

基于消息复用的位置服务隐私保护算法

赵耘华, 白光伟, 沈 航

(南京工业大学 计算机科学与技术系, 江苏 南京 211816)

摘 要: 提出一种基于消息复用的位置服务隐私保护(MSPP)算法,用于提供不同安全性需求的位置服务.首先对源消息集中的数据进行预伪装,将通过预伪装算法成功伪装的消息添加到复用消息集.然后,通过消息复用,形成多层伪装区域来保护用户的位置隐私.仿真结果表明,针对不同安全性需求的消息,该机制提高了匿名成功率,缩短了系统平均响应时间.

关键词: 位置服务; 隐私; k -匿名; 消息复用

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2015)02-0148-05

A Message-Sharing-Based Privacy Preserving Algorithm for Location-Based Service

ZHAO Yun-hua, BAI Guang-wei, SHEN Hang

(Department of Computer Science and Technology, Nanjing University of Technology, Nanjing 210009, China)

Abstract: This paper proposes a message-sharing-based privacy preserving (MSPP) algorithm for providing various security requirements. And first, the data in source message set are cloaked and the cloaked data is put into the message-sharing set. Then, through message sharing, MSPP uses multi-cloaked area to protect users' location privacy. Our simulation results demonstrate that MSPP algorithm achieves high success rate along with low response time.

Key words: location-based service; privacy; k -Anonymity; message sharing

1 引言

基于位置的服务(Location-based Service, LBS)是指移动终端利用各种定位技术获得当前位置信息,再通过无线网络得到某项服务^[1].服务器在提供 LBS 服务时,需使用移动用户的位置信息,位置信息在服务过程中容易泄露或遭到攻击,这极大地限制了 LBS 应用的发展.

为保护 LBS 中移动用户的位置信息,许多位置隐私保护算法已被提出.其中,位置 k -匿名算法^[2-5]是使用较为广泛的位置隐私保护算法,其核心思想是,当用户发出查询请求时,用一个覆盖其他 $k-1$ 个用户的伪装区域来代替用户的真实位置,

从而使服务提供商无法从 k 个查询消息中分辨出用户的准确位置和身份.通过位置 k -匿名,服务器得到的不是用户的准确位置信息,而是包含用户位置的区域.

现有 LBS 研究大多基于位置匿名代理,通用的系统体系结构如图 1 所示.

文献^[2]中提出了一种用于位置服务的中间件架构和算法.对 k -匿名算法用于位置服务的技术可行性进行了分析.但该方法对于移动客户端使用同一个隐私保护等级,不接受用户根据自身需要指定其匿名等级,灵活性较低.

针对上述问题,本文提出了 MSPP,可以根据自身需求指定其期望的匿名水平 k ,从而解决文献^[2]中消

收稿日期: 2014-04-20; 修回日期: 2014-06-10

基金项目: 国家自然科学基金项目(60673185, 61073197); 江苏省自然科学基金项目(BK2010548); 江苏省科技支撑计划(工业)项目(BE2011186); 江苏省六大高峰人才基金资助项目(第八批)

息匿名成功率不高以及隐私水平不可变的问题。

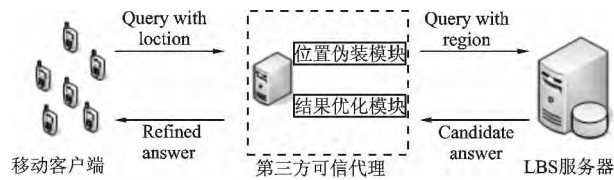


图1 LBS系统架构

2 MSPP 算法描述

2.1 网络模型和符号定义

本文采用如图1所示的系统架构.将每个用户发出的查询请求记为五元组 $m \in S = (id, \{x, y, t\}, \{d_x, d_y, d_t\}, k, C)$, 其中, id 代表提出查询请求的用户编号, $\{x, y, t\}$ 代表用户提出该消息时所处的坐标和提出的时间, $\{d_x, d_y, d_t\}$ 代表用户在寻找其他 $k-1$ 个消息时能接受的最大时间空间误差, 即, 其他 $k-1$ 个消息的时空坐标 $\{x', y', t'\}$ 必须满足 $x' \in (x - d_x, x + d_x)$, $y' \in (y - d_y, y + d_y)$, $t' \in (t - d_t, t + d_t)$. 消息的 k 值指所需的最低匿名水平, $k=1$ 表示该消息不需要匿名, $k>1$ 表示要成功伪装该消息, 必须找到其他至少 $k-1$ 个不同的消息作为伪装同伴. 因此, 较大的 k 值意味着更高层次的隐私安全需求. C 代表消息的文本内容. 同时, 为防止伪装区域过小而失去隐私保护的意义, 本文设置最小伪装区域面积限制参数 $A_{\min}$.

定义1 复用消息集 $w \in W = (ID, R, M)$: 定义已被伪装的消息集为复用消息集. 其中, ID, R, M 分别代表每个复用消息集的唯一编号, 消息集的伪装区域, 复用消息集内所包含消息 id 的集合.

定义2 邻居图 $G_i(V, E)$: $G_i(V, E)$ 是一个无向图, V 代表图中节点, E 代表边, i 代表触发伪装的节点 m_i , 对于一个移动客户端发送的消息 m_i , 若待伪装消息集中存在一个消息 m_j , 满足: (1) $L(m_i) \in b(m_j)$; (2) $L(m_j) \in b(m_i)$; (3) $m_j \neq m_i$, 则将 m_j 添加进 V , 将边 (m_i, m_j) 添加进 E .

定义3 复合约束框 $B(\{m_1, m_2, \dots, m_i, \dots\})$: B 为一个矩形, 矩形的大小为

$$x \in [\max(m_i, x - m_i, d_x), \min(m_i, x + m_i, d_x)],$$

$$y \in [\max(m_i, y - m_i, d_y), \min(m_i, y + m_i, d_y)].$$

2.2 源消息预伪装

现存的 k -匿名算法大多以先来先服务的顺序伪装消息, 寻找能够满足消息 k 值和约束条件的其他 $k-1$ 个

用户, 成功伪装并获取服务后将其删除. 这样的做法容易造成的弊端是, 场景中大量低 k 值消息为配合高 k 值用户的伪装而提高了实际匿名等级, 伪装区域也随之变大, 从而降低了查询结果的精确度.

表1 符号含义表

符号	参数意义
S	源消息集
W	复用消息集
R	伪装区域
$L(m)$	m 的时空位置
$b(m)$	m 的约束框
$B(W)$	W 的复合约束框
G_i	m_i 的完全邻居图
$n(m, G_i)$	m 在图 G_i 中的邻居
$A_{\min}$	最小伪装区域面积

由于 k 值低的用户安全性需求低, 寻找到 $k-1$ 个伪装同伴的难度低, 为了提高用户的伪装效率, 降低用户隐私等级需求和实际伪装等级之差. MSPP 首先对消息进行预伪装, 将消息按 k 值从低到高的顺序伪装, 能够保证消息的伪装等级需求, 也能够缩小较低 k 值消息的伪装集大小和伪装区域大小, 提高查询结果的精确度. 在预伪装阶段成功伪装的消息, 将在消息复用阶段被复用.

MSPP 需确保伪装集内消息的时空约束条件满足, 再考虑消息的预伪装或复用伪装. 因此, 消息预伪装和复用消息伪装阶段都需要先找到消息的完全邻居图, 算法1中描述了如何确定完全邻居图 G_i .

算法1 完全邻居图搜索 NBRGraph

```

1  Input:  $m_i$ 
2  Output:  $G_i$ 
3  foreach  $m_i \in S, m_j \in S, m_j \neq m_i$ 
4    if  $L(m_j) \in b(m_i)$  and  $L(m_i) \in b(m_j)$ 
5      add  $m_j$  into  $G(V, E)$ ;
6      add  $(m_i, m_j)$  into  $G(V, E)$ ;
7    end if
8  end for
9   $G_i(V_i, E_i) \leftarrow$  Subgraph of  $G$  consisting of  $m_i$ ;
10 sort the messages in  $G_i$  in descending order of  $m, k$ ;
11 foreach  $m \in V_i$ 
12   while  $|E_i| < (|V_i| \cdot (|V_i| - 1)) / 2$  do
13     if  $|n(m, G_i)| < |V_i| - 1$ 
14        $V_i \leftarrow V_i \setminus \{m\}$ ;
15        $E_i \leftarrow E_i \setminus \{(m, n(m, G_i))\}$ ;
16     end if
17   return  $G_i(V_i, E_i)$ 
18 end while

```

对于源消息集中不同的两个消息 $m_i (\in S)$ 和 $m_j (\in S)$, 分别考察 m_i 和 m_j 的时空坐标和对方的约束框是否存在包含关系, 若存在, 则在 m_i 和 m_j 之间建立一条边, 得到 G (对应算法1的第3~8行). 当 S 中的消息 m_i 需要伪装时, 取出 G 中包含节点 m_i 的子图 G_i . 由于一个包含 n 个节点的完全图边的个

数为 $(n \cdot (n-1))/2$, 故为保证 G_i 是一个完全图, 逐个判断 G_i 的边数是否小于 $(|V_i| \cdot (|V_i| - 1))/2$, 若小于, 则按 k 值从高到低的顺序删除不符合条件的节点, 直到 G_i 为一个完全图 (对应算法 1 的第 9 ~ 15 行). 最后, 返回完全邻居图 G_i .

得到了 m_i 的完全邻居图 G_i , G_i 中的每个消息都可以保证其时空坐标包含在其余消息的约束框内, 如果这些消息同时满足 k -匿名条件, 则他们可以组成一个伪装集一同被伪装. 算法 2 中给出了 MSPP 中的消息预处理算法, $\max_k$ 表示伪装集 w 中的最大 k 值.

算法 2 消息预处理

```

1  Input: S
2  Output: W
3   $W \leftarrow \emptyset$ 
4  sort the messages in S in increasing order of  $m.k$ ;
5  for the first item  $m_i \in S$ 
6    NBRGraph( $m_i$ );
7     $w \leftarrow V_i$ ;
8  while  $|w| < \max\_k$  and  $MBR(w) \geq A_{\min}$ 
9    drop the message with the highest  $k$  from  $w$ ;
10    $S \leftarrow S \setminus w$ ;
11    $W \leftarrow W \cup w$ 
12   update  $|W|$ ,  $\max\_k$ ;
13 end while
```

算法 2 将消息按 k 值升序排列, 消息 m_i 通过算法 1 找到了完全邻居图 G_i , 接着, 考察 G_i 中节点的 k 值, 若最大 k 值大于节点的个数, 则将节点按 k 值降序排列, 依次删除最高 k 值的节点, 直到 $|w| \geq \max_k$ 且节点集最小边界矩形大于 $A_{\min}$. 这里通过一个案例对算法 2 进行解释. 如图 2(a) 所示, 已知一个消息集 $S = \{m_1, m_2, m_3, m_4\}$, 首先将消息按 k 值升序排列, 取出第一个消息 m_1 , 由图 2(a) 可知, m_1, m_2, m_3 的空间位置点和约束图存在彼此包含关系, 故可形成完全邻居图. 如图 2(a) 所示, 图中节点集即为一个候选伪装集. 继续考察图中节点的 k 值, 由于计算结果 ($|w| = 3, \max_k = 4$), 不满足 k -匿名伪装的条件, 故删除拥有最高 k 值的节点 m_3 , 更新 $\max_k$ 和 w . 此时计算结果 ($|w| = 2, \max_k = 2$), 满足 k -匿名伪装的条件, 故伪装集为 $T = \{m_1, m_2\}$, $MBR(m_1, m_2)$ 为伪装区域.

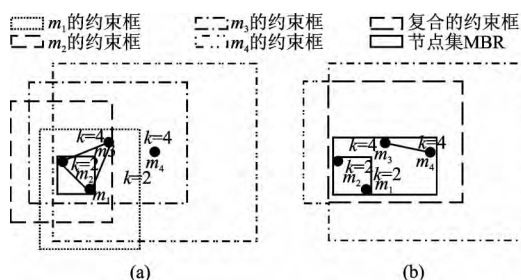


图 2 消息复用伪装机制

通过预伪装算法成功伪装的消息以集合 w 的形式保存在复用消息集 W 中, 未成功伪装的消息保留在源消息集 S 中, 在消息复用阶段, 这些消息将在复用消息集中搜索符合复用条件的消息集, 并与这些消息集一同进行伪装.

2.3 消息复用算法

在传统的位置 k -匿名隐私保护算法中, 成功伪装的消息在获取服务后被系统移除. 而在 MSPP 中, 成功伪装的消息将从 S 移入复用消息集 W 并等待复用. 对于源消息集 S 中未成功伪装的消息 m , 系统将在复用消息 W 中寻找符合条件的消息集 w , 与 m 一起伪装. 由于复用消息集 W 中的消息已被伪装, 消息在 W 中以集合的形式存在, 其准确坐标已无被伪装区域取代, 故源消息 m 寻找复用消息时, 无法在 W 中孤立地寻找单个消息, 只能寻找集合形式的消息. 当寻找到适合复用的消息集时, 直接将 m 与消息集一起伪装.

在协助其他消息伪装时, 被复用的消息已进行了伪装, 并获取了查询服务, 不需要服务器对其进行重复地服务. 因此, 在消息复用伪装阶段, W 只参与消息的伪装, 不获取服务. 对于未伪装的消息而言, 在寻找复用消息时, 只需要考虑复用消息是否在源消息 m 的约束框内, 而不需要考虑该消息是否包含于复用消息的约束框. 即在搜索复用消息集的过程中, 若复用消息中的伪装区域包含在 m 的约束框内, 则该复用消息集能够被用于 m 的伪装.

算法 3 消息复用

```

1  Input: W, S
2  Output:  $\delta$ 
3  foreach  $m_i$  in S
4    NBRGraph( $m_i$ );
5     $\delta \leftarrow V_i$ 
6    foreach  $w \subset W$ 
7      while  $|\delta| < \max\_k$ 
8        if  $MBR(w) \subset B(\delta)$ 
9           $\delta = W \cup \delta$ ;
10       end if
11     end while
12   end for
13   if  $|\delta| < \max\_k$ 
14     return 0;
15   else
16     retrun  $\delta$ 
17    $W = W \cup \delta$ ;
18 end if
19 end for
```

与消息预伪装算法相同, 消息复用算法将源消息集内的消息逐个取出, 针对每一条消息 m_i 找到 G_i , 消息在 G_i 中以节点的形式存在, 将 G_i 中包含的

所有消息添加到 $\delta.W$ 中的复用子集伪装区域必须包含在 δ 中所有消息的复合约束框 $B(\delta)$ 内,若满足,则该复用消息集能够配合 m 的伪装(对应算法3的第7~12行).那么,将该复用消息集添加到 m 的备选伪装集中.算法继续判断下一个复用消息集,直到备选伪装集中的消息个数满足 δ 的 k 值要求.

如图2(b)所示, S 中的消息 m_1, m_2 已在预伪装阶段被成功伪装, m_3, m_4 伪装不成功,仍然保留在 S 中,在消息复用伪装阶段,算法首先选取了先到达的 m_3 作为触发节点,找到完全邻居图 G_3 ,发现 G_3 的点集 $V = \{m_3, m_4\}$,于是找到 $B(\{m_3, m_4\})$.此时,在复用消息集中寻找到 m_1, m_2 所在的消息集,复用消息集 $\{m_1, m_2\}$ 的伪装区域包含在 m_3, m_4 的复合约束框 $B(\{m_3, m_4\})$ 内,于是,消息集 $\{m_1, m_2\}$ 被添加到 m_3, m_4 候选伪装集中.此时,候选伪装集的所包含的消息个数等于 m_3, m_4 的最大 k 值4,满足 k -匿名条件,说明 $\{m_1, m_2, m_3, m_4\}$ 能够被成功伪装,并作为可以被复用的消息添加到复用消息集 W 中.

3 仿真实验与结果分析

3.1 实验设计

本文通过仿真的方法对 MSPP 的性能进行分析,并与 Local- k 查询算法进行比较.实验数据为在一个 1 km^2 的区域中随机生成的 50 000 个消息,其消息分布密度与实际环境接近,与目前大多实验仿真环境消息密度类似.消息的 k 值范围为 $[2, 10]$,该匿名等级取值范围目前较流行.随机选取 k 值时,本实验分别对 $\{5, 4, 3, 2\}$ 和 $\{6, 7, 8, 9, 10\}$ 使用参数为 0.6 的 Zipf 分布随机选取.基于 Zipf 分布, $k \in [5, 7]$ 在源数据中个数最多,而 $k \in [2, 4]$ 和 $k \in [8, 10]$ 在源数据中个数相对较少,这一数据分布符合匿名水平大小的一般规律.实验将时空约束框拆分为空间约束框和时间约束度,空间容忍度用面积表示.由定义1,约束框的面积为 4 dx dy ,时间约束度为 20 s,伪装区域面积的最小限制 $A_{\min}$ 设定为面积的 0.01%,即 100 m^2 .具体实验参数如表2所示.

本实验衡量算法的标准包括平均匿名成功率、平均响应时间、平均匿名等级和平均伪装区域面积.匿名成功率越高,算法的性能越好;平均响应时间越短,算法的效率越高;平均匿名等级越接近最优匿名等级,伪装性能越好;在保证伪装质量的前提下,平均伪装区域面积越小, LBS 在提供询问查询服务时的系统开销越小,查询结果越精确.

表2 实验参数

参数名称	设置
源消息数量	50000
匿名等级 k	$[2, 10]$
平均空间约束框 4 dx dy	400 m^2
平均时间约束度 d_t	20 s
最小面积限制 $A_{\min}$	100 m^2

3.2 结果分析

本实验首先测试了 MSPP 的平均匿名成功率,然后给出了系统平均响应时间,并与 Local- k 算法进行了比较.

图3给出 k 值在不同取值情况下, MSPP 和 Local- k 的伪装成功率结果.两种算法平均匿名成功率随 k 值的增加而下降,这是因为 k 值低的消息比 k 值高的消息更容易寻找到 $k-1$ 个消息,伪装难度相对较低.从图3中可以看出,不论是源消息集所有消息的成功率,还是不同 k 值消息的成功率, MSPP 都高于 Local- k ,这是由于高 k 值消息循环复用了已成功伪装的消息,增加了成功伪装消息的数量.

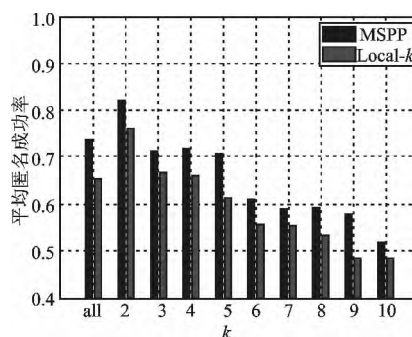


图3 平均匿名成功率对比图

图4给出了 MSPP 和 Local- k 算法处理源消息集的平均响应时间.从图中可以看出,在 k 值小于7时, MSPP 平均响应时间明显低于 Local- k 算法,这是由于本算法是在时间先后的基础上,以 k 值从低到高的顺序对源消息进行伪装,低 k 值的节点不会因为需要配合高 k 值节点的伪装而等待过长时间.而 Local- k 算法只按源消息的时间先后顺序进行伪装,每个消息为寻找伪装同伴或配合伪装同伴消耗了大量时间,导致消息伪装的响应时间变长.

图5中给出了 MSPP 与 Local- k 的平均匿名等级.图中添加了最佳匿名水平线作为平均匿名水平提供参照.从图中可以看出,平均匿名等级随 k 值的增加上升. MSPP 的平均匿名水平低于 Local- k ,且更加靠近最佳匿名水平.这是因为 MSPP 在消息到达时间顺序前提下,优先考虑了 k 值较低的消息,大多数低 k 值($k \in [2, 4]$)消息能够根据自身匿名等级要求寻找伪

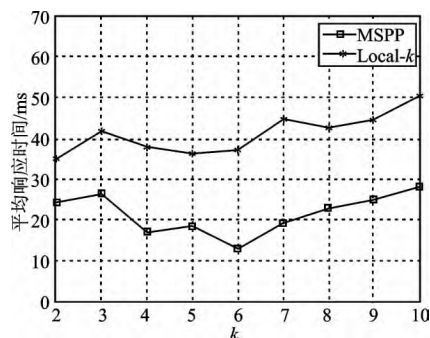


图4 平均系统响应时间对比图

装同伴,而在 Local- k 中,消息往往因为配合中高 k 值 ($k \in [5, 10]$) 消息的伪装导致自身匿名水平提高。MSPP 在平均匿名等级上的优势有效地降低了 Local- k 中可能出现的过度伪装的情况。

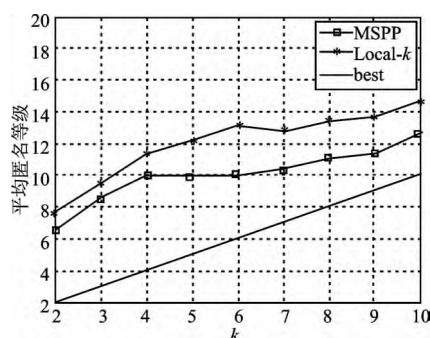


图5 平均匿名等级对比图

图6给出了 MSPP 和 Local- k 的平均伪装区域面积, MSPP 和 Local- k 的平均伪装区域面积都与 k 值成正比,同时,在用户分布密度相同的情况下,MSPP 的平均伪装区域面积小于 Local- k 。产生这种差异的原因为 MSPP 各个匿名等级对应的实际平均匿名等级比 Local- k 对应的平均匿名等级低,在用户密度相等的情况下,伪装集内消息个数的减少有效地缩小了伪装区域的面积,这有利于查询结果精准度的提高。MSPP 通过仿真产生的伪装区域面积是可以被用户接受的,基于该面积所产生的查询目标距离误差也是可以被用户接受的。

4 结束语

本文提出了基于消息复用的隐私保护算法 MSPP,该算法采用了消息复用与 k -匿名算法相结合的策略,将隐私保护算法分为预伪装和消息复用伪装两部分。仿真表明该机制提高了匿名成功率,降低了系统

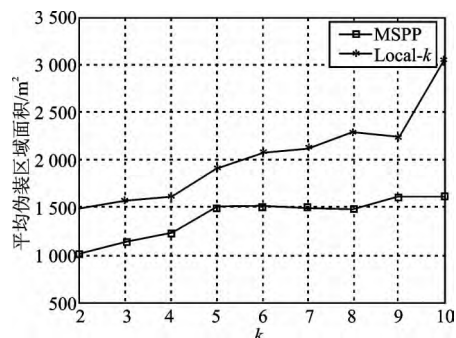


图6 平均伪装区域面积对比图

响应时间,更好地达到了用户位置隐私需求。

参考文献:

- [1] 周傲英, 杨彬, 金澈清, 等. 基于位置的服务: 架构与进展[J]. 计算机学报, 2011, 34(7): 1155-1171.
- [2] Pan X, Xu J, Meng X. Protecting location privacy against location-dependent attacks in mobile services [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1506-1519.
- [3] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking[C] // Proceedings of the 1st ACM international conference on Mobile systems, applications and services, 2003: 31-42.
- [4] Gedik B, Liu L. A customizable k-anonymity model for protecting location privacy[C] // Proceedings of the IEEE International Conference on Distributed Computing System (ICDCS'05). Columbus, Ohio, USA: IEEE, 2005: 620-629.
- [5] Mokbel M F, Chow C Y, Aref W G. The new casper: query processing for location services without compromising privacy[C] // Proceedings of the 32nd international conference on Very large data bases. Endowment: IEEE, 2006: 763-774.

作者简介:

赵耘华 女, (1990-), 硕士. 研究方向为位置服务隐私保护. E-mail: zyh625@126.com

白光伟 男, (1961-), 博士, 教授, 博士生导师. 研究方向为无线传感器网络、移动互联网、网络体系结构和协议、网络系统性能分析和评价、多媒体网络服务质量等.

沈航 男 (1984-), 博士研究生. 研究方向为无线网络编码、移动互联网、无线多媒体通信协议等.