

面向大数据的隐私发布暴露检测方法

柯昌博^{1,2} 黄志球² 吴嘉余¹

(南京邮电大学计算机学院 南京 210003)¹ (南京航空航天大学计算机科学与技术学院 南京 210016)²

摘 要 为了防止云服务非法获取用户的个人敏感隐私信息,提出一种面向大数据的隐私信息发布检测与保护方法。首先,对用户的隐私数据进行分类,分别对隐私数据的相似度和暴露代价进行度量;其次,根据相似度和暴露代价检测云服务所要求用户提供的隐私数据中是否包含暴露链和关键隐私数据;再次,对连续隐私数据集(包含暴露链和关键隐私数据的数据集)进行离散化,同时防止离散的隐私数据集(不包含暴露链和关键隐私数据的数据集)连续化;最后,通过实验对离散的隐私数据集与没有离散的数据集进行隐私数据链的发现,从查准率和查全率上看,Exact 过滤器的查准率低于未被离散的数据集 57%,而查全率低于未被离散的数据集 17%。因此,所提方法达到了保护用户敏感隐私信息的目的。

关键词 大数据,隐私增强,隐私发布检测,隐私暴露链,关键隐私数据

中图法分类号 TP399 文献标识码 A DOI 10.11896/jsj.kx.190100050

Big Data Oriented Privacy Disclosure Detection Method for Information Release

KE Chang-bo^{1,2} HUANG Zhi-qiu² WU Jia-yu¹

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)¹

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016 China)²

Abstract In order to prevent cloud services from illegally acquiring user’s personal sensitive privacy information, this paper proposed a privacy information publishing detection and protection method for big data. Firstly, the user’s privacy data are classified, and the similarity and the disclosure cost of privacy data are measured respectively. Secondly, according to the similarity and the disclosure cost, this method detects whether privacy data required by cloud services contain disclosure chain and key privacy data. Thirdly, continuous privacy datasets, including disclosure chains and key privacy data are decomposed. At the same time, discrete datasets are prevented from being composed into continuous datasets, which do not contain disclosure chains and key privacy data. At last, privacy data chains between discrete privacy datasets and original privacy datasets are discovered by experiment. In terms of precision and recall, the precision of Exact filter is 57% lower than that of non-discrete data sets, while the recall rate is less than 17%. Therefore, the proposed method achieves the purpose of protecting user’s sensitivity privacy information.

Keywords Big data, Privacy enhancement, Privacy release detection, Privacy disclosure chain, Key privacy data

1 引言

大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合,具有数量大(Volume)、种类多(Variety)、实时变化(Velocity)的特征^[1]。据统计,平均每每秒有 200 万用户在使用谷歌搜索,Facebook 用户每天共享的信息超过 40 亿, Twitter 每天处理的推特数量超过 3.4 亿,且每年的数据量以指数形式增长,其中 3/4 的数据是由个体在创建或移动数字文件时贡献的,如一个标准的美国上班族每年能贡献 180 万 MB 的数据量^[2]。而其中包含的大量个人

隐私信息可以被数据代理商挖掘并用于商业用途,如安客诚(Acxiom)公司通过数据处理手段获取了 500 多万名分布在世界各地的消费者的个人信息,并且可以通过数据关联和逻辑推理等技术分析个人的行为及心理倾向。2014 年,美国罗彻斯特大学的 Adam Sadilek 和微软实验室的 John Krumm 通过大数据中的信息来预测一个人未来可能到达的位置,其准确率高达 80%。某知名移动应用由于不注意保护位置大数据,导致犯罪分子根据三角测量方法可以推断出用户的家庭住址等敏感信息,已引发了多起犯罪案件^[3]。而社交网络分析研究表明,可以通过其中的群组特性发现用户的属性,例

到稿日期:2019-01-07 返修日期:2019-04-09 本文受中国博士后基金(2016M591842),江苏省博士后基金(1601198C),国家自然科学基金青年项目(Grant 61602262),江苏省自然科学基金青年项目(Grant BK20150865)资助。

柯昌博(1984—),男,博士,讲师,CCF 会员,主要研究方向为云计算中的隐私保护、基于本体的软件工程等,E-mail:brobo.ke@njupt.edu.cn;黄志球(1965—),男,博士,教授,博士生导师,CCF 高级会员,主要研究方向为软件工程、形式化方法、数据仓库;吴嘉余(1996—),男,硕士生,主要研究方向为云计算中的隐私保护。

如,通过分析用户的 Twitter 信息,可以发现用户的政治倾向、消费习惯等个人偏好。因此,面向大数据,如何保护个人隐私信息已经成为广大学者研究的热点^[4]。

由于大数据的商业价值是潜在的,同时用户泄漏个人隐私信息也是潜在的,大数据未知的用途无法提前告诉用户,企业也无法承担发现大数据的创新性用途后通知每个用户并请求用户同意再进行使用的成本。因此,面向大数据的隐私数据发布检测与保护成为了研究焦点^[5]。

为了满足用户的功能需求,各种传感器或云服务终端必须收集用户的隐私信息,并且在云服务协同组合过程中透明交互,这时采用传统的信息安全技术很难对交互与共享中的隐私数据进行保护^[6];并且,当传感器和云服务终端将用户隐私数据上传到大数据中心后,大数据中心可以利用数据挖掘或逻辑推理的方法获取用户的敏感数据。因此,基于隐私数据的生命周期,本文提出的方法在发布阶段通过对隐私数据的相似度和暴露代价的度量来检测用户发布的或传感器收集的隐私数据中是否存在暴露链,并针对连续隐私数据进行离散化,防止离散隐私数据连续化,从而可以在隐私数据的发布与收集阶段防止其向外暴露。

社交网络(如 Fackbook、Twitter、微信、QQ 等)和移动互联网的快速发展,使得大数据中基于个体的信息量快速增多并实时更新,这给面向个人隐私信息挖掘的数据代理机构提供了机会。他们可能对大数据平台进行攻击,从而对商业隐私数据进行窃取,或者部分社交网络和电子商务机构通过出售用户的个人隐私信息获得商业利润。

为了获取用户的个人敏感隐私信息,通常对商业大数据进行深度挖掘和知识推理,以获取用户的个人偏好、敏感隐私信息(如姓名、地址、电话号码、卡号等),以及个体之间的关系、商业取向,甚至对用户进行行为预测,以使商业机构更好地推销产品和提供服务。而在大多数个人隐私数据没有被授权的情况下,用户不愿意将其暴露给其他商业机构用于商业目的或者其他非法目的而泄漏自己的隐私信息。

因此,本文的研究主要针对用户数据源的发布过程,对个人隐私数据进行离散化、碎片化,使得利用深度挖掘和知识推理来获取用户个人隐私信息的难度加大,以至于通过此方法无法获取用户有价值的隐私信息,从而达到保护个人隐私信息的目的。

文献[7]提出了一种基于动态信任关系的移动人群感知数据隐私保护(DTRPP)机制。该机制将密钥分发与信任管理相结合,根据公钥的支持者数量和信任程度来评估公钥的信任价值。信任值由相遇节点提供的公钥的准确性来估计。实验结果表明,与传统机制相比,该机制有效地保护了数据隐私,在平均延迟、交付速度和加载速度方面都有更好的性能。文献[8]提出了一种新的轨迹隐私保护设计方案。此方案的核心思想是提取个人轨迹检索的查询逻辑,使实际的轨迹元组在服务器端不聚集到任何路由身份或用户身份。查询逻辑存储在用户的客户端设备上,非集群的位置元组存储在后端服务器上,从而保证了客户端的轨迹重构和服务端端的隐私保护。文献[9]在分析现有隐私规则的基础上分别定义了用户和服务提供者的隐私策略,提出了隐私策略自动匹配方法,

该方法能对隐私数据的类型、数据暴露的目的、数据的收集者以及维持时间进行检测。

文献[10]提出一种用户隐私保护框架,该框架假定应用是可信的,并通过扰动的方法来保护隐私数据,但没有对此框架的具体理论和应用实现进行阐述。文献[11]提出了一种混合模式的隐私数据的扰动方法 AENDO,该方法保持相邻数据的属性不变,而对目标隐私数据进行扰动,最终根据相邻数据的属性来得到原隐私数据。这种方法可以提高敏感性隐私数据的安全性。文献[12]提出了一种多级保密的协同过滤系统方案,该方案在将隐私数据提交给服务器前对每个等级的隐私数据进行扰动。实验验证了该方法的有效性,其可以有效地保护用户的隐私。文献[13]采用粗糙集对隐私数据的匿名化过程进行管理,以保证匿名化数据的质量,并给出了有效的度量和表示算法。文献[14]提出了利用人工指纹来保护指纹的隐私和一种兼顾全局和局部指纹定位的模型来生成人工指纹定位。模型参数可以很容易地由一个具有简单约束的用户特定键来引导,并且该模型可以使用另一个密钥生成不同的人工指纹。实验证明了此方法的可行性与有效性。文献[15]提出了动态隐私保护模型,该模型旨在保证移动设备用户的隐私,在用户数据传输中,使用动态规划来产生一个保护用户隐私的最优解决方案,开发了一个 Android 应用程序,并使用它来评估模型的有效性。

在大数据环境下,云计算作为计算平台,传统的信息安全方法针对隐私数据存储(PaaS 层)和链路传输过程(IaaS 层)的保护是非常有效的。但是,云计算软件服务层(SaaS 层)是一种协同计算、透明交互的系统,隐私数据在服务组合过程中进行交互时需要将用户的隐私数据共享,并对服务参与者透明。明文传达所带来的数据隐私问题,使得采用传统的信息安全方法来解决此问题比较困难,只有将信息安全和相应的隐私增强技术相结合,才能更好地解决大数据环境下云计算平台的隐私保护问题。因此,本文提出了一种隐私增强方法,以防止用户的敏感隐私数据被暴露。

2 隐私数据发布时的暴露检测

2.1 隐私数据的分类

首先通过与知识领域本体之间的映射得到隐私属性之间的关系,然后利用本体树进行描述。隐私属性的本体树可以用二元组 $T=(V,E,root(T))$ 表示,其中 V 表示隐私数据,为本体树的节点; E 表示隐私数据之间的语义关系 $\langle pd_i, pd_j \rangle$; $root(T)$ 的根节点为 Thing,是所有隐私属性类的超类。除底层之外,其他各层之间是类与子类的关系,即包含关系,例如, Name 类包含 realName 和 nickName;而只有底层与其上层是属于关系,例如,country,province,city 是属于 address 类。

定义 1(隐私数据 $pd(Privacy Data)$) 根据用户的需求和分层结构关系,对隐私数据进行分类,将其分为离散隐私数据和连续隐私数据,即 $pd=\{dp,cp\}$ 。离散的隐私数据是指隐私数据集中任意两个隐私数据的组合都不会导致用户隐私信息的暴露,可以表示为 $dp=\langle \neg \exists pd(pd_i), pd_j \rangle \rightarrow dc$,其中 $i,j \geq 1$ 。连续的隐私数据是指隐私数据集中存在两个隐私数据的组合可以导致用户隐私信息的暴露,可以表示为 $cp=\langle \exists pd(pd_i), pd_j \rangle \rightarrow dc$,其中 $i,j,r \geq 1$ 。

定义 2(隐私数据属性(Privacy Data Attribute)) 根据隐私数据的属性,可以将隐私数据分为关键隐私数据和非关键隐私数据。关键隐私数据 kp (Key Privacy)是唯一能确定用户身份信息的隐私项,在隐私数据集中存在某个隐私数据与其他任何一个隐私数据的组合都是连续隐私数据,称这种隐私数据为关键隐私数据,可以表示为 $kp = \langle \exists pd(pd_i), pd_k \rangle \rightarrow dc$,其中 $i \geq 1, n \geq k \geq 1$ 且 $i \neq k$ 。非关键隐私数据 nkp (Non-Key Privacy)是指除能确定用户身份信息以外的隐私项,可以表示为关键隐私数据的非,即 $nkp = \neg kp$ 。

定义 3(组合离散数据链 dpc(Composition Discrete Privacy Data Chain)) 离散数据集所组成的链称为组合离散数据链。组合离散数据链需要满足以下两个条件中的任意一个:

(1)在离散数据集中,任意一个离散数据与连续隐私数据或者离散隐私数据进行组合都不会推导出连续的隐私数据,即 $[(\forall dp(dp_i), cp_i) \vdash cp] \vee [(\forall dp(dp_i), dp_i) \vdash cp]$,其中 $cp_i \notin kp$ 。通常称这种离散数据链为外组合离散数据链。

(2)设隐私数据链长为 a ,隐私数据链所组成的集合与关键隐私数据组合可以推导出连续的隐私数据,但集合中的任意元素与关键隐私数据不能推导出连续的隐私数据,即 $[(\langle dp_1, dp_2, \dots, dp_k \rangle, kp) \rightarrow cp] \wedge [(\forall dp(dp_i), kp) \vdash cp]$,其中, $k = a$ 。称这种离散数据链为内组合离散数据链。

2.2 隐私数据相似性度量

假设用户隐私需求本体树和服务隐私描述本体树(下文简称为需求树和描述树)之间具有上下文层次(语义)关系的一致性。即,若需求树中的某个节点 s_q 在描述树中所对应的层次为 i ,则其子类节点或者其属性必须在描述树中的第 $i + \alpha$ 层。本文中的相似度分为 3 类,即概念相似度 $\phi = sim_d(s_q, s_d)$ 、属性相似度 $\mu = sim_T(s_q, s_d)$ 、结构相似度 $\zeta = sim_s(s_q, s_d)$ 和相似度 $\tau = sim(s_d, s_q)$ 。

(1)概念相似度

概念相似度是在分层的基础上定义的,根据两棵树中节点之间的关系,将其分为 3 个层次,记为: $R \overset{T}{\leftrightarrow} D$ 。其中, R 表示需求树中的节点, D 表示描述中的节点, T 表示两节点间的层次关系,即 $T = \{e, su, p\}$ 。由于两棵树有相同的根节点(Thing),因此存在 3 种可能:同层、上下层和下上层。

同层(exact):查询树中的 R 节点与描述树中的 D 节点具有相同的层次数,并且后代节点之间是一一对应的,即 $R_i \overset{e}{\leftrightarrow} D_j \Leftrightarrow (j = i) \wedge (son(R_i) \vdash son(D_j))$ 。

上下层(subsume):查询树中的 R 节点的层次数小于描述树中的 D 节点的层次数,并且查询树中的节点 R 的孩子与描述树节点 D 的后代节点对应,即 $R_i \overset{su}{\leftrightarrow} D_j \Leftrightarrow (j = i + \alpha) \wedge (des(R_i) \vdash des(D_j))$ 。

下上层(plugin):查询树中的 R 节点的层次数大于描述树中的 D 节点的层次数,并且查询树中的节点 R 的孩子与描述树节点 D 的后代节点对应,即 $R_i \overset{p}{\leftrightarrow} D_j \Leftrightarrow (j = i - \alpha) \wedge (des(R_i) \vdash des(D_j))$ 。

本文的相似度计算是基于语义词典 Wordnet 的。在 Wordnet 中,每个节点 s 表示一个概念。Pantel 等^[12]根据 Wordnet 定义了两个概念的相似度:

$$\phi = \begin{cases} \frac{2 \times \log p(s)}{\log p(s_1) + \log p(s_2)}, & \text{if } R_i \overset{T}{\leftrightarrow} D_j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

其中, $p(s) = count(s) / total$ 表示在 Wordnet 中概念节点 s 及其子节点所包含的单词个数在整个词典中所占的比例, $total$ 是 Wordnet 的单词总数,节点 s 是 s_1 和 s_2 的公共祖先节点。

(2)属性相似度

设在两棵树 T_q 和 T_d 中,假如某节点为对象,两个对象分别为 OA 和 OB 。而 OA 和 OB 中的属性分为两类,一类是简单属性,如整型或者是字符型。这种类型的相似度可以直接根据概念相似度(sim_d)求得。另一类为关系类,即可以通过某种关系函数计算得到其相似度。关系函数表示为: $f_r = (ID_A, ID_B, P_A, P_B)$ 。

设两个对象为 $OA = \langle ID_A, C_A, P_A \rangle$ 和 $OB = \langle ID_B, C_B, P_B \rangle$,并且有共同的属性 $P \in T$ (其中 T 为共同的属性集),则属性相似度 sim_p 可以定义为:

$$sim_p(OA, OB) = \begin{cases} sim_d, & \text{int, char} \\ f_r = \diamond(ID_A, ID_B, P_A, P_B), & R \end{cases} \quad (2)$$

其中, f_r 是根据属性之间的特定语义关系所确定的。不同的属性有不同的语义关系,这其中有一对一的关系,也有一对多的关系。例如:对象 OA 指某个人 Tom, OB 指 Jack,假如 Tom 的属性项包含年龄(Age),而 Jack 的属性项包含生日(Birthday),则这两个属性就是关系型的,即 $f_r = \diamond(Age, Birthday)$, $Age = PresentYear - Birthday$;又如: OA 和 OB 分别指两个圆, OA 的属性项包含圆的半径(r),而 OB 的属性项包含面积(S),同理它们为关系型的,即 $f_r = \diamond(r, S)$, $S = \pi r^2$ 。此时,如果所得到的属性值是一致的,则属性相似度为 1,否则为 0。那么,对象节点的属性相似度 sim_T 可以定义为:

$$\mu = \frac{\Delta}{\forall p \in T} sim_p(OA, OB) \mu \quad (3)$$

(3)结构相似度

对于查询本体树 T_q 和描述本体树 T_d ,由于树中有 3 种节点,即类、对象和属性,则应分 3 类情况进行讨论。

1)假若此节点为属性,则可以直接用属性相似度与所属对象的概念相似度进行计算。因此,节点的结构相似度为:

$$\zeta = \frac{1}{2} sim_d(O(s_q), O(s_d)) + sim_p(OA, OB) \quad (4)$$

证明:要证明式(4)为属性节点的结构相似度是合理的,必须分两个方面证明。

①必须要有结构信息,即满足对象到属性的层次关系。由定义可知,该条件是满足的。

②相似度的值必须在 0 到 1 之间。因为 $0 \leq sim_d(O(s_q), O(s_d)) \leq 1$, 且 $0 \leq sim_p(OA, OB) \leq 1$, 所以 $0 \leq \zeta \leq 1$ 。因此,式(4)的结构相似度是合理的。

2)假若此节点为类,则可以根据超类和子类的相似度计算,而超类的相似度为 $h = sim_d(supc(s_q), supc(s_d))$,子类的相似度为 $\lambda = \sum_{subc \in d} sim_d(s_q, s_d)$ 。因此,设 $\bar{\omega} = |subc(s_q)|$, $\theta = |subc(s_d)|$ 和 $\vartheta = |subc(s_q) \cap subc(s_d)|$,节点(类)的结构相似度为:

$$\zeta = \frac{h + \lambda}{\min(\bar{\omega}, \theta, \vartheta) + 1} \quad (5)$$

证明:由对式(4)的证明可知,只需对②进行证明即可。

设 $a = \min(\bar{\omega}, \theta, \vartheta)$, 由于 $0 \leq h \leq 1$ 并且 $0 \leq \lambda \leq h$, 因此 $0 \leq (a) + (b) \leq a + 1$, 故 $0 \leq \zeta \leq 1$. 得证。

3) 若此节点为对象, 节点 s_q 与 s_d 有相似的祖先节点, 并且这两个节点属性相似。这里, 两个节点的属性可以分以下 3 种情况讨论。

① 对于两个节点 s_q 和 s_d , 有 $|T_{s_q}| = |T_{s_d}|$, 并且 $\mu = 1$, 则有 $T_{s_q} = T_{s_d}$ 。

② 对于两个节点 s_q 和 s_d , 有包含关系, 即 $T_{s_q} \subseteq T_{s_d} \vee T_{s_q} \supseteq T_{s_d}$ 。

③ 对于两个节点 s_q 和 s_d , $\exists A: T_{s_q} \subseteq A \wedge T_{s_d} \subseteq A \wedge T_{s_q} \cap T_{s_d} \neq \emptyset$ 。

综上, ①, ② 和 ③ 这 3 种关系称为属性间的相容关系。

若 $\exists S_D: S_D \in \text{supc}(s_d)$, $\exists S_Q: S_Q \in \text{supc}(s_q)$, 使得 $\exists S_D \forall \text{supc}(s_q)(\text{sim}_d(S_D, S_Q)) > \alpha$, 并且 $\mu > \beta$ (其中, α 和 β 分别为概念相似和结构相似的阈值), 此时, 设 $\ell = \sum_{p \in T} \mu$, $\mathfrak{S} = |T_{s_q}|$, $\mathfrak{R} = |T_{s_d}|$, $\mathfrak{N} = |T_{s_q} \cap T_{s_d}|$, 节点的结构相似度为 (证明同上, 略):

$$\zeta = \frac{h + \ell}{\min(\mathfrak{S}, \mathfrak{R}, \mathfrak{N}) + 1} \quad (6)$$

因此, 两棵本体树之间对应节点的总相似度为 (总相似度的阈值为 γ):

$$\tau = \frac{1}{a + b} (a \times \phi + b \times \zeta) \quad (7)$$

2.3 隐私数据暴露检测

定义 4 (隐私敏感度 PS-Degree (Privacy Sensitivity Degree)) 隐私敏感度是用户对个人隐私数据的敏感程度, 设 $p = [p_1, p_2, \dots, p_n]$ 为用户的隐私数据链, 敏感度可以表示为 $sv = [sv_1, sv_2, \dots, sv_n]$, 其中 sv_i 表示数据 p_i 的敏感度, $1 \leq i \leq n$ 。将用户的隐私敏感度分为两类, 一类是用户有隐私需求, 则根据用户的隐私需求, 将用户的隐私信息的敏感程度定义为 $[0, 1]$ 区间上的任意实数, 其中 0 表示敏感度最弱, 1 表示敏感度最强。另一类是用户无隐私需求, 则根据第一类用户对隐私数据的敏感程度将用户的隐私信息分为 5 个等级: A++ 表示非常敏感; A+ 表示比较敏感; A 表示敏感; B+ 表示一般敏感; B 表示不敏感。

定义 5 (隐私数据暴露向量) 隐私数据暴露向量是指用户是否将个人隐私数据集暴露给云服务提供者, 所对应的真值向量为 $dv = [dv_1, dv_2, dv_3, \dots, dv_n]$, 其中 $dv_i \in \{0, 1\}$ 。如果 dv_i 取值为 1, 表示暴露隐私数据对象 p_i ; dv_i 取值为 0 则表示不暴露数据对象 p_i , $1 \leq i \leq n$, $p_i \in p$ 。

定义 6 (隐私数据暴露代价) 隐私数据暴露代价是指用户为了获取功能服务而暴露个人隐私信息的代价。隐私数据暴露代价是隐私数据敏感度与隐私数据暴露向量的函数, 并且暴露代价分别与敏感度和暴露向量成正比, 即隐私数据敏感度越大, 隐私数据暴露代价就越大; 隐私数据暴露得越多, 隐私数据暴露代价也越大。隐私数据暴露代价可以利用敏感度向量矩阵与暴露向量矩阵计算得到, 具体可以表示为:

$$\text{Disp} = [dv_1, dv_2, dv_3, \dots, dv_n] \times [sv_1, sv_2, \dots, sv_n]^T$$

其中, $\text{Disp} \in R^+$, $1 \leq i \leq n$ 。

定义 7 (暴露链 dc (Disclosure Chain)) 暴露代价大于 1 的隐私数据集称为暴露链, 即 $(dc \subseteq pd) \wedge \text{Disp}_{pd} \geq 1$ 。

定义 8 (隐私集 PES (Privacy Element Set)) 隐私集包括两种类型, 第一种类型是服务要求用户暴露的最小隐私数据集, 即 $PES_s = \{pd_{s1}, pd_{s2}, \dots, pd_{si}, \dots, pd_{sk}\}$, 其中 pd_{si} 为云服务提供商要求用户暴露的隐私项, 在集合上为服务输入和前置条件的子集, 即 $pd_{si} \subseteq (P_i, I_i)$, $0 \leq i \leq k$ 。PES 为服务的隐私项集, P 和 I 分别表示服务的前置条件和输入。

第二种类型是当用户向云服务提供商发出服务请求时, 用户隐私需求中愿意暴露的隐私集, 即 $PES_u = \{pd_{u1}, pd_{u2}, \dots, pd_{ui}, \dots, pd_{uk}\}$, 其中 pd_{ui} 为用户隐私需求中愿意暴露的隐私项。

利用隐私信息的相似度和暴露代价来检测云服务提供商要求用户暴露的隐私集中是否包含隐私暴露链或关键隐私数据, 即判断隐私数据集是离散的隐私数据还是连续的隐私数据。如果满足以下两个条件, 即:

$$\sum_{i=1}^{k-1} \text{sim}_i dv_i + \sum_{i=k}^{n-1} \text{sim}_i \geq \varphi \text{ 且 } \text{Disp}_{\text{OrdReq!}} \geq \delta$$

则认为隐私集中包含暴露链或关键隐私数据。其中, sim_i 表示用户隐私数据与云服务所要求的隐私数据之间的相似度, dv_i 表示隐私数据暴露向量, φ 表示相似度的阈值, $\text{Disp}_{\text{OrdReq!}}$ 表示隐私数据暴露代价, δ 表示隐私数据暴露代价的阈值, 由于隐私数据的敏感度的最大值为 1, 因此通常设 δ 值为 1。

3 面向发布的隐私数据保护方法的流程

离散数据的保护就是防止服务提供者要求用户提供的隐私数据集构成连续隐私数据链。由于离散数据链有外离散数据链和内离散数据链之分, 因此, 分两种情况进行讨论。

(1) 外离散数据链为不连续的虚链, 即数据集中某些数据之间存在逻辑关系, 但不包含暴露链。但是当有一个或者多个关键隐私数据进入离散隐私数据时, 原来的不连续的虚链就可能组合成连续的暴露链。随着服务的不断演化, 服务所需的隐私数据也逐渐演化。演化隐私数据的加入, 使得原来不连续的虚链就会变成一条完整的连续的暴露链。此演化过程如图 1 中的 e-dpc 部分所示。

(2) 同理, 内离散数据链在服务组合之初为连续的虚链, 由于内离散数据链具有整体成链才有可能形成暴露链的特点, 因此当关键隐私数据加入后, 即使形成了以关键隐私数据为首元素的连续链, 此链也不包含于暴露链或者不等于暴露链。但由于演化的结果, 也有可能形成完全且完整的暴露链。此演化过程如图 1 中的 i-dpc 部分所示。

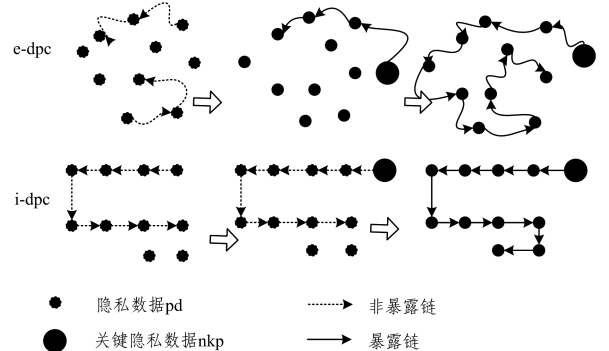


图 1 外离散隐私数据和内离散隐私数据的演化过程

Fig. 1 Evolution of external discrete privacy data and internal discrete privacy data

由于服务组合的过程也是用户隐私数据组合的过程,通过以上的分析可以看出,对于外离散隐私数据,只要防止关键隐私数据参与组合,就可以有效保护用户的隐私数据;对于内离散隐私数据,只要防止关键隐私数据与内离散隐私完整链组合,就可以避免隐私泄露。

根据定义,连续隐私数据包含如下特征:

- (1)必然包含暴露链;
- (2)可能包含关键隐私数据。

连续隐私数据的保护就是将连续隐私数据分解为离散隐私数据。针对特征(1),可以将暴露链中具有内离散数据链的父数据进行分解,将其转化为外离散数据。具体分解步骤如下(如图 2 所示):首先检测连续隐私数据中存在的暴露链,然后以暴露链中的元素为根节点对隐私本体树进行搜索,查找存在的子节点集合;使子节点集合形成内离散数据链 i-dpc;利用内离散数据链 i-dpc 替换隐私数据 dp_i ;删除内离散数据链中的链外数据;将原有的暴露链替换为非暴露链。

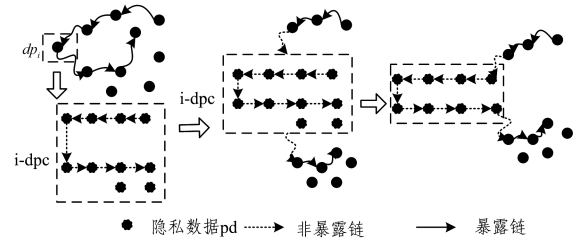


图 2 暴露链中的数据分解为内离散隐私数据

Fig. 2 Data in disclosure chain are decomposed into internal discrete privacy data

针对特征(2),首先以关键隐私数据为根节点对隐私本体树进行遍历,如果关键隐私数据不存在子节点,则检测隐私数据集,查找暴露链,并采用图 2 中的方法进行分解对存在的暴露链;如果关键隐私数据存在子节点,则利用如下步骤进行分解(如图 3 所示):以关键隐私数据 kp 为根节点检索隐私本体树,查找子节点集合;使子节点集合形成内离散数据链 i-dpc;利用内离散数据链 i-dpc 替换关键隐私数据 kp ;删除内离散数据链中的链外数据;将原有的暴露链替换为非暴露链。

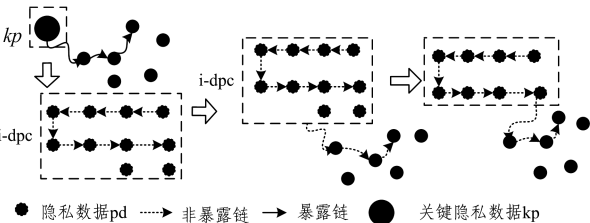


图 3 关键隐私数据分解为内离散隐私数据链

Fig. 3 Key privacy data are decomposed into internal discrete privacy data links

4 实验与数据分析

本文开发了原型系统 DC-Checker,将其部署在亚马逊云计算平台 Amazon EC2 上。实验使用了 OWLS-TC version 4.0 中的 883 个 OWL-S advertisements 和手工修改的 18 个 OWL-S advertisements 与 33 个用户隐私需求,同时利用网络爬虫获取了大约 1 024 terabytes 的干扰数据,并根据实验目

的分别将原隐私数据和离散后的隐私数据混杂在干扰数据中,利用查准率和查全率对数据进行分析。

在隐私属性的匹配中,使用 OWLS-MX 中的 M4 的配置,而对 DC-Checker 中参数的配置分别为: $a=0.8, b=1, \alpha=0.5, \beta=0.5, \gamma=0.5$ 。

对参数的取值做以下说明:

(1)关于 a 与 b 的取值:本文的结构相似度是建立在概念相似度之上的,只有结构相似度能体现概念之间的语义关系。假若某个概念的结构相似度大于阈值,则说明此概念一定满足上下文语义关系,进而可以推断出此节点一定是满足用户需求的节点,因此结构相似度的权值一定大于概念相似度的权值,且 $b=1$,即在公式 $\tau = \frac{1}{a+b} (a \times \phi + b \times \zeta)$ 中, $b \geq a$, 且 $b=1$ 。这一点是可以确定的,所以在实验时取 $a=0.8, b=1$ 。大多数学者对权值的确定采用智能算法进行迭代得到,或者根据前人的研究结果给定一个经验值。由于篇幅的原因并为了突出本文的研究重点,本文取 a 为一个小于 b 的经验值。即 $a=0.8$,且没有对 a 的取值进行计算说明和实验。

(2)关于 α, β 和 γ 的取值:在整个约束条件中,结构相似度的阈值起决定作用,即 β 的取值对匹配结果起关键作用;匹配结果与 α 的取值没有关系, α 可以取 $[0, 1]$ 内的任何值;而 γ 的取值是由 α 和 β 决定的。假设取 $\alpha=0.5$,此时, γ 与 β 满足线性关系,即 $\gamma = \frac{0.8 \times 0.5 + 1 \times \beta}{1 + 0.8}$, γ 是随着 β 的增大而增大的。

本文将实验数据分为两组,其中一组利用半手工的方法对隐私数据进行离散化,并将其混杂在干扰数据中,然后采用 DC-Checker 进行隐私暴露链的发现,并分别从 Exact, Plugin 和 Subsume 3 个层次对两组实验发现暴露链的成功率进行对比,结果如图 4、图 5 所示。

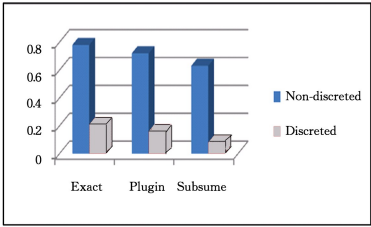


图 4 暴露链查准率的对比

Fig. 4 Comparison of accuracy of disclosure chain

由图 4 可知,对离散后的数据进行隐私暴露链的发现与没有进行离散的数据集发现的查准率相比,在 Exact 过滤器中低 57%,在 Plugin 过滤器中低 56%,在 Subsume 过滤中低 55%。

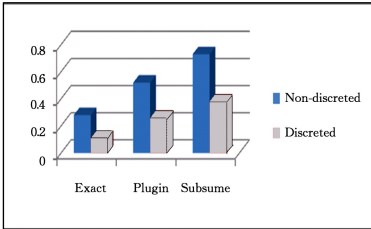


图 5 暴露链查全率的对比

Fig. 5 Comparison of recall rate of disclosure chain

由图 5 可知,对离散后的数据进行隐私暴露链的发现与没有进行离散的数据集发现的查全率相比,在 Exact 过滤器中低 17%,在 Plugin 过滤器中低 26%,在 Subsume 过滤中低 35%。

因此,对发布前的用户隐私数据进行离散化,可以有效地降低隐私数据的暴露,从而达到保护用户隐私信息的目的。

结束语 本文在面向隐私数据的发布与收集阶段,对隐私数据的暴露进行检测,并对数据进行碎片化以达到保护用户个人敏感信息的目的。通过检测隐私数据链的实验可以看出,离散后的隐私数据集比离散前隐私数据集所检测的暴露链查准率和查全率都低,说明所提方法对用户隐私数据的发布有着很好的保护作用。下一步将对离散后的隐私数据进行干扰,以进一步地保护用户的隐私信息。

参 考 文 献

[1] SNIJDERS C, MATZAT U, REIPS U D. “Big Data”: big gaps of knowledge in the field of internet science[J]. International Journal of Internet Science, 2012, 7(1): 1-5.

[2] DU J, JIANG C, CHEN K C, et al. Community-structured evolutionary game for privacy protection in social networks[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(3): 574-589.

[3] HUGUENIN K, BILOGREVIC I, MACHADO J S, et al. A predictive model for user motivation and utility implications of privacy-protection mechanisms in location check-ins [J]. IEEE Transactions on Mobile Computing, 2018, 17(4): 760-774.

[4] PENG H F, HUANG Z Q, FAN D J, et al. Specification and Verification of User Privacy Requirements for Service Composition[J]. Journal of Software, 2016, 27(8): 1948-1963. (in Chinese)

彭焕峰, 黄志球, 范大娟, 等. 面向服务组合的用户隐私需求规约与验证方法[J]. 软件学报, 2016, 27(8): 1948-1963.

[5] LIU X Y, WANG B, YANG X C. Survey on Privacy Preserving Techniques for Publishing Social Network Data[J]. Journal of Software, 2014, 25(3): 576-590. (in Chinese)

刘向宇, 王斌, 杨晓春. 社会网络数据发布隐私保护技术综述

[J]. 软件学报, 2014, 25(3): 576-590.

[6] ZHANG Y X, HE J S, ZHAO B, et al. A Privacy Protection Model Base on Game Theory [J]. Chinese Journal of Computers, 2016, 39(3): 615-627. (in Chinese)

张伊璇, 何泾沙, 赵斌, 等. 一个基于博弈理论的隐私保护模型[J]. 计算机学报, 2016, 39(3): 615-627.

[7] WU D, SI S, WU S, et al. Dynamic trust relationships aware data privacy protection in mobile crowd-sensing[J]. IEEE Internet of Things Journal, 2018, 5(4): 2958-2970.

[8] XIAO Z, YANG J J, HUANG M, et al. QLDS: A Novel Design Scheme for Trajectory Privacy Protection with Utility Guarantee in Participatory Sensing[J]. IEEE Transactions on Mobile Computing, 2018, 17(6): 1397-1410.

[9] KE C B, HUANG Z, LI W, et al. Service Outsourcing Character Oriented Privacy Conflict Detection Method in Cloud Computing [J]. Journal of Applied Mathematics, 2014, 2014: 1-11.

[10] GU L, CHEUNG S C. Constructing and testing privacy aware services in cloud computing environment: Challenges and opportunities[C]//Proceedings of the 1st Asia-Pacific Symposium on Internet ware. New York: ACM, 2009: 1-10.

[11] NI W W, CHONG Z H. Clustering-oriented privacy-preserving data publishing[J]. Knowledge-Based Systems, 2012, 35: 264-270.

[12] POLATIDIS N, GEORGIADIS C K, PIMENIDIS E, et al. Privacy-preserving collaborative recommendations based on random perturbations[J]. Expert Systems with Applications, 2017, 71: 18-25.

[13] YE M, WU X, HU X, et al. Anonymizing classification data using rough set theory[J]. Knowledge-Based Systems, 2013, 43: 82-94.

[14] LI S, ZHANG X, QIAN Z, et al. Key based Artificial Fingerprint Generation for Privacy Protection[J]. IEEE Transactions on Dependable and Secure Computing, 2018, PP(99): 1-1.

[15] GAI K, CHOO K K R, QIU M, et al. Privacy-preserving content-oriented wireless communication in internet-of-things[J]. IEEE Internet of Things Journal, 2018, 5(4): 3059-3067.