incentive mechanism for *K*-anonymity in LBS privacy protection based on credit mechanism[J]. Soft Computing, 2017, 21(14): 3907-3917.

[17] 韦茜, 李星毅. 基于*K*-匿名的快递信息隐私保护应用[J]. 计算机应用研究, 2014, 31(2): 555-557.
WEI Qian, LI Xing-yi. Express information protection application based on *K*-anonymity[J]. Application Research of Computers, 2014, 31(2): 555-557.

[18] OLIVEIRA S R M, ZAIANE O R. Privacy preserving clustering by data transformation[J]. Journal of Information and Data Management, 2010, 1(1): 37-51.

[19] 吕品, 钟珞, 王文兵, 等. MA-Datafly: 一种支持多属性泛化的*K*-匿名方法[J]. 计算机工程与应用，2013，49(4): 138-139.
LÜ Pin, ZHONG Luo, WANG Wen-bing, et al. MA-Datafly: *K*-anonymity approaches for supporting multi-attribute generalization[J]. Computer Engineering & Applications, 2013, 49(4): 138-139.

[20] HUNDEPOOL A, DOMINGOFERRER J, FRANCONI L, et al. Statistical disclosure control[M]. Chichester, UK: John Wiley & Sons Ltd, 2012.

[21] LI Tian-cheng, LI Ning-hui, ZHANG Jian, et al. Slicing: a new approach for privacy preserving data publishing[J]. IEEE Transactions on, Knowledge and Data Engineering, 2012, 24(3): 561-574.

[22] MURPHY P M, AHA D W. University of California Irvine machine learning repository[EB/OL]. (1996-02-15). http://archive.ics.uci. edu/ml/.

[23] 晏华, 刘贵松. 采用熵的多维*K*-匿名划分方法[J]. 电子科技大学学报, 2007, 36(6): 1228-1231.
YAN Hua, LIU Gui-Song. Multidimensional *K*-anonymity partition method using entropy[J]. Journal of University of Electronic Science and Technology of China, 2007, 36(6): 1228-1231.

编 辑 蒋 晓