(s,d)-个性化 K-匿名隐私保护模型

傅鹤岗,杨 波

(重庆大学 计算机学院,重庆 400044)

摘 要:在 K-匿名模型的基础上提出了(s,d)-个性化 K-匿名隐私保护模型,该模型能很好地解决属性泄漏问题,并通过实验证明了该模型的可行性。

关键词:K-匿名;隐私保护;个性化;属性泄漏

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2011)05-0085-03

(s,d)-personalization K-anonymity privacy protection model

Fu Hegang, Yang Bo

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: This paper extended a new model (s,d)-personalization K-anonymity privacy protection model based on K-anonymity. Which would deal with the arrribute disclosure well and it is also showed feasible by experiments.

Key words: K-anonymity; privacy protection; personalization; attribute disclosure

随着互联网技术的飞速发展,基于网络的虚拟社会逐步形成,信息的收集、加工、传播更加快捷。现代社会是信息高度共享的社会,使得数据库安全问题日益突出,其中对数据的窃取、篡改和破坏直接危害着数据库的安全,成为亟待解决的问题。随着数据挖掘技术的兴起,大量的信息如:病人就诊信息、学生学籍信息、员工工资及档案信息等面临着泄漏的风险。对个人、企业甚至国家的危害是不容小觑的,个人信息的泄露容易造成诈骗的发生;企业和国家信息的泄露容易造成国家机密的暴露,直接危害国家安全。

自由保护型的数据库隐私保护处理的隐私信息是对外公开的,所有人都可以使用,主要保护隐私信息和个人的对应关系[1]。即攻击者可以轻松获取数据库中的记录,攻击的目标是某条隐私信息和某个体的一对一关系。典型的攻击方法是链接攻击(Linking Attack)[2]。

1 K-匿名技术

较好地解决链接攻击的方法是参考文献[2]中 Samarati 和 Sweeney 引入的 K-匿名机制。它要求公布后的数据中存在一定数量的不可区分的个体,从而使攻击者无法判断出敏感属性的具体个数,以此达到保护个人隐私的目的。为了使数据表满足 K 匿名性质,需要对原始表在准标识符上进行加工,如采用抑制或者泛化技术。

K-匿名技术通过生成若干等价组,使等价组内 QI 属性和隐私属性不再是一一对应的关系,从而保证了个人隐私信息不被泄露。等价组的概念为:在准标识符上的投影完全相同的、记录组成的记录集合,即等价组内所有的记录在准标识符上的属性值完全相同,但是其他属性可以不同。

定义 1 K-匿名。给定数据表 $A(B_1, B_2, \cdots\cdots B_n)$,QI 是与 A 相关联的准标识符,当且仅当在 A[QI]中出现的每个值序列至少在 A[QI]中出现 K 次,则 A 满足 K-匿名。A[QI]表示 A 表中的元组在 QI 上的投影[3]。

表 1 为原始数据表,其中年龄、性别、地区编码为准标识符,疾病为敏感属性,没有任何可以唯一标识个体身份的属性存在,如身份证号码、姓名等。经过 3-匿名化处理后如表 2 所示,每一条记录都有另外两条记录在准标识符上与其相同。即使攻击者知道某条记录在表 2 中,也无法确定哪条记录与其对应,但这样并不能完全防止隐私泄露。因此参考文献[3]提出了 1-多样性概念,即把等价组内出现频率最高的敏感属性限制在 1/1 以内。p-sensitive K-匿名模型[4]是在 K-匿名模型的基础上要求每个等价组内至少要有 p 个不同的敏感属性值,在一定程度上抵御了属性泄漏问题,但是当 K 值很大的时候就表现得不是很好。(a,k)-匿名模型[5]限制了等价组内敏感属性出现的频率不高于 a,在一定程度上防止

了一致性攻击，但是它对所有敏感属性采用相同的约束，无法达到实用的目的。参考文献[6]提出了一种不基于概括和隐匿的新方法——Anatomy，通过将原始关系的准标志符属性和敏感属性以两个不同的关系发布，利用它们之间的有损连接保护隐私数据的安全。这些模型都没有考虑敏感属性敏感度问题，而且无法抵御背景知识攻击。

表 1 原始数据表

编号	年龄	性别	地区编码	疾病
r1	21	女	14235	流感
r2	33	男	13054	感冒
r3	57	女	14350	心脏病
r4	72	男	14255	肝炎
r5	25	女	14245	肺炎
r6	47	男	13097	心脏病
r7	80	男	14278	肝硬化
r8	68	女	14351	艾滋病
r9	29	女	14256	肝炎
r10	77	女	14355	肝癌
r11	77	男	14289	肺炎
r12	39	男	13024	艾滋病

表 2 3-匿名化数据表

编号	年龄	性别	地区编码	疾病
R_1	2*	女	142**	流感
R_5	2*	女	142**	肺炎
R_9	2*	女	142**	肝炎
R_2	[30-50]	男	130**	感冒
R_6	[30-50]	男	130**	心脏病
R_{12}	[30-50]	男	130**	艾滋病
R_3	[57-77]	女	1435*	心脏病
R_8	[57-77]	女	1435*	艾滋病
R_{10}	[57-77]	女	1435*	肝癌
R_4	[72-80]	男	142**	肝炎
R_7	[72-80]	男	142**	肝硬化
R_{11}	[72-80]	男	142**	肝炎

2 (s,d)-个性化 K-匿名隐私保护模型

K-匿名的主要缺陷：(1)K-匿名没有考虑到匿名后可信属性由于缺乏多样性而导致的隐私泄露问题(同质性攻击)；(2)默认所有属性都有相同的重要性；(3)不能抵御背景知识攻击。

本文介绍的 (s,d)-个性化 K-匿名隐私保护模型就是为了解决这些问题而提出来的。在介绍(s,d)-个性化 K-匿名隐私保护模型前需要给出的定义：s-相似等价组、临界敏感度、高危敏感度、d-非关联约束。

定义 2 s-相似等价组。是指在敏感属性值上相似的至少 s 个记录组成的等价组，在这里相似的定义根据具体的应用会有所不同。例如：如果敏感信息是疾病，则可以将病变器官作为相似划分标准，如胃部疾病，肝部疾病等。

定义 3 临界敏感度。由专家确定或者根据具体应用领域灵活确定的、能够较好体现对敏感属性保护程度的一个数值度量，其值在 0~1 之间。

定义 4 高危敏感度。高危敏感度是指敏感属性值的敏感度大于、等于临界敏感度，其值在 0~1 之间。

定义 5 d-非关联约束。对于 s-相似等价组 E，在 E 中高危敏感度属性值出现的频率 f 不高于 d，即 $|f|/|E|<d$ $(0\leqslant d\leqslant 1)$，其中 d 是用户确定的参数。但其必须满足 d 不能等于 1 且不能过大，即不能太接近 1。

定义 6 (s,d)-个性化 K-匿名隐私保护模型。如果一等价组由位于不同相似组的 s-相似等价组组成，每个 s-相似等价组都满足 d-非关联约束，并且每个等价组至少由 K 条记录组成，如果数据表 T 中的每个等价组都满足以上条件，那么就称数据表 T 满足(s,d)-个性化 K-匿名隐私保护模型。

(s,d)-个性化 K-匿名隐私保护模型就是利用一个等价组中如果包含了多组 s-相似等价组，并且每个 s-相似等价组都满足 d-非关联约束，就可以更加有效地抵御同质性攻击及属性泄漏。另外如果每组相异值包含了多组相似值，可更加有效地抵御背景知识攻击，从而大大降低隐私信息泄漏的风险。本文阐述的(s,d)-个性化 K-匿名隐私保护模型如表 3 所示。

表 3 包含 2 个 2-相似等价组的 4-匿名等价组

工作	年龄	生日	地址	疾病	敏感度
教师	[25-40]	198*	河北省	肺癌	0.90
教师	[25-40]	198*	河北省	肺癌	0.90
工人	[25-40]	198*	河北省	肝癌	0.90
工人	[25-40]	198*	河北省	肝癌	0.90

根据病变器官这一相似性进行 2-相似分组，可以看出该等价组满足 2-相似条件，从 K=4 的匿名表可以看出，由于敏感属性疾病这一列都是高危敏感度属性值，敏感度高达 0.9，即使其满足匿名条件，但是该等价组的隐私信息也已经暴露出来了，攻击者很容易得出该等价组对应的个体患有很严重的疾病，也就造成了属性泄漏。虽然从某种程度上来说还没有造成身份泄露，但这也是人们所不希望的。

根据(s,d)-个性化 K-匿名隐私保护模型的规定，调整如表 4、表 5 所示。

表 4 s=2,d=0.5,K=4 的例子 1

工作	年龄	生日	地址	疾病	敏感度
教师	[25-40]	198*	河北省	肺癌	0.90
教师	[25-40]	198*	河北省	支气管炎	0.45
工人	[25-40]	198*	河北省	肝癌	0.90
工人	[25-40]	198*	河北省	乙肝病毒携带者	0.35

表 5 s=2,d=0.5,K=4 的例子 2

工作	年龄	生日	地址	疾病	敏感度
教师	[25-40]	198*	河北省	肺癌	0.90
教师	[25-40]	198*	河北省	肺炎	0.55
工人	[25-40]	198*	河北省	乙肝小三阳	0.40
工人	[25-40]	198*	河北省	肝癌	0.90

表 4、表 5 中的每个等价组都满足 s=2,d=0.5 (即 sensitivity>0.70 的敏感属性值出现在每个 2-相似等价组中的频率≤0.5)，K=4 条件，但是可以较好地防止属性泄漏问题。从敏感属性敏感度的分布来看，经过调整记录得到的这两个表其实就是减少了每个 2-相似分组中高危属性值的出现频率。表 4 中将癌症的出现频率控

制在了50%以内，表5中也将癌症的出现频率控制在了50%以内。本文提出的(s,d)-个性化K-匿名隐私保护模型，在K-匿名模型基础上做出了改进，有效地解决了由高危属性值出现频率过高而导致的属性泄漏问题，同时能很好地抵御同质性攻击和背景知识攻击。

3 (s,d)-个性化K-匿名隐私保护模型算法

输入：数据表T，对敏感属性的敏感度进行标记$s=\{S_1,S_2\cdots\cdots S_n\}$，敏感属性按相似性分组$g=(g_1,g_2,\cdots,g_n)$，准标识符各个属性的权重$W=(w_1,w_2,\cdots,w_n)$，参数为$s,d,K$。

输出：满足(s,d)-个性化K-匿名隐私保护模型的数据表Ta'。

处理过程：

(1)生成s-相似等价组，并且这些等价组满足d-非关联约束。

(2)对生成的s-相似等价组寻找使泛化信息损失最少的、K/s个不位于相同相似组内的s-相似等价组：

Ta'={}

For(对于Ga中的每一个分组Ga')

Gt={}，在Ga'中取一条记录

If(|Gt|!=K/s)

在Gt中找一分块Gt'，使得Gt'中的记录t'和t的敏感属性值不属于同一个敏感属性组，并且dist(QI[t],QI[t'])最小，Gt=Gt∪Gt'，Ga=Ga/Gt。

End if Ta'=Ta'∪Gt

End for

(3)对生成的满足(s,d)-个性化K-匿名隐私保护模型条件的各等价组进行泛化处理，即对Ta'中的每个分块进行泛化处理。

4 实验

实验所使用的数据集来自UCI机器学习数据库[7]中的adult数据库，该数据库在研究K-匿名应用最多，已经成为该领域事实上的标准测试集。数据库大小为5.5 MB，本文选取其中的30 704条记录及15个属性，其中准标识符数量选择6个，将职业(WORKCLASS)作为敏感属性。敏感属性泄漏分析如表6所示。

表6 敏感属性泄漏分析

K-匿名	模型	敏感属性泄漏个数	K-匿名	模型	敏感属性泄漏个数
4-匿名	4-匿名	36	6-匿名	6-匿名	32
	(2,0.8)-个性化4-匿名	5		(2,0.8)-个性化4-匿名	4
	(2,0.5)-个性化4-匿名	3		(2,0.5)-个性化4-匿名	2

实验软硬件环境：

硬件环境：Intel Pentium (R) Dual-Core CPU，2GB RAM。

操作系统：Microsoft Windows XP。

编程环境：Eclipse+Mysql Server 5.1。

执行时间分析如图1所示。

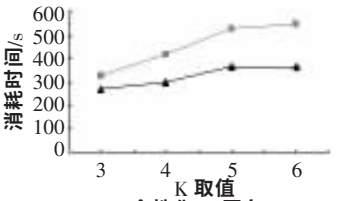

图1 两种模型下执行时间的对比

本文针对K-匿名没有考虑到匿名后可信属性由于缺乏多样性而导致的隐私泄露、默认所有属性都有相同的重要性、不能抵御背景知识攻击等问题，提出了一种新的(s,d)-个性化K-匿名隐私保护模型。该模型通过s-相似分组，并且限制每个s-相似等价组内的高危敏感属性值出现的频率小于d，然后组合不同相似分组内的s-相似分组使其满足K-匿名条件。实验证明该模型能很好地弥补K-匿名的不足，有效地防止了隐私泄露。

参考文献

[1] 刘喻，吕大鹏，冯建华，等.数据发布中的匿名化技术研究综述[J].计算机应用，2007，27(10)：2361-2364.

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

[3] MACHANAVAJJHALA A, GEHRKE J, KIFER D, et al. L-diversity: Privacy beyond K-anonymity [C]// Proc of the 22 nd International Conference on Data Engineering New York: ACM Press, 2006.

[4] TRAIAN T M, BNDU V. Privacy protection: p-sensitive k-anonymity property [C]//.Proc of the 22 nd International Conference on Data Engineering New York: ACM Press, 2006.

[5] WONG R C, Li Jinyong, FU A W, et al. (a,k)-anonymity: an enhanced k-anonymity model for privacy preserving [C]//. Proc of the 12 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York: ACM Press, 2006.

[6] Xiao Xiaokui, Tao Yufei. Anatomy: simple and effective privacy preservation [C]//Proc of the 32 nd International Conference on Very Large Data Bases [SI]:VLDB Endowment, 2006:139-150.

[7] HETTICH C B S, MERZ C. UCI repository of machine learning databases [EB/OL]. (1996-05-01) [2008-04-20]. http://archiveics uci edu/ml/datasets/Adult.

(收稿日期：2010-09-01)

作者简介：

傅鹤岗，男，1950年生，副教授，主要研究方向：电子商务，软件工程。

杨波，男，1985年生，硕士，主要研究方向：数据发布中的隐私保护。