软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(8):1922−1933 [doi: 10.13328/j.cnki.jos.005093]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

时空数据发布中的隐式隐私保护*

王 璐, 孟小峰, 郭胜娜

(中国人民大学 信息学院,北京 100872)

通讯作者: 孟小峰, E-mail: xfmeng@ruc.edu.cn

摘 要: 随着大数据时代的到来,大量的用户位置信息被隐式地收集.虽然这些隐式收集到的时空数据在疾病传播、路线推荐等科学、社会领域中发挥了重要的作用,但它们与用户主动发布的时空数据相互参照引起了大数据时代时空数据发布中新的个人隐私泄露问题.现有的位置隐私保护机制由于没有考虑隐式收集的时空数据与用户主动发布的位置数据可以相互参照的事实,不能有效保护用户的隐私.首次定义并研究了隐式收集的时空数据中的隐私保护问题,提出了基于发现-消除的隐私保护框架.特别地,提出了基于前缀过滤的嵌套循环算法用于发现隐式收集的时空数据中可能泄露用户隐私的记录,并提出基于频繁移动对象的假数据添加方法消除这些记录.此外,还分别提出了更高效的反先验算法和基于图的假数据添加算法.最后,在若干真实数据集上对提出的算法进行了充分实验,证实了这些算法有较高的保护效果和性能.

关键词: 隐式隐私;时空数据;隐私保护

中图法分类号: TP311

中文引用格式: 王璐,孟小峰,郭胜娜.时空数据发布中的隐式隐私保护.软件学报,2016,27(8):1922−1933. http://www.jos.org.cn/1000-9825/5093.htm

英文引用格式: Wang L, Meng XF, Guo SN. Preservation of implicit privacy in spatio-temporal data publication. Ruan Jian Xue Bao/Journal of Software, 2016,27(8):1922−1933 (in Chinese). http://www.jos.org.cn/1000-9825/5093.htm

Preservation of Implicit Privacy in Spatio-Temporal Data Publication

WANG Lu, MENG Xiao-Feng, GUO Sheng-Na

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: In the emerging big data era, in addition to explicit publication of users' locations on geo-social networks, positioning system embedded in mobile phones implicitly records users' locations. Although such implicitly collected spatiotemporal data play an important role in a wide range of applications such as disease outbreak control and route recommendation for life science or smart city, they cause new serious privacy issues when cross-referencing with the explicitly published data from users. Existing location privacy preservation techniques fail to preserve the proposed implicit privacy because they ignore the cross-reference between implicitly and explicitly spatiotemporal data. To tackle this issue, this work for the first time investigates and defines the implicit privacy and proposes the discover and eliminate framework. In particular, this paper proposes prefix filtering based nest loop algorithm and frequent moving object based algorithm to generate dummy data to preserve the proposed implicit privacy. Further, it constructs an improved reverse a priori algorithm and graph based dummy data generation algorithm respectively to make the solution more practical. The results of extensive experiments on real world datasets demonstrate the effectiveness and efficiency of the proposed methods.

Key words: implicit privacy; spatio-temporal data; privacy preserving

* 基金项目: 国家自然科学基金(61532010, 61379050, 91224008); 国家高技术研究发展计划(863)(2013AA013204); 高等学校博士学科点专项科研基金(20130004130001); 中国人民大学科学研究基金(11XNL010)

Foundation item: National Natural Science Foundation of China (61532010, 61379050, 91224008); National High-Tech R&D Program of China (863) (2013AA013204); Research Fund for the Doctoral Program of Higher Education of China (20130004130001); Research Funds of Renmin University of China (11XNL010)

收稿时间: 2015-12-19; 修改时间: 2016-04-14; 采用时间: 2016-06-02

大数据时代,随着定位技术的发展,基于位置的服务越来越普及,用户的时空数据通过各种各样的服务发布出来.在用户通过签到等移动社交网络服务主动发布自己的时空行为的同时,大量记录人们行为的时空数据在人们使用手机进行通话、接发短信息时被移动通信商隐式收集[1].由于手机与移动通信商间的时空数据由手机的信号基站自动收集,这些隐式收集的数据具有数据量大、蕴含人类行为的特征,并且在疾病传播[2,3]、贫困消除[4]、城市规划[5]等重大科学社会问题以及路线推荐[6]、乘车出行[7]等重要生活应用中发挥了关键作用.然而,这些隐式收集到的时空数据通过与用户主动发布的时空数据相互参照,会暴露用户有关个人身份、行动目的、健康状况、兴趣爱好等多方面的敏感隐私信息[8,9].近年来,随着个人隐私观念的增强以及数据发布使用中法律法规的健全,这些隐式收集到的时空数据在供科学研究与数据挖掘前需要首先消除其中可能暴露用户隐私的记录.

为了确保这些个人敏感信息不被泄露,大量针对时空数据隐私保护的工作致力于匿名化可能暴露个人敏感信息的时空数据.例如,位置与轨迹数据上的 k 匿名等方法将用户位于特定时间范围和空间区域的位置记录泛化为相同的空间区域,从而让攻击者不能识别出某个时间范围与空间区域中的特定用户[2].但是,这些方法没有考虑到攻击者可以参照用户主动发布的时空数据从隐式收集到的时空数据中找到暴露用户隐私的记录,因此,经过这些方法保护的时空数据集依然会泄露用户隐私的数据.

例如,假设攻击者看到 Alice 在社交网络上主动广播了自己 5 月 1 日去海南、10 月 1 日去哈尔滨的两次旅行,如果手机运营商收集到的时空数据集中在这两天分别去过海南和哈尔滨的记录都具有相同的用户 ID,那么攻击者就可以推断这条记录中的用户 ID 在现实中对应用户 Alice,从而根据数据库中同样 ID 的时空记录发现 Alice 去过的其他地方,并进而推断出她的兴趣爱好、行为习惯等隐私信息.

最新的研究[10]表明,现有的位置或轨迹隐私保护方法即使将待发布位置泛化为 15 平方公里,时间戳泛化为一个小时,95%以上的用户也可以被 4 个泛化后的时空记录唯一地确定.如果这些用户还通过签到服务主动发布了自己的若干行为,他们就很容易被攻击者识别出来.

在大数据时代,由于攻击者可以通过越来越多的位置社交网站同时收集用户主动发布的时空数据,隐式收集到的时空数据集更容易暴露用户隐私,这对保护这些时空记录在效果、效率和效用这 3 个方面提出了严重挑战:(1) 对时空数据集的保护效果如何不受攻击者收集用户主动发布的时空数据的能力的增强而减弱;(2) 由于数据集中可能暴露用户隐私的记录个数随着时空数据集的增大呈指数级增长,如何高效发现待保护的记录;(3) 为了发布时空数据而对时空点集进行保护时,如何保证较高的数据效用.

为了应对上述挑战,本文首先在隐式收集到的数据集上定义与用户主动发布数据无关的隐式隐私,以保证无论攻击者收集到多少用户主动发布的数据,经过隐私保护后的隐式收集到的时空数据集都不会泄露额外的信息.随后,本文提出发现和消除两个步骤来保护隐式收集到的时空数据.为了高效寻找可能暴露用户隐私的记录,本文提出基于前缀过滤的嵌套循环算法,并针对其可扩展性差的缺陷提出高效的反先验算法.为了避免这些记录暴露用户的隐私,我们设计了基于频繁移动对象的假数据添加算法,并针对其数据扭曲程度高的缺陷提出了基于图的假数据添加算法,以确保经过保护后的数据集具有高数据效用.

(1) 首次提出并定义时空数据中的隐式隐私,并提出保护该隐私的框架.我们首次定义了隐式收集到的时空数据中的(ε,k)隐私问题,并提出了发现-消除框架和相应算法确保经过保护的数据集不因为攻击者收集到更多用户主动发布的时空数据而泄露额外信息.

(2) 高效的发现算法.我们针对框架中的发现步骤提出了基于前缀过滤的嵌套循环的算法,并针对其扩展性差的缺陷设计了基于反先验算法的优化算法,高效地寻找暴露隐式隐私的时空点集合.

(3) 高效的保护算法.我们针对框架中的消除步骤提出了基于频繁移动对象的假数据添加算法,并针对其数据效用差的缺陷设计了基于图的假数据添加方法,大幅提高了数据效用.

(4) 真实数据集上的充分实验.我们使用两个真实数据集验证了我们方法的效果和效率.

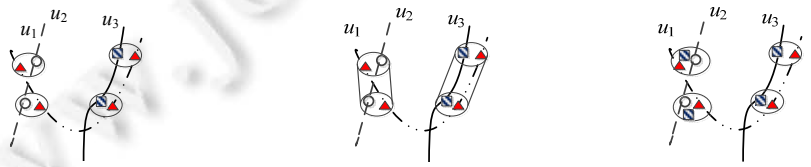
本文第 1 节主要介绍时空数据发布中隐私保护的相关技术.第 2 节介绍隐式隐私问题及其发现-消除保护框架.第 3 节介绍相应的发现,消除算法.第 4 节是实验效果展示.

1 相关工作

匿名方法[11,12]被广泛应用于保护时空数据发布中的敏感信息,并取得了巨大成功.文献[13]对其进行了详细的综述.匿名方法可以分为位置 k 匿名和轨迹 k 匿名.

位置 k 匿名将每个要发布的时空数据从位置和时间两个维度进行泛化使得每个经过泛化后的时空区域包含至少 k 个移动对象[14].它的一些变种使用空间索引组织各个移动对象来提高泛化和查询的性能[15],另一些变种通过考虑每个时空数据在不同泛化区域之间进行移动的最大速度限制让泛化更成功[16].但是,位置 k 匿名不能解决大数据时代隐式位置信息收集与主动位置发布带来的身份泄露.例如,图 1(a)分别包括 3 个用户在 4 个时刻的位置记录,每个大圆圈表示同一时刻的位置数据经过 k 匿名方法的泛化,攻击者在没有任何背景知识的情况下,在这一时刻不能分辨其中的 2 个用户.但是如果 u_1 将自己曾出现在第 1 个和最后一个时空点的信息通过签到服务公开在社交网络中,则攻击者很容易将 u_1 与三角形表示的用户关联起来.

轨迹 k 匿名[17,18]试图针对整条轨迹进行匿名保护,使得每条轨迹满足 k 匿名.然而由于移动对象轨迹的多样性,对所有轨迹完全匿名是不可能的[17,19,20],因此,现有方法只针对不同时间段将轨迹分片,对于分片后的轨迹进行 k 匿名处理.如图 1(b)所示,全部轨迹被分成两片,由于每片都包括两个用户的轨迹,这两个分片满足轨迹 2 匿名.然而,攻击者依然可以将三角形对应的用户和其被隐式收集到的时空数据中的标识关联起来:攻击者可以发现三角形用户同时出现在左右两个分片中,而在位置社交网络的公开数据中,也只有一个用户同时出现在左右两个分片的范围内进行了签到服务.因此,无论是位置 k 匿名还是轨迹 k 匿名,现有的方法均不能抵抗时空数据中的隐式隐私泄露.事实上,我们只要在此基础上增加两个时空假数据,即可消除与公开数据唯一关联的问题,如图 1(c)所示.



(a) 3 个用户在 4 个时刻的位置记录 (b) 轨迹 2 匿名后的位置记录 (c) 添加假数据后的位置记录

Fig.1 Examples for privacy preserving methods

图 1 隐私保护方法举例

此外,文献[21]针对有关联的两个数据集(如同一家商店的消费者位置数据和购物数据)间的相互参照引起的隐私泄漏问题,提出了高效的解决方案.但该方法对攻击者的背景知识做了限定,不能适用于公开数据多来源化的大数据时代.同时,近年来差分隐私因其可针对任意背景知识进行保护而吸引了研究学者的注意[22,23],但差分隐私的方法由于只能基于数据的统计值发布进行保护,并不适用于我们发布整个时空数据集进行分析的需求.另外一些基于云计算或众包的研究,通过将数据发布给不相互通信的攻击者保护准确数据发布中的隐私[24].但这些方法往往采用整数规划在发布数据和隐私保护中寻求平衡,实际上暴露了位置隐私.

2 问题定义和隐私保护框架

本节中,我们在被隐式收集的时空数据集上定义与攻击者的收集信息能力无关的(ε,k)隐式隐私,并证明在隐私保护过程中,最小化数据改变量是 NP-hard 问题.最后介绍发现-消除隐私保护框架.

2.1 问题定义

当某个用户 u 在时间 t 在位置 l 使用手机进行收发短信息、接打电话时,形如三元组$\langle u,l,t\rangle$的时空记录被手机移动运营商隐式收集,这些记录组成时空数据集 D.其中,每个表示用户出现的具体位置和时刻的二元组$\langle l,t\rangle$称为时空点.通过对在相同时空点出现的用户进行聚集,给定位置 loc 和时间 $time$,时空点$\langle loc,time\rangle$能够关联一个

在这个时空点共现的用户集合 $Sp_{loc,time}=\{u|\langle u,l,t\rangle\in,l=loc,t=time\}$.

由于针对同一行为,用户的时空记录根据收集方式的不同可能存在微小差异,攻击者可以使用两个参数ε_1和ε_2来分别表示时空点在时间和位置的差异.当隐式收集的数据集中有若干个(≥1)时空点在时间上相差不超过ε_1、位置上相距不超过ε_2时,攻击者可以将这些时空点视为同一时空点和公开数据集中的时空点相参照.如表1中的时空点Sp_{A,T_1}和$Sp_{A,(T_1+\varepsilon_1/2)}$,可合并为$Sp_{merge}=Sp_{A,((T_1+(T_1+\varepsilon_1/2))/2)}=\{u_1\}\cup\{u_2\}=\{u_1,u_2\}$.在本文中,为了保证隐私保护程度不因攻击者收集公开信息的能力而改变,原有时空点和合并后的时空点均被攻击者用于和公开数据集中时空点相参照,因此被称为有效时空点.

定义 1(有效时空点). 给定隐式收集的时空数据集 D,以及时空阈值ε_1和ε_2,对于任意时空点$\langle l_i,t_i\rangle\in D$,若$\exists D'\subseteq D$,使得$\forall\langle l_j,t_j\rangle\in D'$,有$|t_i-t_j|<\varepsilon_1$且$|l_i-l_j|<\varepsilon_2$,则称$\langle l_i,t_i\rangle\cup\bigcup_{\langle l_j,t_j\rangle\in D'}\langle l_j,t_j\rangle$为有效时空点.

根据定义1,表1共有6个有效时空点,分别是$Sp_{A,T_1}=\{u_1\}$,$Sp_{A,(T_1+\varepsilon_1/2)}=\{u_2\}$,$Sp_{B,T_1}=\{u_3,u_4\}$,$Sp_{C,T_2}=\{u_1,u_3\}$,$Sp_{D,T_2}=\{u_2,u_4\}$和$Sp_{merge}=\{u_1,u_2\}$.其中,前5个时空点由于它们本身的存在成为有效时空点,第6个有效时空点由Sp_{A,T_1}和$Sp_{A,(T_1+\varepsilon_1/2)}$合并而成.为了表述方便,我们将时间阈值与距离阈值合写为$\varepsilon=\{\varepsilon_1,\varepsilon_2\}$.

Table 1 Example for valid spatio-temporal point ($T_2>T_1+2\varepsilon_1$)

表 1 有效时空点示例($T_2>T_1+2\varepsilon_1$)

Userid	Location	Timestamp	Userid	Location	Timestamp
u_1	A	T_1	u_2	A	$T_1+\varepsilon_1/2$
u_3	B	T_1	u_4	B	T_1
u_1	C	T_2	u_3	C	T_2
u_2	D	T_2	u_4	D	T_2

如果一个用户u通过签到服务等方式主动公开了自己位于若干时空点集合$\{\langle l_1,t_1\rangle,\langle l_2,t_2\rangle,\ldots,\langle l_n,t_n\rangle\}$($n\geq1$)的事实,而在隐式收集到的时空数据集$D$中,这些时空点对应的用户集合$Sp_1,Sp_2,\ldots,Sp_n$满足$Sp_1\cap Sp_2\cap\ldots\cap Sp_n=\{u\}$,那么攻击者可以将公开自己位置的用户与隐式收集到的时空数据集中的用户u关联起来.这个攻击过程,要求攻击者能够收集到用户足够多的公开的时空点,我们用k来衡量攻击者收集同一用户公开的时空点个数的能力.

基于将用户发布的时空点与隐式收集到的时空数据集合中的时空点关联的过程,我们给出时空数据隐式收集中的(ε,k)隐私问题.

定义 2((ε,k)隐式隐私). 给定隐式收集的时空数据集合D、时间阈值与距离阈值$\varepsilon=\{\varepsilon_1,\varepsilon_2\}$以及攻击者攻击能力$k$,当存在$1\leq i\leq k$个有效时空点,使得$|Sp_1\cap\ldots\cap Sp_i|=1$时,我们称$\{Sp_1,\ldots,Sp_i\}$暴露了($\varepsilon$,$k$)隐式隐私.

根据定义2,当攻击者收集到了Alice在两个时空点$\langle A,T_1\rangle$,$\langle C,T_2\rangle$的公开数据时,$\{Sp_{A,T_1},Sp_{C,T_2}\}$这个时空点集就暴露了($\varepsilon$,2)隐式隐私.

2.2 隐私保护框架

首先,我们允许用户根据数据集的不同特点设置(ε,k)隐私问题中的参数.具体来说,当公开数据集是包含时空点的时效性较差的微博数据时,时间阈值、距离阈值可以比时效性较强的签到数据大;当待保护用户(群体)经常公开发布自己所在的时空点时,攻击者的攻击能力k可以设置得大一些.

值得注意的是,为了保护(ε,k)隐私,最小化需要更改的时空数据是一个NP-hard问题.

定理 1. 最小化数据改变量来保护(ε,k)隐私问题是NP-hard问题.

证明:由于最小化数据改变量来保护k匿名问题是NP-hard问题,下面我们只需证明k匿名问题可以在多项式时间内规约到(ε,k)隐私保护问题.这对应着在(ε,k)隐私保护问题中,我们将ε设定为{0,0}.这样,任何k匿名问题的解都是ε设定为{0,0}后(ε,k)隐私保护问题的解,任何ε设定为{0,0}后(ε,k)隐私保护问题的解也是k匿名问题的解.这个映射可以在多项式时间内完成.由此可以得出,保护(ε,k)隐私问题是NP-hard问题. □

基于定理1,为了高效保护(ε,k)隐私问题,我们采用发现-消除框架.如图2所示,隐私保护框架分为3个模块:

数据预处理模块、发现违反隐私模块、消除隐私泄露模块.

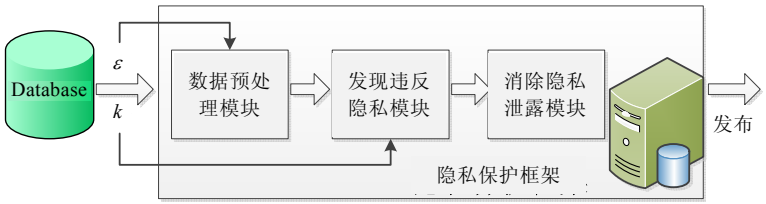


Fig.2 Implicit privacy preservation framework

图 2 隐式隐私保护框架

根据隐式隐私的定义,我们首先在数据预处理模块中根据时间阈值和距离阈值ε构建全部时空点.由于我们事先不能确定用户公开发布了哪些时空点,而大数据时代攻击者通常可以从多个社交网络来源收集到同一用户发布的时空数据.为了确保用户的隐私不被泄露,在发现违反隐私模块中,我们寻找任何由 $i(1\leqslant i\leqslant k)$个时空点组成的时空点组合,并检查攻击者是否可以匹配隐式收集到的时空数据中的某个用户.接下来,消除隐私泄露模块针对发现的时空点集合进行保护,消除隐式收集到的时空数据集对(ε,k)隐私的泄露.这样,无论用户主动发布的时空点是什么,都可以找到并消除导致(ε,k)隐私泄漏的时空数据.

3 算法与分析

本节分别介绍发现以及消除(ε,k)隐私泄露的算法.

3.1 发现(ε,k)隐私泄露

我们首先介绍发现(ε,k)隐私泄露的基于前缀过滤的嵌套循环算法,然后指出它在效率上的不足,并给出更高效的反先验算法.

3.1.1 基于前缀过滤的嵌套循环算法

根据隐私保护参数ε和 k,发现泄露(ε,k)隐私的时空点数据的基本想法是枚举全部 k 个时空点组成的集合,并检查每个组合是否与唯一用户关联.为此,基于前缀过滤的嵌套循环算法(prefix filter based nest loop,简称 PF-NL 算法),通过枚举 k 位不重复的数字实现对时空点的 k 重嵌套循环,并使用前缀过滤的方法在枚举中进行剪枝.首先,我们介绍(ε,k)隐私的重要性质.

性质 1. 记 U_k 为暴露(ε,k)隐私的全部时空点集合,并记 u_k 为大小为 k 的可以唯一关联移动对象的全部时空点集合,则 $U_k=u_1\cup\ldots\cup u_k$.

我们略去了性质 1 的平凡的证明,性质 1 说明,我们可以从大小为 1 的时空点集合开始,直到枚举全部大小不超过 k 的时空点集合.为此,我们把 n 个时空点从 1 开始编号,每个大小为 k 的时空点集合可以看成一个具有 k 位,进制为 n 的数字,且每个相同的时空点集合将 k 个代表时空点的数字按顺序排列后具有唯一的表达方式[25].

这样,大小为 $k(k>1)$的时空点集合具有长度为$\{1,\ldots,k-1\}$的前缀.例如,我们可以把表 1 中的 6 个有效时空点 Sp_{A,T_1},$Sp_{A,(T_1+\varepsilon_1/2)}$,$Sp_{B,T_1}$,$Sp_{C,T_2}$,$Sp_{D,T_2}$ 和 Sp_{merge} 分别用编号 1~6 表示.考虑(ε=0,k=3)隐私,对于大小为 3 的时空点集合{1,2,3},它具有前缀{1,2}.我们知道,$|Sp_1\cap Sp_2|=0$,因此,具有此前缀的时空点集合一定也不会暴露(0,3)隐私.在基于前缀过滤的方法中,我们避免枚举包含这种前缀的时空点集合.

结合上面的基本想法和基于前缀过滤的优化方法,我们提出了发现违反隐式隐私的基本算法 PF-NL.

算法 1. PF-NL 算法.

输入:大小为 n 的时空点集合,隐私需求(ε,k).

输出:全部违反(ε,k)隐私的时空点集合 R.

1. $num=[0,\ldots,k-1]$, $bound=[n-k+1,\ldots,n]$, $R=\varnothing$

2. **while** true **do**
3. **if** *violate*(*num*)==1 **then**
4. *R*.*add*(*num*)
5. **else**
6. *prefix*=$\min_p$*violate*(*num*[1…*p*])=0
7. **endif**
8. $p=\max_{p<prefix} num[p]<bound[p]$
9. **if** *p*<0
10. **break**
11. **endif**
12. *num*[*p*]=*num*[*p*]+1
13. *num*[*p*+1,..,*k*]=[*num*[*p*]+1,..,*num*[*p*]+$\bar{k}p$+1]
14. **endwhile**
15. **return** *R*

在算法 PF-NL 中,我们依次枚举大小为 *k* 的时空点集合 *num*,其每一位对应的最大值用 *bound* 数组表示(第 1 行).当某个时空点集合中包含的各个移动对象的集合相交后大小为 1 时,我们将它加入全部违反隐私需求的时空点集合 *R* 中(第 3 行、第 4 行).这样,产生新时空点集合的过程就是将数组中的某个元素增加的过程(第 8 行~第 13 行).值得注意的是,我们在检查某个时空点是否暴露(*ε*,*k*)隐私时计算它可能存在的前缀(第 6 行),并在产生新的时空点集合时过滤掉此前缀(第 8 行).当不能产生新的时空点时,算法 PF-NL 结束(第 10 行).

3.1.2 反先验算法

算法 PF-NL 利用前缀过滤试图在枚举中过滤掉相交为空的时空点集合,加速了违反(*ε*,*k*)隐私的时空点集合的寻找过程.但由于可以用来过滤的前缀很多,我们不能一一记录,它还是做了很多不必要的工作.下面我们介绍反先验算法,利用广度优先搜索的思想,在搜索更大的时空点集合时避免包含不违反(*ε*,*k*)隐私的较小时空点集合.

算法 2 展示了反先验算法(reverse aprior,简称 RA 算法).其中,我们用 u_i,z_i 和 c_i 分别表示大小为 *i* 的违反、不可能违反和在更大的时空点集合中有可能违反(*ε*,*k*)隐私的时空点集合.

在 RA 算法中,我们首先检查所有时空点,并把违反(*ε*,*k*)隐私的时空点加入到 u_1 中;由于时空点中一定包含移动对象,它们也全部被加入 c_1(第 1 行).

接下来,我们考虑由更多时空点组成的集合.较大的时空点集合由较小的可能违反(*ε*,*k*)隐私的时空点集合组成(第 5 行).由于大小为 *i*+1 的集合有多种组合方式,不同的组合可能导致不同大小的待检验时空点集合,为了降低时间开销,我们选择笛卡尔积最小的两个时空点集合生成它(第 4 行).同时,用来进行检查是否违反(*ε*,*k*)隐私的时空点集合不应包含更小的不可能违反(*ε*,*k*)隐私的时空点集合(第 6 行).最后,violate 将时空点集合中各个时空点包含的移动对象集合求交集,并返回其大小.那些交集大小为 1 的时空点集合违反了(*ε*,*k*)隐私,被加入 u_{i+1}(第 6 行),而交集大小为 0 的时空点集合将不可能违反(*ε*,*k*)隐私,因此加入 z_{i+1},并从 c_{i+1} 中去除(第 8 行、第 9 行).

例 1:针对表 1,我们首先检查其中包含的 6 个有效时空点独自组成的时空点集合是否有违反(*ε*,*k*)隐私的情况.首先,因为 Sp_1 和 Sp_2 各包含一个用户,u_1={1,2},c_1={3,4,5,6}.接下来,我们根据 c_1 与 c_1 的笛卡尔积生成大小为 2 的待检验的时空点集合,即{{3,4},{3,5},{3,6},{4,5},{4,6}…,{5,6}}.经检验,违反(*ε*,*k*)隐私的 4 对时空点集合是 u_2={{3,4},{3,5},{4,6},{5,6}}.而时空点集合{3,6},{4,5}中不包含相同移动对象,因此,包含它们的时空点集合不可能违反(*ε*,*k*)隐私,因此把他们从 c_2 中去掉,最终得到 c_2=∅,z_2={{3,6},{4,5}}.因此,算法 RA 至此结束.

算法 2. RA 算法

输入:隐私需求(*ε*,*k*).

输出:全部违反隐私需求的时空点集合 *R*.

1. Scan database and form u_1 and c_1
2. i=1, R=∅
3. **while** (c_i≠∅ and i≤k)
4. $w=\arg\min_w |c_w|\times|c_{i-w+1}|$
5. c_{i+1}=c_w×c_{i-w+1}
6. c_{i+1}={e∈c_{i+1}|∀f⊆e,f∉z_i,f∉z_{i-w+1}}
7. u_{i+1}={e|e∈c_{i+1},*violate*(e)=1}
8. z_{i+1}={e|e∈c_{i+1},*violate*(e)=0}
9. c_{i+1}=c_{i+1}−z_{i+1}
10. **endwhile**
11. R=u_1∪…∪u_k

值得指出的是,一般来说,时空点集合越大,我们需要检验的时空点集合就越多.但在例 1 中,当时空点集合的大小变大时,需要检查的时空点集合并没有增多,这在直观上说明了 RA 算法的有效性.

3.2 消除违反(ε,k)隐私的时空数据

在本节中,我们介绍添加假数据的方法来消除违反(ε,k)隐私的时空数据.特别地,我们分别介绍了不需要寻找违反(ε,k)隐私的时空点集合的基于频繁移动对象的假数据添加方法和基于图的可以实现较高数据效用的隐私泄露消除方法.

3.2.1 基于频繁移动对象的假数据添加方法

不考虑已找到的违反(ε,k)隐私的时空点集合,一个简单的隐私保护方法是为每个用户在其存在的各个时空点中添加具有相同用户 id 的假数据,比如对表 1 中造成(0,2)隐私泄露的进行保护.这样添加假数据后得到 $Sp_{A,T_1}=\{u_1,u_5\}$,$Sp_{A,(T_1+\varepsilon_1/2)}=\{u_2,u_6\}$,$Sp_{B,T_1}=\{u_3,u_4,u_7,u_8\}$,$Sp_{C,T_2}=\{u_1,u_3,u_5,u_7\}$,$Sp_{D,T_2}=\{u_2,u_4,u_6,u_8\}$ 和 Sp_{merge}={u_1,u_5,u_2,u_6}.这时,针对 u_5~u_7 用户的时空数据添加了 8 条假数据,不再违反(0,2)隐私.但该方法需要增加 100%的假数据.作为它的改进,我们为每个时空点添加出现最频繁的两个移动对象.

算法 3(frequent moving object,简称 FMO 算法)展示了这个基于频繁移动用户添加假数据的过程.给定唯一性隐私参数(ε,k),我们首先找到最频繁出现的两个移动对象,此过程存在大量快速算法[26];其次,我们在每个时空点集合中加入这些移动对象.

算法 3. FMO 算法.

输入:隐私需求(ε,k),时空点集合 D.

输出:满足(ε,k)隐私约束的发布数据 P.

1. P=∅
2. F={{u_1,u_2}|∄u∈{u_3,…,u_n},*count*(u)>min{*count*(u_1),*count*(u_2)}}
3. **for each** ⟨u,l,t⟩∈D
4. P.*add*(⟨u,l,t⟩)
5. P.*add*(⟨u_1,l,t⟩)
6. P.*add*(⟨u_2,l,t⟩)
7. **endfor**

3.2.2 基于图的假数据添加方法

由于没有考虑到第 3.1 节中找到的暴露唯一性隐私的时空点集合,算法 FMO 对时空点不加区分地添加假数据,导致了很多不需要添加的数据.

例 2:对于 4 个时空点:{u_1,u_6},{u_2,u_3},{u_3,u_4},{u_7,u_8},根据算法 FMO 添加假数据得到{u_1,u_6,u_2,u_3},{u_2,u_3},{u_2,u_3,u_4},{u_7,u_8,u_2,u_3},添加了 62.5%的数据.但是,第 1 个时空点显然不需要添加任何假数据.

基于第3.1节中找到的违反(ε,k)隐私的时空点集合,算法4展示了基于图的假数据添加(graph based dummy filling,简称 G-DF 算法)过程.在算法 G-DF 中,我们把位置数据看成一个图,其中每个时空点代表图中的一个节点,而如果两个时空点存在于某个暴露唯一性隐私的时空点集合中,我们就在图中为它们添一条边(第 2 行).我们只针对图中的每个连通分量运行算法 3,即寻找每个连通分量中的最频繁的两个移动对象,把它们加入到此连通分量中的每个节点中(第3行~第5行).

例3:注意到例2中违反(ε,k)隐私的只有大小为2的时空点集合$\{\{u_2,u_3\},\{u_3,u_4\}\}$,图3是例2转化成图后的情况.这样,根据算法G-DF,例2中的数据被转化成$\{u_1,u_6\},\{u_2,u_3\},\{u_2,u_3,u_4\},\{u_7,u_8\}$,相比FMO算法增加了62.5%的数据,G-DF算法只增加了12.5%的数据.

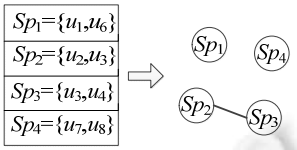


Fig.3 Transform the spatio-temportal point into graph

图3 将时空点转化成图

算法 4. G-DF 算法.

输入:暴露唯一性隐私的时空点集合 R.

输出:满足(ε,k)隐私约束的发布数据 P.

1. $P=\varnothing$
2. Transform D and R into Graph g
3. **for each** connectivity sg in g
4. Perform Algorithm 3
5. **endfor**

4 实 验

我们使用了两个真实数据集对发现-消除框架中的两种发现违反(ε,k)隐式隐私的时空数据的算法和两种消除这种违反的算法进行了性能和效果的比较.特别地,我们将通过实验回答下述问题:

(1) 隐私保护参数(ε,k)如何影响时空数据中的(ε,k)隐私;

(2) 隐私保护参数(ε,k)对违反隐私发现算法性能的影响;

(3) 隐私保护参数(ε,k)对隐私保护算法的效果和性能的影响.

4.1 实验环境与数据集描述

我们使用 Java 1.7 实现本文算法 1~算法 4,实验环境是一台 Linux 服务器,英特尔至强 E5645 2.4Ghz 处理器,128G 内存,1T SATA 硬盘.除了本文的 PF-NL 方法和 RA 方法用于发现违反(ε,k)隐私的时空点集合以及本文的 FMO 和 G-DF 方法用于消除违反(ε,k)隐私的时空点集合以外,其他对比方法包括:

- YCWA[18]:该方法是采用轨迹匿名技术保护时空数据隐私的最新方法,它将轨迹划分成停留点,并通过匿名化这些停留点保护隐私信息.该方法主要侧重轨迹隐私保护的性能.
- SEQ-ANON[19]:该方法侧重轨迹匿名技术的数据可用性,在对轨迹进行匿名化的同时,尽量减少被改变的时空点与原始时空点之间的距离.

我们将两个公开数据集 GeoSocial[27]和 GeoLife[28]当作用户被隐式收集的时空数据集进行实验.它们的数据大小和用户个数见表2.

Table 2 Dataset

表 2 数据集

数据集	记录条数	移动对象个数
GeoSocial	4M	18K
GeoLife	20M	17K

4.2 隐私保护效果对比

图 4 显示了在 GeoSocial 和 GeoLife 数据集上,在较严格的(ε,k)隐私条件下(ε_1=10min,ε_2=1km,k=10)各个方法进行时空数据隐私保护的效果.我们的方法 RA·G-DF 在发现阶段使用性能最好的 RA 方法,在消除阶段使用性能最好的 G-DF 方法.我们可以看到,经过 YCWA 和 SQL-ANON 处理后的轨迹依然有大量违反(ε,k)隐私的时空点集合.这是因为这些方法没有考虑到隐式收集到的时空数据与用户主动发布的时空数据相互参照造成的隐私泄露.

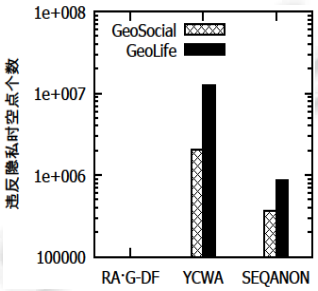


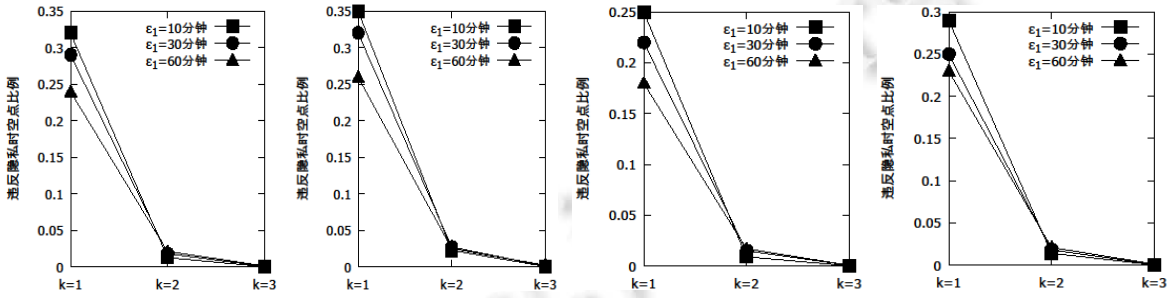
Fig.4 Comparision on effect of privacy preservation

图 4 隐私保护效果对比

4.3 (ε,k)隐私

如图 5 所示,当我们把ε_2固定而调整ε_1时,对 3 个数据集来说,随着 k 的增加,3 个数据集中违反(ε,k)隐私的时空点集合所占比例逐步下降.这种现象为第 3.1.2 节中 RA 算法的高效提供了依据.

另一方面,当ε_2增大时,图 4(b)、图 4(d)显示,两个数据集上暴露(ε,k)隐私的时空点集合的比例逐渐增加.这是因为更多的有效时空点容易和某个用户的公开位置匹配起来,进而违反(ε,k)隐私.



(a) GeoSocial 上ε_2=100m (b) GeoSocial 上ε_2=1km (c) GeoLife 上ε_2=100m (d) GeoLife 上ε_2=1km

Fig.5 Effect of ε and k on three datasets

图 5 3 个数据集上ε,k 的效果

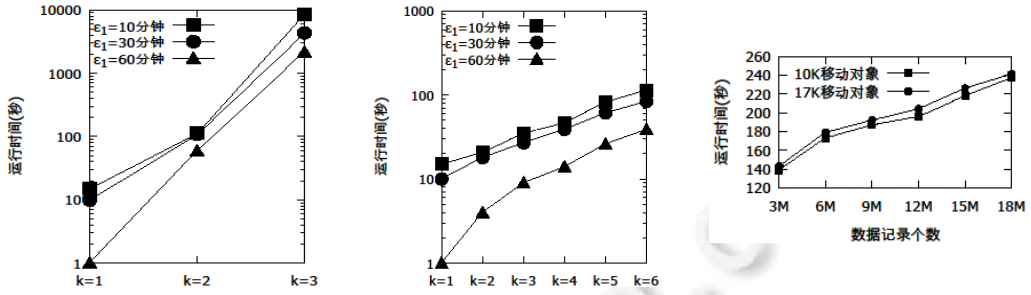
4.4 发现违反(ε,k)隐私算法性能

我们用真实数据集 GeoLife 和 GeoSocial 比较和验证我们提出的两个(ε,k)隐式隐私发现算法的运行效率.

图 6 显示了两种算法在 GeoSocial 数据集,在不同的(ε,k)下寻找出全部违反(ε,k)隐私的运行时间.从图 6(a)

和图 6(b)的比较可以看出,在相同的唯一性隐私参数(ε,k)下,RA 算法在相同隐私参数下比 PF-NL 算法快 1~2 个数量级.特别地,PF-NL 算法甚至不能计算出 k>3 的情形.

针对 RA 算法,我们调整 GeoLife 数据集的大小和移动对象个数对其性能进行了测试.图 6(c)显示,GeoLife 数据集的大小对算法的影响最大,因为它改变了数据集中时空点个数;而移动对象个数对算法的影响不大.



(a) PF-NL 算法在 GeoSocial 数据集上的性能(ε_2=1km)　(b) RA 算法在 GeoSocial 数据集上的性能(ε_2=1km)　(c) RA 算法在 GeoLife 数据集上的性能(k=3)

Fig.6　Performance comparison of algorithm PF-NL and RA

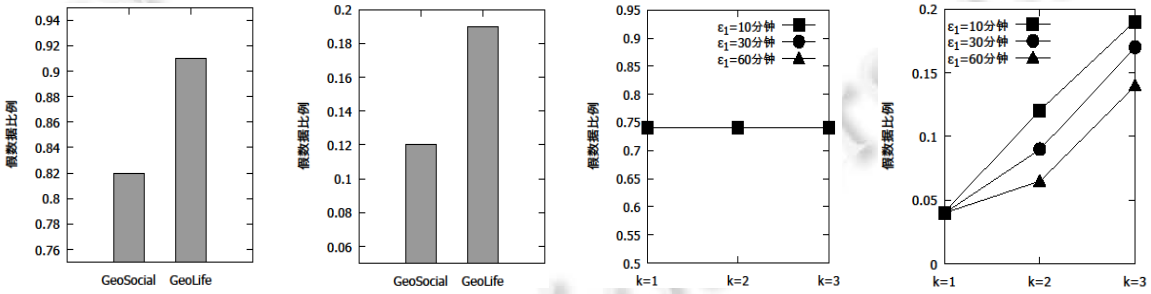
图 6　PF-NL 和 RA 算法性能对比

4.5　(ε,k)隐私保护算法效果

我们不断调整隐私参数ε和 k,查看两个隐私保护算法对数据集进行隐私保护后发布的数据效用.

如图 7(a)、图 7(b)所示,FMO 算法对这两个数据集进行保护后添加的假数据比例都超过了 80%;而 G-DF 算法可以只添加 10~20%的假数据.

图 7(c)说明,隐私参数的变化不影响 FMO 算法的数据效用.这是因为 FMO 算法只为每个时空点添加 2 个假数据.图 7(d)说明,对于 G-DF 算法,当ε_1减小时,需要添加的假数据量变大.这是因为如图 5(a)、图 5(c)所示,ε_1减小时违反(ε,k)隐私的时空点集合变多.我们还可以看出,G-DF 算法的数据效用远好于 FMO 算法.这是因为 G-DF 算法不添加不必要的假数据.



(a) FMO 算法的数据效用　(b) G-DF 算法的数据效用　(c) FMO 算法的数据效用(ε_2=1km)　(d) G-DF 算法的数据效用(ε_2=1km)

Fig.7　Performance comparison of algorithm of FMO and G-DF

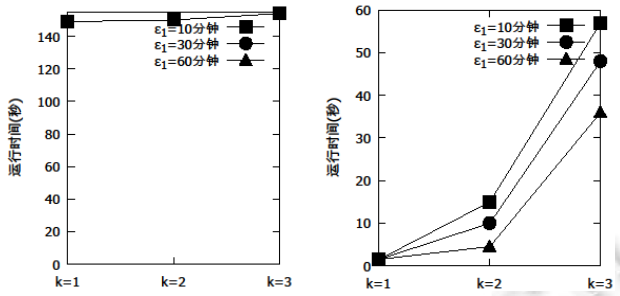
图 7　FMO 和 G-DF 算法性能对比

4.6　(ε,k)隐私保护算法性能

我们使用最大的真实数据集 GeoLife 比较两个(ε,k)隐私保护算法的运行效率.

图 8(a)展示了 FMO 对(ε,k)隐私的保护性能.可以看出,给定数据集后,隐私参数对运行时间没有影响.图 8(b)中隐私保护算法的运行时间的趋势和图 6 中违反(ε,k)隐私的时空点集合比例的变化趋势相似.这是因为算法

G-DF 需要对所有找出的违反(ε,k)隐私的时空点集合进行处理,因此其处理时间和在确定(ε,k)隐私参数(ε,k)下违反隐私的时空点集合个数正相关.



(a) FMO 算法的性能 (b) G-DF 算法的性能

Fig.8 Performance comparison of algorithm of FMO and G-DF under GeoLife dataset (ε_2=1km)

图 8 Geolife 数据集上 FMO 和 G-DF 算法性能对比(ε_2=1km)

5 结束语

本文首次针对用户主动发布的数据集与被隐式收集到的时空数据集相互参照的情况提出了时空数据中(ε,k)隐式隐私的定义,并提出了发现-消除隐私保护框架.特别地,本文提出了两种高效算法用于发现违反(ε,k)隐私的时空点集合.此外,本文提出了添加假数据的匿名保护方法.为了提高数据效用,本文进一步提出了基于图的假数据添加方法.真实数据集上的充分实验证明,本文算法是高效的.在未来的工作中,我们将进一步改进本文方法的性能.

References:

[1] Nathan E, Alex P. Reality mining: Sensing complex social systems. Journal of Personal and Ubiquitous Computing, 2006,10(4): 255−268. [doi: 10.1007/s00779-005-0046-3]

[2] Le MA, Tatem AJ, Cohen JM, Hay SI, Randell H, Patil AP, Smith DL. Travel risk, malaria importation and malaria transmission in Zanzibar. Scientific Reports, 2011,1(7364):271−275. [doi:10.1038/srep00093]

[3] Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. Quantifying the impact of human mobility on malaria. Science, 2012,338(6104):267−270. [doi: 10.1126/science.1223467]

[4] Hill S, Banser A, Berhan G, Eagle N. Reality mining Africa. In: Proc. of the AAAI Spring Symp. on Artificial Intelligence for Development. 2010. http://www.seas.upenn.edu/~ngns/docs/References/Hill%202010%20realityminingafrica.pdf

[5] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Yang Q, Agarwal D, Pei J, eds. Proc. of the KDD. New York: ACM Press, 2012. 186−194. [doi: 10.1145/2339530.2339561]

[6] Zheng K, Shang S, Yuan J, Yang Y. Towards efficient search for activity trajectories. In: Jensen CS, Jermaine CM, Zhou XF, eds. Proc. of the ICDE. Washington: IEEE Computer Society, 2013. 230−241. [doi: 10.1109/ICDE.2013.6544828]

[7] Yuan NJ, Zheng Y, Zhang L, Xie X. T-Finder: A recommender system for finding passengers and vacant taxis. IEEE Trans. on Knowledge & Data Engineering, 2013,25(10):2390−2403. [doi: 10.1109/TKDE.2012.153]

[8] Wicker SB. The loss of location privacy in the cellular age. Communications of the ACM, 2012,55(8):60−68. [doi: 10.1145/2240236.2240255]

[9] Wang L, Meng XF, Information SO. Location privacy preservation in big data era: A survey. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):693−712 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4551.htm [doi: 10.13328/j.cnki.jos.004551]

[10] Montjoye YAD, Hidalgo CA, Verleysen M, Blondel VD. Unique in the Crowd: The privacy bounds of human mobility. Open Access Publications from Université Catholique De Louvain, 2013,3(6):776−776.

[11] Bu GG, Liu L. A customizable *k*-anonymity model for protecting location privacy. In: Proc. of the Icdcs. 2004. 620−629.

[12] Cicek AE, Nergiz ME, Saygin Y. Ensuring location diversity in privacy-preserving spatio-temporal data publishing. VLDB Endowment, 2014,23(4):609−625. [doi: 10.1007/s00778-013-0342-x]

[13] Fung BCM, Wang K, Chen R, Yu PS. Privacy-Preserving data publishing: A survey of recent developments. ACM Computing Surveys, 2010,42(4):2623−2627. [doi: 10.1145/1749603.1749605]

[14] Sweeney L. *K*-Anonymity: A model for protecting privacy. Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008,10(5):557−570. [doi: 10.1142/S0218488502001648]

[15] Mokbel MF, Chow CY, Aref WG. The new casper: Query processing for location services without compromising privacy. In: Dayal U, Whang KY, *et al*., eds. Proc. of the VLDB. New York: ACM Press, 2009. 763−774.

[16] Pan X, Xu J, Meng X. Protecting location privacy against location-dependent attack in mobile services. IEEE Trans. on Knowledge & Data Engineering, 2011,24(8):1506−1519. [doi: 10.1109/TKDE.2011.105]

[17] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases. In: Alonso G, Blakeley J, Chen ALP, eds. Proc. of the ICDE. Washington: IEEE Computer Society, 2008. 376−385. [doi: 10.1109/ICDE.2008.4497446]

[18] Huo Z, Meng X, Hu H, Huang Y. You can walk alone: Trajectory privacy-preserving through significant stays protection. In: Lee SG, Peng ZY, *et al*., eds. Proc. of the DASFAA. Berlin: Springer-Verlag, 2012. 351−366. [doi: 10.1007/978-3-642-29038-1_26]

[19] Poulis G, Skiadopoulos S, Loukides G, Gkoulalas-Divanis A. Apriori-Based algorithms for k^m-anonymizing trajectory data. Trans. on Data Privacy, 2014,7(2):165−194.

[20] Domingo-Ferrer J, Trujillo-Rasua R. Microaggregation- and permutation-based anonymization of movement data. Information Sciences, 2012,208(21):55−80. [doi: 10.1016/j.ins.2012.04.015]

[21] Hu H, Xu J, On ST, Ng JKY. Privacy-Aware location data publishing. ACM Trans. on Database Systems, 2010,35(3):53−56. [doi: 10. 1145/1806907.1806910]

[22] Dwork C. Differential privacy. In: Bugliesi M, Preneel B, *et al*., eds. Proc. of the ICALP. Berlin: Springer-Verlag, 2006. 1−12. [doi: 10.1007/11787006_1]

[23] Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. VLDB Endowment, 2009,3(1):66−69. [doi: 10.14778/1920841.1920970]

[24] Rekatsinas T, Deshpande A, Machanavajjhala A. SPARSI: Partitioning sensitive data amongst multiple adversaries. VLDB Endowment, 2013,6(13):1594−1605. [doi: 10.14778/2536258.2536270]

[25] Knuth D. The art of computer programming. Vol.4, Fascicle 2: Generating all Tuples & Permutations. Addison-Wesley Professional, 2008.

[26] Metwally A, Agrawal D, El Abbadi A. Efficient computation of frequent and top-*k* elements in data streams. In: Eiter T, Libkin L, eds. Proc. of the ICDT 2005. Berlin: Springer-Verlag, 2005. 398−412. [doi: 10.1007/978-3-540-30570-5_27]

[27] Geo H, Tang J, Liu H. Addressing the cold-start problem in location recommendation using geo-social correlations. Data Mining & Knowledge Discovery, 2015,29(2):299−323. [doi: 10.1007/s10618-014-0343-4]

[28] Zheng Y, Xie X, Ma WY. GeoLife: A collaborative social networking service among user, location and trajectory. Bulletin of the Technical Committee on Data Engineering, 2010,33(2):32−39.

附中文参考文献:

[9] 王璐,孟小峰.位置大数据隐私保护研究综述.软件学报,2014,25(4):693−712. http://www.jos.org.cn/1000-9825/4551.htm [doi: 10.13328/j.cnki.jos.004551]

王璐(1986－),女,河北邢台人,博士,CCF专业会员,主要研究领域为位置隐私保护.

郭胜娜(1990－),女,硕士生,主要研究领域为位置隐私保护.

孟小峰(1964－),男,博士,教授,博士生导师,CCF会士,主要研究领域为Web数据管理,移动数据管理,XML 数据管理,云数据管理.