面向数据库应用的隐私保护研究综述

周水庚[1),2)] 李 丰[1)] 陶宇飞[3)] 肖小奎[4)]

[1)](复旦大学计算机科学技术学院 上海 200433)

[2)](复旦大学上海市智能信息处理重点实验室 上海 200433)

[3)](香港中文大学计算机科学与工程系 香港)

[4)](美国康乃尔大学计算机科学系 美国)

摘 要 随着数据挖掘和数据发布等数据库应用的出现与发展,如何保护隐私数据和防止敏感信息泄露成为当前面临的重大挑战.隐私保护技术需要在保护数据隐私的同时不影响数据应用.根据采用技术的不同,出现了数据失真、数据加密、限制发布等隐私保护技术.文中对隐私保护领域已有研究成果进行了总结,对各类隐私保护技术的基本原理、特点进行了阐述,还详细介绍了各类技术的典型应用,并重点介绍了当前该领域的研究热点:基于数据匿名化的隐私保护技术.在对已有技术深入对比分析的基础上,指出了隐私保护技术的未来发展方向.

关键词 数据库应用;隐私保护;数据挖掘;数据发布;随机化;多方安全计算;匿名化

中图法分类号 TP309 **DOI 号**: 10.3724/SP.J.1016.2009.00847

Privacy Preservation in Database Applications: A Survey

ZHOU Shui-Geng[1),2)] LI Feng[1)] TAO Yu-Fei[3)] XIAO Xiao-Kui[4)]

[1)](*School of Computer Science and Technology, Fudan University, Shanghai* 200433)

[2)](*Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai* 200433)

[3)](*Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, China*)

[4)](*Department of Computer Science, Cornell University, USA*)

Abstract As the emergence and development of database applications such as data publishing and data mining, a challenge to the database community is to preserve data privacy and prevent sensitive information from disclosure. Privacy-preserving techniques should be conducive to the applications while preserving data privacy. Based on different principles, various privacy-preserving techniques are developed, such as distortion, encryption and limited distribution. This paper surveys the state of the art of privacy preservation techniques for database applications. The mechanisms and characteristics of various techniques are described, while focus is put on data anonymization, which is a hot topic in the field. Following a comprehensive comparison and analysis of existing techniques, future research directions are highlighted.

Keywords database applications; privacy preservation; data mining; data dissemination; randomization; secure multi-party computation; anonymization

1 引 言

数据挖掘和数据发布是当前数据库应用的两个重要方面.一方面,数据挖掘与知识发现在各个领域都扮演着非常重要的角色.数据挖掘的目的在于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].传统的数据挖掘技术在发现知识的同时,

收稿日期:2008-07-02;最终修改稿收到日期:2009-03-08.本课题得到国家自然科学基金项目(60873070)资助.**周水庚**,男,1966 年生,教授,博士生导师,主要研究领域包括数据库、P2P 计算、生物信息学等. E-mail:sgzhou@fudan.edu.cn. **李 丰**,男,1983 年生,硕士研究生,研究方向为数据库、隐私保护. **陶宇飞**,男,1978 年生,香港中文大学计算机科学与工程系助理教授,主要研究方向为时间数据库、空间数据库和隐私保护. **肖小奎**,男,1981 年生,美国康乃尔大学计算机科学系博士后,主要研究方向为隐私保护.

也给数据的隐私带来了威胁.例如,疾病控制中心需要收集各医疗机构的病例信息,以进行疾病的预防与控制.在这个过程中,传统数据挖掘技术将不可避免地暴露敏感数据(如"病人所患疾病"),而这些敏感数据是数据所有者(医疗机构、病人)不希望被揭露的.另一方面,数据发布是将数据库中的数据直接地展现给用户.而在各种数据发布应用中,如果数据发布者不采取适当的数据保护措施,将可能造成敏感数据的泄漏,从而给数据所有者带来危害.譬如企业发布的产品信息,或者上市公司发布的财务年报,如果不对发布的数据进行仔细甄别,就会给商业上的竞争者以可乘之机.所以,如何在各种数据库应用中保护数据的隐私,成为近年来学术界的研究热点[2-9].

隐私保护技术[3-4]的出现就是为了解决上述问题.具体地说,实施数据隐私保护主要是考虑以下两个方面:(1)如何保证数据应用过程中不泄露隐私;(2)如何更有利于数据的应用.当前,隐私保护领域的研究工作主要集中于如何设计隐私保护原则和算法更好地达到这两方面的平衡.

本文立足于数据库应用,对隐私保护技术的最新进展进行综述.一方面对隐私的基本定义、度量和研究方向等研究背景进行介绍;另一方面,对该领域的主要技术进行分类阐述,在具体应用中对该技术的优缺点、适用范围等进行分析,其中重点介绍研究热点:"数据匿名化"(data anonymization)①.而后在这两方面的基础上对隐私保护技术进行综合对比与分析.目前隐私保护技术在数据库中的应用主要集中在"数据挖掘"和"匿名发布"两大领域,因此本文在介绍特定隐私技术的同时,将重点突出该技术在相关领域的应用.

本文第 2 节简单介绍隐私的概念、度量;第 3 节和第 4 节分别对隐私保护的研究方向和技术分类进行介绍;第 5～7 节对三大类隐私保护技术:基于数据失真的隐私保护技术、基于加密的隐私保护技术和匿名化技术进行阐述;第 8 节对各类技术进行对比、分析,并指明未来的研究方向.

2 隐私与隐私度量

2.1 隐私的定义

简单地说,隐私就是个人、机构等实体不愿意被外部世界知晓的信息.在具体应用中,隐私即为数据所有者不愿意被披露的敏感信息,包括敏感数据以及数据所表征的特性.通常我们所说的隐私都指敏感数据,如个人的薪资、病人的患病记录、公司的财务信息等.但当针对不同的数据以及数据所有者时,隐私的定义也会存在差别的.例如保守的病人会视疾病信息为隐私,而开放的病人却不视之为隐私.一般地,从隐私所有者的角度而言,隐私可以分为两类[5]:

(1)个人隐私(individual privacy)[10]:任何可以确认特定个人或与可确认的个人相关,但个人不愿被暴露的信息,都叫做个人隐私,如身份证号、就诊记录等.

(2)共同隐私(corporate privacy):共同隐私不仅包含个人的隐私,还包含所有个人共同表现出但不愿被暴露的信息.如公司员工的平均薪资、薪资分布等信息.

2.2 隐私的度量

数据隐私的保护效果是通过攻击者披露隐私的多寡来侧面反映的.现有的隐私度量都可以统一用"披露风险"(disclosure risk)来描述.披露风险表示为攻击者根据所发布的数据和其它背景知识(background knowledge)可能披露隐私的概率.通常,关于隐私数据的背景知识越多,披露风险越大.

若 s 表示敏感数据,事件 S_k 表示"攻击者在背景知识 K 的帮助下揭露敏感数据 s",则披露风险 $r(s,K)$ 表示为

$$r(s,K)=P_r(S_k).$$

对数据集而言,若数据所有者最终发布数据集 D 的所有敏感数据的披露风险都小于阈值 α,$\alpha\in[0,1]$,则称该数据集的披露风险为 α.例如,静态数据发布原则 l-diversity[7]保证发布数据集的披露风险小于 $1/l$,动态数据发布原则 m-Invariance[9]保证发布数据集的披露风险小于 $1/m$.

特别地,不做任何处理所发布数据集的披露风险为 1;当所发布数据集的披露风险为 0 时,这样发布的数据被称为实现了完美隐私(perfect privacy)[11-13].完美隐私实现了对隐私最大程度的保护,但由于对攻击者先验知识的假设本身是不确定的,因此实现对隐私的完美保护也只在具体假设、特定场景下成立,真正的完美保护并不存在.

3 主要研究方向与国内研究现状

3.1 隐私保护的主要研究方向

隐私保护问题是伴随着数据应用而提出的.在

① 据我们所知,目前还没有文章对包括数据匿名化在内的隐私保护技术进行过评述.

统计领域,隐私保护问题最先受到关注[14].当前,隐私保护的主要研究方向如表 1 所示.

表 1　隐私保护研究方向

研究方向	示例
通用的隐私保护技术	Perturbation[14]、Randomization[4,15]、Swapping[16]、Encryption[17]
面向数据挖掘的隐私保护技术	Association Rule Mining[18-20]、Classification[21-22]、Clustering[23]
基于隐私保护的数据发布原则	k-anonymity[24-25]、l-diversity、m-Invariance、t-Closeness[26]
隐私保护算法	Anonymized Publication[27-30]、Anonymization with High Utility[31]

隐私保护的研究问题是由实际应用中不同的隐私保护需求决定的.通用的隐私保护技术致力于在较低应用层次上保护数据的隐私,一般通过引入统计模型和概率模型来实现;而面向数据挖掘的隐私保护技术主要解决在高层数据应用中,如何根据不同数据挖掘操作的特性,实现对隐私的保护;基于隐私保护的数据发布原则是为了提供一种在各类应用可以通用的隐私保护方法,进而使得在此基础上设计的隐私保护算法也具通用性.

3.2　国内研究现状

作为新兴的研究热点,隐私保护技术不论在理论研究还是实际应用方面,都具有非常重要的价值.在国内(特指大陆地区),对隐私保护技术的研究亦受到学术界的关注与重视,包括复旦大学、中国科学技术大学、北京大学、东北大学、华中科技大学等在内的多个课题组也开展了相关的研究工作.

国内关于隐私保护技术的研究目前主要集中于基于数据失真或数据加密技术方面的研究,如基于隐私保护分类挖掘算法[32]、关联规则挖掘[33-34]、分布式数据的隐私保持协同过滤推荐[35]、网格访问控制[36]等.文献[37]亦对基于限制发布的隐私保护技术进行了研究,提出了支持多约束的 k-匿名化方法.

总的来说,国内关于隐私保护技术的研究还处于起步阶段,具有广阔的发展空间.

4　隐私保护技术的分类与性能评估

4.1　隐私保护技术的分类

没有任何一种隐私保护技术适用于所有应用.本文将隐私保护技术分为 3 类:

(1)基于数据失真(distorting)的技术.它是使敏感数据失真但同时保持某些数据或数据属性不变的方法.例如,采用添加噪声(adding noise)、交换(swapping)等技术对原始数据进行扰动处理,但要求保证处理后的数据仍然可以保持某些统计方面的性质,以便进行数据挖掘等操作.

(2)基于数据加密的技术.它是采用加密技术在数据挖掘过程中隐藏敏感数据的方法,多用于分布式应用环境中,如安全多方计算(Secure Multiparty Computation,SMC)[38-39].

(3)基于限制发布的技术.它是根据具体情况有条件地发布数据.如不发布数据的某些域值、数据泛化(generalization)[24-25]等.

另外,对于许多新方法[12-13,40-41],由于其融合了多种技术,很难将其简单地归到以上某一类,但它们在利用某类技术优势的同时,将不可避免地引入其它的缺陷.基于数据失真的技术,效率比较高,但却存在一定程度的信息丢失;基于加密的技术则刚好相反,它能保证最终数据的准确性和安全性,但计算开销比较大;而限制发布技术的优点是能保证所发布的数据一定真实,但发布的数据会有一定的信息丢失.

在接下来的 3 节,本文将对这三类隐私保护技术及其应用进行深入阐述.

4.2　隐私保护技术的性能评估

隐私保护技术需要在保护隐私的同时,兼顾对应用的价值以及计算开销.通常从以下三方面对隐私保护技术进行度量:

(1)隐私保护度.通常通过发布数据的披露风险来反映,披露风险越小,隐私保护度越高.

(2)数据缺损.是对发布数据质量的度量,它反映通过隐私保护技术处理后数据的信息丢失:数据缺损越高,信息丢失越多,数据利用率(utility)越低.具体的度量有:信息缺损(information loss)的程度[21-22,42-43]、重构数据与原始数据的相似度[31]等.

(3)算法性能.一般利用时间复杂度对算法性能进行度量.例如,采用抑制(suppression)实现最小化的 k-匿名问题已经证明是 NP-hard 问题[28];时间复杂度为 $O(k)$ 的近似 k-匿名算法[44],显然优于复杂度为 $O(k\log k)$ 的近似算法[28].均摊代价(amortized cost)是一种类似于时间复杂度的度量,它表示算法在一段时间内平均每次操作所花费的时间代价.除此之外,在分布式环境中,通信开销(communication cost)也常常关系到算法性能,常作为衡量分布式算法性能的一个重要指标.

5 基于数据失真的隐私保护技术

数据失真技术通过扰动(perturbation)原始数据来实现隐私保护.它要使扰动后的数据同时满足:

(1)攻击者不能发现真实的原始数据.也就是说,攻击者通过发布的失真数据不能重构出真实的原始数据.

(2)失真后的数据仍然保持某些性质不变,即利用失真数据得出的某些信息等同于从原始数据上得出的信息.这就保证了基于失真数据的某些应用的可行性.

当前,基于数据失真的隐私保护技术包括随机化、阻塞(blocking)[45-47]、交换、凝聚(condensation)[48]等.一般地,当进行分类器构建[4,49]和关联规则挖掘[18],而数据所有者又不希望发布真实数据时,可以预先对原始数据进行扰动后再发布.

5.1 随机化

数据随机化即是对原始数据加入随机噪声,然后发布扰动后数据的方法.需要注意的是,随意对数据进行随机化并不能保证数据和隐私的安全[50],因为利用概率模型进行分析常常能披露随机化过程的众多性质.随机化技术包括两类:随机扰动(random perturbation)和随机化应答(randomized response).

5.2 随机扰动

随机扰动采用随机化过程来修改敏感数据,从而实现对数据隐私的保护.一个简单的随机扰动模型[4]如图 1(a)所示.

对外界而言,只可见扰动后的数据,从而实现了对真实数据值的隐藏.但扰动后数据仍然保留着原始数据分布 X 的信息,通过对扰动后的数据进行重构(图 1(b)所示),可以恢复原始数据分布 X 的信息.但不能重构原始数据的精确值 $x_1,x_2,\cdots,x_n$.

输入	1. 原始数据为 $x_1,x_2,\cdots,x_n$,服从于未知分布 X; 2. 扰动数据为 $y_1,y_2,\cdots,y_n$,服从特定分布 Y
输出	随机扰动后的数据:$x_1+y_1,x_2+y_2,\cdots,x_n+y_n$

(a) 随机扰动过程

输入	1. 随机扰动后数据:$x_1+y_1,x_2+y_2,\cdots,x_n+y_n$ 2. 扰动数据的分布 Y
输出	原始数据分布 X

(b) 重构过程

图 1 随机扰动与重构过程

随机扰动技术可以在不暴露原始数据的情况下进行多种数据挖掘操作.由于通过扰动数据重构后的数据分布几乎等同于原始数据的分布,因此利用重构数据的分布进行决策树分类器训练后,得到的决策树能很好地对数据进行分类[4].在关联规则挖掘中,通过往原始数据注入大量伪项(false item)来对频繁项集进行隐藏[18],再通过在随机扰动后的数据上估计项集支持度,从而发现关联规则.除此之外,随机扰动技术还可以应用到 OLAP 上实现对隐私的保护[41].

5.3 随机化应答

随机化应答[15]的基本思想是:数据所有者对原始数据扰动后发布,使攻击者不能以高于预定阈值的概率得出原始数据是否包含某些真实信息或伪信息.虽然发布的数据不再真实,但在数据量比较大的情况下,统计信息和汇聚(aggregate)信息仍然可以较为精确地被估算出.随机化应答技术与随机扰动技术的不同之处在于敏感数据是通过一种应答特定问题的方式间接提供给外界的.

随机化应答模型有两种:相关问题模型(related-question model)和非相关问题模型(unrelated-question model).相关问题模型是通过设计两个关于敏感数据的对立问题,如:

(1)我含有敏感值 A;

(2)我没有敏感值 A.

数据所有者根据自己拥有的数据随机选取一个问题进行应答,但不让提问者知道回答的具体问题.当大量数据所有者进行回答后,通过计算可以得出含有敏感值的应答者比例和不含敏感值应答者的比例.假设应答者随机选取问题 1 的概率为 θ,则有以下等式成立:

$$P^*(A=\text{yes})=P(A=\text{yes})\cdot\theta+P(A=\text{no})\cdot(1-\theta);$$

$$P^*(A=\text{no})=P(A=\text{no})\cdot\theta+P(A=\text{yes})\cdot(1-\theta).$$

其中 $P^*(A=\text{yes})$ 是回答中 yes 的比例,$P(A=\text{yes})$ 是含有敏感值 A 的数据所有者的比例.通过以上两个等式,联合对所有应答进行估计得出的 $P^*(A=\text{yes})$ 和 $P^*(A=\text{no})$,可以得到含有(或不含有)敏感值 A 的数据所有者比例 $P(A=\text{yes})$(或 $P(A=\text{no})$).

在这整个过程中,由于不能确定与应答者回答的相关问题,因此不能确定其是否含有敏感数据值.由于基于随机化应答技术采用应答模式提供信息,因此多用于处理分类数据(categorical data).

MASK[51](Mining Associations with Secrecy Konstraints)是一种基于随机化应答技术的布尔关联规则挖掘算法.它利用预先定义的分布函数产生

随机数并对原始数据进行扰动,数据使用者基于扰动数据,结合应答信息对数据进行重构,在此基础上,估计出项集的支持度从而找出频繁项集.

5.2 阻塞与凝聚

随机化技术一个无法避免的缺点是:针对不同的应用都需要设计特定的算法对转换后的数据进行处理,因为所有的应用都需要重建数据的分布.基于此,文献[48]提出了凝聚技术:它将原始数据记录分成组,每一组内存储着由 k 条记录产生的统计信息,包括每个属性的均值、协方差等.这样,只要是采用凝聚技术处理的数据,都可以用通用的重构算法进行处理,并且重构后的记录并不会披露原始记录的隐私,因为同一组内的 k 条记录是两两不可区分的.

与随机化技术修改数据、提供非真实数据的方法不同,阻塞技术采用的是不发布某些特定数据的方法,因为某些应用更希望基于真实数据进行研究.阻塞技术具体反应到数据表中,即是将某些特定的值用一个不确定符号代替.例如通过引入除{0,1}外的代表不确定值的符号"?"可以实现对布尔关联规则的隐藏.由于某些值被"?"代替,那么对某些项集的计数则为一个不确定的值,位于一个最小估计值和最大估计值范围内.于是,对于敏感关联规则的隐藏即是设计一种算法,在阻塞尽量少的数据值情况下将敏感关联规则可能的支持度和置信度控制在预定的阈值以下[45-46].类似于对关联规则的隐藏,利用阻塞技术还可以实现对分类规则的隐藏[47].

6 基于数据加密的隐私保护技术

在分布式环境下实现隐私保护要解决的首要问题是通信的安全性,而加密技术正好满足了这一需求,因此基于数据加密的隐私保护技术多用于分布式应用中,如分布式数据挖掘、分布式安全查询、几何计算、科学计算等.在分布式下,具体应用通常会依赖于数据的存储模式和站点(site)的可信度及其行为.

分布式应用采用两种模式存储数据:垂直划分(vertically partitioned)的数据模式和水平划分(horizontally partitioned)的数据模式.垂直划分数据是指分布式环境中每个站点只存储部分属性的数据,所有站点存储的数据不重复;水平划分数据是将数据记录存储到分布式环境中的多个站点,所有站点存储的数据不重复.

对分布式环境下的站点(参与者),根据其行为,可分为:准诚信攻击者(semi-honest adversary)和恶意攻击者(malicious adversary):准诚信攻击者是遵守相关计算协议但仍试图进行攻击的站点;恶意攻击者是不遵守协议且试图披露隐私的站点.一般地,假设所有站点为准诚信攻击者.

6.1 安全多方计算

众多分布环境下基于隐私保护的数据挖掘应用都可以抽象为无信任第三方(trusted third party)参与的 SMC 问题,即怎样使两个或多个站点通过某种协议完成计算后,每一方都只知道自己的输入数据和所有数据计算后的最终结果.

以在分布式下计算集合的并为例:假设有 N 个独立站点 $S_1,S_2,\cdots,S_N$,站点 S_i 拥有数据 D_i,这 N 个站点在不暴露每个站点具体数据情况下,计算出 $\bigcup_{i=1}^{N} D_i$.具体过程如图 2 所示[39].

输入	站点 $S_1,S_2,\cdots,S_N$,数据 $D_1,D_2,\cdots,D_N$
输出	$\bigcup_{i=1}^{N} D_i$
过程	1. $\forall i$, $i\in[1,N]$, 站点 i 产生私有密钥 K_i,且任意两个站点 i,j 的密钥满足可交换:$E_{K_i}(E_{K_j}(D))=E_{K_j}(E_{K_i}(D))$,其中 D 为任意数据; 2. 对任意站点 i: 2.1. 加密数据 D_i 并将 $E_{K_i}(D_i)$ 传递给其它站点; 2.2. 将其它站点传递给自己的加密数据用 K_i 加密后传递给其它站点; 3. 任意数据 D_i 都被所有站点的密钥加密后,为 $E_{K_x}(\cdots E_{K_i}(D_i)\cdots)$.直到全部数据都被所有站点的私有密钥加过密后,去掉重复的数据项,将剩余的数据项再依次传递给所有站点解密; 4. 最后对数据进行解密的站点将得到 $\bigcup_{i=1}^{N} D_i$.

图 2　集合并集安全计算流程

可以证明[39],由于采用了可交换加密技术的顺序无关性,在整个求集合并集的过程中,除了集合交集的大小和最终结果被披露外,没有其它私有信息泄露,所以该计算集合并的方法是安全的.

由于多数 SMC 基于"准诚信模型"假设之上,因此应用范围有限.SCAMD(Secure Centralized Analysis of Multi-party Data)协议[52]在去除该假设基础上,引入准诚信第三方实现当站点都是恶意时进行安全多方计算;文献[6]提出抛弃传统分布式环境下对站点行为约束的假设,转而根据站点的动机,将站点分为弱恶意攻击者和强恶意攻击者,用可交换加密技术解决在分布环境下的信息共享问题.

当前,关于 SMC 的主要研究工作集中于降低计算开销、优化分布式计算协议[53]以及以 SMC 为工具解决问题等.

6.2 分布式匿名化

匿名化即是隐藏数据或数据来源.因为对大多数应用而言,首先需要对原始数据进行处理以保证敏感信息的安全;然后再在此基础上,进行数据挖掘、发布等操作.分布式下的数据匿名化都面临在通信时,如何既保证站点数据隐私又能收集到足够的信息来实现利用率尽量大的数据匿名.

以在垂直划分的数据环境下实现两方的分布式 k-匿名为例.两个站点 S_1 和 S_2,它们拥有的数据分别为 $\{ID, A_{1_1}, A_{1_2}, \cdots, A_{1_{n_1}}\}$,$\{ID, A_{2_1}, A_{2_2}, \cdots, A_{2_{n_2}}\}$.其中 A_{i_j} 为 S_i 拥有数据的第 j 个属性.利用可交换加密在通信过程中隐藏原始信息,再构建完整的匿名表判断是否满足 k-匿名条件来实现[54](图 3 所示).

输入 站点 S_1,S_2,数据$\{ID, A_{1_1}, A_{1_2}, \cdots, A_{1_{n_1}}\}$,$\{ID, A_{2_1}, A_{2_2}, \cdots, A_{2_{n_2}}\}$

输出 k-匿名数据表 T^*

过程

1. 2 个站点分别产生私有密钥 K_1 和 K_2,且满足:$E_{K_1}(E_{K_2}(D))=E_{K_2}(E_{K_1}(D))$,其中 D 为任意数据;
2. 表 T^*←Null;
3. while T^* 中数据不满足 k-匿名条件 do
4. 站点 $i(i=1$ 或 $2)$:
   4.1. 对$\{ID, A_{i_1}, A_{i_2}, \cdots, A_{i_{n_i}}\}$进行一次泛化,为:$\{ID, A_{i_1}^*, A_{i_2}^*, \cdots, A_{i_{n_i}}^*\}$,其中 $A_{i_j}^*$ 表示 A_{i_j} 泛化后的值;
   4.2. $\{ID, A_{i_1}, A_{i_2}, \cdots, A_{i_{n_i}}\}$←$\{ID, A_{i_1}^*, A_{i_2}^*, \cdots, A_{i_{n_i}}^*\}$;
   4.3. 用 K_i 加密$\{ID, A_{i_1}^*, A_{i_2}^*, \cdots, A_{i_{n_i}}^*\}$并传递给另一站点;
   4.4. 用 K_i 加密另一站点加密的泛化数据并回传;
   4.5. 根据两个站点加密后的 ID 值对数据进行匹配,构建经 K_1 和 K_2 加密后的数据表 T^*:$\{ID, A_{1_1}^*, A_{1_2}^*, \cdots, A_{1_{n_1}}^*, A_{2_1}^*, A_{2_2}^*, \cdots, A_{2_{n_2}}^*\}$;
5. end while.

图 3 分布式 k-匿名算法流程

在水平划分的数据环境中,可以通过引入第三方,利用满足以下性质的密钥来实现数据的 k-匿名化[55]:每个站点加密私有数据并传递给第三方,当且仅当有 k 条数据记录的准标志符属性[24]值相同时,第三方的密匙才能解密这 k 条数据记录.

更一般地,不考虑数据的具体存储模式,一种能确保分布式环境下隐私安全的模型是 k-TTP(k-Trusted Third Party)[56].k-TTP 利用信任第三方,确保了当且仅当至少有 k 个站点的信息改变时,所有站点的相关统计信息才能被披露.k-TTP模型的约束,使我们不能揭露少于 k 个站点的统计信息.

由于分布式固有的复杂性,当前实现分布式数据匿名化的主要挑战是解决数据分散、站点自治、安全通信等之间的矛盾和冲突.

6.3 分布式关联规则挖掘

在分布式环境下,关联规则挖掘的关键是计算项集的全局计数,加密技术能保证在计算项集计数的同时,不会泄露隐私信息.

例如,在数据垂直划分的分布式环境中,需要解决的问题是:如何利用分布在不同站点的数据计算项集(item set)计数,找出支持度大于阈值的频繁项集.此时,计算项集计数的问题被简化为在保护隐私数据的同时,在不同站点间计算标量积的问题.已有计算标量积的方法包括引入随机向量[39]进行安全计算或用随机数代替真实值[19],然后用代数方法进行计算等.

6.4 分布式聚类

基于隐私保护的分布式聚类的关键是安全地计算数据间的距离,有以下两种常用模型:

(1) Naïve 聚类模型.各个站点将数据用加密方式安全地传递给信任第三方,由信任第三方进行聚类后返回结果.

(2) 多次聚类模型.首先各个站点对本地数据进行聚类并发布结果,再通过对各个站点发布的结果进行二次处理,实现分布式聚类[40,57].

不论哪种分布式聚类模型,都利用了加密以实现信息的安全传输.当然,还有基于隐私保护的其它分布式聚类方法,如在任意划分数据的环境下的 k-mean 聚类算法[58],通过引入随机数来保证安全传输的最大期望(expectation maximization)聚类算法[39]等.

基于数据加密的隐私保护技术除用于以上应用场合外,还可以用于解决分布式决策树生成[59]、贝叶斯网络构建[60]等问题.

7 基于限制发布的隐私保护技术

限制发布即是有选择地发布原始数据、不发布

或者发布精度较低的敏感数据,以实现隐私保护.当前此类技术的研究集中于“数据匿名化”:即在隐私披露风险和数据精度间进行折中,有选择地发布敏感数据及可能披露敏感数据的信息,但保证对敏感数据及隐私的披露风险在可容忍范围内.数据匿名化研究主要集中在两个方面:一是研究设计更好的匿名化原则,使遵循此原则发布的数据既能很好地保护隐私,又具有较大的利用价值.另一方面是针对特定匿名化原则设计更“高效”的匿名化算法.本节内容将围绕这两方面展开.

值得一提的是,随着数据匿名化研究的逐渐深入,如何实现匿名化技术的实际应用[61],成为当前研究者关注的焦点:例如如何采用匿名化技术,实现对数据库的安全查询,以保证敏感信息无泄漏等[12,62-64].

数据匿名化一般采用两种基本操作:

(1)抑制.抑制某数据项,亦即不发布该数据项;

(2)泛化.泛化是对数据进行更概括、抽象的描述.譬如,对整数 5 的一种泛化形式是[3,6],因为 5 在区间[3,6]内.

7.1　数据匿名化原则

数据匿名化所处理的原始数据,如医疗数据、统计数据等,一般为数据表形式:表中每一条记录(或每一行)对应一个个人,包含多个属性值.这些属性可以分为 3 类:

(1)显式标识符(explicit identifier).能唯一标识单一个体的属性,如身份证号码、姓名等.

(2)准标识符(quasi-identifiers).联合起来能唯一标识一个人的多个属性,如邮编、生日、性别等联合起来则可能是准标识符.

(3)敏感属性(sensitive attribute).包含隐私数据的属性,如疾病、薪资等.

例如,表 2 为一原始医疗数据,每一条记录对应一个唯一的病人,其中{“姓名”}为显式标识符属性,{“年龄”,“性别”,“邮编”}为准标识符属性,{“疾病”}为敏感属性.

表 2　原始数据[8]

姓名	年龄	性别	邮编	疾病
Andy	4	M	12000	胃溃疡
Bill	5	M	14000	消化不良
Ken	6	M	18000	肺炎
Nash	9	M	19000	支气管炎
Alice	12	F	22000	流感
Betty	19	F	24000	肺炎

7.1.1　k-匿名

Samarati 和 Sweeney 提出的 k-匿名原则即是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[24].我们称不能相互区分的 k 条记录为一个等价类(equivalence class).这里的不能区分只对非敏感属性项而言.一般 k 值越大,对隐私的保护效果更好,但丢失的信息越多.表 3 为匿名化表 2 中原始数据后的结果,其满足 2-匿名的原则,即等价类中每一条数据不能和另一条数据相区分.

表 3　匿名化数据

年龄	性别	邮编	疾病
[1,5]	M	[10k,15k]	胃溃疡
[1,5]	M	[10k,15k]	消化不良
[6,10]	M	[15k,20k]	肺炎
[6,10]	M	[15k,20k]	支气管炎
[11,20]	F	[20k,25k]	流感
[11,20]	F	[20k,25k]	肺炎

k-匿名的缺陷在于没有对敏感数据做任何约束,攻击者可以利用一致性攻击(homogeneity attack)和背景知识攻击(background knowledge attack)来确认敏感数据与个人的联系[7],导致隐私泄露.(α,k)-匿名原则[65]在此基础上进行了改进,其在保证发布的数据满足 k-匿名化原则的同时,还保证发布数据的每一个等价类中,与任一敏感属性值相关的记录的百分比不高于 α.

7.1.2　l-diversity

l-diversity 保证每一个等价类的敏感属性至少有 l 个不同的值.l-diversity 使得攻击者最多以 $1/l$ 的概率确认某个体的敏感信息.同样,表 3 发布的数据也是满足 2-diversity 的:每一个等价类中至少有 2 个不同的敏感属性值.另外,l-diversity 还有两种其它的形式:

(1)基于熵的 l-diversity:如果每个等价类的熵 $Entropy(E)>\log l$,那么所发布的数据满足基于熵的 l-diversity.其中,等价类的熵定义为 $Entropy(E)=-\sum_{s\in S}p(E,s)\log p(E,s)$,$p(E,s)$为等价类 E 中敏感属性值为 s 的记录的百分比.熵越大,表示等价类的敏感属性值分布越均匀,攻击者揭露个人的隐私就越困难.

(2)递归(c,l)-diversity:如果每个等价类都满足 $r_1<c(r_l+r_{l+1}+\cdots+r_m)$,那么就说所发布的数据满足递归($c$, l)-diversity.这里,r_i 表示该等价类中第 i 频繁的敏感属性值的个数.递归(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①.

7.1.3 t-Closeness

t-Closeness 在 l-diversity 基础上,考虑了的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[26].

定义 1. 令 $P=\{p_1,p_2,\cdots,p_m\}$,$Q_i=\{q_1,q_2,\cdots,q_m\}$分别表示各敏感值的全局分布和在等价类 C_i中的分布.对任意等价类 C_i,若 P 与 Q_i的距离 $D[P,Q_i]$满足:$D[P,Q_i]<t$,则发布的数据满足匿名化原则 t-Closeness.其中:阈值 $t\in[0,1]$;度量距离可采用可变距离:$D[P,Q_i]=\sum_{i=1}^{m}\frac{1}{2}|p_i-q_i|$或 KL 距离:$D[P,Q_i]=\sum_{i=1}^{m}p_i\log\frac{p_i}{q_i}$.

除以上匿名化原则外,文献[8]提出了个性化隐私保护(personalized privacy preservation)的匿名化原则,以满足不同个人隐私保护的要求和级别,并克服了统一匿名化所造成的数据"过分"保护和保护"不足".

一般遵循 k-匿名、l-diversity 等匿名化原则发布数据都采用泛化技术,这在很大程度上降低了数据的精度和利用率.一种高精度的数据发布方法是 Anatomy[66]:首先利用原始数据产生满足 l-diversity 原则的数据划分,然后将结果分成两张数据表发布,一张表包含每个记录的准标识符属性值和该记录的等价类 ID,另一张表包含等价类 ID、每个等价类的敏感属性值及其计数.这种将结果"切开"发布的方法,在提高准标识符属性数据精度的同时,保证了发布的数据满足 l-diversity 原则,对敏感数据提供了较好的保护.

7.2 数据匿名化算法

大多数匿名化算法致力于解决根据通用匿名原则,怎样更好地发布匿名数据.另一部分工作致力于解决在具体应用背景下,如何使发布的匿名数据更有利于应用.近年,出现了采用聚类思想进行匿名化的算法,能在发布数据精度和计算开销间达到较好的平衡,将在本节的最后进行介绍.

7.2.1 基于通用原则的匿名化算法

不同情况下,实现 k-匿名的算法有多种度量可采用,如:等价类所包含的平均记录条数[29]、数据的信息缺损[22]、实现数据匿名的操作数、可识别度量[67](discernability metrics)等.通常采用泛化(抑制)技术来实现最优化的 k-匿名原则的算法,对泛化空间(抑制策略)的搜索直接影响到了算法的性能.然而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-hard②,因此,很大一部分实现 k-匿名的算法研究着眼于设计高效的近似算法.

如图 4 所示,基于通用原则的匿名化算法常包括泛化空间枚举、空间修剪、选取最优泛化、结果判断与输出等步骤.例如,最早提出的 MinGen 算法[25]采用的就是每一步都完全搜索泛化空间,选出最优的泛化操作,一直进行这样的操作直到数据满足 k-匿名原则.但 MinGen 算法由于采用完全搜索,时间复杂度高,因此并不实用.Datafly 算法[25]在 MinGen 算法的基础上,引入抑制与启发式泛化指导原则对效率进行了提升.

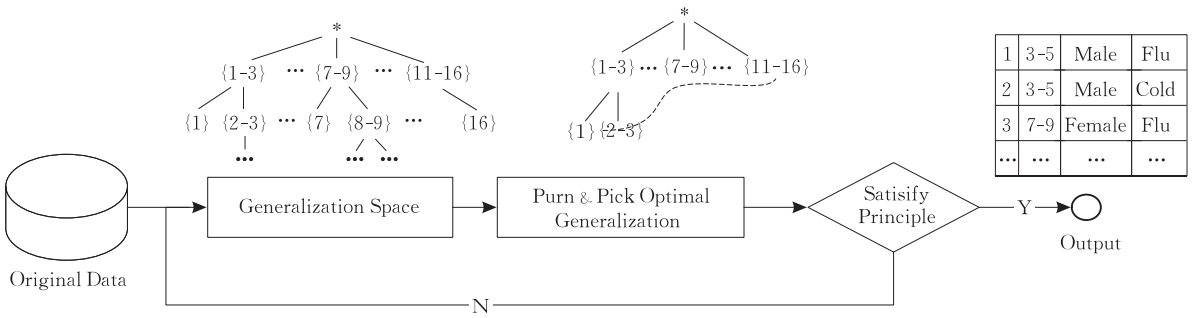

图 4 匿名算法流程

一种广泛应用的 k-匿名算法是 Incognito[30],它首先构建包含所有全域泛化(一种全局重编码技术)方案的泛化图(generalization graph),然后自底向上对原始数据进行泛化,每次选取最优泛化方案前,预先对泛化图进行修剪以缩小搜索范围,不断进行以上操作直到数据满足 k-匿名原则.其它优化的 k-匿名算法[67]基本上也是采用修剪泛化空间来提升性能.

多维 k-匿名算法[29]能够发布精度较高的数据,

① 递归(c,l)-diversity 的一种进化形式是"npd-递归(c_1,c_2,l)-diversity"[7],它还同时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 如采用抑制实现最优化的 k-匿名问题[28,44]、采用全局重编码技术和可识别度量的最优化 k-匿名问题[29]以及采用局部重编码和 MAXSIZE 度量的最优化 k-匿名问题[68].

它将原始数据映射到一个多维空间,k-匿名问题即转换为在空间中对多维数据进行最优化划分的问题.

实现其它匿名化原则的算法[7,26,65-66],大多是基于k-匿名算法,不同之处在于判断算法结束的条件,而泛化策略、对搜索空间的修剪等都是基本相同的.因此,本文将不再做具体介绍.

7.2.2　面向特定目标的匿名化算法

在特定的应用场景下,通用的匿名化算法可能不能满足特定目标的要求.因此需要设计具有针对性的匿名化算法.例如,考虑到数据应用者需要利用发布的匿名数据构建分类器,那么设计匿名化算法时就需要考虑在保护隐私的同时,怎样使发布的数据更有利于分类器的构建.并且采用的度量指标要能直接反映出对分类器构建的影响.已有的自底向上的匿名化算法[21]和自顶向下的匿名化算法[22]都采用了信息增益(information gain)作为度量.因为发布的数据信息丢失越少,构建的分类器的分类效果将越好.自底向上的匿名化算法通过每一次搜索泛化空间,采用使信息丢失最少的泛化方案进行泛化,重复执行以上操作直到数据满足匿名原则的要求.自顶向下的匿名化算法的操作过程与之相反.

类似地,针对以发布数据利用率最大化为特定目标的应用,文献[31]提出了 Anonymized Marginals 信息发布方法;针对以防止关联规则推导为首要目标的应用,需要采用抑制,不发布能最大化降低关联规则支持度和置信度的属性值,从而破坏关联规则推导攻击[68-69];当发布的信息是多个视图时,文献[62]提出了保证发布的信息满足k-匿名原则的算法.

7.2.3　基于聚类的匿名化算法

基于聚类的匿名化算法将原始记录映射到特定的度量空间中,再对空间中的点进行聚类来实现数据匿名.类似于k-匿名,算法保证每个聚类中至少有k个数据点.

根据度量的不同,文献[70]提出了r-gather 和r-cellular 这两种聚类算法.以r-gather 算法为例,它以所有聚类中的最大半径为度量,需要达到的目标是:对所有数据点进行聚类,在保证每个聚类至少包含k个数据点的同时,也使所有聚类中的最大半径越小越好.表 5 是采用 2-gather 算法对表 4 原始数据聚类后发布的结果.由于发布的结果只包含聚类中心、半径以及相关的敏感属性值,同一个等价类中的记录不可区分,因此对个人的敏感信息实现了隐藏.

表 4　原始数据

年龄	地址	疾病
a	b	胃炎
$a+2$	b	消化不良
c	$d+3$	感冒
c	d	肺炎
c	$d-3$	感冒

表 5　聚类后的数据

年龄	地址	记录数	疾病
$a+1$	b	2	胃炎 消化不良
c	d	3	感冒 肺炎 感冒

基于聚类的匿名化算法主要面临两个挑战:(1)怎样对原始数据的不同属性进行加权?因为对属性的度量越准确,那么聚类的效果就越好;(2)怎样将不同性质的属性统一映射到同一个度量空间中[71].

7.3　动态环境下的数据匿名化算法

前面所提到数据匿名化算法,都是针对静态数据而言,未考虑数据动态变化时带来的挑战.在动态环境下,数据通常会随时间的推移增加或减少,数据发布要求亦会不相同.在动态环境下直接应用基于静态数据的匿名化算法,虽然在某一时刻发布的匿名化数据能很好地保护隐私,但攻击者通过利用多个时刻发布的数据进行联合攻击,很容易披露敏感信息.

考虑到现实生活中很多情况是数据不断地增加(如医院所拥有的病例信息),文献[72]提出并解决了基于动态递增数据的多次发布问题.假设原始数据为T,关于T的一系列增量更新为$\Delta T_1,\Delta T_2,\cdots$.令根据$T$与前$i$次的增量更新发布的数据$T_i^*$为$T_i^*=f_i(T\cup\Delta T_1\cup\cdots\cup\Delta T_i)$,其中$f_i$为匿名算法.一系列发布$T_1^*,T_2^*,\cdots$对数据实现了$k$-匿名隐藏,当满足:

(1)T^*是k-匿名化的:$T^*=f(T)$;

(2)$\forall i\geq 1,T_i^*$是k-匿名化的;

(3)对每个非空整数集$\{i_1,i_2,\cdots,i_n\}$,推导表[73]$I(f_{i_1}(T_1),f_{i_2}(T_2),\cdots,f_{i_n}(T_n))$亦是$k$-匿名化的.

问题的复杂性在于:不仅要保证每一次单独发布数据的匿名化,而且要保证即使通过联合多次发布的数据进行攻击,隐私仍然能够得到保护.文献[42]提出了另一种基于"攻击检测与防止"的方法:

首先对当前的数据进行匿名化处理,然后再检测是否有攻击联合先前发布的数据而披露隐私;直到没有攻击能够披露数据隐私时,则停止对数据的进一步匿名化.

只有增量更新的数据集被称为"准动态"数据集,同时有数据增加和减少的数据集则称为动态数据集.文献[9]提出了一种在动态环境下保护隐私的匿名化原则 m-Invariance.假设 $T^*(1),T^*(2),\cdots,T^*(n)$ 是在动态环境下先后发布的一系列数据,我们称这一系列发布的数据满足 m-Invariance 匿名化原则,当且仅当同时满足:

(1)对 i 时刻发布的数据 $T^*(i)$,其每一个等价类中都至少有 m 条记录且这些记录都有不同敏感属性值;

(2)如果某条记录出现在不同时刻的多次发布中,那么每一次发布这条记录所在的等价类包含的敏感属性值形成的集合须相等.

条件(1)保证了每个时刻发布的数据的隐私披露风险不会高于 $1/m$,同时两个条件联合起来保证攻击者利用多次发布的数据进行攻击时,不会披露新增加和已经减少的数据的隐私.一种满足 m-Invariance 匿名化原则的数据发布算法是:① 首先将前后两次共有的数据分配到包含相同的敏感属性值集合的等价类中;② 然后尝试将新增加的记录分配到这些等价类中,同时保证剩下的未分配的数据满足可以形成条件(1)的等价类;③ 为剩下未分配的数据建立新的等价类;最后对过大、可以分裂的等价类进行调整.

除了数据的插入、删除会引起数据的动态变化外,每条记录属性值的更新,同样会导致数据动态变化.文献[73]首先对包含 QI 属性值和敏感属性值更新的动态数据匿名发布问题进行了研究.其假设敏感属性值由永远不变和随机动态变化两种组成:对后者而言,由于其随时间是随机变化的,对其进行多次发布不会带来新的威胁;而对始终不变(permanent)的敏感属性值而言,如果在多次发布中不考虑它不变的特点,将变得不再安全.因此,该问题的关键在于如何实现不变敏感属性值的匿名发布.文中提出了一种针对此种情况的匿名化原则 l-scarcity,并通过基于角色的划分(role-based partition)和基于 Cohort 的划分(cohort-based partition)两种技术来实现满足 l-scarcity 原则的发布.

7.4 小 结

数据匿名化由于能处理多种类型的数据,并发布真实的数据,能满足众多实际应用的需求,因此受到广泛关注.图 5 示例了数据匿名化场景及相关因素.可以看到,数据匿名化是一个复杂的过程,需要同时权衡原始数据、匿名数据、背景知识、匿名化技术、攻击等众多因素.

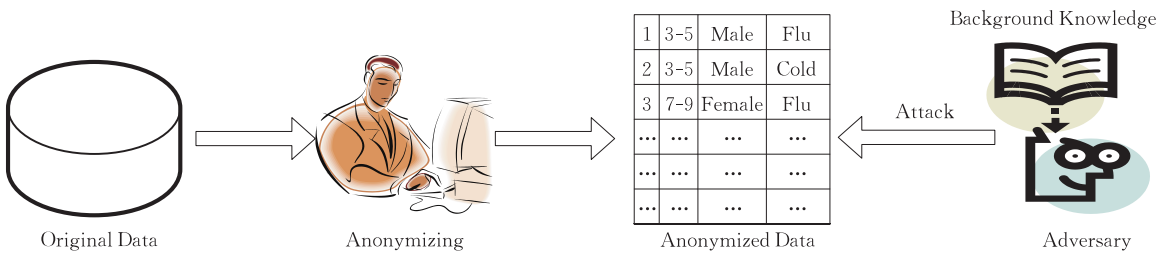

图 5 数据匿名化场景

8 总结与展望

隐私保护技术在诸多领域都有广泛的应用,是近年来学术界新兴的研究课题.本文侧重数据库应用,对隐私保护技术的研究现状进行综述.首先给出了隐私及其度量的定义,然后在对已有的隐私保护技术进行分类的基础上,介绍了基于失真、加密和匿名化的三大类隐私保护技术,特别是,对当前隐私保护领域的研究热点"基于数据匿名化的隐私技术"进行了比较详尽的阐述与分析.

容易看出,每类隐私保护技术都有不同的特点,在不同应用需求下,它们的适用范围、性能表现等不尽相同.从表 6 可以看出,当针对特定数据实现隐私保护且对计算开销要求比较高时,基于数据失真的隐私保护技术更加适合;当更关注于对隐私的保护甚至要求实现完美保护时,则应该考虑基于数据加密的隐私保护技术,但代价是较高的计算开销(在分布式环境下,还会增加通信开销).而数据匿名化技术在各方面都比较平衡:能以较低的计算开销和信息缺损实现对隐私保护.表 7 对隐私保护技术进行了进一步的对比分析.

表 6　隐私保护技术的性能评估

	隐私保护度	计算开销	数据缺损	数据依赖性	通信开销
基于数据失真的隐私保护技术	中	低	高	高	低
基于数据加密的隐私保护技术	高	高	低	低	高
数据匿名化	高	中	中	低	低

表 7　隐私保护技术的对比分析

	主要优点	主要缺点	代表技术	典型应用
基于数据失真的隐私保护技术	计算开销小 实现简单	数据失真 严重依赖于数据,不同数据需设计不同的算法	随机扰动 随机化回答 阻塞 凝聚	各种数据挖掘操作,如 ・关联规则挖掘 ・关联规则隐藏 ・决策树分类器构建等
基于数据加密的隐私保护技术	数据真实、无缺损 高隐私保护度	计算开销、通信开销大 部署复杂,实际应用难度较高	SMC 分布式下实现隐私保护的关联规则挖掘算法[20]、数据匿名化算法[55]等	分布式下的各种数据挖掘与发布操作,如 ・分布式关联规则挖掘 ・分布式数据匿名发布 ・分布式聚类 ・分布式安全计算等
数据匿名化	适用于各类数据、众多应用,算法通用性高 能保证发布数据的真实性 实现简单	存在一定程度的数据缺损 存在一定程度的隐私泄露 实现最优化的数据匿名开销较大	匿名化原则: ・k-匿名 ・l-diversity ・m-invariance 匿名化算法: ・Mondrian ・Incognito ・r-cellular	发布匿名化数据.基于发布的数据可进行各类数据挖掘操作,如 ・关联规则挖掘 ・决策树分类器构建等 ・聚类等

随着信息不断膨胀、信息获取渠道越来越多样化,数据库的应用无处不在,不论是在理论研究还是实际应用领域,对隐私保护技术进行研究都具有非常重要的意义.但由于隐私保护技术涉及多学科交叉且发展时间较短,还存在许多问题有待进一步研究:

(1)分布式和 Web 环境下的隐私保护研究

随着分布式数据库以及 Web 应用的发展和普及,众多已有的针对集中式数据库应用的隐私保护技术不能满足分布式环境下的新需求.由于分布式环境下各站点相对独立、数据异构的特点,通信、数据协同等其它操作将更加频繁.而这些操作,有意或无意间,都对敏感数据和隐私信息构成了威胁.

如何在分布式情况下,实现多点高效协同工作的同时,保证频繁的信息交互、数据传输行为过程中,不会给隐私信息、敏感数据带来威胁? 如何在保护各独立站点私有隐私的同时,还实现对整个分布式系统的共同隐私的保护?如何使得隐私保护策略或算法在有效的同时,对分布式查询、存储以及网络拓扑结构的负面影响尽量小?分布式数据库和 Web 具有巨大的潜力和广阔的应用前景,虽然在分布式环境下进行隐私保护的相关研究,将面临一系列新的问题和挑战,但相关问题的解决,将无疑对各种应用起到巨大的推动作用.

(2)特定应用背景下专有隐私保护技术的研究

虽然数据库在所有领域都有广泛的应用,但是不同领域的应用场景却千差万别.不仅数据的表现形式、存储方式、数量、更新频率等都各不相同,而且隐私信息的表现形式、数量往往也是不同的.因此,众多领域和现实应用,都急切需要一种符合其实际情况和特点、针对性强、效率/效果优的隐私保护方法.

以交通监管和定位服务(LBS)应用为例,由于面对的设备(汽车、手机等)种类、数量都很多,还会频繁的移动,那么应用将面对海量、频繁更新的位置数据,并且这些数据常常是非连续、有缺失的.在统计相关信息或为用户提供查询服务时,如何使得这个过程不会暴露单个设备及其所有人的隐私信息,使得返回给用户的查询结果更加有效,都是潜在的研究课题.

(3)基于动态数据的隐私保护技术研究

大部分现有隐私保护技术都是基于静态数据集的,而现实世界中,数据库中的数据却是无时无刻不在变化,包括数据表现形式的改变、属性的增减、新数据的加入、旧数据的删除等.并且,数据库数据的这种变化,一般都不是完全随机、独立的,数据与数

据之间,数据与数据变化之间,都是相互关联的.因此,怎样在这种更加复杂的环境下同时实现对动态数据的利用和隐私保护,是一个更具挑战的难题.

以针对动态数据集的匿名发布为例,虽然已有解决方案能处理数据具有动态插入和删除的情况[9],但是是在假设"各个插入、删除数据之前完全独立相关的"的前提下进行的.而在现实生活中,该假设往往不成立.如从足球队数据库中删除一条后卫球员的数据(表示该球员被解雇),那么紧接着插入的新数据将很有可能也是一名后卫球员的数据(表示聘请一名新球员).在这种情况下,利用数据行为之间的关系,文献[9]提出的解决方法将不能保证敏感数据的安全,而需要提出一种崭新的、更严格的匿名发布方法.

致 谢 在此,作者向对本文的工作给予支持和建议的同行表示感谢!

参 考 文 献

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

[2] Agrawal D, Aggarwal C C. On the design and auantification of privacy preserving data mining algorithms//Proceedings of the Symposium on Principles of Database Systems (PODS). Santa Barbara, California, USA, 2001: 247-255

[3] Verykios V S, Bertino E, Fovino I N, Provenza I N, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, 2004, 3(1): 50-57

[4] Agrawal R, Srikant R. Privacy preserving data mining//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Dallas, Texas, 2000: 439-450

[5] Clifton C, Kantarcioglu M, Vaidya J. Defining privacy for data mining//Proceedings of the National Science Foundation Workshop on Next Generation Data Mining. Baltimore, MD, USA, 2002: 126-133

[6] Zhang N, Zhao W. Distributed privacy preserving information sharing//Proceedings of the 31st Very Large Data Bases (VLDB) Conference. Trondheim, Norway, 2005: 889-900

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

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

[9] Xiao X, Tao Y. *m*-Invariance: Towards privacy preserving re-publication of dynamic datasets//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Beijing, China, 2007: 689-700

[10] Directive 95/46/ec of the European parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities, 1995, No I. (281): 31-50

[11] Deutsch A, Papakonstantinou Y. Privacy in database publishing//Proceedings of the 10th International Conference on Database Theory (ICDT). Edinburgh, Scotland, 2005: 230-245

[12] Miklau G, Suciu D. A formal analysis of information disclosure in data exchange//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Maison de la Chimie, Paris, France, 2004: 575-586

[13] Machanavajjhala A, Gehrke J. On the efficiency of checking perfect privacy//Proceedings of the Symposium on Principles of Database Systems (PODS). Chicago, Illinois, USA, 2006: 163-172

[14] Adam N, Wortmann J C. Security-control methods for statistical databases: A comparison study. ACM Computing Surveys, 1989, 21(4): 515-556

[15] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. The American Statistical Association, 1965, 60(309): 63-69

[16] Fienberg S E, McIntyre J. Data swapping: Variations on a theme by Dalenius and Reiss//Proceedings of the Privacy in Statistical Databases (PSD). Barcelona, Spain, 2004: 14-29

[17] Pinkas B. Cryptographic techniques for privacy preserving data mining. ACM SIGKDD Explorations, 2002, 4(2): 12-19

[18] Evfimievski A, Srikant R, Agrawal A, Gehrke J. Privacy preserving mining of association rules//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Madison, Wisconsin, 2002: 217-228

[19] Vaidya J S, Clifton C. Privacy preserving association rule mining in vertically partitioned data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Edmonton, Alberta, Canada, 2002: 639-644

[20] Kantarcioglu M, Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1026-1037

[21] Wang K, Yu P S, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection//Proceedings of the IEEE International Conference on Data Mining (ICDM). Brighton, UK, 2004: 249-256

[22] Fung B C M, Wang K, Yu P S. Top-down specialization for information and privacy preservation//Proceedings of the 21st International Conference on Data Engineering (ICDE). Tokyo, Japan, 2005: 205-216

[23] Vaidya J, Clifton C. Privacy preserving *K*-means clustering over vertically partitioned data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Washington, DC, USA, 2003: 206-215

[24] Sweeney L. *k*-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

[25] 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

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

[27] Aggarwal C C. On *k*-anonymity and the curse of dimensionality//Proceedings of the 31st Very Large Data Bases (VLDB) Conference. Trondheim, Norway, 2005: 901-909

[28] Meyerson A, Williams R. On the complexity of optimal *k*-anonymity//Proceedings of the Symposium on Principles of Database Systems (PODS). Paris, France, 2004: 223-228

[29] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional *k*-anonymity//Proceedings of the 22nd International Conference on Data Engineering (ICDE). Atlanta, Georgia, USA, 2006: 25-35

[30] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full domain *k*-anonymity//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Baltimore, Maryland, 2005: 49-60

[31] Kifer D, Gehrke J. Injecting utility into anonymized datasets//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Atlanta, Georgia, USA, 2006: 217-228

[32] Ge Wei-Ping, Wang Wei, Zhou Hao-Feng, Shi Bo-Le. Privacy preserving classification mining. Journal of Computer Research and Development, 2006, 43(1): 39-45(in Chinese)
(葛伟平, 汪卫, 周皓峰, 施伯乐. 基于隐私保护的分类挖掘. 计算机研究与发展, 2006, 43(1): 39-45)

[33] Zhang Peng, Tong Yun-Hai, Tang Shi-Wei, Yang Dong-Qing, Ma Xiu-Li. An effective method for privacy preserving association rule mining. Journal of Software, 2006, 17(8): 1764-1774(in Chinese)
(张鹏, 童云海, 唐世渭, 杨冬青, 马秀莉. 一种有效的隐私保护关联规则挖掘方法. 软件学报, 2006, 17(8): 1764-1774)

[34] Luo Yong-Long, Huang Liu-Sheng, Jing Wei-Wei, Yao Yi-Feng, Chen Guo-Liang. An algorithm for privacy-preserving Boolean association rule mining. Acta Electronica Sinica, 2005, 33(5): 900-903(in Chinese with English abstract)
(罗永龙, 黄刘生, 荆巍巍, 姚亦飞, 陈国良. 一个保护私有信息的布尔关联规则挖掘算法. 电子学报, 2005, 33(5): 900-903)

[35] Zhang Feng, Chan Hui-You. Research on privacy preserving collaborative filtering recommendation based on distributed data. Chinese Journal of Computers, 2006, 29(8): 1487-1495(in Chinese)
(张锋, 常会友. 基于分布式数据的隐私保持协同过滤推荐研究. 计算机学报, 2006, 29(8): 1487-1495)

[36] Qiang Wei-Zhong, Zou De-Qing, Jin Hai. Research on privacy preservation mechanism for credentials and policies in grid computing environment. Journal of Computer Research and Development, 2007, 44(1): 11-19(in Chinese)
(羌卫中, 邹德清, 金海. 网格环境中证书和策略的隐私保护机制研究. 计算机研究与发展, 2007, 44(1): 11-19)

[37] Yang Xiao-Chun, Liu Xiang-Yu, Wang Bin, Yu Ge. *K*-anonymization approaches for supporting multiple constraints. Journal of Software, 2006, 17(5): 1222-1231(in Chinese)
(杨晓春, 刘向宇, 王斌, 于戈. 支持多约束的 *K*-匿名化方法. 软件学报, 2006, 17(5): 1222-1231)

[38] Yao A C. How to generate and exchange secrets//Proceedings of the 27th IEEE Symposium on Foundations of Computer Science (FOCS). Toronto, Canada, 1986: 162-167

[39] Clifton C, Kantarcioglou M, Lin X, Zhu M Y. Tools for privacy preserving distributed data mining. ACM SIGKDD Explorations, 2002, 4(2): 28-34

[40] Merugu S, Ghosh J. Privacy-preserving distributed clustering using generative models//Proceedings of the IEEE International Conference on Data Mining (ICDM). Melbourne, Florida, USA, 2003: 211-218

[41] Agrawal R, Srikant R, Thomas D. Privacy preserving OLAP//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Baltimore, Maryland, 2005: 251-262

[42] Byun J, Sohn Y, Bertino E, Li N. Secure anonymization for incremental datasets//Proceedings of the 3rd VLDB Workshop on Secure Data Management (SDM). Seoul, Korea, 2006: 48-63

[43] Xu J, Wang W, Pei J, Wang X, Shi B, Fu A W. Utility-based anonymization using local recoding//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, PA, USA, 2006: 785-790

[44] Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A. Anonymizing tables//Proceedings of the 10th International Conference on Database Theory (ICDT). Edinburgh, Scotland, 2005: 246-258

[45] Chang L W, Moskowitz I S. An integrated framework for database inference and privacy protection//Proceedings of the 4th Annual IFIP WG 11.3 Working Conference on Database Security. Schoorl, The Netherlands, 2000: 161-172

[46] Saygin Y, Verykios V S, Elmagarmid A K. Privacy preserving association rule mining//Proceedings of the 12th International Workshop on Research Issues in Data Engineering (RIDE). San Jose, USA, 2002: 151-158

[47] Moskowitz I S, Chang L W. A decision theoretical based system for information downgrading//Proceedings of the 5th Joint Conference on Information Sciences (JCIS). Atlantic City, NJ USA, 2000: 82-89

[48] Aggarwal C C, Yu P S. A condensation approach to privacy preserving data mining//Proceedings of the 9th International Conference on Extending Database Technology (EDBT). Heraklion, Greece, 2004: 183-199

[49] Du W, Zhan Z. Using randomized response techniques for privacy-preserving data mining//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Washington, DC, USA, 2003: 505-510

[50] Kargupta H, Datta S, Wang Q, Sivakumar K. On the privacy preserving properties of random data perturbation techniques//Proceedings of the IEEE International Conference on Data Mining (ICDM). Melbourne, Florida, 2003: 99-106

[51] Rizvi S, Haritsa J R. Maintaining data privacy in association rule mining//Proceedings of the 28th Very Large Data Bases (VLDB) Conference. Hong Kong, China, 2002: 682-693

[52] Malin B, Airoldi E, Edoho-Eket S, Li Y. Configurable security protocols for multi-party data analysis with malicious participants//Proceedings of the 21st International Conference on Data Engineering (ICDE). Tokyo, Japan, 2005: 533-544

[53] Goethals B, Laur S, Lipmaa H, Mielikäinen T. On private scalar product computation for privacy-preserving data mining//Proceedings of the 7th Annual International Conference in Information Security and Cryptology (ICISC). Seoul Korea, 2004: 104-120

[54] Jiang W, Clifton C. A secure distributed framework for achieving k-anonymity. The International Journal on Very Large Data Bases, 2006, 15(4): 316-333

[55] Zhong S, Yang Z, Wright R N. Privacy-enhancing k-anonymization of customer data//Proceedings of the Symposium on Principles of Database Systems (PODS). Baltimore, Maryland, USA, 2005: 139-147

[56] Gilburd B, Schuster A, Wolff R. K-TTP: A new privacy model for large-scale distributed environments//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Seattle, WA, USA, 2004: 563-568

[57] Jagannathan G, Pillaipakkamnatt K, Wright R N. A new privacy-preserving distributed k-clustering algorithm//Proceedings of the 2006 SIAM International Conference on Data Mining (SDM). Bethesda, Maryland, 2006: 492-496

[58] Jagannathan G, Wright R N. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Chicago, IL, USA, 2005: 593-599

[59] Du W, Zhan Z. Building decision tree classifier on private data//Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining. Maebashi City, Japan, 2002: 1-8

[60] Kardes O, Ryger R S, Wright R N. Implementing privacy-preserving Bayesian-net discovery for vertically partitioned data//Proceedings of the ICDM Workshop on Privacy and Security Aspects of Data Mining. Los Alamitos, 2005: 26-34

[61] Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Privacy: Theory meets practice on the map//Proceedings of the 24st International Conference on Data Engineering (ICDE). Cancun, Mexico, 2008: 277-286

[62] Yao C, Wang X S, Jajodia S. Checking for k-anonymity violation by views//Proceedings of the 31st Very Large Data Bases (VLDB) Conference. Trondheim, Norway, 2005: 910-921

[63] Xiao X, Tao Y. Dynamic anonymization: Accurate statistical analysis with privacy preservation//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Vancouver, BC, Canada, 2008: 107-120

[64] Dalvi N, Suciu D. Answering queries from statistics and probabilistic views//Proceedings of the 31st Very Large Data Bases (VLDB) Conference. Trondheim, Norway, 2005: 805-816

[65] Raymond Chi-Wing Wong, Li J, Ada Wai-Chee Fu, Wang K. (α, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, PA, USA, 2006: 754-759

[66] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation//Proceedings of the 32nd Very Large Data Bases (VLDB) Conference. Seoul, Korea, 2006: 139-150

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

[68] Du Y, Xia T, Tao Y, Zhang D, Zhu F. On multidimensional k-anonymity with local recoding generalization//Proceedings of the 23rd International Conference on Data Engineering (ICDE). Istanbul, Turkey, 2007: 1422-1424

[69] Wang K, Fung B C M, Yu P S. Handicapping attacker's confidence: An alternative to k-anonymization. Knowledge and Information Systems: An International Journal (KAIS), 2006, 11(3): 345-368

[70] 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 Systems (PODS). Chicago, Illinois, USA, 2006: 153-162

[71] Li J, Wong R C W, Fu A W C, Pei J. Achieving *k*-anonymity by clustering in attribute hierarchical structures//Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK). Krakow, Poland, 2006: 405-416

[72] 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

[73] Bu Y, Fu A, Wong R C W, Chen L, Li J. Privacy preserving serial data publishing by role composition//Proceedings of the 34th Very Large Data Bases (VLDB) Conference. Auckland, New Zealand, 2008

ZHOU Shui-Geng, born in 1966, professor, Ph. D. supervisor. His research interests include database, P2P computing, bioinformatics etc.

LI Feng, born in 1983, master candidate. His research interests are database and privacy preservation.

TAO Yu-Fei, born in 1978, assistant professor. His research interests include temporal databases, spatial databases, and privacy preservation.

XIAO Xiao-Kui, born in 1981, postdoctoral researcher. His research interest is privacy preservation.

Background

Data publishing and data mining are two important areas of database applications. On one hand, data mining aims at extracting potential valuable knowledge (patterns or rules) out of the massive data. However, the traditional knowledge discovery technologies always accompany with the threat to data privacy and security. For example, Centre of Disease Control (CDC) needs to collect patients' information for the purpose of disease control. The sensitive data, which the owner hopes to keep secret for long, will be exposed inevitably if adopting traditional data mining technologies. On the other hand, data publishing is to release data directly from database to the public, which may lead to the disclosure of sensitive information and harm to its owners if no protection measure is taken. This situation makes preserving data privacy in various database applications a hot research topic in database community.

Privacy preservation techniques are proposed to tackle the threats to sensitive information. Specifically, two main issues should be considered: (1) how to preserve privacy in the data application; (2) how to enhance the data utility in the application.

This paper surveys the cutting-edge privacy preserving technologies and demonstrates its applications in database field. Not only the traditional privacy solutions (such as perturbation, randomization, encryption etc.) are discussed, but also the recent advances, e. g. anonymization techniques, are presented in detail.