

面向数据发布和挖掘的隐私保护研究进展

王姣¹, 范科峰², 王勇³

(1. 桂林电子科技大学电子工程与自动化学院, 广西 桂林 541004;

2. 中国电子技术标准化研究院, 北京 100007;

3. 桂林电子科技大学计算机科学与工程学院, 广西 桂林 541004)

摘要: 随着计算机技术的迅速发展, 数据越来越多, 为了从这些大量数据中获取有用信息, 需要对其进行挖掘, 然而, 在此过程中不免会造成数据相关者隐私的泄露, 如何提高数据的安全性、保护有用信息不被外泄变得尤为重要。分析了在数据发布和挖掘过程中若干现有数据隐私保护技术的方法, 简述了 JTC1 制定的隐私保护相关国际标准, 并根据其不同应用领域提出了未来可能的研究方向。为信息安全领域相关的人员提供了一定参考基础。

关键词: 隐私保护; 大数据; 数据挖掘; 标准

中图分类号: TP309

文献标识码: A

doi:10.11959/j.issn.2096-109x.2016.00021

Progress of research on privacy protection for data publication and data mining

WANG Jiao¹, FAN Ke-feng², WANG Yong³

(1. School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China;

2. China Electronics Standardization Institute, Beijing 100007, China;

3. School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: With the rapid development of the computer technology, there are more and more data in the society. In order to acquire knowledge from the large amounts of data, collecting and data mining is necessary. However, the privacy information will inevitably be disclosed during the process. So it is particularly important to improve the security of data and protect the useful data to avoid being disclosed. Several methods of data privacy preserving technology were analyzed when data was processed and briefly discussed the international standards which were made by JTC1 about privacy protection. According to its different application fields, the possible future research directions was proposed. Certain reference foundation could be provided for people who were in the field of information security.

Key words: privacy protection, big data, data mining, standard

1 引言

随着社会的进步和信息通信技术的迅猛发展, 数据量越来越多, Google 公司每月处理的数

据量超过 400 PB; 百度每天大约要处理几十 PB 数据; Facebook 每天生成 300 TB 以上的日志数据^[1]。在信息化时代的今天, 数据除了呈现上述海量性之外, 类型也变得繁多起来, 以 Web2.0

收稿日期: 2015-10-27; 修回日期: 2016-01-08。通信作者: 范科峰, kefengfan@163.com

基金项目: 国家自然科学基金资助项目 (No.61172053)

Foundation Item: The National Natural Science Foundation of China (No.61172053)

技术为基础的新型社交网络,以及云计算、物联网的兴起,使得越来越多的数据呈现半结构化,甚至非结构化特性,信息社会已然步入大数据(big data)^[2]时代,大数据时代的数据存在多源异构、分布广泛、动态增长等特点^[3],这些数据价值量大,但价值密度低,在对其进行有效分析过程中,在得到想要结果时,人们普遍将关注点集中在如何保证自己的信息不被泄露上。

数据大多涉及到个人隐私,如病人的病情、用户的信用卡收支记录、顾客的花费记录等,通常由于某些原因,人们并不希望自己的信息被他人知晓。在保护隐私的前提下得到数据分析的有用结果变得至关重要,然而这不仅需要技术的不断进步,还需要法律法规和相关标准的完善。这对很多企业来说既是机遇也是挑战,他们投入巨大的资金希望得到更及时和有用的信息来满足增长和盈利需求,例如,作为大数据分析的探路者和领导者,IBM自2005年至2012年,投资160亿美元进行了30次与大数据有关的收购^[1],在2014年初,IBM又投入10亿美元组建独立的Watson部门,率先于业界开展前瞻认知计算实践^[4]。

2 隐私保护关键技术

隐私是伴随着人类社会的形成而产生的,对于不同国家、地域和对象,概念也会有所不同。因此,隐私权也作为一项有关隐私的基本人权,逐渐在各个国家的法律和相关政策条款中出现。最早涉及隐私权的法律政策文件源于1890年Warren和Brandeis的“The right to privacy”^[5],其中提出“隐私权是个人独处的权利,此权利是宪法规定的人所共享的自由权利的重要组成部分”。随着信息技术不断推动人类社会的发展,一些数据信息被悄无声息地保存在不同的地方,并且被不正当地使用,进而产生了隐私和安全问题,人们对隐私保护的呼声越来越高,除了相关政策的出台,越来越多的人致力于技术层面的研究,从而产生了大量的方法,隐私保护技术的发展也逐渐趋于多元化。

多年前就有不少人致力于对隐私保护技术的研究。1989年Adam等提出了扰动方法^[6];2000年,Agrawal等提出了随机化方法^[7];2002年,

Clifton等提出了安全多方计算(SMC, Secure multi-party computation)技术^[8];2004年,Fienberg提出了交换方法^[9];随后基于博弈论^[10]的隐私保护方法的提出也为其注入了新鲜的血液。

目前,在数据发布过程中,对原始数据采用失真、匿名、加密等技术,以实现隐私保护;在数据挖掘过程中,针对关联规则、分类、聚类等,研究高效的隐私保护的挖掘算法来减少由挖掘所带来的隐私风险^[11]。基于以上两个层面,本节主要介绍数据隐私保护的关键技术。

2.1 失真技术

数据失真技术,就是对原始数据进行扰动,基本思想是隐藏真实数据,只呈现出数据的统计学特征^[12]。失真后的数据仍然保持原本的某些特性不变,但攻击者是不能根据发布的失真数据重构出真实的原始数据的。失真技术主要包括随机化、阻塞、变形、交换等,以此来隐藏关联规则。

2.1.1 随机化

随机化技术是在原始数据中加入随机噪声,从而保护敏感数据不被发现。例如,在原始数据中注入大量伪项,隐藏频繁项集。然而任意地对数据进行随机化,并不能保证数据和隐私的安全,文献[13]为此提供了一种基于随机矩阵的数据过滤技术。同时,文献[14]也提出了一种新的数据随机处理方法,即部分隐藏的随机化回答方法,此方法是将数据干扰和查询限制相结合对原始数据进行变换和隐藏。

随机化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,比如,在随机扰动后的数据上估计项集支持度,从而发现规则^[15];或者通过对随机干扰数据的重构,设计高效的分类挖掘算法,利用重构数据的分布进行决策树分类器训练,最终得到的决策树可以很好地对数据进行分类^[16]。

2.1.2 阻塞

阻塞对于原始数据的修改并不引入虚假的噪声数据,而是对其进行泛化模糊处理^[17]。基本原理是:将数据表中的某些特定数值换成“?”,使支持度或置信度处于某个区间范围内,当此区间范围的下界取值小于设定的阈值时,即可实现关联规则的隐藏。文献[18]提出了通过使用未知

值来代替部分敏感的原始数据,文献[19]也提出了针对特定的敏感规则对原始数据进行隐藏。阻塞虽然可以使一部分敏感信息得到很好的保护,但由于所提供的所有数据都是真实的原始数据,所以对整个数据集的隐私保护程度并不是很高。

2.1.3 变形

变形类似于阻塞,不同的是用布尔矩阵表示数据库中的数据,将敏感事务对应的数值进行取反操作,同时修改和过滤原有事务的属性,使敏感规则的支持度和置信度低于设定的阈值,从而达到关联规则的隐藏。

但是一个无法避免的问题是,对原始数据进行阻塞和变形之后,都需要重建数据的分布,它们必须针对不同的应用需要设计特定的算法来对转换后的数据进行处理。对此,文献[20]提出了一种凝聚技术,将原始数据记录分成组,每组有 k 条记录产生的统计信息,同一组内的 k 条记录是两两不可区分的,因此重构后的记录并不会泄露原始数据的隐私。

2.1.4 交换

交换是在记录之间交换数据值来平衡隐私和数据挖掘的一种技术^[21],其核心是:在原始数据中,交换不同记录的某些属性值,但前提必须要保证不改变其统计特征,最后发布交换后的数据,这样便能提高数据的不确定性。在文献[22]中,为了有效访问以加密形式存储的数据,使用3个独立的服务器来管理数据,不断对所访问的数据进行重写和重加密。文中数据交换意味着通过3个服务器间的信息交换来改变所访问的数据的物理地址,并阐明了如何将交换技术应用于3个服务器中来保护隐私。

2.2 匿名技术

数据匿名化的两种主要方法是抑制和泛化^[17],顾名思义,抑制是不发布某些数据项,泛化是对数据进行概括与抽象描述。对于数据匿名化,其研究的重难点在于,如何在既能保护隐私又具有较大价值的前提下设计更好的匿名化原则和算法。

2.2.1 k -匿名

k -匿名隐私保护模型由L.Sweeney在1998年首次正式提出^[23],他在文献[24]中提出了 k -匿名原则,即保证所发布数据集中的每一条记录与其

他 $k-1$ 条记录不能区分,因此数据挖掘者不能辨别出隐私信息所属的具体个体,从而起到隐私保护的作用, k 值越大,隐私保护效果越好,然而数据丢失也越为严重。 k -匿名数据是不确定数据中的一种,有效解决了链接攻击^[18]问题,但其主要是针对单一约束条件进行处理,而在实际应用中,会涉及到大量约束条件,对此,文献[25]提出了多约束 k -匿名方法Classfly⁺,此算法的核心是先将约束集划分为 M 个独立的约束子集,然后再将独立约束子集中的约束按照匿名度低优先原则进行排序,若独立约束子集含一个约束,则采用Classfly算法进行匿名化处理,若含多个约束,则采用多约束概括过滤进行匿名化处理。

2.2.2 l -多样性

l -多样化模型^[26]是对 k -匿名的扩展,此模型要求每个等价类的敏感属性至少有 l 个不同的值,增加了敏感值与所属个体的连接难度,防止了 k -匿名易受同质性攻击和重标志攻击的缺陷。然而不同个体对隐私保护有不同的需求,对此,文献[27]通过设置敏感属性的保护属性来实现个体与敏感值之间关联关系的个性化保护需求,提出了一种面向个体的个性化扩展 l -多样性隐私匿名模型,与此同时,为实现该模型,还提出了一种个性化扩展 l -多样性逆聚类(PELI-clustering)的算法,此算法首先从数据集中任意选取一个元组作为聚类质心,根据其敏感属性集得到该质心相应的匿聚类候选集,再形成满足扩展 l -多样性的匿聚类等价类,重新计算质心,并将离此质心最远的元组作为下一个聚类的质心,重复此过程直到全部元组归入相应的匿聚类等价类或不满足聚类的条件为止。

2.2.3 t -闭合

上述提到的 l -多样性存在两个问题,一是当 l 值较小,数据记录值过大时,等价类数量会相当庞大;二是对单个敏感属性而言,如果两个敏感属性值差异过大,很难确定敏感属性值的敏感度^[28]。对此,在 k -匿名和 l -多样性基础上,文献[29]提出了 t -闭合方法,此方法要求所有等价类的敏感属性值分布与该属性的总体分布差异小于 t ,文中给出了等价类和表满足 t -闭合的条件,若一个等价类的敏感属性值分布和整个表的属性值分布差异不超

过阈值 t , 则这个等价类符合 t -闭合条件, 若一个表中的所有等价类都符合 t -闭合, 则整个表符合 t -闭合。

2.3 加密技术

加密技术多用于分布式数据应用之中, 通过对原始数据进行加密以实现隐私保护。任何一种普通的计算都可转化为无可信第三方参与的安全多方计算 (SMC, secure multi-party computation) 的框架^[30], SMC 主要用于两个或多个互不信任的参与方之间进行隐私保护的协同计算。

2.3.1 安全多方计算

安全多方计算的概念最初是由 Yao 在 1982 年提出的^[31], 确保输入的独立性、计算的正确性, 同时不能向参与计算的其他成员泄露输入值。一个 SMC 模型主要由参与方、安全性定义、通信网模型、信息论安全与密码学安全 4 个方面组成, 其应用领域涉及到电子选举、投票、拍卖等。虽然安全多方计算相对来说比较安全和准确, 但涉及到的加密技术计算开销、通信开销也较高, 因此 SMC 是以牺牲费用为前提提高隐私保护度的。目前, 对于安全多方计算的研究主要集中于降低计算开销、优化分布式计算协议等^[12]。

2.3.2 同态加密技术

同态加密 (homomorphic encryption) 作为 SMC 的核心技术之一, 其概念最初由 Rivest 等在 1978 年提出^[32], 是一种允许直接对密文进行操作的加密变换技术^[33], 它既能实现基本的加密操作, 也能实现密文间的多种计算功能。同态加密算法包括能实现一种同态性的半同态加密算法和可以满足所有同态性质的全同态加密算法^[34], 不过由于全同态加密算法的计算复杂性, 目前还没得到广泛的应用。

满足乘法同态性的 RSA 算法^[35], 设 p, q 是两大素数, 由于大整数分解较为困难, 因此 $n = pq$ 难以在有限时间内进行分解。RSA 算法也存在一些问题。一是在公私钥生成之后, 同一个明文加密后的密文总相同, 这就对其安全性提出了挑战, 对此, 文献[36]提出了满足加法同态性的 Paillier 算法, 因同一明文两次加密会产生不同密文, 相对 RSA 算法提高了方案的安全性。二是倘若攻击者尝试所有可能的密钥进行蛮力攻击

或对大数因式分解进行数字攻击, RSA 算法的安全性也会受到威胁, 为此, 文献[37]提出了一种改进型的 RSA 算法 (MREA, modified RSA encryption algorithm), MREA 是一种非对称密钥密码体系, 公钥只用来加密, 私钥只用来解密, 因此通过加密签名是不能识别身份的。文献[38]详细介绍了两成员和多成员情况下同态加密技术的过程, 并分别对其正确性、复杂度和隐私性进行了分析。此外, 还出现了一类异或同态加密算法, 文献[39]采用概率编码方法和一个同态按位异或计算的密码系统, 构造了两种安全协议, 这不同于以往基于安全算术和的运算, 而是基于安全的按位异或运算。

同态加密技术的优点如下: 1) 可以先对多个密文进行计算后再解密, 减少计算代价; 2) 可以实现无密钥方对密文的计算, 密文计算无须经过密钥方, 减少通信代价; 3) 可以实现让解密方只能获知最后的结果, 而无法获得每一个密文的消息, 可以保证信息的安全性^[40]。近年来, 同态加密技术的突破性进展为云计算的安全保护提供了新的契机, 研究高效、实用的全同态加密方案^[41], 并将其应用到云计算服务上, 具有重要的现实意义。

2.3.3 数字信封技术

数字信封技术使用两层加密体系, 结合了对称加密和非对称加密的优点, 保障信息传输安全。对称加密, 即加密和解密的密钥相同; 非对称加密, 即加密密钥和解密密钥不同。数字信封技术过程如下:

1) 发送方用对称密钥加密信息, 并用接收方的公开密钥将此对称密钥加密 (这部分称为数字信封), 形成消息密文和密钥密文, 将二者发送给接收方;

2) 接收方用相应的私有密钥打开数字信封, 得到对称密钥, 然后用此对称密钥打开加密信息。

数字信封技术可满足数据交换的高保密性要求, 应用较为广泛。例如, 将同态加密技术和数字信封技术相结合, 并应用于数据挖掘决策树分类^[42]的隐私保护之中。

2.3.4 Shamir 秘密共享技术

秘密共享是一种将秘密分割存储的密码技术, 但其关键是如何设计更好的分割和恢复。

Shamir 秘密共享技术可以有效预防共谋攻击并且可以在不违背隐私保护的前提下进行多方计算^[43], 其基本思想是将一个密钥分解成 n 个部分, 只有知道了其中的至少 $k(k \leq n)$ 个部分才能恢复出原来的秘密信息。

假定 v_s 是隐私信息, P 是分配隐私信息的 $P_1, P_1, \dots, P_n$ 组成的集合, k 是重建隐私信息至少需要的股份数。Shamir 秘密共享算法简述如下:

- 1) 随机选择一个多项式 $q(x) = a_{k-1}x^{k-1} + \dots + a_1x^1 + v_s$, 其中 $a_{k-1} \neq 0$, $q(0) = v_s$;
- 2) 选择 m 个不同公开的随机数 $x_1, x_2, \dots, x_m$, 其中 $x_i \neq 0$, $m \geq n$;
- 3) 计算 $share_i = q(x_i)$;
- 4) 将 $share_i$ 发送给 P_i 。

因为多项式 $q(x)$ 中有 k 个未知量, 为了得到隐私信息 v_s , 至少需要构建 k 个方程。因此即便有 $k-1$ 个隐私分配者串通也不能得到这个隐私的任何信息。

文献[43]详细分析了此算法的正确性、复杂度和安全性。通过分析方程组的形式, 得知 k 个部分参与的计算即使有 $k-1$ 个部分勾结也不能计算出这个隐私信息, 因此该方法可以达到隐私保护的目的。秘密共享技术有诸多优点, 例如保证密钥的安全性和完整性, 防止权力过分集中被滥用, 增加系统的可靠性等。因此, 将其与数字签名、身份认证等技术结合可形成具有广泛应用价值的密码学算法和安全协议。

2.4 基于聚类算法的隐私保护技术

聚类是根据数据间不同和相似的特性, 将数据分成不同的类别, 最后使同一聚簇中的个体差别尽可能小, 而不同聚簇之间个体差异尽可能大。聚类是一个无监督的分类, 它没有任何先验知识可用^[44]。在此, 主要介绍基于 EM 和 K -means 聚类算法的隐私保护技术。

2.4.1 基于 EM 算法聚类的隐私保护

EM 算法 (expectation maximization algorithm), 即最大期望算法, 是一种迭代算法, 主要用于计算不完全数据的极大似然估计, 大大降低了极大似然估计的计算复杂度。EM 算法的每一步迭代中包括一个 E 步 (expectation step) 即

期望步和一个 M 步 (maximum likelihood step) 即极大似然步, 如此迭代下去, 直至满足某个收敛条件为止。由于 EM 算法收敛的优劣很大程度上取决于其初始参数, 因此如何初始化 EM 参数^[45,46]是一个关键的问题, 一般采用随机中心、层次聚类、 K -means 和 Binning 等方法。

虽然人们在不断地改进 EM 算法, 但基于隐私保护的安全聚类协议并不是很多。对于水平分布的数据, 文献[47]给出了一种 EM 混合模型下的安全算法, 基本思想是在每次迭代中, 每个参与者只从数据对象中生成一个局部模型, 并根据上次迭代结果计算全局信息, 然后将自己的局部模型和其他参与者的局部模型合并成全局模型。但这种方法至少需要 3 个参与方, 因为倘若只有两方的话, 便可以根据全局模型和自己的局部模型得到另一方的局部模型。针对这种情况, 文献[48]提出了一种只有两个参与者的基于 EM 聚类的隐私保护算法, 讨论了在高斯混合模型 (GMM, Gaussian mixture model) 里, 如何在不共享各自信息的同时安全计算高斯分布的期望 μ_i 、协方差矩阵 $\sum_i$ 、类 i 的概率 π_i 。

2.4.2 K -means 聚类的隐私保护算法

K -means 算法也是基于聚类算法中的一个典型算法, 同样也是一种迭代算法。其基本思想是找出 K 个聚类中心, 使每个数据点与其最近的聚类中心的平方距离和最小。基本过程如下:

- 1) 从 n 个点中随机选取 k 个点作为中心;
- 2) 分别测量其他每个点到 k 个中心的距离, 并将其归到最近的中心, 得到 k 个类;
- 3) 重新计算 k 个类的中心点;
- 4) 若新中心点和原中心点相同或小于提前设定的阈值, 则算法结束, 否则继续步骤 2) 和步骤 3)。

对于 K -means 聚类的隐私保护, 关键是对聚类均值的隐私保护, 但是在算法的每一步迭代中, 参与方是知道均值的^[49]。为了解决这一问题, 文献[48]提出了一个协议, 即在不揭露聚类均值的前提下允许每个参与方计算到聚类中心点的距离。对于垂直分布的数据聚类, 既要得到划分的效果, 又达到不能泄露各方对象属性的个数和各方类的平均值的隐私保护目的, 文献[50]在

K -means 算法基础上,结合安全多方计算和同态加密算法,提出了一种对于 K -means 聚类的隐私保护方案,假定有 r 个参与方, n 个公共实体,每个参与方对于同一实体集有不同的属性,最后的结果是每个参与方只知道对应于他们自己属性的均值以及实体的划分。

3 隐私保护相关标准

国际标准化组织/国际电工委员会 (ISO/IEC, International Organization for Standardization/International Electrotechnical Commission) 的第一联合技术委员会 (JTC1) 是一个信息技术领域的国际标准化委员会,它推进了国际信息技术标准化的进程。SC27 是 JTC1 下属的专门负责信息安全技术领域的分技术委员会,SC27 下设 5 个工作组,其中第五工作组 WG5 主要负责研究和制定身份管理与隐私保护领域的信息安全国际标准。鉴于隐私保护相关的标准涉及范围很广,本节主要介绍 SC27 WG5 制定的与隐私保护相关的国际标准。

3.1 ISO/IEC 29100《信息技术 安全技术 隐私框架》

ISO/IEC 29100《信息技术 安全技术 隐私框架》^[51]为信息与通信技术 (ICT, information and communication technology) 系统内个人可识别信息 (PII, personally identifiable information) 提供了一个高层次的框架。此框架定义了一个通用的隐私术语;介绍了处理 PII 过程中的成员和它们各自的角色;描述了隐私保护需考虑的事项;并且根据现有的隐私规则提供了一些参考规则。该标准提供的隐私框架可以作为制定其他隐私标准的基础。

3.2 ISO/IEC 27018《信息技术 安全技术 在公有云中 PII 处理者的 PII 实用规则》

ISO/IEC 27018 这一标准^[52]根据 ISO/IEC 29100 中的隐私规则建立了通用的可接受的控制对象、措施和指南,使得在公有云计算环境下保护 PII。该标准帮助公有云服务提供商作为 PII 处理者时履行适用的义务;使公有云 PII 处理者在相关方面浅显易懂,因此云服务消费者可以选择较好的基于云管理的 PII 处理服务;帮助云服务消

费者和公有云 PII 处理者达成合约共识;在物理和逻辑网络安全风险增高的情况下,为云服务消费者提供一个执行审计、合法权利和责任的机制。

3.3 ISO/IEC 29190《信息技术 安全技术 隐私保护能力评估模型》

ISO/IEC 29190 标准^[53]试图向组织提供一个关于如何评估其隐私保护能力水平的高层次指南。特别地,它规定了评定隐私能力的评估步骤;设定了隐私能力评估级别;在隐私能力评估的关键功能区域提供了指南;提供了执行评估过程的指南,并且提供了如何将隐私能力评估融入到组织运作中的指南。

3.4 ISO/IEC 29134《信息技术 安全技术 隐私影响评估 方法学》

ISO/IEC 29134 标准^[54]为隐私影响评估 (PIA, privacy impact assessment) 的进行提供了指导方针,并给出了一个隐私保护框架和具体的隐私影响评估方法,解释了如何管理在 PII 处理过程中产生的隐私风险。此外,该标准还描述了 PIA 报告的结构和内容,PIA 是一种用来评估在某个项目、技术和服务等方面隐私影响的工具,并与利益相关者协商采取补救措施来避免或减小不利影响。

本节主要介绍的是 SC27 WG5 制定的与隐私保护相关的国标。然而由于不同国家的管理机制和理念各不相同,不同实体涉及到的隐私方面的差异也较大,应根据我国实际情况,通过研究和跟进国际国外相应标准及其发展趋势,不断推进和深化我国的隐私保护标准。

4 展望

目前,数据量呈指数式增长,数据结构和数据格式日趋复杂化,安全与隐私已成为大数据时代各行各业关注的焦点。无论是在技术层面还是标准制定层面,都需要进行进一步研究和完善。

随着无线传感器网络、移动社交网络以及云计算相关应用的逐步展开,其安全问题也备受关注。如何将隐私保护技术和它们进一步结合,设计具有针对性且性能良好的算法,提高保护度和结果精确度,降低算法复杂度以及减少能量消耗等都需要进行深入研究。

在无线传感器网络中,研究主要集中于数据

聚集、查询和访问控制过程中的隐私保护,由于这是资源受限的分布式自组织多跳网络,前面介绍的隐私保护关键技术并不能直接应用于此网络。在隐私保护数据聚集,可以使用逐跳加密机制、端到端加密机制、非加密策略等。逐跳加密机制是聚集节点将子节点上传来的加密数据解密后进行聚类,然后加密上传给父节点,由于中间节点都需进行加解密,因此计算代价较高;而端到端加密机制,聚集节点不解密只加密的方法减少了加解密的计算代价;非加密策略就是在不加密的情况下,加入与真实数据不可区分的伪装数据以实现隐私保护,此方法可支持非线性聚类,但其隐私保护能力较弱。在隐私保护数据查询中,目前大多采用范围查询、Top- k 查询、基于类型的查询等方法:范围查询可通过桶模式和加密技术实现,再添加验证编码进行正确性和完整性验证;Top- k 查询是采用扰动和安全比较等技术实现隐私保护;在基于类型的查询中,传感器节点采集特定类型的数据,使用椭圆曲线多项式技术将敏感数据的类型和内容隐藏,应对共谋攻击。在隐私保护访问控制中,盲签名技术隐私保护度较强,但通信代价较高,环签名技术则与其相反。如何根据其特点,优化数据管理,设计保护协议,将隐私保护技术和传感器网络技术有效结合是这一领域需要进一步研究的方向。

在移动社交网络中,各种移动定位设备的涌现产生了大量的位置和轨迹数据,对其如何保护是迫切需要解决的问题,目前,社交网络隐私保护技术主要集中于基于 k -匿名、Markov 链、聚类、随机化等思想,轨迹隐私保护技术主要集中于假数据、泛化和抑制等,除了对发布的数据进行一定的处理外,还要考虑数据发布时间之间的联系,此外,即使用户可以控制自己发布的内容,也无法控制朋友发布涉及自己的内容,这也给相关人员带来了巨大的挑战,如何降低隐私泄露程度并且提高数据可用性成为了研究的重点。因此,需要设计多样化的社会网络隐私保护模型,目前已初步尝试将关系数据中的差分隐私应用到其中,不过由于大数据的规模和结点之间高度的相关性,可能导致数据差分隐私的复杂度较高。

在云计算相关应用中,各种大量资源都链接

在一起,形成一个巨大的虚拟资源共享池,它以便利、经济、高可扩展性等一系列优势吸引了越来越多的企业和公司,然而其安全问题是制约云计算发展的关键因素。近年来,研究者不断致力于对虚拟机安全、数据外包安全、可信计算环境等相关方面的研究,为保护用户数据的隐私,用户在对数据加密后交给云服务器存储,当用户进行查询时,也需对查询条件进行加密,这对云服务器的要求很高,必须能够根据加密的查询条件在加密的数据上进行查询,如何真正实现相关技术应用于云计算中,形成支撑云计算安全的技术体系,为用户提供安全可靠的保障是未来需要解决的实质性问题。目前,基于 ORAM 的可搜索加密技术能达到较高的安全保障,但需付出很大的计算代价;基于对称加密的可搜索技术是一种无交互密文搜索方法,但较易遭受统计攻击;较为符合云计算环境下隐私保护实际需求的方法是安全排名查询,此方法是系统根据某种准则进行查询将结果返回给用户,系统适用性较强,不过仍需进一步研究。

除了隐私保护相关领域技术层面的研究,还需要通过法律法规和标准对其进行规范。然而由于各国情况不同,实际应用和管理需求不同,必须结合实际情况进行法律法规和标准的制定。我国尚缺乏一部专门用于信息通信技术 (ICT, information communication technology) 系统的隐私保护法律,可参考国际国外的相关法规和标准,结合我国实际情况,尽快出台相关政策。

5 结束语

本文从数据发布和挖掘的角度出发,介绍了几种典型的隐私保护技术方法,以及 JTC1 制定的与隐私保护相关的标准,并分析了其未来可能的发展方向。总体上说,对于隐私保护的相关研究,还需要进一步努力,制定合理的政策法规,并在此基础上加强技术方面的探索,才能更好地让数据为我所用,使隐私更好地得以保护。

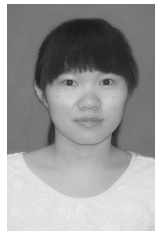
参考文献:

- [1] 李国杰,程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院

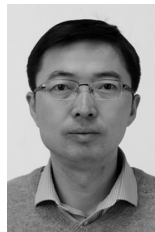
- 刊, 2012, 27(6).
- LI G J, CHENG X Q. Big data study: a major strategic area-the research status of big data and scientific reflection[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6).
- [2] CLIFFORD L. Big data[J]. Nature, 2008, 455(7209):1-136.
- [3] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1):146-169.
- MENG X F, CI X. Big data management: conception, technology and challenge[J]. Computer Research and Development, 2013, 50(1): 146-169.
- [4] 官思发, 朝乐门. 大数据时代信息分析的关键问题、挑战与对策[J]. 图书情报工作, 2015, 59(3):12-18.
- GUAN S F, CHAO L M. The problem, challenge and countermeasure of the big data information analysis[J]. Library and Information Service, 2015, 59(3):12-18.
- [5] WARREN S D, BRANDEIS L D. The right to privacy[J]. Harvard Law Review, 1973, 4(6):193-220.
- [6] ADAM N R, WORTMANN C. Security control methods for statistical databases: a comparative study[J]. ACM Computing Surveys, 1989, 21(4): 515-556.
- [7] AGRAWAL R, SRIKANT R. Privacy-preserving data mining[J]. Sigmod Record, 2000, 29(2):439-450.
- [8] CLIFTYON C, KANTARCIOGLU M, VAIDYA J, et al. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations, 2003, 4(2).
- [9] MCINTYRE S E, MCLNTYRE J. Data swapping: variations on a theme by dalenius and reiss[J]. Lecture Notes in Computer Science, 2004:14-29.
- [10] KARGUPTA H, DAS K, LIU K. Multi-party, privacy-preserving distributed data mining using a game theoretic framework[C]//The 11st European conference on Principles and Practice of Knowledge Discovery in Databases. Berlin Heidelberg: Springer-Verlag, c2007: 523-531.
- [11] XU L, JIANG C X, WANG J, et al. Information security in big data: privacy and data mining[J]. IEEE Access, 2014, 2:1-28.
- [12] 李晓晔, 孙振龙, 邓佳宾, 等. 隐私保护技术研究综述[J]. 计算机学报, 2013, 40:199-202.
- LI X H, SUN Z L, DENG J B, et al. Review of privacy protection[J]. Computer Science, 2013, 40: 199-202.
- [13] KARGUPTA H, DATTA S, WANG Q, et al. On the privacy preserving properties of random data perturbation techniques[C]//IEEE International Conference on Data Mining. c2003:99.
- [14] 张鹏, 童云海, 唐世渭, 等. 一种有效的隐私保护关联规则挖掘方法[J]. 软件学报, 2006, 17(8):1764-1774.
- ZHANG P, TONG Y H, TANG S W, et al. An effective method of digging privacy protection association rule[J]. Journal Software, 17(8): 1764-1774.
- [15] ELMAGARMID A K, VERYKIOS V S, SAYGIN Y. Privacy preserving association rule mining[C]// Twelfth International Workshop Research Issues in Data Engineering: Engineering E-Commerce/ E-Business Systems. c2002:151-158
- [16] AGRAWAL R, SRIKANT R. Privacy preserving data mining[J]. ACM Sigmod Record, 2000, 29(2): 439-450.
- [17] 张海涛, 黄慧慧, 徐亮, 等. 隐私保护数据挖掘研究进展[J]. 计算机应用研究, 2013, 30(12):3529-3535.
- ZHANG H T, HANG H H, XU L, et al. Progress of private protection data mining[J]. Application Research of Computers, 2013, 30(12): 3529-3535.
- [18] SAYGM Y I, VERYKIOS V S, CLIFTON C. Using unknowns to prevent discovery of association rules[J]. ACM Sigmod Record, 2001, 30(4):45-54.
- [19] OLIVEIRA S R M, ANE O. Privacy preserving frequent itemset mining[C]//IEEE International Conference on Privacy, Security and Data Mining. c2002:43-54.
- [20] AGGARWAL C C, YU P S, et al. A condensation approach to privacy preserving data mining[M]//9th International Conference on Extending Data Mining Technology, Heraklion, Crete. Berlin Heidelberg: Springer, c2004: 183-199.
- [21] ESTIVILL V, BRANKOVIC L. Data swapping: balancing privacy against precision in mining for logic rules[J]. Lecture Notes in Computer Science, 1999:389-398.
- [22] FORESTI S, PARABOSCHI S, PELOSI G, et al. Protecting access confidentiality with data distribution and swapping[C]//Big Data and Cloud Computing. c2014:167-174.
- [23] SAMARATI P, SWEENEY L. Protecting privacy when disclosing information: K -anonymity and its enforcement through generalization and suppression[C]//IEEE Symposium on Research in Security and Privacy, Chicago. c1998.
- [24] SWEENEY L. K -anonymity: a model for protecting privacy[J]. International Journal of Uncertainty Fuzziness & Knowledge Based Systems, 2002, 10(5):557-570.
- [25] 杨晓春, 刘向宇, 王斌, 等. 支持多约束的 K -匿名化方法[J]. 软件学报, 2006, 17(5):1222-1231.
- YANG X C, LIU X Y, WANG B, et al. K -anonymous method of multiple constraints supported[J]. Journal of Software, 2006, 17(5): 1222-1231.
- [26] MACHANAVAJJHALA A, GEHRKE J, KIFER D, et al. L -diversity: privacy beyond k -anonymity[C]//IEEE International Conference on Data Engineering. c2006:24.
- [27] 王波, 杨静. 一种基于逆聚类的个性化隐私匿名方法[J]. 电子学报, 2012, 40(5):883-890.
- WANG B, YANG J. An anonymity privacy method based on inverse clustering[J]. Acta Electronica Sinica, 2012, 40(5):883-890.
- [28] 刘英华, 杨炳儒, 马楠, 等. 分布式隐私保护数据挖掘研究[J]. 计算机应用研究, 2011, 28(10):3606-3610.
- LIU Y H, TANG B R, MA N, et al. Study of distributed privacy protection[J]. Application Research of Computers, 2011, 28(10): 3606-3610.
- [29] LI N, LI T, VENKATASUBRAMANIAN S. T -closeness: privacy beyond k -anonymity and l -diversity[C]//IEEE International Conference on Data Engineering. c2007:106-115.
- [30] 汤琳, 何丰. 隐私保护的数据挖掘方法的研究[J]. 计算机技术与发展, 2011, 21:156-159.
- TANG L, HE F. Study of data mining method based on privacy protection[J]. Computer Technology and Development, 2011, 21: 156-159.
- [31] YAO A C. How to generate and exchange secrets[C]//IEEE Symposium on Foundations of Computer Science. c1986:162-167.
- [32] RIVEST R L, ADLEMAN L, DERTOUZOS M L. On data banks and privacy homomorphisms[J]. Foundations of Secure Computa-

- tions, 1978:169-179.
- [33] 钱萍, 吴蒙. 同态加密隐私保护数据挖掘方法综述[J]. 计算机应用研究, 2011, 28(5):1614-1617.
- QIAN P, WU M. Review of privacy protection data mining based on homomorphic encryption[J]. Application Research of Computers, 2011, 28(5):1614-1617.
- [34] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//Annual ACM Symposium on Theory of Computing. c2009: 169-178.
- [35] RIVEST R L, SHAMIR A, ADLEMAN L. A method for obtaining digital signatures and public-key cryptosystems[J]. Communications of the ACM, 1978, 21:120-126.
- [36] PAILLER P. Public-key cryptosystems based on composite degree residuosity classes[J]. Advances in Cryptology -Eurocrypt, 1999, 547(1):223-238.
- [37] DHAKAR R S, GUPTA A K, SHARMA P. Modified RSA encryption algorithm (MREA)[C]//IEEE International Conference on Advanced Computing and Communication Technologies, c2012:426 - 429.
- [38] ZHAN J, MATWIN S, CHANG L. Privacy-preserving collaborative association rule mining[J]. Journal of Network & Computer Applications, 2007, 30(3):1216-1227.
- [39] ZHANG Y, CHEN Q, ZHONG S. Efficient and privacy-preserving min and k-th min computations in mobile sensing systems[J]. IEEE Transactions on Dependable & Secure Computing, 2015:1.
- [40] 夏超. 同态加密技术及其应用研究[D]. 合肥: 安徽大学, 2013.
- XIA C. Study of homomorphic encryption technology and application [D]. Hefei: Anhui University, 2013.
- [41] 陈智翌, 王箭, 宋新霞. 全同态加密研究[J]. 计算机应用研究, 2014, 31(6):1624-1630.
- CHEN Z Z, WANG J, SONG X X. Study of homomorphic encryption[J]. Application Research of Computers, 2014, 31(6):1624-1630.
- [42] ZHAN J. Using homomorphic encryption for privacy-preserving collaborative decision tree classification[C]//Computational Intelligence and Data Mining. c2007:637-645.
- [43] GE X, YAN L, ZHU J, et al. Privacy-preserving distributed association rule mining based on the secret sharing technique[C]//The 2nd International Conference on IEEE Software Engineering and Data Mining. c2010:345-350.
- [44] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1):48-61.
- SUN J G, LIU J, ZHAO L Y. Study of clustering algorithm[J]. Journals of Software, 2008, 19(1):48-61.
- [45] BIERNACKI C. Initializing EM using the properties of its trajectories in Gaussian mixtures[J]. Statistics & Computing, 2004, 14(3): 267-279.
- [46] BIERNACKI C, CELEUX G, GOVAERT G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models[J]. Computational Statistics & Data Analysis, 2003, 41(3/4):561-575.
- [47] LIN X, CLIFTON C, ZHU M. 2005. Privacy-preserving clustering with distributed EM mixture modeling[J]. Knowledge & Information Systems, 2005, 8(1):68-81.
- [48] DUNG L T, BAO H T. Privacy preserving EM-based clustering[C]//International Conference on IEEE Computing and Communication Technologies. 2009:1-7.
- [49] JHA S, KRUGER L, MCDANIEL P. Privacy Preserving Clustering[C]//10th European Symposium on Research in Computer Security, Milan. Berlin Heidelberg: Springer, 2005:397-417.
- [50] VAIDYA J, CLIFTON C. Privacy-preserving k -means clustering over vertically partitioned data[C]//Ninth ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. c2003: 206-215.
- [51] ISO/IEC JTC1/SC27. Information technology-security techniques-privacy framework[S].
- [52] ISO/IEC JTC1/SC27. Information technology-security techniques-code of practice for PII protection in public clouds acting as PII processors[S].
- [53] ISO/IEC JTC1/SC27. Information technology-security techniques-Privacy capability assessment model[S].
- [54] ISO/IEC JTC1/SC27. Information technology-security techniques-privacy impact assessment-methodology[S].

作者简介:



王姣(1990-), 女, 河北石家庄人, 桂林电子科技大学硕士生, 主要研究方向为工业大数据的安全测评。



范科峰(1978-), 男, 陕西礼泉人, 中国电子技术标准化研究院信息安全研究中心副主任、高级工程师, 主要研究方向为信息技术、信息安全领域关键技术及标准化。



王勇(1964-), 男, 四川阆中人, 博士, 桂林电子科技大学教授, 主要研究方向为信息安全。