**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 9 | 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化处理，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的，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Group ID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，造成隐私泄露的风险升高。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。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，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：

假设用户待发布的数据表为T{A1，A2，A3，···，Ap，S1，S2，···Sd}，其中Ai（1ip）是待发布数据T的准标识属性（QI），p代表准标识属性的个数。Sj（1jd）是待发布数据T的敏感属性，d代表敏感属性的个数。设T中记录个数为n，即|T|=n，那么发布数据表中每条记录记为ti（1in），另t[X]标识记录t在X属性上的取值，其中X{A1，A2，A3，···，Ap，S1，S2，···Sd}。

（1）复合敏感属性.。待发布数据表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，指该敏感属性的所有取值，|Si|为D（Si）的基数，指该敏感属性取值的个数。

（2）复合敏感属性向量[2]. 待发布数据表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[4]. 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，T中所有记录的分组记为GT{G1，G2，G3，···，Gm}，其中m为最终分组数，并且 =T，并且=（1ijm）。

（4）单敏感属性L-多样性[5]。对于一个分组G，G中只包含单敏感属性的记录，假设v为G中出现频度最大的敏感属性取值，且c(v)为v在G中出现的次数，如果满足，|G|为G 中记录的个数。

（5）复合敏感属性L-多样性分组. 对于一个包含复合敏感属性的分组G，如果G中的任一一维敏感属性Si（1id）都满足单敏感属性L-多样性，则该分组满足复合敏感属性满足l-多样性。那么对于T中所有分组GT{G1，G2，G3，···，Gm}，如果其中每个分组Gi（1im）都满足复合敏感属性L-多样性性质，则称GT为T上的复合敏感属性L-多样性分组。

有了以上定义以后，我们知道待发布数据表T若存在多个敏感属性，仅仅满足单敏感属性的的L-多样性原则无法保证所有敏感属性的隐私保护需求，针对多敏感属性的数发布，需要满足多敏感属性中每个属性都满足L-多样性原则，即需要得到待发布数据表T的复合敏感属性L-多样性分组。杨晓春等人提出基于有损连接技术的支持多敏感属性的隐私数据发布多维桶分组技术，目标是找到具有多敏感属性的待发布数据表T的分组方案，使得到的分组均满足复合敏感属性L-多样性。

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。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）），即含的记录数。以表 3-1-1 的原始数据为例，建立多维桶如表3-1-4 所示。

表 3-2-1 由表3-1-1构造的d维桶（d=2）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Flu | Pneumonia | HIV | Gastritis | Cancer |
| John | {t1} | {t2} |  |  | {t3} |
| Bob | {t4} | {t5} |  |  |  |
| Anne |  |  |  | {t6,t7} |  |
| Hugo |  |  | {t8} |  |  |
| Marry | {t9} |  |  |  |  |

在得到表3-1-4的多维桶后，分别采取MSB的贪心算法，包括最大桶优先算法，最大单维桶优先算法和最大多维桶优先算法来得到最后的满足复合敏感属性L-多样性分组的发布数据T’。

基于多维桶的分组算法为复合敏感数据发布的隐私保护提供了分组方案，并解决了多敏感属性下的隐私泄露问题，但仍然存在一些不足之处：

（1）多维桶分组技术在实现分组算法的时候，只考虑了最后分组对原数据的覆盖率问题，以贪心策略尽可能的得到更多的满足复合敏感属性L-多样性的分组，没有考虑到在实际数据发布中，每一维敏感属性的取值中，可能存在敏感度高低的问题，例如医疗数据发布表当中，个体敏感属性Disease取值中，HIV的敏感度肯定高于Flu。这就导致，多维桶分组算法在得到待发布数据表的所有分组后，可能会存在某个分组中，某个敏感属性的取值均是敏感属性度较高的情况，出现敏感信息分布切斜的现象，从而导致个体的敏感信息泄露。

（2）多维桶分组技术考虑到实际应用中某些元组包含着更重要的信息，需要尽量保留在发布的数据中，从而提出加权多维桶分组技术。加权多维桶分组技术只是考虑到部分重要数据的权值，在分组时先将权值高的记录先加入分组，这种方式虽然达到了保留重要数据的目的，但是由于重要数据一般都是敏感属性较高的数据，这种加权分组方式在实际应用中很有可能加剧敏感信息倾斜，造成个体隐私泄露。为解决这一问题，我们应该在保留高权值敏感信息的情况下同时保证高敏感度信息在分组中分布的均匀性。

3.2.2 基于L-覆盖性聚类分组的隐私数据发布方法

金华等人分析多维桶分组算法存在分组效率低，可能会存在由于每次分组选取桶的顺序问题造成大量不必要的数据记录遭到隐匿的问题，进一步对多敏感属性的医疗数据发布方法进行研究，提出了基于有损连接技术和相同敏感属性集的L-覆盖性聚类分组算法[6]。在多维桶提出的多敏感属性L-多样性的基础上给出以下定义：

（1）移除。对于一个分组G，若移除G中任意一条记录中的某一敏感属性值t[Si]（tG，1id）,需要将G中所有包换t[Si]的记录都删除。

（2）多敏感属性L-覆盖性。在一个分组G中，若至少需要移除L个敏感属性值，才能将G中所有记录移除，则称分组G满足多敏感属性L-覆盖性。

（3）相同敏感属性集。对于待发布数据表T中，包含同一敏感属性取值的所有记录组成的集合称为相同敏感属性集，记为SID（v）。待发布数据表T中，记录t的所有敏感属性值t[Si]的相同敏感属性集的并集称为记录相同敏感属性集，记为t.TSID，t.TSID=(t[Si])。对于一分组，中所有记录的记录相同敏感属性集的并集称为分组相同敏感属性集，记为.GSID,.GSID=。

（4）平均概率泄露度。由于需要处理完所有剩余记录，数据集T中肯定会存在仅仅满足多敏感属性L-覆盖性而不满足复合敏感属性L-多样性的分组。设得到的发布数据分组为GT{G1，G2，G3，···，Gm},||L，G中所有敏感属性的不同取值个数为n，每个敏感属性的取值为（1in）,在的出现的频率记为，则分组的概率泄露度定义为,则数据集的平均概率泄露度为=，其中|T|标识待发布数据集T中记录总数。

L-覆盖性聚类分组方法的主要思想是通过聚类的方法将满足L-覆盖性的记录进行分组，采用聚类思想，首先在数据集中顺序选取L个满足L-覆盖性的记录构成分组。然后对剩余记录分两步处理：第一步处理剩余记录中可以添加到其他分组而且仍满足多敏感属性L-多样性性质的记录；第二步是处理第一步剩余的记录，将剩余记录均匀地添加到分组较小的分组中，以降低平均概率泄露度。我们以表3-1-1为原始数据表为例，利用L-覆盖性聚类分组算法进行分组。另L=3，首先将数据集T中的每一条记录看做是一个分组，并计算他们各自的GSID。根据GSID的定义我们可以得到G1.GSID={t1,t2,t3,t4}，G2.GSID={t1,t2,t3,t5,t9}，G3.GSID={t1,t2,t3}，G4.GSID={t1,t4,t5,t9}，G5.GSID={t2,t4,t5}，G6.GSID={t6,t7}，G6.GSID={t6,t7}，G7.GSID={t6,t7}，G8.GSID={t8}，G9.GSID={t1,t4,t9}。按照L-覆盖性聚类算法顺序选取规则，首先选取G1，然后选择不在G1.GSID中含有的记录分组G6，得到G1,6.GSID={t1,t2,t3,t4,t6,t7}，然后选取不在G1,6中含有的记录分组G8，此时已得到一个分组{t1,t6,t8}，然后移除记录t1，t6，t8。算法依次循环进行得到另一个分组{t2,t7,t9}，剩余记录{t3,t4,t5}无法满足L-覆盖性（L=3），t5能够加入到第一个分组中得到{t1,t5,t6,t8}满足复合敏感属性L-多样性（L=3），

t3与t4加入到第二个分组得到{t2,t3,t4,t7,t9}，满足L-覆盖性（L=3）,最终得到的发布结果如表3-2-2所示。

表 3-2-2 L-覆盖性积累分组算法发布数据结果

|  |  |  |
| --- | --- | --- |
| Tuple ID | QIs | Group ID |
| t1 | ··· | G1 |
| t2 | ··· | G2 |
| t3 | ··· | G2 |
| t4 | ··· | G2 |
| t5 | ··· | G1 |
| t6 | ··· | G1 |
| t7 | ··· | G2 |
| t8 | ··· | G1 |
| t9 | ··· | G2 |

|  |  |  |
| --- | --- | --- |
| Group ID | Physician | Disease |
|  | John | Flu |
| Bob | Pneumonia |
| Anne | Gastritis |
| Hugo | HIV |
| G2 | John | Pneumonia |
| John | Cancer |
| Bob | Flu |
| Anne | Gastritis |
| Marry | Flu |

L-覆盖性分组算法解决了多敏感属分组后存在的剩余记录的问题，并提出了隐私平均概率泄露度的概念，但是含有敏感属性的数据发布中，最重要的就是保护个体记录的隐私不被泄露，所以该算法的不足之处也显而易见：

（1）满足L-覆盖性的复合敏感属性分组，并不一定满足复合敏感属性L-多样性，特别是在剩余记录存在大量相似敏感属性的时候，由于L-覆盖性算法在分组完成后为了将剩余记录全部加入已分组当中，会造成部分分组中敏感属性相同的记录出现的概率大与1/L。

（2）L-覆盖性分组算法总是顺序选取记录来进行分组，而且没有考虑敏感属性的度量问题，该算法虽然解决了剩余记录问题，但却在个性化分组和分组效率上存在比较大的局限性，分组效果通常不理想，在剩余记录较多的情况下，容易造成个体记录的隐私泄露。

3.3 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
4. 孙岚, 郭旭东, 王一蕾, 吴英杰. 个性化隐私保护轨迹发布算法[J]. 系统工程与电子技术.2014,36(12):2550-2555
5. Machanavajjhala A，Gehrke J，and Kefer D．／-diversity：Privacy beyond k-anonymity//Proceedings of the 22nd International Conference on Data Engineering．Atlanta，Georgia,
6. 面向多敏感属性医疗数据发布的隐私保护技术 – 金华 刘善成等人