面向查询服务的数据隐私保护算法

朱 青 赵 桐 王 珊

(中国人民大学数据工程与知识工程教育部重点实验室 北京 100872)

(中国人民大学信息学院计算机系 北京 100872)

摘 要 个性化信息服务提高了 Web 查询精度,但同时也带来数据隐私保护的问题.尤其在面向服务的架构(SOA)中,部署个性化应用时,如何解决隐私保护,这对于个性化服务是一个挑战.随着隐私安全成为微数据发布过程中越来越重要的问题,好的匿名化算法就显得尤为重要.论文总结了前人研究中考虑到准标识符对敏感属性影响的 k-匿名算法,提出了直接通过匿名化数据计算准标识符对敏感属性效用的方法以及改进的效用矩阵,同时为了更好地衡量匿名化数据的信息损失,论文中提出了改进的归一确定性惩罚的评价指标,从匿名化数据隐私安全的角度进行分析,实现了改进 L-diversity 算法,即基于信息损失惩罚的满足 L-diversity 的算法.它是准标识符对不同敏感属性效用的、并具有较好隐私安全的改进算法.

关键词 隐私保护;k-匿名;L-差异;SOA;服务计算

中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2010.01315

Privacy Preservation Algorithm for Service-Oriented Information Search

ZHU Qing ZHAO Tong WANG Shan

(*Institute of Data and Knowledge Engineering of Ministry of Education, Renmin University of China, Beijing* 100872)

(*Information School, Renmin University of China, Beijing* 100872)

Abstract Personalized information services offer a promising way to improve the accuracy of Web search, but they bring about additional requirements related to data privacy preservation. Nevertheless, current SOA usually have one of the main barriers for deploying personalized search applications, and how to do privacy-preserving personalization is a great challenge. Privacy becomes a more and more serious concern in service-oriented information search, so good algorithms are in need to be designed. In this paper, the authors considered the previous research on k-anonymity involving the influence of Quasi-identifier on sensitive attribute Bottom-up k-anonymity and present a method for calculating the influence of Quasi-identifier on sensitive attribute Bottom-up k-anonymity through microdata directly and improved utility matrix. To better evaluate the information loss of the anonymity data, the authors also present a quality metric, both the two major factors: data utility and privacy guarantee are well preserved, Improved Normalized Certainty Penalty (INCP). To achieve better privacy protection, the authors present a method based on the utility of Quasi-identifier which is L-diversity satisfied.

Keywords privacy preserving; k-anonymity; L-diversity; SOA; service computing

---

收稿日期:2010-06-11.本课题得到上海市高可信计算重点实验室开放课题(07dz22304)资助.**朱 青**,女,1963 年生,副教授,研究方向为可信数据管理与可信服务计算、数据的隐私保护、高性能数据库.E-mail: zq@ruc.edu.cn. **赵 桐**,男,1988 年生,研究方向为数据的隐私保护. **王 珊**,女,1944 年生,教授,博士生导师,研究领域为高性能数据库、可信数据管理.

1 引 言

网络数据查询服务中,搜索引擎为信息检索提供了方便,尤其是个性化信息服务为提升搜索引擎查询结果的高质量服务(QoS)提供了保障.但是,个性化服务需要收集和集成大量的用户个人信息,精确地描述用户的个性特征和个性模型,开放的网络环境同时带来个性化数据隐私保护的巨大挑战.如何在成功进行高质量查询服务的同时保证隐私数据不被泄露是一个重要的课题.再则,传统的数据挖掘技术旨在从大量的数据中抽取出潜在的、有价值的知识(模型或规则)[1],数据挖掘技术在发现知识、信息获取的同时也对数据隐私保护构成了威胁.

面向查询服务的数据隐私保护中常用的个人属性标识有显示标识符:能唯一标识单一个体的属性,如身份证号码、姓名等;准标识符(Quasi-identifier):组合起来能唯一标识一个人的多个属性,如邮编、性别、生日等的联合表示;敏感属性:包含敏感数据的属性,尤其是涉及隐私的,如疾病、个人薪资、病人患病记录、单位财务信息等.数据的隐私保护将保护数据的敏感属性.

实例 1. 医疗控制中心原始医疗数据[10],每一条记录对应一个唯一的病人{姓名,年龄,性别,邮编,疾病},其中{"姓名"}为显示标识符属性,{"年龄","性别","邮编"}为准标识符属性,{"疾病"}为敏感属性.当用户发出医疗数据查询服务时,为了保护数据隐私,在数据发布的过程中,原始数据表中的显示标识符会被去掉,所以作为隐私保护原则,隐去用户的姓名,如图 1 所示.{~~姓名~~,年龄,性别,邮编,疾病}.在公开含有隐私信息的数据时,为了保护隐私,常将一些能够进行个人定位的属性隐藏不予公开,例如:姓名;但仅仅隐藏这部分属性是不够的,尤其在开放的网络环境下,其它属性的组合也可能起到定位效果.例如:选民登记数据服务,可根据用户请求列出{姓名,年龄,性别,邮编}(如图 2 所示),根据两次查询服务,用户很容易推断出:Betty 有艾滋病(如图 3 所示),所以,Betty 的个人隐私受到侵害.这种属性组合称为准标识符,利用它们的组合来将隐私信息与某个个人联系起来,从而造成隐私泄露的攻击方法就被称为重标识攻击.

年龄	性别	邮编	疾病
25	F	12300	艾滋病
29	F	14000	肺炎
38	M	13500	支气管炎
37	M	13010	流感
40	M	13400	支气管炎
26	M	12600	流感

图 1 隐藏姓名的医疗数据

姓名	年龄	性别	邮编
Betty	25	F	12300
Linda	35	F	13000
Bill	21	M	12000
Sam	35	M	27000
John	28	M	14000

图 2 选民登记数据服务

姓名	年龄	性别	邮编	疾病
Betty	25	F	12300	艾滋病

图 3 个人隐私数据泄漏

个性化查询服务中,基于限制发布的隐私保护技术是一类通过有选择地发布原始数据、不发布数据或者发布精度较低的敏感数据以实现隐私保护的技术.其中,"数据匿名化"是研究的焦点."数据匿名化"是通过将原始数据进行匿名化处理,使得数据在隐私披露风险和数据精度之间进行折衷,从而兼顾数据的可用性和数据的隐私安全性.

常用的数据匿名化,在不同的应用中,不同的准标识符对敏感属性会有不同的效用.例如,考虑匿名化一个患者的数据集来做疾病分析.假设为了满足 k-匿名,可以把一个 5 位的邮编泛化为一个有 4 位有效数字的邮编(例如从 12300 泛化成 1230*).或者,也可以把病人的年龄泛化成一个区间(例如从 25 泛化成[20,30]).在多种情况下,准确的年龄信息对于疾病的分析非常重要,而邮编精度下降 1 位则通常是可以被接受的(一个 5 位邮编的前 4 位仍然表示一个相对较小的区域).所以年龄信息尽可能保持准确.

本文提出一种面向查询服务的数据隐私保护模型,如图 4 所示.对于用户提出的查询请求,采用数据隐私保护改进的 k-匿名算法,在满足用户查询服务的同时,有效地保护了数据隐私.弥补了当前大多数 k-匿名算法存在的不足.尤其是采用准标识符效用的 k-匿名算法(即准标识符对敏感属性影响的方法)在提高数据精度的同时,显著地降低了匿名化数据对一致性攻击的抵御能力.使得算法在面向查询服务的数据隐私保护方面的性能仍不能令人满意.如何平衡 k-匿名算法与效用计算是面向查询服务面临的挑战.

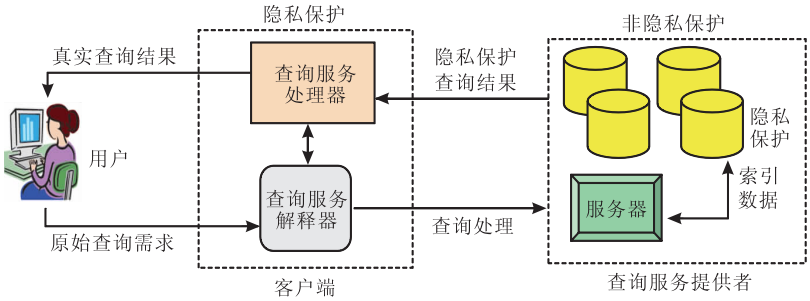

图 4　面向查询服务的数据隐私保护模型

因此，面向查询服务的数据隐私保护算法的设计目标是：提供给用户的查询结果达到质量保障、安全可信、高效服务三者的协调与统一. 质量保障是指最终能够提供给用户的查询服务结果是准确的满足(QoS)服务质量保证的结果；安全可信是指对隐私数据提供了足够的保护；高效服务是指系统开销、响应时间和时间复杂度小. 这三方面的要求在现实信息服务中往往相互矛盾，追求平衡和兼顾是算法研究的关键.

本文主要贡献：提出了反映不同准标识符对不同敏感属性值的全新效用矩阵；给出了反映匿名化数据信息损失相对完善的保证指标；提出了充分考虑准标识符效用的并满足 L-diversity 匿名化原则的算法；使其权衡几方面因素：即服务质量、信息损失、运行时间和数据隐私保护等，经实验验证，综合效果表现较好.

本文第 2 节介绍相关工作；第 3 节给出数据隐私保护术语与模型；第 4 节描述改进的效用匿名化算法；第 5 节为实验和算法性能评测；第 6 节为总结与未来工作展望.

2　相关工作

面向互联网各种开放式查询服务，在提供共享信息的同时，个人隐私的保护是一个挑战. 自 k-匿名被提出，为了达到更好地保护数据隐私的目的，已有多种改进型的变种数据匿名化原则和数据匿名化算法相继推出. 如 L-差异(L-diversity)匿名化、T-closeness 匿名化等. k-匿名方法主要作用于准标识符，针对重标识攻击(Re-identification). L-差异(L-diversity)匿名化原则[3]，考虑了同一个等价类中敏感属性多样性的问题，要求等价类中至少有 L 个不同的敏感属性；T-closeness 匿名化原则[4]在 L-diversity 匿名化原则的基础上考虑了等价类内敏感属性的分布问题，要求等价类中敏感属性及其值的分布差异不超过 T；在匿名化算法方面，由于在很多简单限制条件下的最优化 k-匿名问题已经被证明是 NP-hard，因此，多数关于 k-匿名的算法研究的焦点是设计高效的近似算法[2]. 早期提出的 MinGen 算法[5]采用完全搜索，时间复杂度高. Datafly 算法[6]在 MinGen 算法的基础上，引入了抑制与启发式泛化指导原则对效率进行了提升. Aggarwal 等[7]提出了基于聚类的匿名化算法. Pei 和 Xu 等[8]提出并解决了基于动态递增数据的多次发布问题. 但是这些算法都没有考虑到准标识符对敏感属性的效用.

Xu 和 Wang 等[9]提出了一种基于效用的匿名化算法，其考虑到在不同的应用中，不同准标识对敏感属性的效用. 给不同准标识符赋予不同的权重，从而降低了匿名化数据的信息损失. 同时提出了一种自底向上的基于效用的 k-匿名算法(Bottom-up k-anonymity)，它的主要思想是在整个数据集上依据最小化归一确定性惩罚的原则合并等价类. Li 等[10]提出的算法在文献[9]的基础上进一步考虑了不同准标识符对不同敏感属性值的效用，考虑到准标识符效用的算法，为本文中提出基于信息损失惩罚的满足 L-diversity 的算法起到了借鉴作用. 文献[10]中的 k-ACLUDK 算法，它的主要思想是首先基于效用矩阵将相似的敏感属性值聚类到一起，然后对于聚类结果中的每一类，采用随机合并等价类.

3　数据隐私保护术语与模型

3.1　准标识符对敏感属性的效用

为了降低数据对连接攻击的敏感性，从而保护数据隐私，Samarati 和 Sweeney 提出了 k-匿名原则. k-匿名原则要求所发布的数据表中的每一条记录不能区分于其它 $k-1$ 条记录[12]. 一般 k 值越大，对隐私的保护效果越好，但丢失的信息越多，常常会导致信息效用与隐私保护之间的不平衡.

k-匿名形式化描述为：给定一个参数 k 和准标

识符$(A_{i1},\cdots,A_{il})$,一个表 T 是 k-匿名的要求对于每一条记录 $t\in T$,存在至少其它 $k-1$ 条记录 $t_1,\cdots,t_{k-1}$ 满足这 k 条记录在准标识符上有相同的投影,即

$$t_{(A_{i1}\cdots A_{il})}=t_{1(A_{i1}\cdots A_{il})}=\cdots=t_{k-1(A_{i1}\cdots A_{il})},$$

记录 t 和全部其它在准标识符上不能和 t 相互区分的记录构成一个等价类.

不同准标识符对不同敏感属性值的效用也往往是不同的.考虑实例 1 的数据集,对于很多遗传性疾病,性别比邮编对疾病分析更为有用;而对于地方性疾病,例如"地方性甲状腺肿",邮编则比性别有更大的效用.因此,考虑不同准标识符对不同敏感属性值的效用有助于降低匿名化数据的信息损失.

数据发布者寻求发布数据不仅安全而且有用.效用矩阵测量数据中的有用信息,针对数据发布衡量不同候选的效用.效用矩阵定义为

$$\boldsymbol{U}=(u_{ij})_{m\times n}=(P_{s_iq_j})_{m\times n}.$$

效用矩阵中的元素 u_{ij} 表示准标识符中的第 j 个属性对敏感属性取第 i 个值的效用,它等于 $P_{s_iq_j}$,即准标识中第 j 个属性 q_j 对第 i 个敏感属性 s_i 的取值概率.对于数值型数据,$P_{s_iq_j}$ 为对应于 s_i 时,q_j 的取值范围与在整个数据集上 q_j 的取值范围之比.对于分类数据,$P_{s_iq_j}$ 为对应于 s_i 时,q_j 的取值个数与在整个数据集上 q_j 的取值个数之比.

采用概率作为效用.概率越低,认为效用越明显,概率越高,效用越不明显.

3.2 基于信息损失惩罚 L-diversity 算法

考虑不同准标识符对不同敏感属性值的影响,提出基于信息损失惩罚的满足 L-diversity 的算法(ILP-based L-diversity).下面先给出几个概念.

敏感属性值间的距离. 给定 2 个不同的敏感属性值 s_l 和 s_k,那么它们之间的距离为

$$D_{lk}=\sum_{j=1}^{n}|u_{lj}-u_{kj}|,$$

其中 n 是准标识符中属性的个数.敏感属性值间的距离反映了敏感属性值间的相似程度.

信息损失惩罚(Information Loss Penalty,ILP). 按准标识符为数值型和分类型分别进行计算

(1)数值型准标识符

设 T 是一个表,其准标识符为$(A_1,\cdots,A_n)$,设 $A_1,\cdots,A_n$ 均为数值型属性.假设一个记录 $t=(x_1,\cdots,x_n)$经过匿名化后变为

$$t'=([y_1,z_1],\cdots,[y_n,z_n]),$$

其中 $y_i\leq x_i\leq z_i(1\leq i\leq n)$.对于数值型准标识符属性 A_1,ILP 的定义[9]为

$$ILP_{A_i}(t)=\sum_{j=1}^{m}w_{ij}\frac{z_i-y_i}{|A_i|},$$

其中$|A_i|=\max\limits_{t\in T}\{t.A_i\}-\min\limits_{t\in T}\{t.A_i\}$,是所有记录在准标识符属性 A_i 上的取值范围,m 为敏感属性值的个数.对于每一对准标识符属性 A_i 和敏感属性 S_j 的组合,权重 w_{ij} 反映准标识符属性 A_i 对敏感属性 S_j 的效用.因此对于一条记录的全部数值型准标识符,ILP 定义为

$$\begin{aligned}ILP_n(t)&=\sum_{j=1}^{m}\sum_{i=1}^{n}(w_{ij}\cdot ILP_{A_i}(t))\\&=\sum_{j=1}^{m}\sum_{i=1}^{n}\left(w_{ij}\cdot\frac{z_i-y_i}{|A_i|}\right),\end{aligned}$$

其中 n 为数值型准标识符属性的个数,m 为敏感属性值的个数.

(2)分类型准标识符

假设表 T 有分类型准标识符 A,记录 t 在属性 A 上的值为 v.设当记录 t 被匿名化后,记录 t 在属性 A 上的值被泛化为集合$\{v_1,\cdots,v_l\}$,其中 $v_1,\cdots,v_l$是记录 t 所在的等价类中所有记录在属性 A 上的取值.对于分类型准标识符属性 A,ILP 定义为

$$ILP_A(t)=\sum_{j=1}^{m}w_{ij}\frac{size(u)}{|A|},$$

其中$|A|$是所有记录在准标识符属性 A 上取的不重复值的个数,m 为敏感属性值的个数.$size(u)$的定义如下,设 $v_1,\cdots,v_l$ 是一棵层次树的所有叶节点,u 是层次树中的一个结点,它是 $v_1,\cdots,v_l$ 的祖先并且 u 不存在任何后代也是 $v_1,\cdots,v_l$ 的祖先,那么称 u 为 $v_1,\cdots,v_l$ 的最近共同祖先(closet common ancestor),记为 $ancestor(v_1,\cdots,v_l)$.以 u 为祖先并且为叶结点的结点个数就定义为 $size(u)$.因此对于一条记录的全部分类型准标识符,ILP 定义[9]为

$$ILP_c(t)=\sum_{j=1}^{m}\sum_{i=1}^{n}(w_{ij}\cdot ILP_{A_i(t)}),$$

其中 n 为数值型准标识符属性的个数,m 为敏感属性值的个数.

综合数值型和分类型的准标识符,一条记录的 ILP 定义为

$$ILP(t)=ILP_n(t)+ILP_c(t).$$

全表的 ILP 即是表中所有记录的 ILP 之和,定义为

$$ILP(T)=\sum_{t\in T}ILP(t)=\sum_{t\in T}\sum_{j=1}^{m}\sum_{i=1}^{n}(w_{ij}\cdot ILP_{A_i(t)}),$$

其中 n 为数值型准标识符属性的个数,m 为敏感属性值的个数.

信息损失惩罚这个指标可以较全面地反映匿名

化数据的信息损失.

4　改进的效用匿名化算法

4.1　基于信息损失惩罚算法设计

基于信息损失惩罚的满足 L-diversity 的算法考虑了不同准标识符对不同敏感属性值的效用.算法首先根据待匿名的数据计算出效用矩阵,然后通过效用矩阵对敏感属性值进行聚类,使相似的值处于同一个类中.之后先将数据集中的每条记录初始化为等价类,然后依据敏感属性值的聚类结果将等价类进行聚类.最后在数据记录的每个类中,以最小化信息损失惩罚为原则,采用自底向上的方式以贪心法对等价类进行合并,生成等价类.算法最后对所有等价类中的记录选择性地进行数据失真操作,从而使等价类满足 L-diversity 匿名化原则.

系统模型整体结构分 5 步:(1)根据表 U 计算效用矩阵.可以通过数据集计算,或直接从专家处获取效用矩阵.(2)对敏感属性值进行聚类.设计并实现 Clustering 算法.(3)将数据集中的每条记录初始化为等价类,然后根据敏感属性值的聚类结果将所有等价类依据其敏感属性值进行聚类.(4)采用自底向上的贪心算法,同时采用数据失真的方式使得匿名化数据满足 L-diversity,但是在保证数据隐私的同时降低匿名化数据的数据真实性和准确性,为权衡效用和匿名之间的关系,算法的设计可根据具体情况处理.

4.2　根据效用矩阵聚类敏感属性值

对敏感属性值进行聚类.聚类个数为 m/L,其中 m 为敏感属性值的个数,L 为 L-diversity 的参数 L,当聚类中心收敛后,算法对聚类的结果进行检查,如果聚类结果中存在敏感属性值的个数小于 L 的类,则重新进行聚类.如果重新聚类的次数超过阈值,仍不能满足所有类中敏感属性值的个数均不小于 L,则将聚类的个数减 1,然后再重新聚类.这样做保证了匿名化数据的等价类中的敏感值个数不小于 L,是匿名化数据满足 L-diversity 的前提.

算法先“根据效用矩阵将敏感属性值进行聚类”,作为子程序单独先写出来,为后面的程序整体提供被调用的子函数算法:

依据效用矩阵对敏感属性值进行聚类

Clustering 算法.

输入:敏感属性值,效用矩阵,参数 l

输出:敏感属性值的聚类结果

```
{ 根据敏感属性值的个数和参数 l 计算聚类的初始个数 N;
  随机选取 N 个敏感属性值作为 N 个聚类的初始中心;
初始化聚类次数计数器;
while 存在敏感属性值个数小于 l 的聚类 do
{ while 存在中心未收敛的聚类 do
    for 全部敏感属性 do
    { 计算敏感属性值到每个聚类的距离;
      将敏感属性值分配到距离最近的聚类;}
for 全部聚类 do
{ 更新聚类中心; } }
聚类次数++
if 聚类次数超过限制 then
{ 聚类的个数 K 减少 1;
    重置聚类次数计数器;}
}
```

4.3　基于信息损失惩罚的 L-diversity 算法

基于信息损失惩罚的满足 L-diversity 的算法;根据表 T 计算效用矩阵.除了通过数据集计算外,效用矩阵也可以直接从专家处获取.

$$\boldsymbol{U}=(u_{ij})_{m\times n}=(P_{s_iq_j})_{m\times n}.$$

准标识符对敏感属性不同值的效用由矩阵 $\boldsymbol{U}$ 表示,其中 m 是敏感属性可以取值的个数,n 是准标识符中的属性个数.在矩阵中 $s_i(1\leq i\leq m)$ 是敏感属性的第 i 个取值,$q_j(1\leq i\leq m)$ 是准标识符中的第 j 个属性.矩阵中的元素 t_{ij} 表示准标识符中的第 j 个属性对敏感属性取第 i 个值的效用.

$$\boldsymbol{U}=(U_{ij})_{m\times n}=\begin{bmatrix} & q_1 & \cdots & q_n \\ s_1 & 0.8 & \cdots & 0.6 \\ \vdots & \vdots & \cdots & \vdots \\ s_m & 0.25 & \cdots & 1.0 \end{bmatrix}.$$

采用概率作为效用,在效用矩阵中 $U_{11}=0.8$,表示准标识符中第 1 个属性 q_1 对第 1 个敏感属性值 S_1 的取值概率为 0.8.概率越低,效用越明显,概率越高,效用越不明显.论文通过匿名化数据直接计算效用矩阵,矩阵元素是连续型的数据.改进后的算法见文献[10],效用矩阵由专家给出,即效用被分为 5 个级别:非常高、高、一般、低、非常低.

算法考虑准标识符对敏感属性,尤其是不同敏感属性的影响,使得具有一个敏感值的等价类的比例更高.为了解决这一问题,算法在满足 k-匿名的基础上进一步满足 L-diversity 原则,从而极大地提高了匿名数据对一致性攻击的抵抗力.算法采用贪心法,在等价类的聚类结果的每一个类中合并等价类.对每一趟处理,算法依据ILP最小化的原则合

并等价类，直到所有等价类的大小均不小于 k. 在等价类合并的过程中，如果等价类的大小已经大于等于 k，则这个等价类将不会进行下一轮的合并，最终所有等价类的大小均在 k 至 $2k$ 之间，等价类合并算法的时间复杂度是 $O(\log_2 k)$.

改进的 L-diversity 算法.

输入：表 T，参数 k，参数 l

输出：表 T'

```
{ 根据表 T 计算效用矩阵；
  根据效用矩阵对敏感属性值进行聚类；
  将表 T 中的每条记录初始化为 1 个等价类；
  根据敏感属性值的聚类结果对等价类进行聚类；
  for 全部记录的聚类 do
  { while 存在记录个数小于 k 的等价类 do
      for (全部记录个数小于 k 的等价类 G)) do
      { 在聚类中搜索等价类 G',
        使得 ILP(GUG')最小，合并 G 和 G' }
    while 存在敏感属性值小于 l 的等价类 do
    for (全部敏感属性值小于 l 的等价类) do
    { 对等价类进行数据失真操作，
      增加 1 种敏感属性值；}
  } 输出表 T 的匿名化结果表 T';
}
```

算法对所有的等价类进行检查，如果等价类中只有 1 个敏感属性值，我们认为这个等价类对于一致性攻击非常敏感. 算法对等价类进行的数据失真操作是：随机选择等价类中的一条记录，将其敏感属性值随机地改变为同一个敏感属性值聚类结果中的其它敏感属性值. 这样做人为地改变了记录和等价类的敏感属性值，降低了数据真实性，但是由于只是将敏感属性值改变为同一个敏感属性值聚类结果中其它的敏感属性值，所以数据失真程度相对较低.

基于信息损失惩罚的满足 L-diversity 算法的主要的时间开销在其第 4 部分，即采用自底向上的贪心法对等价类进行合并的这一部分. 在经过 k 轮等价类合并后，如果第 k 轮不是最后一轮，那么每个等价类中的记录数为 2^k，如果第 k 轮是最后一轮，那么每个登记类中的记录数在 2^k 和 2^{k-1} 之间. 因此经过至多 $\lceil \log_2 k-1 \rceil$ 轮，每个等价类将有至少 k 条记录. 而每一轮将在表 T 的一部分上进行，其平均大小为 T_1/m. 因此对于每个类，时间复杂度为 $O\left(\lceil \log_2 k-1 \rceil \left|\frac{T_1}{m}\right|^2\right)$，类的个数为 $\left\lfloor \frac{m}{l} \right\rfloor$，$0 \leq \left|\frac{l}{m}\right| \leq 1$，所以算法的时间复杂度为 $O(\lceil \log_2 k \rceil |T|^2)$.

5 实验与测试

5.1 实验环境设置

下面通过实验比较 Bottom-up k-anonymity[9]、KACLUDK[10] 和 ILP-Based l-diversity 这 3 种算法. 实验环境：Intel Core i3 2.13GHz 处理器，4.00GB 内存，Window 7 操作系统. 全部的算法使用 Python 2.6.4 实现，编译器是 PythonWin 2.6.4.

实验数据集采用 UCI 机器学习数据中的真实数据集 Adults 作为本实验的数据集，这一数据集是测试 k-匿名的基准数据集. 按照文献[13]中的方式对数据集进行预处理. 属性值有缺失的记录被去掉. 最后用于实验的数据集共包含 30162 条记录.

实验选用 {*age*, *education number*, *marital status*, *native country*, *race*, *salary class*, *sex*, *work class*} 这 8 个属性作为准标识符属性，其中 {*age*, *education number*} 为数值型属性，{*marital status*, *native country*, *race*, *salary class*, *sex*, *work class*} 为分类型属性. 在分类型属性中 {*marital status*, *work class*} 这 2 个属性具有多层结构，{*native country*, *race*, *salary class*, *sex*} 这 4 个属性具有 2 层结构. 实验选用 {*occupation*} 作为敏感属性，{*occupation*} 共有 14 个值.

5.2 实验及结果评估

实验中 Bottom-up k-anonymity 算法的准标识符的权重我们是根据效用矩阵计算出来的. 对于每个准标识符，其对所有敏感属性的权重之和经过单位化处理后即为这个准标识符在 Bottom-up k-anonymity 算法中的权重，即第 i 个准标识符 A_i 的权重 w_i 定义为

$$w_i = \frac{\sum_{j=1}^{m} w_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{n} w_{ij}},$$

其中 n 为准标识符属性的个数，m 为敏感属性值的个数. 对权重进行单位化使得准标识符属性的所有权重之和为 1，从而使得 3 种算法对于 ILP 指标具有可比性. 实验中设定 KACLUDK 算法将敏感属性值聚为 3 类.

衡量指标，实验用 4 个评价指标来衡量匿名化数据的质量：可辨别性惩罚、信息损失惩罚、运行时间以及一致性攻击最敏感比例.

实验 1. 运行时间测试.

KACLUDK 算法的实际运行时间远远小于

ILP-Based L-diversity 算法和 Bottom-up k-anonymity 算法，当 k 大于 5 时 ILP-Based L-diversity 算法的实际运行时间大约为 20s，Bottom-up k-anonymity 算法的实际运行时间大约为 50s. 如图 5 所示，KACLUDK算法的时间复杂度为 $O(\lceil \log_2 k \rceil |T|)$，而 ILP-Based L-diversity 算法和 KACLUDK 算法的时间复杂度为 $O(\lceil \log_2 k \rceil |T|^2)$，同时由于 ILP-Based L-diversity 算法是对每个类独立进行等价类的合并而 Bottom-up k-anonymity 算法是在全表上进行等价类的合并，因此 ILP-Based L-diversity 算法比 Bottom-up k-anonymity 算法在时间复杂度上具有更小的系数. 所以 ILP-Based L-diversity 算法的运行时间小于 KACLUDK 算法.

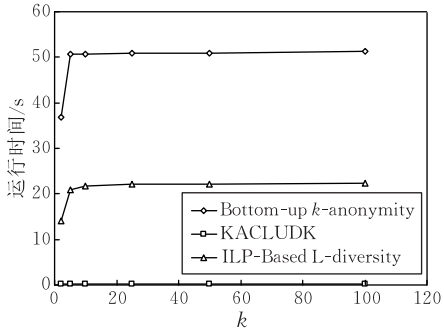

图 5　运行时间测试图

实验 2.　信息损失惩罚(ILP).

评价指标是信息损失惩罚. ILP-Based L-diversity 算法和 Bottom-up k-anonymity 算法的曲线具有基本相同的走势，但是前者的信息损失惩罚略大一些. 而 KACLUDK 算法的信息损失明显大于前两种算法. 如图 6 所示，Bottom-up k-anonymity 算法由于采用在全表上进行等价类合并，因此其结果最好. KACLUDK 算法是在聚类结果的每一个类中独立地进行等价类的随机合并，因为没有充分考虑准标识符对敏感属性值的效用，因此它在这个评价指标上表现最差. ILP-Based L-diversity 算法是在聚类结果的每一个类中以最小化信息损失惩罚为原则进行等价类的合并. 由此可见对敏感属性值的聚类在某些情况下增加了匿名化数据的信息损失. 采用的评价指标是可辨别性惩罚. 可辨别性惩罚是一个用于衡量匿名化数据信息损失的基准评价指标. 可辨别性惩罚为所有等价类大小的平方和，其定义为

$$DP=\sum |E|^2.$$

3 个算法在可辨别性惩罚这个评价指标上表现

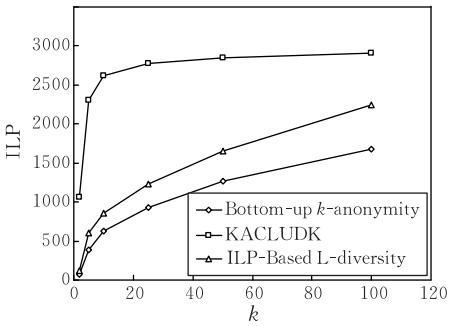

图 6　信息损失惩罚图

基本相同，这是由于这 3 个算法均采用自底向上的贪心法，并且每一轮等价类的合并过程基本相同. 这就导致了最终匿名化数据的等价类大小的平方和基本相同.

实验 3.　一致性攻击敏感比(HASR).

采用的评价指标是一致性攻击最敏感比例，它指的是匿名化数据中敏感属性值的个数为 1 的等价类个数与等价类总数的比值，这个评价指标反映了敏感数据对一致性攻击的敏感程度. 其定义为

$$R_{\mathrm{HASR}}=\frac{N_2}{N}\times 100\%,$$

其中 N_2 表示敏感属性值的个数为 1 的等价类个数，N 表示等价类总数. 如图 7 所示，表明了 3 个算法在一致性攻击最敏感比例这个评价指标上的表现，实验结果表明当 $k>10$ 时，对一致性攻击最敏感的等价类比例基本为 0，而当 $k<5$，尤其是当 $k=2$ 即 k 可取的最小值时，一致性攻击最敏感比例非常高. 由于 KACLUDK 和 ILP-Based L-diversity 算法均首先采用聚类算法对敏感属性进行了聚类，导致了每个类中的敏感属性的个数减少，以致出现等价类中的敏感属性个数仅有 1 个的情况，在这样的等价类中的记录的隐私非常不安全. KACLUDK 算法在对敏感属性进行聚类时没有限制每个类中敏感属性的个数，而 ILP-Based L-diversity 算法则对其依据 L 值进行了限制.

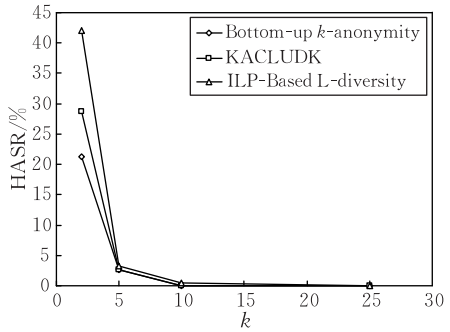

图 7　一致性攻击敏感比(HASR)

Bottom-up K-anonymity 算法和 ILP-Based L-diversity算法是基于与信息损失相关的评价标准来进行等价类合并的,这也会导致相似的记录会被合并到同一个等价类,从而算法在降低匿名化数据信息损失的同时不可避免地提高一致性攻击最敏感比例.因此,专注于降低匿名化数据信息损失的算法都应该格外注意这一现象.

实验 4. 一致性攻击比例与数据失真比例定义为

$$R_{\text{man}}=\frac{N_{\text{man}}}{N}\times 100\%,$$

其中 N_{man} 表示敏感属性值被改变的记录总数,N 表示数据集中的记录总数.

如图 8,3 条曲线分别表示:比例 1: ILP-Based L-diversity 算法未实现 L-diversity 时,匿名化数据中等价类的一致性攻击最敏感比例.比例 2:ILP-Based L-diversity 算法实现 3-diversity 时,匿名化数据中等价类的一致性攻击最敏感比例.比例 3:由于 ILP-Based L-diversity 算法实现 3-diversity 而导致的数据失真的比例.

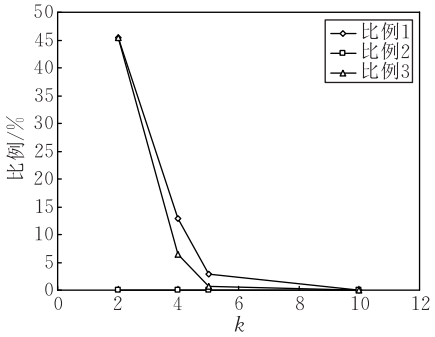

图 8 一致性攻击比例与数据失真比例

当 ILP-Based L-diversity 算法实现 L-diversity 匿名化原则时,匿名化数据的一致性攻击最敏感比例降低为 0,但是算法所采用的数据失真的方法在保护匿名化数据隐私的同时也牺牲了数据的真实性与准确性.因此应该谨慎权衡实际应用中匿名化数据的隐私安全性和数据的准确性来选用不同的匿名原则和匿名算法.正如 k-匿名原则对匿名化数据隐私的保护虽然不如 L-diversity 充分但是它也使得数据保留了更高的准确性.

ILP-Based L-diversity 算法结合了 K-ACLUDK 算法和 Bottom-up k-anonymity 算法的优点并充分考虑了匿名化数据对一致性攻击敏感的问题,在信息损失方面、实际运行时间方面和对匿名化数据隐私保护方面的综合表现较好.

6 总结与展望

研究了面向查询服务的数据隐私保护算法,提出基于信息损失惩罚的满足 L-diversity 的算法.研究基于不同准标识符对不同敏感属性效用的计算,并且满足 L-diversity 的匿名化算法.基于信息损失惩罚的满足 L-diversity 的算法,通过改进原始 L-diversity的算法,并从匿名化数据隐私安全的角度进行了完善,较好地平衡了信息损失、实际运行时间和对匿名化数据的隐私保护,其综合性能表现较好.

数据服务技术是数据库管理技术的进一步发展.未来的研究将涉及数据库、IR、Web 数据管理、数据挖掘、服务计算以及智能推理等综合技术和理论.对于面向查询服务的数据隐私保护算法的研究,如何进一步提高匿名化数据每个等价类中敏感属性的多样性同时尽可能地保证数据的精度仍是一个非常值得研究的问题.如何在考虑不同准标识符对不同敏感属性效用的情况下,研究时间复杂度更低的算法具有重要意义.

参 考 文 献

[1] Han J, Kamber M. Data Mining: Concepts and Techniques. 2nd Edition. San Francisco: Morgan Kaufmann Publishers, 2006

[2] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao Xiao-Kui. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5): 847-861(in Chinese)
(周水庚,李丰,陶宇飞,肖小奎. 面向数据库应用的隐私保护研究综述. 计算机学报, 2009, 32(5): 847-861)

[3] Machanavajjhala A, Gehrke J, Kifer D, Venlita- Subramaniam M. l-diversity: Privacy beyond k-anonymity//Proceedings of the 22nd International Conference on Data Engineering(ICDE). Atlanta, Georgia, USA, 2006: 24-35

[4] Li N, Li T. t-closeness: Privacy beyond k-anonymity and l-diversity//Proceedings of the 23rd International Conference on Data Engineering (ICDE). Istan Buttom-up k-anonymity, Turkey, 2007: 106-115

[5] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571-588

[6] Xiao X, Tao Y. Personalized privacy preservation//Proceedings of the ACM SIGMOD Conference on Management of Data(SIGMOD). Atlanta, Georgia, USA, 2006: 229-240

[7] Aggarwal G, Feder T, Kenthapadi T, Khuller S, Panigrahy R, Thomas D, Zhu Z. Achieving anonymity via clustering//Proceedings of the Symposium on Principles of Database System (PODS). Chicago, Illinois, USA, 2006: 153-162

[8] Pei J, Xu J, Wang Z, Wang W, Wang K. Maintaining K-anonymity against incremental updates//Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSDBM). Banff, Canada, 2007: 5-14

[9] Xu Jian, Wang Wei, Pei Jian, Wang Xiaoyuan, Shi Baile, Fu Ada Wai-Chee. Utility-based anonymization using local recoding//Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, PA, USA, 2006: 785-790

[10] Li Taiyong, Tang Changjie, Wu Jiang, Luo Qian, Li Shengzhi, Lin Xun, Zuo Jie. k-anonymity via clustering domain knowledge for privacy preservation//Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Jinan, Shandong, China, 2008, 4: 697-701

[11] Machanavajjhala Ashwin, Kifer Daniel et al. Privacy: From theory to practice on the map//Proceedings of the 21st International Conference on Data Engineering ICDE. Cancun, Mexico, 2008: 277-286

[12] Sweeney L. K-anonymity: A model for protection privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002, 10(5): 571-588

[13] Bayardo R J, Agrawal R. Data privacy through optimal k-anonymization//Proceedings of the 21st International Conference on Data Engineering (ICDE'05). Tokyo, Japan, 2005: 217-228

[14] Zhu Qing, Wang Shan, Ding Bo-Lin et al. Service-oriented search algorithm on data grid. Chinese Journal of Computers, 2006, 29(7): 1234-1240(in Chinese)
(朱青，王珊，丁博麟等. 基于数据网格面向服务的查询算法. 计算机学报，2006，29(7)：1234-1240)

ZHU Qing, born in 1963, associate professor. Her research interests include high performance database and privacy preservation algorithm, grid computing, distributed system, semantic Web and service computing.

ZHAO Tong, born in 1988. His research interests focus on privacy preservation algorithm design.

WANG Shan, born in 1944, professor, Ph. D. supervisor. Her research interests include high performance database, data warehouse and knowledge engineering.

Background

This work is supported by the Key Lab Found of High Trusted Computing in Shanghai. A new research is information searching for database on Web environment. The project aims to build a service platform and provide privacy preservation algorithm for trusted IR service. The authors have made researches on data integrating, keyword searching, and trusted privacy preservation on Web service computing. Currently, they have made some progress in keyword searching on SOC, Web service composition, service discovery, data integrating and data mapping.