基于 OLA 的 K 匿名算法的改进

胡翔天，宫秀军，陈海亮
（天津大学 计算机科学与技术学院，天津 300072）

摘 要：主要对数据匿名化中的一种重要方法 K-匿名进行了研究和分析，重点对 K 匿名算法中的一种较高效的算法最优泛化格 OLA（Optimal Lattice Anomy-zation）进行了介绍，并针对 OLA 为取得最优结果计算节点过多、时间过长的问题进行了进一步研究，在 OLA 算法的基础上提出一种基于节点度积优先（度积为父节点数与子节点数的乘积）的算法，该算法相较于 OLA，需要计算的节点数和时间都显著减少，对 OLA 算法有明显的改进。

关键词：K 匿名；最优泛化格；数据匿名化
中图分类号：TP391 **文献标识码**：A **文章编号**：1674-7720(2011)22-0068-04

Enhaning method based on OLA K-anoymity alogrithm

Hu Xiangtian, Gong Xiujun, Chen Hailiang
(School of Computer Science & Technology, Tianjin University, Tianjin 300072, China)

Abstract: Through researching and analyzing the K-anoymity, this paper focused on an efficient algorithm of the K-anoymity, named Optimal Lattice Anomy-zation(OLA). Then analyzing the defects of the OLA, found the OLA calculates too much nodes and costing too much time. To solve these problems, this paper raised a degree product priority algorithm based on the OLA. Through the experiments, we found the new alogrithm had much improvement.

Key words: K-anoymity; OLA; data anoymity

随着网络信息技术的发展，信息资源的共享大大提高了信息资源的利用价值。大量信息的共享在给统计研究带来方便的同时也对个人隐私带来了威胁。因此，在发布数据时要尽量保护数据中的隐私。

数据匿名化是发布数据时保护个人隐私的一种有效手段。数据匿名化常用的处理手段源于统计数据库中的数据处理方法，主要是通过以发布数据中的属性值的信息损失为代价，换取通过这些属性值再标识某些个体的准确性，同时尽可能保证发布数据的可用性，在发布数据的准确性和隐私保护之间达到一种平衡，与传统的保证发布数据整体趋势而牺牲单个数据记录准确性的隐私保护方法相比，为发布数据提供了更好的可用性。通常做法是数据收集者通过隐藏或改变数据中的部分信息，使得攻击者无法通过发布出去的数据唯一地推导出敏感信息所属的个体，从而实现对个体隐私的保护。K-匿名算法是一种重要的数据匿名化方法。K-匿名算法中的一种比较高效的算法叫做最优格匿名算法 OLA（Optimal Lattice Anomy-zation），此算法使用一种叫做格（Lattice）的结构，通过遍历该结构中的节点从而最后得到最优的节点。然而 OLA 遍历节点的顺序并不能够最大程度上减少需要计算的 Lattice 的个数。本文在 OLA 算法的基础上提出了一种度优先的节点遍历方式，即通过节点的度积大小来遍历节点，从而显著减少最优结果的计算时间。

1 K-匿名

K-匿名是一个典型的微数据发布模型。微数据定义为一条表达和描述个体信息的数据记录，为个体信息的载体。这些信息包括个体的标识信息（如姓名、身份证号等）、敏感信息（如病史等）以及一些非敏感信息（如性别）。每个信息都是以个体属性和相应的属性值匹配的方式作为微数据（记录）的某个分量[1]。K-匿名就是通过匿名化原始数据中的某些属性值以导出满足一定匿名要求的匿名数据集并用于发布，为保证数据的有效性，这些被泛化的属性一般是非敏感属性，对于敏感属性一般不进行匿名化，因为发布数据中的敏感属性通常是所研究的主要内容，如医院患者就诊记录中的疾病信息，

泛化该属性将导致发布数据失去意义。同时 K-匿名保证敏感属性值不对应到具体的个体。通常 K-匿名要求对应于任意一条投影到这些属性上的值行，该 k 条记录组成一个等价组，从而使个体隐藏在 k 条数据之中，而无法确定 k 条数据中具体哪一条记录是该个体对应的记录，从而达到对自由访问数据型数据隐私保护的目的。对于敏感属性这些对统计数据库统计结果相对重要的属性则保证数据的精确性，以属性值的部分损失换取隐私属性值的被保护。

为准确描述 K-匿名的概念，一般将发布数据表中的个体记录的属性分为标识符、准标识符、敏感属性三类。

标识符：标识符属性是指能够直接标识出个体身份的属性，如姓名、身份证号码、社会保险号码等，通过这些属性值能够直接确定具体的个体。

准标识符 QI(Quasi-Indentifiers)：也叫做类标识符属性，同时存在于发布数据表和外部数据源表中，利用此两种数据表进行连接的推演来表示个人隐私信息的一组属性[2]。不同的发布数据表可以根据不同的情况划分为不同的准标识符属性，通常准标识符由专家选择，而非用户随便选取。一般情况下可以以年龄、教育程度、性别、地区等作为准标识符。

敏感属性 SA(Sensitive-Attributes)：个人隐私属性。发布数据中，个体不希望其他用户知道的信息属性。例如个人的工资水平、患者就诊记录中的所患疾病。

等价组：在准标识符上的投影完全相同的记录组成的组。等价组中所有的记录在准标识符上的属性值完全相同，其他的属性值可以不同。

K-匿名准确描述：给定数据表 $T[A_1,A_2,\cdots,A_n]$，QI 是与 T 相关联的准标识符，当且仅当在 T[QI]中出现的每个值序列至少在 T[QI]中出现 K 次，则 T 满足 K-匿名。T[QI]表示 T 表元组在 QI 上的投影。

2 最优格匿名算法 OLA

OLA 算法是一种全局最优的 K-匿名算法[3]，它是在 Incognito[4]和 Datafly[5]的基础上进行改进而得到的一种方法。OLA 算法的主要步骤如下：

2.1 泛化格(Lattice)的建立

选取准标识符，并按照一定的标准进行泛化，可以得到各个属性的泛化层次，如图 1 所示为选取年龄为准标识符，根据年龄建立的泛化层次，图 2 为根据所属地区建立的泛化层次。

根据各个属性相应的泛化方式可以建立泛化格。令 $T_i(A_1,\cdots,A_k)$和 $T_j(A_1,\cdots,A_k)$是两个不同的表(即两者为 Lattice 中不同的节点，$(A_1,\cdots,A_k)$为数据的 k 个属性，A_i 为第 i 个属性的泛化等级或泛化高度)。这两个表为对同一数据的各个属性进行不同程度泛化的结果，它们构成泛化格中的两个节点，每个表都是对数据的一种泛化策略。

图 1 年龄的泛化层次

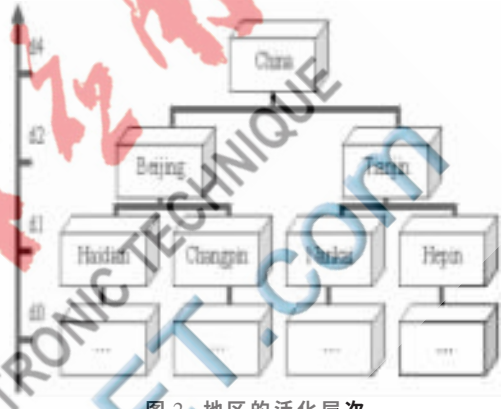

图 2 地区的泛化层次

泛化向量：$L(a_i,\cdots,a_k)$，其中 a_i 表示节点每个属性的泛化等级(或者泛化高度)。

泛化等级：$Level=\sum_{i=0}^{k}a_i$ 表示该节点在 Lattice 中的泛化等级(或者泛化高度)。

距离矢量：$\boldsymbol{DV}_{ij}=[d_1,\cdots,d_k]$，计算公式为：$d_i=(T_{jk}-T_{ik})$，其中，$d_i$ 为泛化等级中属性间路径长度。

两个或多个属性进行不同等级的泛化得到的结果构成属性泛化序列，这些序列构成基于准标识符的泛化等级序列，称为泛化格。图 3 为根据年龄和地区建立的一种泛化格。(i,j)中 i 表示年龄的泛化层次，j 表示地区的泛化层次。

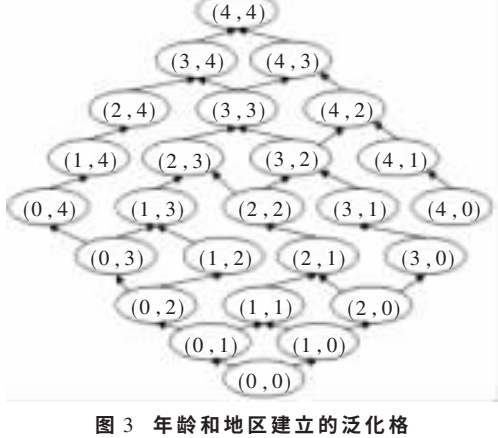

图 3 年龄和地区建立的泛化格

2.2 泛化格的遍历

建立完成泛化格后，需要对泛化格进行遍历以找出最优的泛化方式，OLA 在遍历时使用了 Datafly 的性质：(1)在一个泛化格中，若某一个节点 v 满足 K-匿名，则比 v 高的节点也满足 K-匿名；(2)若某个节点 v 不满足 K-匿名，则比 v 低的节点均不满足 K-匿名。通过这个性质遍历泛化格，可以对已遍历的节点进行标记，同时可以推测与之相关的节点是否满足 K-匿名，加快寻求 K-匿名节点的速度。

具体遍历方式如下：

(1)对于建立的泛化格，使用二分顺序遍历法，找到所有满足 K-匿名的节点。二分顺序遍历法是首先取泛化等级的最高值 L_{max} 和最低值 L_{min}，令 $L_{mid}=(L_{max}+L_{min})/2$，对于泛化等级为 L_{mid} 的节点依次判断是否满足 K-匿名，若满足，则将该节点的祖先节点标记为 K 匿名；如不满足，将该节点的子孙节点标记为不满足 K 匿名。然后以该节点为最低节点，递归地使用二分顺序遍历的方法，直到标记完所有节点。

(2)对于找到的满足 K-匿名的节点，根据单调性只保留高度最低的距离向量。例如：对于两个节点(2,3)、(2,2)都满足 K-匿名，因为节点(2,2)在节点(2,3)的下面，所以只保留节点(2,2)。

(3)如此得到一个最小的满足 K-匿名的节点的集合 k-minimal，计算该集合中每个节点的信息损失量。在各种文献中，有许多衡量信息损失的定义，Domingo-Ferrer[6]提到可以通过比较源数据和处理后数据的相似度来得到信息损失，参考文献[7]也给了类似的定义。本文采用的信息损失量的计算方式如下：

$$InfoLoss=\frac{\sum_{i=1}^{N}\frac{h_i}{DGH_i}}{N} \quad (1)$$

其中，N 表示元组集中的属性个数，DGH_i 表示第 i 个属性的最高泛化等级，h_i 表示属性 i 的当前泛化等级。由式(1)可知泛化程度越高，信息损失量越大；泛化程度越低，信息损失量越小。将信息损失量最小的节点作为最后的结果，这个结果即最优结果。

OLA 算法中最消耗时间的两个步骤是：判断一个节点是否为 K-匿名节点和比较 k-minimal 中所有节点的信息损失量。因此本文以尽量减少需要进行 K-匿名判断的节点的数量作为切入点对其进行改进。

3 算法的改进

OLA 采取的二分遍历法，将会遍历较多的节点，为此本文采取一种度优先的方法对泛化格中的节点进行遍历。把 Lattice 中一个节点的父节点数和子节点数分别叫做该节点的出度和入度，定义一个节点的度积为该节点出度和入度的乘积。改进后的算法的简要步骤如下：

(1)数据预处理：建立泛化格(Lattice)的步骤与 OLA 建立泛化格的情况相同。

(2)最优节点选择算法：

①首先计算 Lattice 中所有节点的度积。

②从 Lattice 中找到度积最大的节点。

(a)判断该节点是否满足 K-匿名。如果该节点满足 K-匿名，可知该节点的所有父节点都为 K-匿名节点。从 Lattice 中删除该节点及其所有祖先节点；然后查找已保存的 k-minimal 的集合，看该集合中是否有该节点的祖先，若有，则从 k-minimal 集合中将其删除；若无，则不操作。最后把该节点保存到 k-minimal 中。

(b)如果该节点不满足 K-匿名，则可知该节点的所有子孙节点都不是 K-匿名节点。从 Lattice 中删除该节点及其所有的子孙节点。

(c)比较所有保存在 k-minimal 集合中节点的信息损失量。信息损失量最小的那个节点，即为所查找的全局最优节点。

该算法的流程图如图 4 所示。

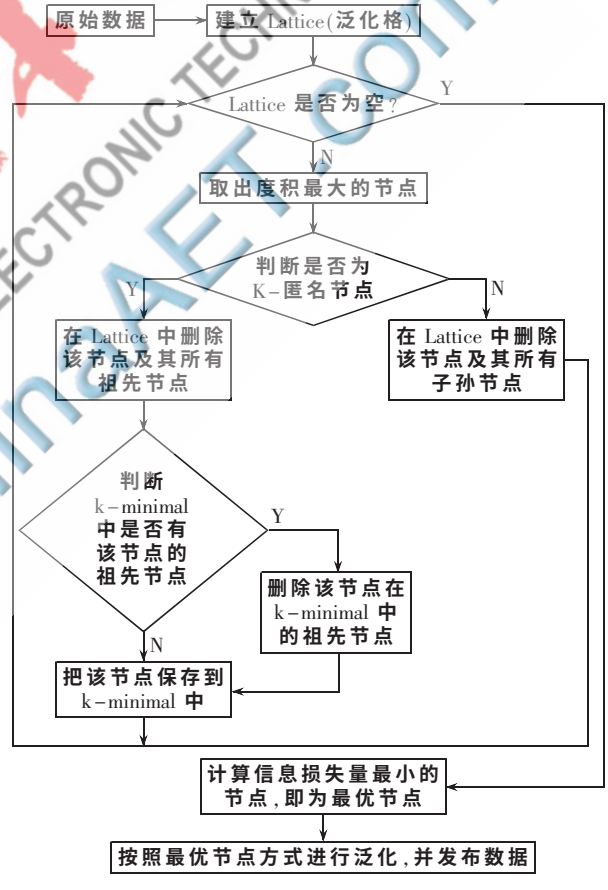

图 4 改进 OLA 算法流程图

4 实验采用的数据及结果

实验使用的数据如表 1 所示。这个数据集为公共数据集，该数据来自 UC Irvine 机器学习储藏室，是美国人口普查中抽出的数据，该数据集已经被很多类似的研究使用过[5,8]。实验时，从数据集中将标识符(姓名、身份证号等)属性和隐私属性去掉，留下准标识符，对准标识符根据其语义建立泛化层次。数据集的准标识符的选取以

及泛化高度如表1中第二列所示。第三列是数据的条数，第四列是建立的Lattice的节点的数目。

表1 实验使用的数据集

数据集	准标识符的选取	行数	节点数
Adult	Age(3), Profession(2), Education(2), Maritalstatus(2), Position(2), Sex(1), Race(1), Gender(2)	30 162	5 184

将OLA和度优先均用于这个数据集，然后将运行的结果加以比较。图5、图6为实验结果。

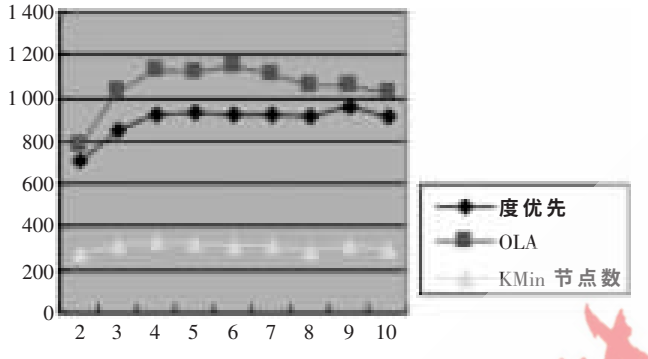

图5 计算满足K-匿名节点数量

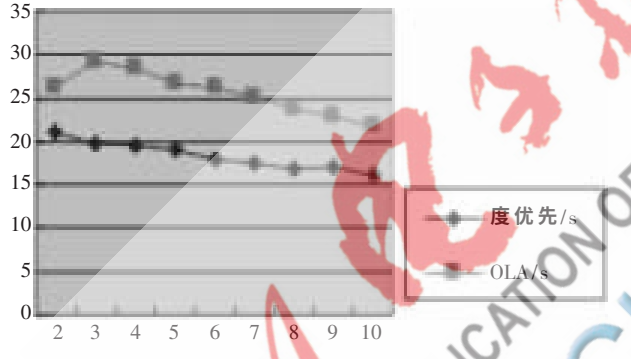

图6 两种算法的运行时间

从两个方面评定算法的执行效率，一方面通过读取源数据判断节点的数量，另一方面是算法的运行时间。图5为两种算法需要计算的节点数量的比较，最下面的折线为最小K-匿名节点的数量。从中可以看出度优先需要计算节点数比OLA要少。图6为两个算法计算完成时间的对比，明显可以看出度优先运行的时间比OLA要少，可见度优先计算K-匿名的算法比OLA要好。

本文介绍了隐私保护中K-匿名的相关概念，简单叙述了K-匿名的一种较好的算法OLA，并针对OLA在遍历Lattice格计算节点过多这一问题进行改进，提出了度积优先的遍历算法。通过OLA和度优先算法对相同数据的实验，可以看出度积优先的算法相对OLA有明显提高。取得最优结果后，按照该结果的泛化方式处理数据，可以得到最终发布的数据。

参考文献

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

[2] DALENIUS T. Finding a needle in a haystack-or identifying anonymouscensus record[J]. Journal of Official Statistics, 1986, 2(3): 329-336.

[3] EMAM K, DANKAR F, ISSA R J, et al. A globally optimal K-anonymity method for the de-identification of health data [J]. J Am Med Inform Assoc, 2009, 16(5): 670-82.

[4] SWEENEY L. Achieving K-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, 10(5): 18.

[5] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Incognito: EfficientFull domain K-anonymity Proc[C]. ACM Management of Data, Baltimore, USA: ACM, 2005: 49-60.

[6] DOMINGO-FERRER J, TORRA V. Risk assessment in statistical microdata protec-tion via advanced record linkage [J]. Journal of Statistics and Computing, 2003, 13(4).

[7] XU J, WANG W, PEI J, et al. Utility-based anonymization using local recoding[C]. 12th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, USA: ACM, 2006: 785-790.

[8] BAYARDO B, AGRAWAL R. Data privacy through optimal K-anonymity[C]. In Proc. of the 21st Int'l Conference on Data Engineering. IEEE CS, 2005: 217-228.

(收稿日期：2011-08-23)

作者简介：

胡翔天，男，1986年生，硕士研究生，主要研究方向：数据挖掘，生物信息学。

宫秀军，男，1972年生，副教授，硕士生导师，主要研究方向：数据挖掘，生物信息学，网格计算。

陈海亮，男，1985年生，硕士研究生，主要研究方向：数据隐私保护。