匿名化隐私保护技术研究进展 *

王平水[1], 王建东[2]
(1. 安徽财经大学 信息工程学院, 安徽 蚌埠 233041; 2. 南京航空航天大学 信息科学与技术学院, 南京 210016)

摘 要: 匿名化是目前数据发布环境下实现隐私保护的主要技术之一。阐述了匿名化技术的一般概念和基本原理,并从匿名化原则、匿名化方法和匿名化度量等方面对匿名化技术进行了总结,最后指出匿名化技术的研究难点以及未来的研究方向。

关键词: 数据发布; 隐私保护; 匿名化; *k*-匿名
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2010)06-2016-04
doi:10.3969/j.issn.1001-3695.2010.06.004

Progress of research on anonymization privacy-preserving techniques

WANG Ping-shui[1], WANG Jian-dong[2]
(1. *College of Information Engineering, Anhui University of Finance & Economics, Bengbu Anhui* 233041, *China*; 2. *College of Information Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016, *China*)

Abstract: Anonymization is one of the primary techniques realizing privacy protection in data dissemination environment. This paper described the general concepts and basal principles of the anonymization techniques, and summarized the anonymization principles, anonymization methods and anonymization measures. Finally discussed the present problems and directions for future research.

Key words: data dissemination; privacy-preserving; anonymization; *k*-anonymity

0 引言

Internet 技术、大容量存储技术的迅猛发展以及数据共享范围的逐步扩大使得数据的自动收集和发布越来越方便。然而,在数据发布过程中隐私泄露问题也日益突出,因此隐私保护问题就显得尤为重要。数据发布中隐私保护对象主要是对用户敏感数据与个体身份之间的对应关系。一般通过删除标志符的方法发布数据是无法阻止隐私泄露的,攻击者可以通过链接攻击获取个体的隐私数据。匿名化技术可有效地解决链接攻击所带来的隐私泄露问题。自从 1998 年 Samarati 等人[1]首次提出匿名化概念以来,国内外专家学者们对匿名化技术开展了广泛深入的研究工作以寻求防止或减少隐私泄露的有效方法,取得了一系列相关研究成果[2~22]。然而,目前如何通过匿名化技术高效实现发布数据的隐私保护仍然是业界研究的热点问题。本文旨在总结和分析数据发布隐私保护中匿名化的基本原理和实现技术现状,以启示相关领域研究人员对匿名化技术的进一步深入研究,促进匿名化隐私保护技术的实际应用。

1 相关概念

1.1 属性划分

数据发布中微数据集可视为包含 n 条记录的数据表文件,其中每条记录包含个体的 m 个属性,属性按其功能可被分成互不相交的四种[2]:

a)标志符(identifiers)。惟一标志个体身份,如身份证号、社会保险号、姓名等。

b)准标志符(quasi-identifiers)。与其他数据表进行链接以标志个体身份,如性别、出生日期、邮政编码等。准标志符的选择取决于进行链接的外部数据表,如图 1 中,准标志符(简称 QI) = {race, birth, sex, zip}。

c)敏感属性(sensitive attributes)。发布时需要保密的属性,如薪金、信仰、健康状况等。

d)非敏感属性(non-sensitive attributes)。可以公开的属性,又称普通属性。

1.2 链接攻击

链接攻击是从发布的数据中获取隐私数据的常见方法。其基本思想为:攻击者通过对发布的数据和其他渠道获取的数据进行链接操作,以推理出隐私数据,从而造成隐私泄露。例如[2],通过将医疗信息表与选民登记表进行链接(图 1),几乎可以惟一确定就诊病人的医疗诊断结果,然而,病人的医疗诊断结果正是需要保护的隐私数据。数据匿名化技术可以在一定程度上解决发布数据的隐私泄露问题,但同时也给发布数据的可用性带来了限制,即产生一定量的信息损失。

2 匿名化原则

匿名化是用于解决因链接攻击所造成的隐私泄露问题的主要技术之一,以下对目前提出的基本匿名化原则进行分析和

收稿日期: 2009-11-06; **修回日期**: 2010-01-14 **基金项目**: 国家"863"计划资助项目(2006AA12A106);安徽省教育厅自然科学基金资助项目(KJ2009B075Z, KJ2009B128Z)

作者简介:王平水(1972-),男,安徽蚌埠人,副教授,博士研究生,主要研究方向为数据挖掘、信息安全(pshwang@163.com);王建东(1945-),男,教授,博导,主要研究方向为机器学习、人工智能、数据挖掘.

总结,并指出其可能存在的问题。

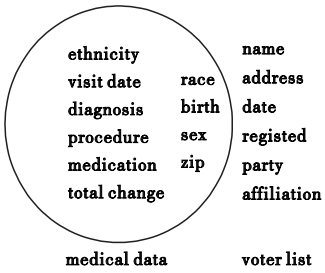

图1　链接攻击示例

2.1　*k*-anonymity

为解决链接攻击所带来的隐私泄露问题,Sweeney 等人[1~3]首次提出并发展了 *k*-匿名(*k*-anonymity)方法。

定义 1　*k*-anonymity。设原始数据表为 PT(A_1,…,A_n),匿名化后数据表为 RT(A_1,…,A_n),QIRT 是与其对应的准标志符。称数据表 RT 满足 *k*-匿名,如果 RT[QIRT]中的每个序列值在 RT[QIRT]中至少出现 *k* 次($k>1$)。数据表 RT 中具有相同准标志符的若干记录称为一个等价类,即 *k*-匿名实现了同一等价类中记录之间无法区分(敏感属性值除外)。如表 2 是表 1 的 2-匿名化表。

表 1　医疗信息表

name	race	birth	sex	zip	disease
Alice	black	1965-3-18	M	02141	Flu
Bob	black	1965-5-1	M	02142	Cancer
David	black	1966-6-10	M	02135	Obesity
Helen	black	1966-7-15	M	02137	Gastritis
Jane	white	1968-3-20	F	02139	HIV
Paul	white	1968-4-1	F	02138	Cancer

表 2　2-匿名化表,QI = {race,birth,sex,zip}

race	birth	sex	zip	disease
black	1965	M	0214 *	Flu
black	1965	M	0214 *	Cancer
black	1966	M	0213 *	Obesity
black	1966	M	0213 *	Gastritis
white	1968	F	0213 *	HIV
white	1968	F	0213 *	Cancer

k-匿名通常可以防止敏感属性值的泄露,因为每个个体身份被准确标志的概率至多为 1/*k*。然而,数据表在匿名化过程中并未对敏感属性作任何约束,这也可能带来隐私泄露。如同一等价类内敏感属性值较为集中,甚至完全相同(可能形式上,也可能语义上),这样即使满足 *k*-匿名要求,也很容易推理出与指定个体相应的敏感属性值。除此之外,攻击者也可通过自己掌握的足够的相关背景知识以很高的概率来确定敏感数据与个体的对应关系,从而导致隐私泄露。因此,*k*-匿名容易受到同质性攻击(homogeneity attack)和背景知识攻击(background knowledge attack)。

2.2　*l*-diversity

为解决同质性攻击和背景知识攻击所带来的隐私泄露,Machanavajjhala 等人[4]在 *k*-匿名基础上提出了 *l*-多样性(*l*-diversity)原则。

定义 2　*l*-diversity。称匿名数据表 RT(A_1,…,A_n)是 *l*-diversity 的,如果 RT(A_1,…,A_n)满足 *k*-匿名,且同一等价类中的记录至少有 *l* 个较好表现(well-represented)的值。其中较好表现有多种解释,如:

a) Distinct *l*-diversity。同一等价类中至少出现 *l* 个不同的敏感属性值。

b) Entropy *l*-diversity。同一等价类中敏感属性值的信息熵至少为 log *l*。等价类 *E* 的敏感属性的信息熵定义为 $H(E)=-\sum_{s\in S}P(E,s)\log P(E,s)$。其中 *S* 为敏感属性值域,$P(E,s)$为敏感属性值 *s* 在等价类 *E* 中出现的概率。

c) Recursive (c,l)-diversity。每个等价类均满足 $r_1<c(r_l+r_{l+1}+\cdots+r_m)$。其中 *m* 表示等价类中不同敏感属性值的个数,r_i 表示该等价类中第 $i(1\leqslant i\leqslant m)$ 频繁的敏感属性值的个数。Recursive (c,l)-diversity 保证了等价类中频率最高的敏感属性值不至于出现频度太高。

d) Recursive $(c1,c2,l)$-diversity。除保证等价类中频率最高的敏感属性值不至于出现频度太高,同时还保证了等价类中频率最低的敏感属性值不至于出现频度太低。

2.3　*p*-sensitive *k*-anonymity

发布的数据满足 *k*-匿名化原则的同时($k>1,p\leqslant k$),还要求同一等价类中的记录至少出现 *p* 个不同的敏感属性值,这与 distinct *l*-diversity 具有基本相同的设计思想[5]。该匿名化原则在某些数据集上可能会带来很大的信息可用性损失,也不足以抵抗敏感属性值的偏斜性攻击(skewness attack)和相似性攻击(similarity attack)。

2.4　(α,k)-anonymity

发布的数据满足 *k*-匿名化原则的同时,还要求同一等价类中任何一个敏感属性值出现的概率不大于 $\alpha(0<\alpha<1)$,即 $|(E,s)|/|E|\leqslant\alpha$[6]。偏斜性攻击和相似性攻击还可能会发生,匿名化过程中的高信息损失依然存在。

2.5　(k,e)-anonymity

该匿名化原则主要针对数值型敏感属性[7],它要求在等价类中敏感属性值的区间范围至少为 *e*。(k,e)-anonymity 试图通过最小 *e* 值克服针对敏感属性值的相似性攻击,但可能会造成高信息损失,也无法抵抗敏感属性值的偏斜性攻击。

2.6　*t*-closeness

发布的数据满足 *k*-匿名化原则的同时,还要求等价类内敏感属性值的分布与敏感属性值在匿名化表中的总体分布的差异不超过 *t*[8]。*t*-closeness 在 *l*-diversity 基础上,考虑了敏感属性的分布问题,它要求所有等价类中敏感属性值的分布尽量接近该属性的全局分布。为度量等价类与匿名化数据表中敏感属性值的分布差异,文中引入了一种独特的距离度量方式 EMD(earth mover's distance),该距离度量方式对数值型敏感属性值和类别型敏感属性值均定义了相应的计算方式。*t*-closeness 解决了针对敏感属性值的偏斜性攻击和相似性攻击,但是匿名化的结果是降低了发布数据的可用性,提高发布数据可用性的惟一办法是增大阈值 *t*。

2.7　个性化匿名

之前的匿名化原则仅提供表级别的保护粒度,对表中所有敏感属性值提供相同程度的保护,并未考虑其相应的语义关系,造成大量不必要的信息损失。文献[9]提出了个性化匿名(personalized anonymity)的概念,并给出个性化匿名的一般方法。所谓个性化匿名是指对数据表中不同敏感属性值提供不同粒度的隐私保护程度,从而减少了因统一匿名化所带来的信

息损失。文献[10]给出了个性化的(α,k)-anonymity 模型,进一步减少了信息损失,并且提高了隐私保护程度。

2.8　动态数据匿名化

目前大部分匿名化原则都是针对静态数据的,并未考虑数据记录动态更新后重发布的隐私保护问题。数据的动态更新在现实中是极为常见的,然而如果还按照原有的方法对更新后的数据集进行匿名化并重发布,很可能在多个不同的发布版本间存在推理通道,从而造成隐私泄露。文献[11]基于推迟发布、新建 l-diversity 等价类思想提出了一种支持记录插入操作的动态重发布匿名方案,能够在一定程度上阻止推理攻击所造成的隐私泄露。文献[12]给出了一种同时支持记录插入和记录删除操作的 m-invariance 匿名方案,通过保证同时出现在不同发布版本中的记录所在的等价类具有完全相同的敏感属性值集合,有效解决了不同版本间的推理通道所造成的隐私泄露问题。文献[13]提出了一种基于划分的增量数据重发布隐私保护 k-匿名算法。文献[14]提出一种(k,c)匿名方案以支持增量数据的重发布。动态数据重发布所引起的隐私泄露问题已引起了研究者的广泛关注。

3　匿名化方法

目前提出的匿名化方法的共同特征主要是通过泛化(generalization)和抑制(suppression)操作实现,该技术不同于一般的扭曲、扰乱和随机化等方法,它们能保持发布前后数据的真实性和一致性。

3.1　泛化

泛化的基本思想是用更一般的值取代原始属性值[3]。通常,泛化可分为以下两种类型:

a)域泛化,是指将一个给定的属性域泛化成一般域。例如,属性 zip 原始域 Z_0 = {02138, 02139,02141,02142} 被泛化成 Z_1 = {0213 *,0214 *},以便在语义上表达一个较大的范围,如图 2 所示。经过连续多次泛化形成的域泛化层次结构称为域泛化层,记为 DGH_A。

b)值泛化,是指原始属性域中的每个值直接泛化成一般域中的惟一值,如图 3 所示。值泛化关系同样决定了值泛化层的存在,记为 VGH_A。

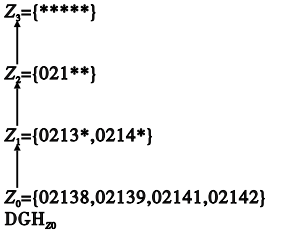

图2　包含抑制的域泛化层

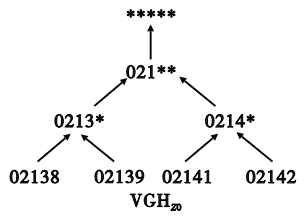

图3　包含抑制的值泛化层

给定数据表 PT(A_1,…,A_n),针对每个属性 A_i 给定其域泛化层 DGH_{Ai},则属性级上的所有可能的泛化数目为 $\prod_{i=1}^{m}(|DGH_{Ai}|+1)$,数据单元级上的所有可能的泛化数目为 $\prod_{i=1}^{m}(|DGH_{Ai}|+1)^n$。其中 m 为数据表中的属性个数,n 为记录个数。

3.2　抑制

抑制是指用最一般化的值取代原始属性值。如图 3 值泛化层 VGH_{Z0} 中处于顶层的最大值即为该属性每个值抑制操作的结果。在 k-匿名化过程中,若某些记录无法满足 k-匿名要求,则一般采取抑制操作,被抑制的相应属性值所在记录要么从数据表中删除,要么相应属性值用若干“ * ”填充,以保持有关统计特性。

3.3　其他方法

除通过泛化和抑制进行匿名化外,文献[15]提出了基于聚类的匿名化方法,将原始数据表先通过聚类技术聚成若干至少包含 k 条记录的簇,然后对每个簇再进行匿名化操作。还有基于数据交换的匿名化方法等,在此不再赘述。

关于匿名化方法,文献[3]给出了最小泛化匿名的概念,文献[16]证明了基于泛化和抑制的匿名化技术的最优解求解是一个 NP 难问题。目前文献中提出的主要是一些启发式算法,力图在合理的时间内找到近似最优解,以减少信息损失为首要优化目标。

4　匿名化度量

在实际应用中,没有一个统一的标准来衡量所有的 k-匿名算法。为处理好隐私保护与信息损失之间的关系以及衡量匿名算法的优劣,以下从不同角度给出与匿名化技术度量相关的参考标准,供选择或设计新的匿名方法时寻求隐私保护和信息损失之间的平衡。

4.1　精度度量

给定原始数据表 PT(A_1,A_2,…,A_{Na}),匿名化数据表 RT(A_1,A_2,…,A_{Na})。其中,N_a 为数据表属性个数,N 为记录个数,DGH_{Ai}是属性 A_i 的域泛化层,则匿名化表 RT 的精度可按如下方式计算[21]:

$$\mathrm{prec}(\mathrm{RT}) = 1 - \frac{\sum_{i=1}^{Na}\sum_{j=1}^{N}\frac{h}{|\mathrm{DGH}_{Ai}|}}{|\mathrm{PT}| \cdot |Na|}$$

其中:h 表示属性 A_i 泛化后在域泛化层的高度,$\frac{h}{|DGH_{Ai}|}$是指每个数据单元匿名化后的信息丢失量。显然,对于任一属性 A_i,泛化层越高,精度越小,信息损失越大。

4.2　可用性度量

可用性度量又称可辨别性度量。Bayardo 和 Agrawal 定义了 k-匿名的可用性度量方案如下[22]:Bayardo 和 Agrawal 认为可通过泛化和抑制操作所要花费的代价来衡量匿名化表的可用性大小。设等价类 E 中的每个记录的泛化代价为该等价类的大小$|E|$($|E|\geqslant k$),即包含记录个数,抑制一条记录的代价为$|D|$,即数据库的大小。于是获取该匿名化表的总代价为

$$C = \sum_{|E|\geqslant k}|E|^2 + \sum_{|E|<k}|D||E|$$

显然,等价类越大及抑制记录越多时,匿名化代价越高,相应地,匿名化表的可用性越小。

4.3　距离度量

可以通过距离方式来度量匿名表中等价类中敏感属性值的分布与其在匿名表中的总体分布的差异。文献[23]提出了 MD(manhattan distance)距离度量方式,文献[24]提出了 KLD(Kullback-Leibler distance)距离度量方式,文献[8]提出了 EMD(earth mover's distance)距离度量方式。以下给出 MD 和 KLD 两种分布间距离度量的计算方式,有关 EMD 距离的具体计算细节参见文献[8]。

给定敏感属性值两个概率分布:P = (p_1,p_2,…,p_n) 和 Q =

$(q_1,q_2,\cdots,q_m)$,则

$$\mathrm{MD}:D[P,Q]=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(p_i-q_i)$$

$$\mathrm{KLD}:D[P,Q]=\sum_{i=1}^{m}p_i\log\frac{p_i}{q_i}=H(P)-H(P,Q)$$

其中:$H(P)=\sum_{i=1}^{m}p_i\log p_i$ 为分布 P 的熵,$H(P,Q)=\sum_{i=1}^{m}p_i\log q_i$ 为分布 P 和 Q 的交叉熵。

很显然,分布距离越大,等价类中敏感属性值的分布与其在匿名表中的总体分布的差异越大,越容易受到偏斜性攻击。

5　结束语

匿名化技术由于能够在数据发布环境下防止用户敏感数据被泄露,同时又能保证发布数据的真实性,在实际应用领域受到广泛关注。但由于对该技术的研究起步较晚,其中还存在不少问题值得深入探讨。以下给出今后该技术的热点研究方向:

a)隐私性与可用性间的平衡问题。目前的研究主要集中于减少信息损失,如何找到一个合理的平衡点达到发布数据隐私性和可用性的折中是需要进一步深入研究的问题。

b)匿名化技术的执行效率问题。目前采用的匿名化方法多为贪婪式算法,执行效率不高,因此需要研究高效的匿名化算法以应对日益剧增的超容量数据的发布问题。

c)度量和评价标准问题。目前还没有统一的匿名化技术度量和评价标准,因此需要致力于该项研究,给匿名化技术一种更为客观合理的评价。

d)动态重发布数据的匿名化问题。目前研究主要集中于对静态数据匿名化,动态更新数据发布的匿名化技术研究不多,因此需要多加关注。

e)多维约束匿名问题。目前研究主要针对单一敏感属性数据表进行匿名化处理,该技术不能通过简单移植方式来解决多敏感属性数据表的匿名化问题,否则将会因属性间存在的函数信赖关系导致敏感信息的泄露。因此需要研究有效的多维约束匿名化算法。

此外,如何高效实现个性化匿名,如何根据实际应用准确选择数据表的准标志符,如何解决分布式环境下多数据表的匿名化等都是值得深入思考和研究的问题。

参考文献:

[1] SAMARATI P, SWEENEY L. Generalizing data to provide anonymity when disclosing information[C]//Proc of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. New York: ACM Press,1998:188.

[2] SWEENEY L. *K*-anonymity:a model for protecting Privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems,2002,10(5):557-570.

[3] 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.

[4] MACHANAVAJJHALA A, GEHRKE J, KIFER D. *L*-diversity: privacy beyond *k*-anonymity[J]. ACM Trans on Knowledge Discovery from Data,2007,1(1):3.

[5] TRUTA T, VINAY B. Privacy protection: *p*-sensitive *k*-anonymity property[C]//Proc of the 22nd International Conference on Data Engineering Workshops. Washington DC: IEEE Computer Society, 2006:94-103.

[6] WONG R, LI J, FU A, *et al*. (*a*, *k*)-anonymity:an enhanced *k*-anonymity model for privacy-preserving data publishing[C]//Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006:754-759.

[7] ZHANG Qing, KOUDAS N, SRIVASTAVA D, *et al*. Aggregate query answering on anonymized tables[C]//Proc of the 23th International Conference on Data Engineering. Pisctaway, NJ: IEEE Press, 2007:116-125.

[8] LI Ning-hui, LI Tian-cheng. *T*-closeness: privacy beyond *k*-anonymity and *l*-diversity[C]//Proc of the 23rd International Conference on Data Engineering. Pisctaway, NJ: IEEE Press, 2007:106-115.

[9] XIAO Xiao-kui, TAO Yu-fei. Personalized privacy preservation [C]//Proc of ACM SIGMOD Conference on Management of Data. New York: ACM Press,2006:229-240.

[10] YE Xiao-jun, ZHANG Ya-wei, LIU Ming. A personalized (*a*,*k*)-anonymity model[C]//Proc of the 9th International Conference on Web-Age Information Management. Piscataway,NJ:IEEE Press,2008:341-348.

[11] BYUN J, SOHN Y, BERTINO E, *et al*. Secure anonymization for incremental datasets[C]//Proc of the 3rd VLDB Workshop on Secure Data Management. Berlin:Springer-Verlag,2006:48-63.

[12] XIAO Xiao-kui, TAO Yu-fei. *M*-invariance: towards privacy preserving re-publication of dynamic datasets[C]//Proc of ACM SIGMOD Conference on Management of Data. New York: ACM Press, 2007: 689-700.

[13] 吴英杰,倪巍伟,张柏礼,等. *K*-APPRP:一种基于划分的增量数据重发布隐私保护 *k*-匿名算法[J]. 小型微型计算机系统,2009,30(8):1581-1587.

[14] BYUNA J, LI Tian-cheng, BERTINO E, *et al*. Privacy-preserving incremental data dissemination[J]. Journal of Computer Security, 2009,17(1):43-68.

[15] AGGARWAL G, FEDER T, KENTHAPADI T, *et al*. Achieving anonymity via clustering[C]//Proc of Symposium on Principles of Database Systems. New York:ACM Press,2006:153-162.

[16] MEYERSON A, WILLIAMS R. On the complexity of optimal *k*-anonymity[C]//Proc of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York: ACM Press, 2004:223-228.

[17] LEFEVRE K, DEWITTD J, RAMAKRISHNAN R. Incpgnoto: efficient full-domain *k*-anonymity[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005:49-60.

[18] FUNG B, WANG Ke, YU P. Top-down specialization for information and privacy preservation[C]//Proc of the 21st IEEE International Conference on Data Engineering. Washington DC: IEEE Computer Society,2005:205-216.

[19] WANG Ke, YU P, CHALRABORTY S. Bottom-up generalization: a data mining solution to privacy protection[C]//Proc of the 4th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society,2004:249-256.

[20] BYUN J, KAMRA A, BERTINO E, *et al*. Efficient *k*-anonymization using clustering techniques[C]//Proc of International Conference on Database Systems for Advanced Applications. Berlin:Springer-Verlag, 2007:188-200.

[21] BAYARDO R, AGRAWAL R. Data privacy through optimal *k*-anonymization[C]//Proc of the 21st International Conference on Data Engineering. Los Alamitos: IEEE Computer Society, 2005:217-228.

[22] KIFER D, GEHRKE J. Injecting utility into anonymized datasets [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press,2006:217-228.

[23] LIN Jian-hua. Divergence measures based on the shannon theory[J]. IEEE Trans on Information Theory,1991,37(1):145-151.

[24] KULLBACK S, LEIBLER R. On information and sufficiency[J]. The Annals of Mathematical Statistics,1951,22(1):79-86.

[25] 周水庚,李丰,陶宇飞,等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009,32(5):847-861.