**数据发布中个人敏感属性数据隐私保护方法研究**

**摘要**

**英文摘要**

**目录**

1. **绪论**

1.1数据发布研究背景及意义

1.2数据发布中隐私研究保护研究现状

1.3数据发布中的隐私规则

1.4本文主要研究内容与组织架构

1. **数据发布中的隐私保护**

2.1 隐私保护中的相关理论

2.2 隐私保护中数据发布匿名模型介绍

2.3 隐私保护原则与信息度量

2.4 本章小结

1. **多维敏感属性数据发布中的隐私泄露**

3.1 多敏感属性数据发布中的问题研究分析

在对数据发布中的隐私保护技术研究初期，大多数敏感数据发布方法都是针对单一敏感属性的保护。但是，在实际的应用中，发布的数据大多数都会涉及到多个敏感属性，特别是这些敏感属性在某些情况下会存在一些关联关系，一些属性虽然对于发布个体不是直接的敏感属性，但是这些属性却和个体的敏感属性有着明显的特定关系，所以这样的属性也应该归类到个体的敏感属性被保护。例如表3-1-1 为将要发布的原始医疗信息，从表中可以看出，敏感属性主治医(Physician) 和疾病(Disease)之间存在着关联性，通过背景知识我们也可以知晓，一个主治医生专治哪几种疾病，其具体的关联性如表3-1-2 所示。

表3-1-1 原始医疗信息表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tuple ID | Name | Age | Sex | Race | Zipcode | Physician | Disease |
| t1 | Sam | 23 | M | White | 821071 | John | Flu |
| t2 | Anne | 44 | F | Black | 821023 | John | Pneumonia |
| t3 | Mike | 56 | F | White | 821045 | John | Cancer |
| t4 | Lily | 35 | M | Black | 821123 | Bob | Flu |
| t5 | Harry | 25 | F | White | 821031 | Bob | Pneumonia |
| t6 | Mona | 30 | M | Black | 821035 | Anne | Gastritis |
| t7 | Tony | 40 | F | White | 821110 | Anne | Gastritis |
| t8 | Lucy | 37 | M | Black | 821115 | Hugo | HIV |
| t9 | Tim | 60 | M | White | 821134 | Marry | Flu |

3-1-2 属性间关联表

|  |  |
| --- | --- |
| Physician | Disease |
| John | Flu,Pneumonia,Cancer |
| Bob | Flu,Pneumonia |
| Anne | Gastritis |
| Hugo | HIV |
| Marry | Flu |

现在加入我们需要对3-1-1表数据内容进行数据发布，并以l-diversity 匿名算法为例进行。我们对表3-1-1原始医疗信息进行匿名话处理，并通过分组算法得到数据发布表，结果如表3-1-3 所示。由表3-1-3 可以看出，由于医院中医生的主治哪些疾病是可以很容易获取，也就是攻击者能够很容易获得表3-1-2内容的背景知识，若攻击者得知该个体的Physician 属性值为“John”再联合攻击者掌握的个体的准标识属性确定个体属于GroupID为3的分组。此时，攻击者推测出该个体的Disease 属性值的概率将高于1/3，这就违反了L多样多样性原则，造成隐私泄露风险升高。这正是由于敏感属性Physician和Disease 间存在着关联性，即使数据发布表中的敏感属性满足l-diversity 匿名模型，依然存在隐私泄露风险。因此针对单敏感属性的隐私保护技术并不能直接用于多敏感属性数据发布，否则会给个体隐私数据的保护带来很大的挑战，所以对于多敏感属性的数据发布仍存在隐私泄漏风险。因此，为了适应实际应用中的数据发布，面向多敏感隐私属性的数据发布方法的研究应受到重视。

表3-1-3 医疗发布数据3-diversity匿名表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group ID | Age | Sex | Race | Zipcode | Physician | Disease |
| t1 | [20,30] | M | White | 8210\*\* | John | Flu |
| t5 | [20,30] | F | White | 8210\*\* | Bob | Pneumonia |
| t6 | [20,30] | M | Black | 8210\*\* | Anne | Gastritis |
| t8 | [31.40] | M | Black | 821\*\*\* | Hugo | HIV |
| t4 | [31,40] | M | Black | 821\*\*\* | Bob | Flu |
| t7 | [31,40] | F | White | 821\*\*\* | Anne | Gastritis |
| t3 | [41,] | F | White | 821\*\*\* | John | Cancer |
| t2 | [41,] | F | Black | 821\*\*\* | John | Pneumonia |
| t9 | [41,] | M | White | 821\*\*\* | Marry | Flu |

3.2 多敏感隐私属性的数据发布方法

因为多敏感属性数据隐私保护有一些特殊的要求，为了防止由于缺少整体性而造成失去隐私敏感属性的连锁攻击，许多学者提出了针对多敏感属性数据的隐私保护模型和方法。文献[3]提出了多敏感属性l-diversity 概念，并对其进行如下定义：

多敏感属性l-diversity. 设数据表T 中有若干个准标识符属性QI 和敏感属性SA，从QI 中任意选取一个属性将其作为唯一的敏感属性，其余的准标识符属性和敏感属性均作为准标识符属性，则此时若数据表T 均满足l-diversity，则说明数据表T 满足多敏感属性l-diversity。

由以上定义可知，多敏感属性l-diversity 规则要求每个敏感属性上的每一个敏感值与所有其他敏感属性上对应的敏感值的个数不少于l 个[12]，这一概念可以很好的解决多敏感属性数据发布的隐私保护问题。但是，当敏感属性个数增加时，每个等价类为了满足多敏感属性l-diversity 规则就必须包含更多的记录，这必然会导致数据表泛化程度加剧，从而造成大量的信息损失。

3.2.1基于多维桶分组技术的隐私数据发布方法

2008 年，杨晓春等人[2] 首次提出了以l-diversity 模型为基础的多维桶分组技术来解决多敏感属性数据发布的隐私保护问题。

多维桶分组技术的基本思路是[17]，首先，将多个敏感属性看成一个高维复合敏感属性向量，也就是说，一个敏感属性对应一维；其次，使用多维桶的向量模型，将数据表中的记录映射到多维桶上；最后，按照某种方法在构造的多维桶上进行分组操作，使分组中的记录尽可能是在各维度上取值都不相同的桶中提取出来的。

多维桶分组技术重点是将多个敏感属性作为高维复合敏感属性来构造桶，并提出以下定义：

（1）复合敏感属性[17]. 数据表T 中所有的敏感属性构成一个复合敏感属性，记作S。其中第i 个敏感属性可看作复合敏感属性的第i 维，记为Si，D（Si）为Si 的取值范围，|Si|为D（Si）的大小。

（2）复合敏感属性向量[17]. 设数据表T，T 中任意记录t 的全部敏感属性取值构造成向量模式<t[s1]，t[s2]，…，t[sd]>，称这样的向量模式为复合敏感属性向量。

（3）分组[45]. 一个分组是T 中记录的子集。T 中每一个记录属于且仅属于一

个分组，关系T 的分组记为GT{G1，G2，……，Gm}， 并且（1  i  j  m）。

（4）单敏感属性l-多样性[17] . 对于一个由只包含单敏感属性的记录组成的分组G，假设v 为G 中出现频度最大的敏感属性取值，如果则G 满足l-多样

性性质，c(v)为v 出现的次数，|G|为G 中记录的个数。

（5）复合敏感属性l-多样性[17]. 对于一个包含复合敏感属性的分组G，如果G

中的每一维敏感属性上都满足单敏感属性l-多样性，则可称分组G 对于复合敏感属性满足l-多样性性质。

多维桶的构造原理如下：复合敏感属性的每个维度对应多维桶的一维，将数据表T中的数据记录根据其复合敏感属性向量每一维的值分别映射到相应的桶中[17]。设数据表T 的敏感属性个数为d，构造的d 维桶记为BUK（S1，S2，……，Sd），其中每个桶记为buk（s1，s2，……，sd），每个桶的大小记为size（buk（s1，s2，……，sd）），即包含的记录数。以表2-1 的原始数据为例，建立多维桶如表3-4 所示。最后，根据约束条件采用某种策略提取记录进行分组。

3.2 面向多敏感属性的个性化隐私保数据发布的方法

1. **面向多维敏感属性的数据发布算法**

4.1 多敏感属性隐私数据发布

4.1.1 有损连接发布—Anatomy

4.1.2 相关定义

为比便于描述和理解，我们首先引入一下相关定义。

4.2 基于二部图匹配的安全数据分组算法

4.3 实验结果及分析

4.4 本章小结

1. **面向多敏感属性的（L，a）-diversity 个性化匿名模型**

5.1 多维敏感属性数据的个性化发布

5.2 (l，a)-diversity 个性化匿名模型

5.3 最大权边二部图匹配分组算法

5.4 L-拆分二部图匹配算法

5.4 实验结果对比及分析

5.5 本章小结

1. **总结与展望**

**参考文献**

1. Anatomy: Simple and Effective Privacy Preservation
2. 数据发布中面向多敏感属性的隐私保护方法
3. Machanavajjhala A,Kifer D,Gehrke J. et al. l-diversity: privacy beyond k-anonymity[C]// Proceedings of the 22nd International Conference on Data Engineering(ICDE’06). New York: ACM,2006:24-35