数据挖掘中匿名化隐私保护研究进展

谭 瑛

云南财经大学信息学院,昆明 650221

摘要　随着信息技术的发展,如何在保证数据高可用性的同时,不泄露数据主体的隐私信息,已日益引起国内外研究者的高度关注。隐私保护技术主要有数据加密、数据失真以及数据匿名化技术,其中匿名化技术是数据挖掘中隐私保护的最主要技术手段。围绕匿名技术的研究,国内外学者提出了多种匿名隐私保护模型,通过对其中 4 种主要模型,即 *k*-匿名模型、*l*-多样性模型、(α,*k*)-匿名模型和 *t*-closeness 模型的分析比较,指出每种匿名模型的特点及优、缺点,并归纳了常用的匿名技术,总结了当前主要的匿名化质量的度量方法。未来匿名化技术作为数据挖掘中隐私保护的主要手段,还将面临着需要进一步解决的问题,对数据挖掘中匿名隐私保护的下一步研究方向进行了展望。

关键词　数据挖掘;隐私保护;*k*-匿名

中图分类号　TP309.2　　**文献标志码**　A　　**doi**　10.3981/j.issn.1000-7857.2013.01.013

Progress in Anonymous Privacy-Preserving in Data Mining

TAN Ying

Department of Information, Yunnan University of Finance and Economics, Kunming 650221, China

Abstract　With the development of information technology, an important issue is to ensure the high usability of the data and to protect the privacy. The privacy-preserving technology is related with the data encryption, the data distortion and the data anonymity in data mining. Among them, the most primary technology is the anonymous privacy-preserving technology. In that respect, a various privacy preserving models were proposed. This paper focuses on the *k*-anonymous model, the *l*-diversity model, the (α, *k*)-anonymous model and the *t*-closeness model, and it is pointed out that each anonymous model has its, advantages and disadvantages. The commonly used anonymity technology and the major anonymous quality measurement methods are reviewed. In the future, the anonymity technology will face new problems, and the privacy-preserving will be further considered in data mining.

Keywords　data mining; privacy-preserving; *k*-anonymity

0　引言

在信息时代,个人信息不再是无价值的数据,大量的个人信息被组织或个人收集、处理、分析、共享、发布,并从中挖掘出有用知识以供商业利用或科学研究,而在这数据挖掘的过程中势必会造成大量个人敏感信息的泄露。如何在保证数据高可用性的情况下,不泄露数据主体的隐私信息已日益引起国内外研究人员的关注。1998 年,Samarati 等[1]首次提出匿名的概念;2002 年,Sweeney 等[2-3]提出 *k*-匿名隐私保护模型,围绕匿名技术的研究已成为近年来的探讨热点,匿名技术也成为解决数据挖掘领域的数据高效用及信息安全领域的隐私保护之间矛盾的主要技术之一。

1　基本概念

1.1　属性分类

在对数据进行匿名化的过程中,所处理的原始数据为包含若干条记录的数据表文件(表 1),而每条记录又包含若干个属性,这些属性根据其功能分为 4 类:(1) 个体标识属性(Individually Identifier Attribute,ID),用来唯一标识个体身份的属性,如表 1 中的姓名;(2) 准标识属性(Quasi Identifier attribute,QI),与其他外部表联合起来能唯一标识个体的一组属性,如表 1 中的年龄、性别、邮政编码;(3) 敏感属性(Sensitive Attribute,SA),描述个体隐私、需要保密的属性,如表 1 中的疾病;(4) 非敏感属性(Non-sensitive Attribute,NA),可以

收稿时间:2012-05-10;修回时间:2012-11-02

作者简介:谭瑛,讲师,研究方向为信息安全、计算机应用,电子信箱:tanyingty@163.com

公开的属性。

表 1　原始数据表

Table 1　Original data table

序号	姓名	年龄	性别	邮政编码	疾病
1	Rose	67	M	650012	Cancer
2	Mike	74	F	650015	Cancer
3	Alice	54	F	650018	Hepatitis
4	Steven	58	M	650024	Flu
5	Kate	31	M	650024	Fever
6	John	37	M	650028	HIV

1.2　隐私数据常受到的攻击

(1) 链接攻击(linkage attack),又称背景知识攻击(background knowledge attack),是发布数据过程中获取隐私数据的主要攻击方法之一。通常组织或个人在发布数据之前会将直接且唯一标识个体身份的属性,即 ID 删除,但是攻击者仍能利用剩下的属性信息,通过与其他数据表进行匹配链接,从而推理获得隐私信息。例如,在发布表 1 时仅仅简单删除了姓名属性,攻击者仍可以通过将表 1 与选民登记表或人口信息登记表进行链接,通过生日、性别和邮政编码这些属性的取值,推理甚至可以确定出 Disease 敏感属性所对应的个体,而这正是需要保护的个人隐私数据。因此虽然在发表数据之前删除了姓名,但并没有达到匿名的效果。如果能切断疾病与姓名之间的联系,即切断敏感信息与个人身份之间的联系,才实现了真正的匿名。

(2) 同质性攻击(homogeneity attack),数据表中具有相同准标识属性取值的若干条记录称为一个等价类 E,同质性攻击是攻击者利用 k-匿名数据中存在某等价类的敏感信息基本相同,结合准标识属性确定等价类的所属个体,对个体隐私造成威胁[4]。

(3) 相似性攻击(similarity attack),在数值型敏感属性数据发布中,即使同一等价类 E 中敏感属性值不同,但是如果敏感属性值相近,攻击者可以利用各种手段以获得敏感属性在某一较小范围内的信息,从而造成隐私信息的泄露,这即为相似性攻击。

2　国内外相关研究工作

针对匿名隐私保护,国内外很多专家学者做了大量的研究,并取得了一定成果。2002 年,Sweeney 等[2,3]提出 k-匿名隐私保护模型。2006 年,Machanavajjhala 等[5]在 k-匿名基础上于提出了 l-多样性模型;Truta 等[6]提出 p-Sensitive k-anonymity 匿名模型;Wong 等[7]提出(α,k)-匿名模型。2007 年,Li 等[8]在 k-匿名模型及 l-多样性模型的基础上,提出 t-closeness 模型;金华等[9]提出基于敏感性分组的(α,k)-匿名模型;Xiao 等[10]提出个性化匿名模型;Aggarwal [11] 提出基于聚类的匿名模型;Wang 等[12]提出自底向上的匿名方法;Fung 等[13]提出自顶向下的匿名方法;Pei 等[14]提出单调递增匿名方法等。当然,基于匿名隐私保护的研究远不止文中所叙述,而且各种新的匿名模型及匿名方法还在不断被提出,它正成为业界的研究热点,在这里不再一一赘述,其中 k-匿名模型、l-多样性模型、(α,k)-匿名模型和 t-closeness 模型是最典型的模型。

3　4 种典型匿名模型及其比较

3.1　k-匿名模型

2002 年,Sweeney 等[2-3]提出 k-匿名隐私保护模型,要求匿名后数据表中的任何一条记录,至少有 k-1 条记录与其具有相同的准标识属性值,即数据表中任何一等价类 E 的记录数不小于 k,如表 2 为表 1 的 2-匿名化表。

表 2　表 1 的 2-匿名化表

Table 2　Anonymous table

分组编号	年龄	性别	邮政编码	疾病
1	[65~75]	*	65001*	Cancer
1	[65~75]	*	65001*	Cancer
2	[50~60]	*	6500**	Hepatitis
2	[50~60]	*	6500**	Flu
3	[30~40]	M	65002*	Fever
3	[30~40]	M	65002*	HIV

但由于没有对敏感数据做任何约束,攻击者可以通过掌握的背景知识攻击个体隐私,或者当同一等价类 E 中敏感数据相近甚至相同时,通过结合准标识属性从而推断出与敏感数据相对应的个体,致使隐私泄露,因而 k-匿名模型无法抵制同质性攻击和链接攻击。同时,当 k 的取值越大,等价类 E 中的记录数就越多,需要泛化的属性值也就越多,导致数据损失的就越多,数据的可用性就越差,反而对隐私的保护却越好;反之,当 k 的取值越小,等价类 E 中的记录数就越少,需要泛化的属性值也就越少,数据损失的就越少,数据的可用性就越好,但是对隐私的保护效果不好。可见 k-匿名模型中 k 的取值十分关键,Sweeney 等[2]认为 k 的取值一般不超过 5 或 6,文献[15]指出 k 的取值应在 3~10 区间范围内,宋金玲等[16]给出了在普通数据质量标准下 k 的优化选择算法,然而这些都不能理想地解决隐私泄露和数据质量之间的平衡关系,k 的合理取值仍将是未来研究的重点。

3.2　l-多样性模型

2006 年 Machanavajjhala 等[5]在 k-匿名基础上于提出了 l-多样性模型,要求每个等价类 E 中敏感属性值至少有 l 个“较好表现”,“较好表现”有如下解释:(1) 同一个等价类 E 中敏感属性至少有 l 个不同取值,如表 3 是表 1 的 2-匿名 2-多样化表;(2) 同一等价类 E 中敏感属性值的信息熵(Entropy)至少为 lg$l$,其中熵定义为 $Enproty(E)=-\sum_{s\in S}P(E,s)\lg p(E,s)$。式中,$S$ 为敏感属性值域,$P(E,s)$为敏感属性值 s 在等价类 E 中

出现的频率;(3) 频率出现最大的敏感值在等价类 E 中出现的频率不能超过一定值，即保证等价类 E 中不会有某一敏感值出现频率太高。虽然这在一定程度上解决了因敏感属性缺少多样性而面临的同质性攻击和链接攻击,但如果敏感属性相近时则不足以抵制相似性攻击，并且由于没有对等价类 E 中敏感属性值出现的频率给予约束,也无法抵制概率攻击。

表 3　表 1 的一个 2-匿名 2-多样化表

Table 3　Anonymous and diversified table

分组编号	年龄	性别	邮政编码	疾病
1	[50~75]	*	6500**	Cancer
1	[50~75]	*	6500**	Cancer
1	[50~75]	*	6500**	Hepatitis
1	[50~75]	*	6500**	Flu
2	[30~40]	M	65002*	Fever
2	[30~40]	M	65002*	HIV

3.3　(α,k)-匿名模型

在 k-匿名模型的基础上 Wong 等[7]提出(α,k)-匿名模型，要求同一等价类中每个敏感属性值出现的频率不大于 α,即通过控制每个等价类中敏感信息出现的频率来抵制概率攻击,以实现隐私保护。但其对所有敏感值设定统一的 α 约束，适应性较差，并且未考虑敏感属性不同取值间的敏感性差异,因此不能很好地抵御同质性攻击,同时又有 k 和 α 的双重约束,信息损失也更大。

3.4　t-closeness 模型

针对 l-多样性的不足,2007 年 Li 等[8]在 k-匿名模型及 l-多样性模型的基础上,结合了敏感属性值的分布,提出了 t-closeness 模型,要求敏感属性值在每个等价类 E 中的分布与其在匿名化表中的全局分布差异小于 t，这能有效地解决相似性攻击。虽然匿名化的结果有效地提高了隐私保护的程度,但对每个敏感属性值的要求太过苛刻,同时它仍是以牺牲数据可用性为代价的，并且当等价类中的记录数较少时，考虑敏感属性值的分布没有什么意义。

3.5　4 种模型比较

表 4 是对以上 4 种模型进行的对比，可见 4 种模型各有优劣,每一种模型都不能在完全抵制各种攻击以防止隐私信息泄露的同时又保证数据的高质量,因此匿名新技术将面临很多挑战。

表 4　4 种模型对比

Table 4　Comparison of four models

模型名称	主要特点	优点	缺点
k-匿名模型	数据表中任何一等价类的记录数不小于 k	一般情况下可防止敏感信息泄露	无法抵制同质性攻击和链接攻击
l-多样性模型	每个等价类中敏感属性值至少有 l 个“较好表现”	一定程度上可抵制同质性攻击和链接攻击	无法抵制相似性攻击和概率攻击
(α,k)-匿名模型	同一等价类中每个敏感属性值出现的频率不大于 α	抵制概率攻击	无法抵制同质性攻击,数据可用性差
t-closeness 模型	敏感属性值在每个等价类中的分布与其在匿名化表中的全局分布差异小于 t	可有效抵制相似性攻击	数据可用性差

4　匿名技术

4.1　泛化技术

泛化是指用一般值替代原始数据或是对原始数据的概括，例如图 1 中的对年龄的泛化。泛化在实现上有两种方法:全局泛化和局部泛化。如果每个准标识属性取值从底层开始一同向上泛化，直到满足隐私保护要求时同时停止泛化,即泛化到同一层次为全局泛化,如图 2 所示。如果泛化到不同层次则为局部泛化,如表 1 中邮政编码,有的泛化为 65001*,有的则为 6500**。泛化技术实现简单,但是有时过度的泛化会导致更多的数据损失。

在泛化技术的基础上，不断有新的改进的泛化技术被提出,如 Sweeney 提出了最小泛化的思想[2,3],Aggarwal[11]提出聚类泛化,即将相似的 k 条记录归为一簇,再对每个簇进行泛化操作等。

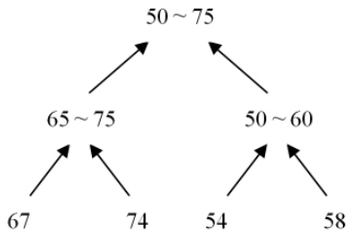

图 1　年龄的泛化

Fig. 1　Generalization of AGE

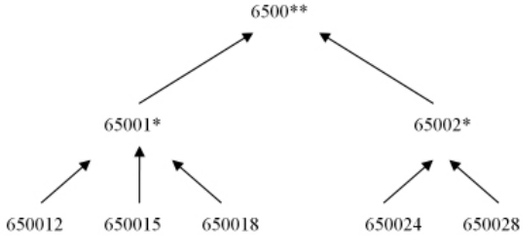

图 2　邮政编码的全局泛化

Fig. 2　Global generalization of ZIP

4.2 抑制技术

抑制技术,又称为隐藏技术,即抑制/隐藏某些数据,有选择地发布原始数据,攻击者不能看到被抑制的数据,被抑制的数据或者被删除,或者用"*"替代。因此抑制技术实现也很简单,对于隐私数据的保护度很高,但是数据失真得很严重,并且当抑制的数据太多时,数据的可用性将大大降低。

4.3 扰乱技术

扰乱是指在数据发布前通过加入噪声等手段对敏感数据进行扰乱,以实现对原始数据的扭曲、改变,但要求扰乱后的数据仍保留着原始数据分布特征。

4.4 有损连接

有损连接是分别发布 2 个表,准标识属性表和敏感属性表,2 个表使用共同的一个等价类分组号关联起来,由于分组号不是主键,因此 2 个表之间的关联形成一种有损连接,通过连接的损失来实现隐私的保护。

除上所述的这些技术外,还有诸如数据变换技术、数据阻塞技术等,然而既要保证隐私保护度最大,同时又要获得最高的数据可用性是不易的,Meyerson 等[17]证明了获得最优 k-匿名化(k>2)是非确定性多项式(Non-deterministic Polynomial,NP),因此力求很好地平衡数据隐私与数据可用性之间的关系将是匿名技术的主要任务。

5 匿名化质量的度量方法

匿名化后的数据质量如何,隐私保护的效果如何,目前还没有统一的衡量标准,不少专家和学者从不同的角度给出了不同的度量方法。

5.1 可辨别度量法

2005 年,Bayardo 等[18]提出了可辨别度量法,其基本思想是为等价类中每条记录分配一个惩罚(penalty)数,惩罚数为该记录所在等价类的记录个数,整个匿名化表的信息损失为所有惩罚数之和:总信息损失=未被完全泛化的记录惩罚值+完全被泛化的记录惩罚值。可见,可辨别度量法是基于等价类的大小来评估匿名化数据的质量,等价类越大,信息损失就越多,数据质量就越差。然而,可辨别度量法未考虑一种特殊情况,即如果某个等价中所有记录均未被泛化,此时仍将对所有记录赋予惩罚数。2006 年 LeFevre 等[19]提出了标准化平均等价类 E 大小度量法,该法度量信息损失度主要有 3 个关键因素,即整个表的记录数、等价类数和 k-匿名的系数 k。可辨别度量法和标准化平均等价类 E 大小度量法都是基于等价类的大小来评估匿名化数据的质量的。

5.2 匿名表效用度量法

2006 年,Xu 等[20]提出了度量匿名表效用的计算方法 NCP,该方法对于在匿名过程中的数值属性和类别属性分别给出了惩罚(penalty)计算方法。

数值属性:$NCP(t)=\sum_{i=1}^{d} w_i \cdot NCP_{A_i}$,其中 w_i 为属性变量的权重,z_i、y_i 分别为属性变量 A_i 泛化后区间的左、右界,$|A_i|$为属性变量 A_i 在数据集中的最大取值和最小取值的差。

类别属性:$NCP(t)=\text{size}(u)/|A|$,其中 size($u$)为泛化后的类别属性变量值集合中值的个数,$|A|$为泛化属性变量的所有可能取值的个数。

每个等价类 G 的惩罚计算公式:$NCP(\text{G})=\sum_{i=1}^{d} w_i \cdot NCP_{A_i}(\text{G})$,其中 d 为准标识属性变量数,为各准标识属性变量的权重。匿名表的信息损失为所有等价类的惩罚之和。

5.3 泛化层次度量法

2006 年,Li 等[21]提出了通过准标识属性泛化层次高度来计算信息损失程度的方法,该方法通过定义加权层次距离(Weighted Hierarchical Distance,WHD)描述匿名扰乱程度,并以此计算出信息损失程度,得出结论,当准标识属性泛化的层次越高,记录的扰乱程度越明显,原始信息损失得越多。

5.4 分布距离度量法

通过距离计算每个等价类中敏感属性值的分布与其在匿名化表中的全局分布的差异,Li 等[8]提出 EMD 距离度量(Earth Mover's Distance)方法,Lin 等[22]提出 MD 距离度量(Manhattan Distance)方法等,并得出结论:分布距离越大,等价类中敏感属性值的分布与其在匿名化表中的全局分布的差异越大,抵制相似性攻击等的能力越差。

6 展望

目前围绕隐私保护研究主要有三大类方向,即基于数据失真的隐私保护技术、基于数据加密的隐私保护技术和基于数据匿名化的隐私保护技术。其中基于数据失真的隐私保护技术拥有计算开销(占用资源)小、实现简单的优点,但是对隐私的保护度并不高,且数据完整性和真实性在较大程度上遭到破坏;基于数据加密的隐私保护技术对隐私的保护度高,数据无缺损,但是代价是计算开销大,实现难度高;而基于数据匿名化的隐私技术与它们两者相比较,平衡了它们的优缺点,适用于各类数据的隐私保护,以较小的数据缺损、较低的计算开销实现较高的隐私保护。因此基于数据匿名化的隐私保护技术在隐私保护中占据着重要的地位,并具有广泛的使用范围,目前它还是一个较新的、面临诸多挑战的研究热点。

(1) 在数据匿名化的过程中,如何保证数据可用性最高,同时隐私保护度最大,这似乎是矛盾的,如何平衡数据可用性及隐私保护之间的关系,做到尽可能获得高质量的数据,同时还能有效地保护隐私,是未来研究的重点。

(2) k-匿名模型中 k 值的选取是十分关键的,它直接影响着隐私保护度和数据的可用性,因此对于 k 的取值研究是具有重要价值的。

(3) 目前基于 k-匿名模型的算法大多都是针对一次发布的数据,如果数据动态连续增加、插入、删除、更新,针对这

些情况的隐私保护问题将是进一步研究的方向。

(4) 现有的研究主要是针对单一敏感属性的数据进行匿名化处理,但是现实中的数据往往涉及多个敏感属性,如果直接将现有方法用于多敏感属性数据的匿名化,将会导致隐私信息的泄露,针对多敏感属性数据发布中的隐私保护问题是下一步探讨的重点工作。

(5) 上文中列出了目前常用的一些匿名化质量的度量方法,但是目前仍缺乏统一的标准,因此制定出有效的、可行的评价体系将是十分必要的。

参考文献(References)

[1] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information[C]//Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. New York, NY: ACM Press, 1998: 188.

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

[3] Sweeny L. k-anonymity: A model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[4] 刘英华, 刘永彬, 李广原, 等. 一种增强的个性化匿名隐私保护模型[J]. 微电子学与计算机, 2011, 28(8): 5.
Liu Yinghua, Liu Yongbin, Li Guangyuan, et al. Microelectronics & Computer, 2011, 28(8): 5.

[5] Machanavajjhala A, Kifer D, Gehrke J et al. l-diversity: Privacy beyond k-anonymity[C]. The 22nd International Conference on Data Engineering, Atlanta, GA, USA, April 3-7, 2006: 24-35.

[6] Truta T M, Vinay B. Privacy protection p-sensitive k-anonymity property [C]. The 22nd International Conference on Data Engineering Workshops, Atlanta, GA, USA, April 3-7, 2006: 94.

[7] Wong R C, Li J, Fu A W, et al. (α, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing [C]. The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006.

[8] Li N H, Li T C, Venkatasubramanian S. t-Closeness-privacy beyond k-anonymity and l-diversity [C]. IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, April 15-20, 2007: 106-115.

[9] 金华, 张志祥, 刘善成, 等. 基于敏感性分级的(α, k)-匿名隐私保护[J]. 计算机工程, 2011, 37(14): 12-17.
Jin Hua, Zhang Zhixiang, Liu Shancheng, et al. Computer Engineering, 2011, 37(14): 12-17.

[10] Xiao X K, Tao Y F. Personalized privacy preservation [C]. ACM Conference on Management of Data (SIGMOD 2006), Chicago, IL, USA, June 27-29, 2006: 229-240.

[11] Aggarwal G, Feder T, Kenthapadi K, et al. Achieving anonymity via clustering [C]. SIGMOD/PODS'06 International Conference on Management of Data and Symposium on Principles Database and Systems, Chicago, IL, USA, June 27-29, 2006: 153-162.

[12] Wang K, Yu P S, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection [C]. ICDM'04. The 4th IEEE International Conference on Data Mining, Brighton, UK, November 1-4, 2004: 249-256.

[13] Fung B, Wang K, Yu P S. Top-down specialization for information and privacy preservation [C]. ICDE 2005. The 21st IEEE International Conference on Data Engineering, Tokyo, Japan, April 5-8, 2005: 205-216.

[14] Pei J, Xu J, Wang Z, et al. Maintaining k-anonymity against incremental updates [C]. SSBDM'07. The 19th International Conference on Scientific and Statistical Database Management, Banff, Alta, Canada, July 9-11, 2007: 5.

[15] Winkler W E. Using simulated annealing for k-anonymity[R]. Research Report 2002-07, Washington DC: Statistical Research, Division, US Bureau of the Census, 2002.

[16] 宋金玲, 刘国华, 黄立明, 等. k-匿名隐私保护模型中 k 值的优化选择算法[J]. 小型微型计算机系统, 2011, 32(10): 1987-1993.
Song Jinling, Liu Guohua, Huang Liming, et al. Journal of Chinese Computer Systems, 2011, 32(10): 1987-1993.

[17] Meyerson A, Williams R. On the complexity of optimal k-anonymity[C]. The 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Paris, France, June 14-16, 2004: 223-228.

[18] Bayardo R J, Agrawal R. Data privacy through optimal k-anonymity[C]. ICDE2005 The 21st International Conference on Data Engineering, Tokyo, Japan, April 5-8, 2005: 217-228.

[19] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity [C]. ICDE'06 The 22nd International Conference on Data Engineering, Atlanta, GA, USA, April 3-7, 2006: 1-25.

[20] Xu J, Wang W, Pei J, et al. Utility-based anonymization using local recoding [C]. The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, USA, August 20-23, 2006: 785-790.

[21] Li J, Wong R C, Fu A W, et al. Achieving k-anonymity by clustering in attribute hierarchical structures [C]//Tjoa A M, Trujillo J. Lecture Notes in Computer Science: Data Warehousing and knowledge Discovery. Heidelberg: Springer-Verlag, 2006, 4081: 405-416.

[22] Lin J H. Divergence measures based on the Shannon theory [J]. IEEE Transactions on Information Theory, 1991, 37(10): 145-151.

(责任编辑　岳臣)