文章编号:1000-5641(2015)05-0096-08

保持数据可用性的细粒度轨迹隐私保护方案

熊胜超, 吴 瑕, 彭智勇
(武汉大学 计算机学院, 武汉 430072)

摘要: 轨迹数据的隐私保护近年来越来越受到重视,现有的工作很少考虑不同的隐私敏感位置之间的区别,也较少考虑不同的轨迹应用之间的区别(例如保险推销和紧急救助). 鉴于轨迹数据用途的多样性以及用户个性化的隐私需求,本文提出了一种细粒度的基于标签的轨迹数据隐私保护方案,此方案能让用户够灵活自主地控制不同隐私敏感的轨迹片段对不同轨迹应用的访问授权. 此外,考虑到大部分的隐私敏感位置都与轨迹停留相关,为了合理地隐藏轨迹中不可见的采样点,本文提出了一种将不可见的隐私敏感轨迹片段中的位置采样点,合理散布到周围频繁访问的多个位置中的方法. 实验结果表明,本文提出的方法能够在有效保护轨迹隐私的同时只引入较小的额外计算负担.

关键词: 轨迹隐私; 位置可见性; 细粒度
中图分类号: TP391 **文献标识码**: A **DOI**:10.3969/j.issn.1000-5641.2015.05.008

Fine-grained privacy-preserving framework while ensuring data usability in trajectory databases

XIONG Sheng-chao, WU Xia, PENG Zhi-yong
(*Computer School*, *Wuhan University*, *Wuhan* 430072, *China*)

Abstract: The privacy of trajectories has aroused a wide concern. In previous works, rarely have the differences between different sensitive locations been discussed, nor the differences between different applications (eg: for advertising and for emergencies). While in fact, some sensitive locations are more important and some applications ought to be granted the access. In this paper, to meet different privacy requirements and data utility requirements, we propose a fine-grained privacy-preserving framework which allows the users to specify which locations are visible to some applications and invisible to others at the same time. In addition, since most sensitive locations are relevant to stay points and a significant stay in a sensitive place may last longer than the ordinary places, we also propose an efficient approach to distribute invisible location samples along the nearby popular visit sequences. Experiment results indicate that our framework per-

收稿日期:2015-06
基金项目:武汉市创新研究团队项目(20140705040202 37)
第一作者:熊胜超,男,硕士研究生,研究方向为数据库与数据挖掘.
E-mail: shengchaoxiong91@whu.edu.cn.
第二作者:吴 瑕,女,博士研究生,研究方向为数据管理. E-mail: xiawu@whu.edu.cn.
通信作者:彭智勇,男,教授,博士生导师,研究方向为数据管理. E-mail: peng@whu.edu.cn.

forms efficiently without introducing significant performance penalties.

Key words: trajectories privacy; location visibility; fine-grained

0 引　　言

全球定位系统 GPS(Global Positioning System)、无线射频识别 RFID(Radio Frequency Identification)和无线通信设备的发展普及，让人们有可能收集到大量的车辆、人群等的移动轨迹数据（简称轨迹数据）. 这些数据在交通管理、流动性分析、路线和位置推荐等很多领域有着重要的应用. 由于移动对象的出行轨迹隐含着部分隐私敏感信息，如家庭住址、政治信仰、健康状况等，数据的采集者有责任和义务保护这些个体的用户隐私不被侵犯[1]. 因而轨迹数据的隐私保护已成为当前的一个研究热点.

移动对象的出行轨迹包含着用户不愿公开的某些敏感信息，对轨迹数据的分享可能导致这部分隐私信息的泄露. 由于不同的敏感位置包含着不同的用户隐私，用户在请求基于轨迹的服务时，可能希望轨迹中的某些隐私敏感片段(例如包含家庭住址信息)对于特定轨迹应用是可见的(例如紧急救助)，而对于其他的应用是不可见的(例如保险推销). 实际上，针对紧急援助的轨迹应用通常需要被特殊授权，因为一旦用户需要紧急援助时，保证能够被找到应该是首要的. 此外，对于包含多个隐私敏感片段的轨迹，也存在需要授权部分隐私敏感片段而保留其他敏感片段的情况. 例如对于广告推荐类的轨迹应用，用户可能既希望接收一些住址附近的商场折扣信息，但又不希望自己的健康状况被获悉.

考虑到用户个性化的轨迹隐私保护需求与轨迹数据可用性需求，本文提出了一种细粒度(能够精确到每一个隐私敏感位置)的基于标签的轨迹隐私保护方案：采用灵活的分散自主授权思想，允许用户自主地给不同的敏感位置和轨迹应用打上不同的隐私标签和隐私可见性标签，进而设置不同轨迹片段对不同轨迹应用的访问授权. 此外，鉴于大部分的隐私敏感位置都与轨迹停留相关，而移动目标在敏感位置的停留时间又通常比较长，本文还提出了一种将停留时间段内的轨迹位置采样合理分散到周围频繁访问序列中的方法.

1 相关研究

数据库中传统的基于角色的访问控制 RBAC(Role-Based Access Control)[2] 由用户集合、角色集合、对象集合、操作集合和授权集合组成. 角色被授权给用户，而对多个对象的操作授权集合被分配给角色，用户一旦获得了角色的授权，就自动获取相应对象操作的所有权限集合. 对 RBAC 的扩展包括考虑时间因素的 TRBAC[3] 和考虑地理位置因素的 GEO-RBAC[4]，以及二者结合的 LoT-RBAC[5]. 时间约束决定角色何时可以被激活而空间约束决定角色能在哪里激活. 当进行细粒度的访问控制时，大多都会面临角色爆炸的困扰.

轨迹的隐私保护是当前的一个研究热点，主要的技术手段通常可以归纳为 3 类：假轨迹干扰[6]、抑制发布[7]以及轨迹 K-匿名[8]. 为了进行数据发布，文献[6]提出了一种旋转真实轨迹，生成大量的假轨迹来干扰攻击者的方法，假轨迹的运动模式与真实轨迹非常接近，并且轨迹之间存在尽可能多的交叉. 文献[7]认为不同的攻击者可能知晓移动对象不同且不相交的轨迹片段，当轨迹片段的发布使得隐私泄露的风险高于阈值时，轨迹片段就不能发布. 由于 GPS 定位存在一定的精度误差，文献[8]提出了轨迹 (K,δ)-匿名的概念，其中

K 表示匿名区用户数量，δ 为定位精度误差，作者通过轨迹聚簇和空间转换，实现所提出的 (K,δ)-匿名模型，这是当前轨迹隐私保护中最为流行的做法. 然而，隐私保护是一个十分个性化的问题. 本文的目的在于针对多种轨迹应用以及轨迹中不同的隐私敏感片段，提出一种用户自主可控的访问授权机制，使得轨迹中部分隐私敏感片段在对某些应用可见的同时，对于其他应用不可见.

2 轨迹隐私保护方案

2.1 轨迹数据与轨迹可见性

对于移动对象 O_i，其轨迹 T_i 是一系列离散的位置采样序列的集合 $T_i=\{ID_i,(x_1,y_1,t_1),(x_2,y_2,t_2),\cdots,(x_n,y_n,t_n)\}$，其中 ID_i 为移动对象唯一标识，(x_i,y_i) 为移动对象在采样时刻 t_i 时所处在的位置，(x_i,y_i,t_i) 为轨迹 T_i 中的一个位置采样. 本文基于标签系统来实现对轨迹隐私敏感片段的访问控制[9]. 假设 Γ 是一个无限大的透明令牌的集合，标签 tag 是集合 Γ 中的一个元素，标签自身不含任何隐私敏感信息，但通常可用来与用户的某种隐私相关联，例如稳私标签 b 可以用来关联用户 Bob 的某一隐私敏感位置 H. 如果 Bob 轨迹中的位置采样点处于隐私敏感位置 H 的范围内，这些位置采样点都会自动地加上隐私标签 b. Label是标签集和 Γ 的子集，根据集合中的元素是否为另一集合的子集，不同的标签集之间可以形成偏序关系.

为了实施轨迹数据的隐私保护，本文定义了轨迹数据中一种新的授权，称作“可见性”. 用户可以为现实生活中的每一个隐私敏感位置都指定一个隐私标签 tag，用户轨迹中位置采样点只要位于隐私敏感位置的范围内，都会自动获得对应的隐私标签，这样每条轨迹 T_i 都与一个隐私标签集 ST_i 相关联，ST_i 中的元素为轨迹 T_i 中所有位置采样点隐私标签的并集. 用户为可能会用到的轨迹应用 A_j 指定隐私可见性授权标签集 SA_j，如果与隐私敏感位置关联的隐私标签在集合 SA_j 中，那么轨迹中与隐私敏感处相关的轨迹片段是可见的；反之则不可见.

定义 1 位置可见性：如果轨迹 T_i 中的位置采样点 l_{ik} 的隐私标签集 $SL_{ik}\subseteq SA_j$，那么位置采样点 l_{ik} 对轨迹应用 A_j 是可见的，否则，位置采样点 l_i 不可见.

定义 2 轨迹可见性：如果轨迹的隐私标签集 $ST_i\subseteq SA_j$，那么轨迹 T_i 对轨迹应用 A_j 是完全可见的，否则，轨迹 T_i 中含有对轨迹应用 A_j 不可见的轨迹片段.

如果轨迹 T_i 对轨迹应用 A_j 完全可见，表明应用 A_j 可以直接获得含有隐私敏感信息的原始轨迹；如果轨迹 T_i 对轨迹应用 A_j 不完全可见，则需要从原始轨迹 T_i 中隐藏不可见的轨迹片段，将不含不可见位置采样的修改轨迹 T_i^{prime} 返还给轨迹应用. 例如，图 1 中 Bob 的轨迹 T_0 只有 5 个位置采样点，但其中 3 个都含有某种隐私敏感信息(实心点表示)，分别是家庭住址(Home)、健康状况(Clinic)、工作地点(Work)，Bob 分别用 p,q,r 进行标注，整条

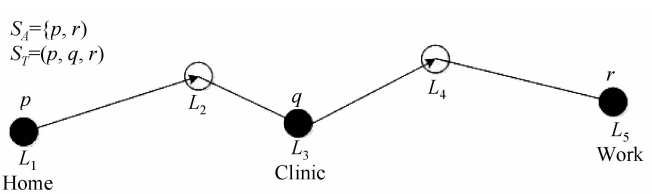

图 1 含有隐私敏感信息的轨迹片段

Fig. 1 Bob's trajectory with sensitive locations

轨迹的隐私标签集 $S_T=\{p,q,r\}$（将 ST_0 简记为 S_T）；现在 Bob 由于需要使用轨迹应用 D 来向他推荐上下班线路上商场的折扣信息，但又不希望泄漏自己的健康状况，于是将 D 的隐私可见性标签集设为 $S_A=\{p,r\}$（将 SA_D 简记为 S_A）；由于 $p\in S_A$ 和 $r\in S_A$，轨迹位置采样 L_1 和 L_5 对轨迹应用 D 都是可见的，但由于 $S_T\not\subset S_A$，轨迹对于应用 D 来说并不完全可见，因此需要将轨迹中不可见的采样点 L_3 隐藏.

为了从轨迹中隐去这些不可见的轨迹片段，简单的做法是删掉或移走这些隐私敏感的位置采样点. 但正如文献[10]中所述，用户轨迹中大部分的隐私敏感信息都与轨迹中的停留点相关，停留点往往与现实生活中的兴趣点 POI(Point of Interest)（如购物中心、俱乐部、餐厅等）相关联. 由于每个 POI 都有相对独特的访问特征，如访问频度、访问持续时长以及开放时间，移动目标在隐私敏感处的停留时长很可能远长于周围普通 POI 的访问时长，意味着需要将轨迹隐私敏感片段中的位置采样点散布到周围的多个建筑中去，而这可能会导致轨迹中语义冲突的出现. 举例来说，用户不会在白天访问只在夜晚营业的酒吧，不会长久地停留在便利店，也不会在刚吃饭的情况下前往另一个餐厅. 总结起来，轨迹修改中要遵从的语义冲突约束有如下 3 条.

约束一：修改轨迹在每个 POI 的停留时长不得长于其正常访问持续时长.

约束二：修改轨迹在每个 POI 的访问时间不能与其开放时间冲突.

约束三：修改轨迹的轨迹语义不存在明显的上下文语义冲突.

当给定一条轨迹和周围的 POI 集合 $P=\{p_1,p_2,\cdots,p_n\}$ 时，可以用 Voronoi 图来简单地提取轨迹语义[11]. 如果轨迹中的位置采样点 l 位于 p_i 所在的单元格，意为对于所有的 $p_j\in P(j\neq i)$，都有 $\mathrm{dist}(l,p_i)<\mathrm{dist}(l,p_j)$，那么就认为轨迹经过了 p_i. 为了简化起见，本文假定如果轨迹在单元格 p_i 的停留时长大于预先设置好的时间阈值，即可认为轨迹访问了 p_i，并且发生了相关的活动. 例，如图 2 所示中的轨迹经过的位置有 $\{p_1,p_2,p_4,p_6\}$，其中只有 p_4 是轨迹访问过的位置，如果 p_4 对应的是一个餐馆，就认为用户在这个餐馆吃过饭.

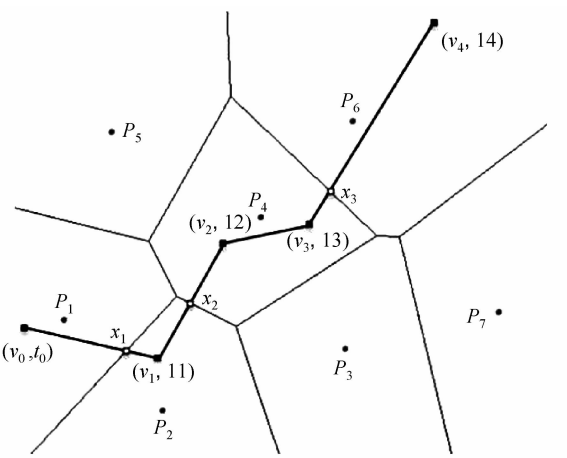

图 2　轨迹 t 的访问序列

Fig. 2　Visit sequence of trajectory t

2.2　隐私敏感位置的隐藏与标签审计

本文提出了一种将不可见轨迹片段中的多个位置采样点合理分散到周围多个普通 POI 的方法，每个 POI 都有一个合理的访问持续时长，这样就可以将轨迹中持续时间较长的多个隐私敏感片段分散到周围的多个 POI 中，同时避免语义冲突. 对应的可用隐藏路径递归

搜索过程的算法描述如下.

算法 1 可用隐藏路径的搜索算法 AFPSearch

输入:附近区域内的可用地物 place[0,…,n];隐私敏感位置 invisibleplace.
输出:可利用的频繁访问位置序列.

```
1: start ← previous place before entering the sensitive zone;
2: end ← following place after leaving the sensitive zone;
3: flag ← false;
4: advance(start,end);
5: if flag then distribute the invisible location samples along bestpath;
6: else output "No solution";
递归过程: advance(start, end)
1: if start or end exceeds the range then
2:       return;
3: end if
4: find all frequent patterns starting from start and ending at end;
5: calculate the total stayDuration of each path;
6: maxStayDuration ← the maximum stay duration among all paths;
7: if maxStayDuration < requiredStay then
8:       start = start->prev; endplace = end->next;
9:       advance(start,end);
10: else
11:       flag ← true;
12:       select a candidate path that satisfies the semantic constraints;
13:       set it as bestpath and return;
14: end if
```

路径的搜索算法 AFPSearch. 首先将轨迹进入隐私敏感区域前后经过的位置设为搜索的起始点,然后查找有没有经过前后起始点的频繁访问位置序列:如果存在,则计算频繁序列上所有 POI 的正常访问持续时间之和,并与轨迹在隐私敏感位置处的停留时长进行比较;如果没有找到对应的频繁序列,或是找到序列的持续时长不符合要求,则迭代更新搜索起始点,直到找到满足要求的可用路径,或是起始点超出界限. 由于算法只在没有找到可用路径的前提下才更新搜索的起始点,消除隐私敏感点所带来的轨迹修改能够有效地最小化.

接下来,本文将介绍上层的标签审计算法 LabelsAudit. LabelsAudit 算法将轨迹中所有对应用不可见的轨迹片段分散地隐藏到周围的普通地物中,这些不可见轨迹片段中位置采样点的特征是携带的标签都在集合 $S_E = S_T - S_A$ 中;此过程生成的修改轨迹不含对应用不可见的隐私敏感片段.

算法 2 标签审计算法 LabelsAudit

输入:原始的携带隐私敏感片段的轨迹 t_o;原始轨迹的隐私标签集 S_T;轨迹应用的隐私可见性标签集 S_A.
输出:修改后的不含不可见隐私敏感片段的轨迹 t_m.

```
1: t_m = t_o;
2: if S_T ⊆ S_A then
3:       return t_m;
4: else
5:       S_E = S_T - S_A;
6:       for each location sample with tag ∈ SE do
7:          derive all invisible location samples within the same sensitive place;
8:          call AFPSearch to find a path and distribute the location samples;
9:          update t_m;
10:      end for
11:      return t_m;
12: end if
```

使用索引机制可以加快可用频繁访问位置序列的查找. 图 3 所示介绍了本文提出的一种简单的轨迹频繁位置序列索引机制. 沿着每一个向下的分支，序列的支持度依次递减而访问持续时长之合依次递增. 当给定频繁序列的搜索起点时，算法将会沿着搜索起点所在的分支向下层次遍历搜索，直到找到匹配终止节点或叶节点. 由于频繁访问位置序列的支持度和持续时间总和都可以从索引项中获得，从而避免了对轨迹数据库的遍历操作.

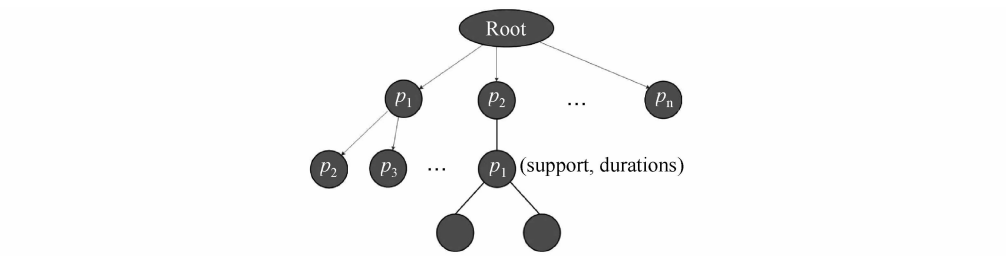

图 3　一种频繁访问位置序列索引

Fig. 3　An available index of frequent visit patterns

2.3　隐私保护与数据可用性分析

所有隐私保护面临的一个基本问题就是在用户的隐私安全和数据可用性中做出合适的折衷. 对用户的隐私进行保护无疑会给数据的可用性带来副作用，但本文实现的轨迹细粒度隐私保护能够在向用户提供灵活且个性化的隐私保护的同时，有效地将对数据可用性的影响降至最低. 现有的文献在进行轨迹隐私保护时，很少对不同的轨迹应用进行区别讨论，一旦涉及用户隐私，轨迹中所有的隐私敏感片段都要进行空间变换或删除，给轨迹数据的应用带来了较大的影响，并且也很难满足用户个性化的隐私需求.

基于分散自主授权的思想，本文所实施的隐私保护策略都是由用户定制，能够较好地满足用户的个性化隐私需求. 对于可信的轨迹应用，即使是带有用户隐私信息原始轨迹，也可以被获取，因为用户确信这些应用不会被滥用或是泄露其隐私，这种情况下的隐私策略对数据可用性基本没有影响；对于不太受信的轨迹应用，只有用户设置其不可见的轨迹隐私片段才会进行保护，再加上隐私保护时采用的贪心搜索策略，对数据可用性的影响能够降到最低. 此外，鉴于所有的轨迹修来都是基于真实轨迹而来，隐私保护对于一些只是需要聚合信息的轨迹应用来说几乎不存在影响；而对于需要使用具体的用户轨迹来提供服务的轨迹应用，隐私保护带来的副作用可能会更为明显，但相对于用户的隐私安全则是必须的.

3　实验评估

本文所采用的轨迹数据集为北京市移动用户的真实出行轨迹[11]，其数据分布如图 4 所示. 首先，通过对实验数据集进行频繁访问位置序列的挖掘，得出在相对支持度为 10% 的前提下，轨迹数据集中的可用频繁访问位置序列的数量与轨迹数据集大小的关系，如图 5 所示. 从图 5 中可以看出，随着数据集中轨迹数量的增加，频繁位置序列的数量逐渐减少直至稳定，而频繁序列挖掘的时间代价则随之递增.

通过对经过某一地区的 1 000 条轨迹进行频繁访问位置序列的挖掘，涉及 25 个地物，

且相对支持度阈值设置为 20%时,我们构建出了如图 6 所示的轨迹频繁访问位置序列索引. 针对数据集中的某条化简后的轨迹〈(P_{15}:26),(P_{14}:191),(P_{13}:38),(P_{14}:166),(P_{15}:8)〉,其含义为〈(位置:访问持续时长)〉. 假定位置 P_{13} 是一个用户隐私敏感位置,在进行轨迹隐私保护时,P_{14} 与 P_{14} 分别设置为频繁序列的搜索起点与终点;由于 P_{14} 刚好是轨迹索引中的一个频繁项,进行隐私保护后的修改轨迹将为〈(P_{15}:26),(P_{14}:395),(P_{15}:8)〉.

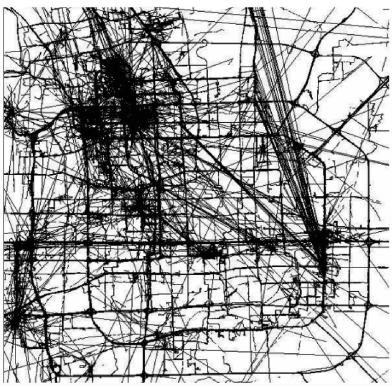

图 4 轨迹实验数据分布

Fig. 4 Snapshot of the distribution

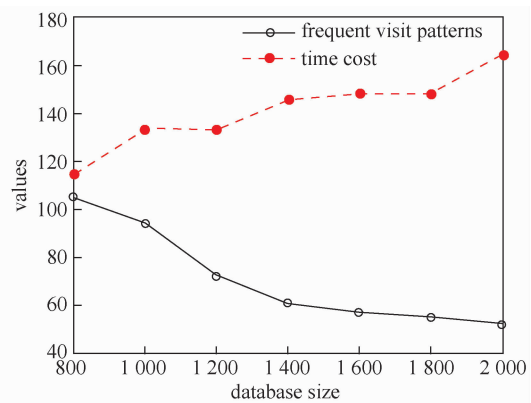

图 5 轨迹实验数据分布

Fig. 5 Patterns of various database sizes

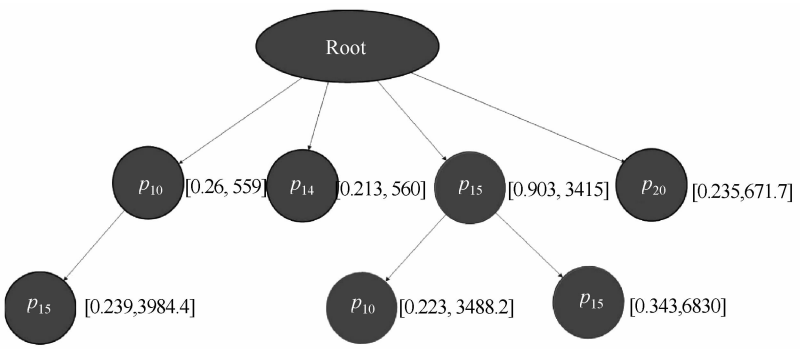

图 6 频繁访问位置序列索引实例

Fig. 6 Real example of frequent visit pattern index

4 结 论

轨迹数据的隐私保护已成为当前一个热门的研究领域,现有的轨迹隐私保护手段很难满足用户个性化的隐私需求,也很少对不同类型的轨迹应用进行区别讨论. 本文提出了一种灵活的基于标签的细粒度隐私保护方案,用户可以自主地控制不同轨迹应用对不同的隐私敏感轨迹片段的访问授权;本文还提出了一种有效隐藏轨迹中这些不可见的轨迹片段的方法,并进行了实验评估. 实验结果表明,本文提出的轨迹隐私保护方案能够在有效保护用户隐私的同时,只带来较小的额外计算负担.

[参　考　文　献]

[1] PELEKIS N, GKOULALAS-DIVANIS A, VODAS M, et al. Privacy-aware querying over sensitive trajectory data[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 895-904.

[2] FERRAIOLO D F, SANDHU R, GAVRILA S, et al. Proposed NIST standard for role-based access control[J]. ACM Transactions on Information and System Security (TISSEC), 2001, 4(3): 224-274.

[3] BERTINO E, BONATTI P A, FERRARI E. TRBAC: A temporal role-based access control model[J]. ACM Transactions on Information and System Security (TISSEC), 2001, 4(3): 191-233.

[4] BERTINO E, CATANIA B, DAMIANI M L, et al. GEO-RBAC: A spatially aware RBAC[C]//Proceedings of the 10th ACM Symposium on Access Control Models and Technologies. ACM, 2005: 29-37.

[5] CHANDRAN S M, JOSHI J B D. LoT-RBAC: A location and time-based RBAC model[M]//Web Information Systems Engineering. Berlin: Springer, 2005: 361-375.

[6] YOU T H, PENG W C, LEE W C. Protecting moving trajectories with dummies[C]//Proceedings of the 2007 International Conference on Mobile Data Management. IEEE, 2007: 278-282.

[7] TERROVITIS M, MAMOULIS N. Privacy preservation in the publication of trajectories[C]//Proceedings of the 9th International Conference on Mobile Data Management. IEEE, 2008: 65-72.

[8] ABUL O, BONCHI F, NANNI M. Never walk alone: Uncertainty for anonymity in moving objects databases [C]//Proceedings of the IEEE 24th International Conference on Data Engineering. IEEE, 2008: 376-385.

[9] KROHN M, YIP A, BRODSKY M, et al. Information flow control for standard OS abstractions[J]. ACM SIGOPS Operating Systems Review, 2007, 41(6): 321-334.

[10] HUO Z, MENG X, HU H, et al. You can walk alone: trajectory privacy-preserving through significant stays protection[M]//Database Systems for Advanced Applications. Berlin: Springer, 2012: 351-366.

[11] XIE K, DENG K, ZHOU X. From trajectories to activities: a spatio-temporal join approach[C]//Proceedings of the 2009 International Workshop on Location-Based Social Networks. ACM, 2009: 25-32.

(责任编辑　李　艺)