文章编号:1008-1542(2021)03-0222-09　　开放科学(资源服务)标识码(OSID):

基于差分隐私的 LBS 用户位置隐私保护方案

于乃文[1],杨少杰[1],陈振国[2],张光华[1,3]

(1.河北科技大学信息科学与工程学院,河北石家庄　050018;2.华北科技学院河北省物联网监控工程技术研究中心,河北廊坊　065201;3.西安电子科技大学综合业务网理论及关键技术国家重点实验室,陕西西安　710071)

摘　要:为了兼顾共享位置数据的可用性和隐私保护需求,针对第三方收集的用户共享位置信息,提出了一种基于差分隐私的 LBS 用户位置隐私保护方案。首先,对共享位置数据集进行预处理,使用字典查询方式构建位置事务数据库,采用 Trie 树结构存储位置数据和频率,提高查询效率,减少加噪次数;其次,在 Trie 树上进行频繁位置选取,并使用差分隐私下的拉普拉斯机制扰动位置频率;最后,基于向上后置处理和一致性约束后置处理 2 种技术对扰动后的数据进行优化,并通过理论证明所提方案满足 ε-差分隐私。结果表明,与已有方法相比,差分隐私保护方案提高了数据处理效率,泛化了敏感位置同时具有较高的精确率和较低的拒真率。所提方案能够有效保护用户的位置隐私,同时在共享位置数据可用性方面具有一定的参考价值。

关键词:数据安全与计算机安全;位置数据;隐私保护;差分隐私;可用性

中图分类号:TP309.2　　**文献标识码**:A　　**doi**:10.7535/hbkd.2021yx03003

Location privacy protection scheme for LBS users based on differential privacy

YU Naiwen[1], YANG Shaojie[1], CHEN Zhenguo[2], ZHANG Guanghua[1,3]

(1. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China; 2. Hebei Engineering Technology Research Center for IoT Data Acquisition & Processing, North China Institute of Science and Technology, Langfang, Hebei 065201, China; 3. State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, Shaanxi 710071, China)

Abstract: In order to take into account the availability of shared location data and privacy protection requirements, aiming at the shared location information collected by the third party, a location privacy protection scheme of LBS users was proposed based on differential privacy. Firstly, the shared location data set was preprocessed, the dictionary query mode was used to

收稿日期:2021-03-29;修回日期:2021-04-30;责任编辑:王淑霞
基金项目:国家自然科学基金(62072239);河北省科技厅重点研发计划项目(18210803D)
第一作者简介:于乃文(1983—),女,河北石家庄人,硕士研究生,主要从事计算机网络管理方面的研究。
通讯作者:张光华教授。E-mail:xian_software@163.com

于乃文,杨少杰,陈振国,等.基于差分隐私的 LBS 用户位置隐私保护方案[J].河北科技大学学报,2021,42(3):222-230.
YU Naiwen, YANG Shaojie, CHEN Zhenguo, et al. Location privacy protection scheme for LBS users based on differential privacy [J]. Journal of Hebei University of Science and Technology, 2021, 42(3): 222-230.

build the location transaction database, and the trie tree structure was adopted to store the location data and frequency, so as to improve the query efficiency and reduce the number of noise. Secondly, frequent location selection was carried out in the Trie tree, and Laplacian mechanism under differential privacy was used to disturb the location frequency. Finally, the perturbed data was optimized based on the two techniques of upward post-processing and consistency constrained post-processing, and theoretically prove that the proposed scheme satisfies ε-differential privacy. The experimental results show that, compared with the existing methods, this scheme improves the efficiency of data processing, generalizes the sensitive locations, and has higher accuracy and lower rejection rate. This scheme has a certain reference value in user location privacy protection and shared location data availability.

Keywords: data security and computer security; location data; privacy protection; differential privacy; availability

随着 GPS、无线通信等技术的飞速发展，基于位置服务（LBS，location-based service）[1]给人们的生活起居、外出工作和社交活动带来了巨大变化。用户通过智能设备从应用商店可以下载基于位置的应用程序，如百度地图、谷歌地图等，在提交的位置信息和查询内容基础上获取所需的服务数据[2]。功能多样化的 LBS 应用遍及人们的日常生活，其应用场景主要有导航服务、社交网络服务、兴趣点推荐服务。然而，位置服务给用户带来诸多便利的同时，各种隐私泄露问题也层出不穷[3-4]。服务器获取了用户所使用的位置信息，进一步分析用户的居住地点、身体状况和日常行为规律，甚至跟踪用户，并将用户信息发布给第三方[5]。相对于 LBS 应用提供的便利，用户更关注个人的隐私安全。个人隐私保护水平影响用户使用 LBS 应用的积极性，因此设计强有力的位置隐私保护机制显得尤为重要。

首先，现有的 LBS 用户位置隐私保护机制大都通过 K 匿名模型[6]的扩展达到保护位置隐私的目的。然而匿名是从外部保护用户的身份而非用户位置信息本身，所以在背景知识攻击下会造成用户个人隐私的泄露[7]。差分隐私[8]给出了一种严格可证明的数据隐私保护模型，其并不关心攻击者拥有的背景知识，而是通过对输出结果进行随机扰动在保护数据隐私安全的同时，又可兼顾数据的可用性。其次，用户使用位置服务时需授权 LBS 应用以获取位置信息，这些被收集的用户位置信息经过第三方处理并发布用于数据分析[9]。常用的处理方法是匿名或使用扩展技术对位置数据集进行处理，然而此类技术难以抵御背景知识的攻击，从而造成用户隐私泄露[10]。DWORK[11]最早提出了差分隐私概念，其优势在于不受背景知识和某条数据变化的影响。此外，将差分隐私运用到位置隐私保护场景中，可通过隐私预算控制添加噪声的大小[12]。再次，待发布的共享位置数据集为 LBS 用户的频繁位置数据集。对于此类频繁模式数据挖掘问题，采用差分隐私技术合理分配隐私预算来兼顾频繁数据的可用性和隐私保护[13]。文献[14]基于广义差分隐私技术添加噪声以扰动用户的位置信息，实现了用户敏感渐进不可区分。文献[15]和文献[16]提出了一个 2 阶段的隐私预算分配策略保护频繁数据的隐私。此类方案单独处理数据效率过低，且破坏了事务数据间的联系，而频繁位置数据的项集之间存在独立且相联的数据特性。最后，频繁模式树[17]、Trie 树[18]等数据结构在处理频繁数据挖掘中，能有效维持数据项间的关系，常用于频繁数据挖掘。文献[19]使用 Trie 树存储事务数据，基于压缩感知技术对事务数据的支持度计数添加满足差分隐私约束的噪声。文献[20]在文献[19]的基础上运用差分隐私设计了工业物联网中位置大数据的隐私保护方案。文献[21]将差分隐私引入兴趣点推荐系统，保护用户位置隐私的同时，又提供了较好的兴趣点推荐效果。上述方案虽然能够兼顾位置数据的隐私保护和可用性，但均存在一个问题，即没有对扰动后的位置数据支持度计数进行后置处理，而第三方发布的 LBS 用户共享位置数据集中位置数据的支持度计数应为整数。

综上所述，第三方发布的共享位置数据集，经差分隐私后能够兼顾 LBS 用户位置数据的可用性和隐私保护，但未考虑频繁位置数据的项集之间存在独立且相联的数据特性。目前，虽然已有将 Trie 树结构引入隐私保护以维持事务数据独立且相联的特性，但在位置隐私保护方案中构建位置搜索树的效率不高，且因未对扰动后的位置数据支持度计数进行后置处理使得可用性降低。本方案引入 Trie 树结构存储位置数据和频率，基于字典的查询方式提高处理位置数据的效率。基于一致性约束后置处理对扰动后的数据进行优化，提高位置数据的可用性，使得共享位置数据更加安全。

1　基于差分隐私的位置保护方案

1.1　总体框架

近年来，互联网和通信技术的飞速发展，使得 LBS 普遍应用于人们的日常生活中。用户在使用这些应

用时产生的位置信息被LBS应用收集，用于数据分析，如频繁位置挖掘、出行规律分析等，甚至发布给第三方获取额外收益。本研究基于差分隐私设计了一种LBS用户位置隐私保护方案，实现用户敏感位置保护。

所提方案中，定义用户被收集的地理位置信息为共享位置数据，方案总体框架如图1所示。首先，对用户共享位置数据集进行预处理，匿去用户身份、查询内容等信息，将位置信息进行编号并构建位置事务数据库。位置事务数据库是包含标志符、位置项集和频繁次数共3列的表格形式，行表示不同位置属性的集合。其次，基于位置事务数据库建立Trie树，并完成频繁位置的选取，其中频繁位置表示用户此阶段经常使用的位置信息，包含用户的敏感位置。再次，采用差分隐私下的拉普拉斯机制对频繁位置的支持度进行扰动，泛化用户的敏感位置达到隐私保护的效果。最后，对扰动后的数据进行后置处理以提高位置数据的可用性，并返回处理后的数据进行查询分析。

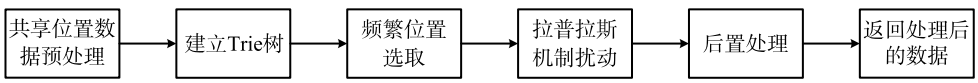


图1 总体框架

Fig. 1 Overall structure

1.2 方案设计

设计一种基于差分隐私的LBS用户位置隐私保护方案，在保护共享位置数据隐私安全的同时较好地维持了数据可用性。将位置数据预处理后构建Trie位置树，并使用指数机制进行频繁位置选取；对选取的频繁位置采用拉普拉斯机制，扰动其真实支持度，泛化敏感位置信息；对扰动后的数据进行后置处理，增强数据的可用性。

1.2.1 基于指数机制的频繁位置选取

共享位置数据集包含用户的身份、访问时间、当前位置信息、经纬度和查询条件等内容，LBS应用在用户授权下对位置信息进行分析处理或向第三方发布，需要考虑用户的隐私安全和位置数据的可用性。本方案兼顾隐私保护和位置数据可用性，首先对共享位置数据集进行预处理，为集合中的位置数据分配编号，将共享位置数据集转化为只包含编号和位置数据的集合。将此集合内容转化为位置事务数据库，并用Trie数据结构存储位置事务数据库的内容，位置事务数据库如表1示。

表1 位置事务数据库

Tab.1 Location transaction database

标志符	位置项集	频繁次数
F1-F20	{1}	20
F21-F40	{2}	20
F41-F50	{3}	10
F51-F60	{1,2}	10
F61-F70	{1,3}	10
F71-F80	{2,3}	10

当用户的位置处于字典中时，其频繁次数上升一次；当字典中查找不到该位置时，将该位置添加进字典中，并令该位置频繁次数取值为1。之后，只需扫描一次即可构建位置事务数据库，避免了在原有数据库中对位置信息的多次查找，提高了处理效率。基于表1构建位置数据的Trie树如图2示。

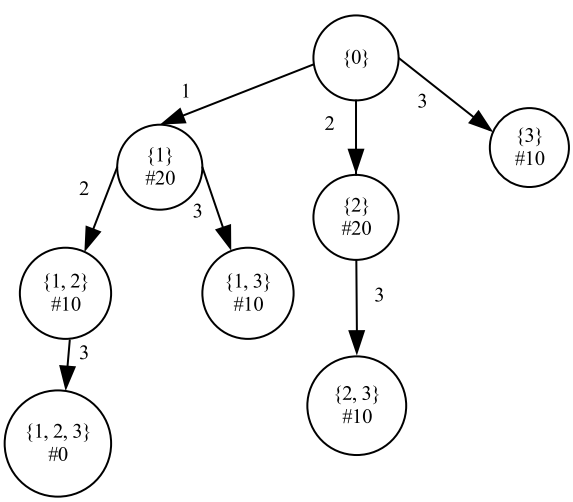


图2 位置数据的Trie树

Fig. 2 Trie tree for location data

图2包含了位置事务数据库的全部项集，节点表示位置数据，节点中的支持度表示频繁位置次数。位置Trie树维持了事务数据的相联特性，在Trie树上进行频繁位置选取效率更高。此外，拉普拉斯机制扰动的是节点的频繁次数而非共享数据集中的单个位置数据，可有效降低拉普拉斯加噪的频率并提高数据的处理效率。在频繁位置选取中，以频繁次数划分用户的敏感位置。实际应用中可融入多个参数，如使用当前位置的时间按照用户敏感时间段分配不同的权值，从而更好地区分敏感位置。基于差分隐私指数机制的频繁位置选取过程如下。

令按层遍历选出的频繁次数不小于 m 的位置数据为 N 个，其构成频繁位置集 $S=\{F_{L_1},F_{L_2},\cdots,$

$F_{L_N}\}$。差分隐私下的指数机制隐私预算为 ε_1,则前 k 个频繁位置中每次选取所分配的隐私预算为 ε_1/k。本方案中,影响位置敏感性的因素仅由位置频繁次数决定,故指数机制中打分函数可定义为位置频繁次数,如式(1)所示:

$$u(S,F_{L_i})=C_{L_i}(F_{L_i})。\tag{1}$$

式中:$u(S,F_{L_i})$ 为指数机制的打分函数;$C_{L_i}(F_{L_i})$ 表示位置 F_{L_i} 的频繁次数。

打分函数 $u(S,F_{L_i})$ 的敏感度 Δu 为 1。而每次选取所分配的隐私预算为 ε_1/k,指数机制如式(2)所示:

$$Pr(F_{L_i})\propto\exp(\frac{\varepsilon u(S,F_{L_i})}{2\Delta u})=\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})。\tag{2}$$

根据式(2)计算位置数据的归一化概率,如式(3)所示:

$$Pr(F_{L_i})=\frac{\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})},\tag{3}$$

式(3)是位置数据归一化后的指数机制输出概率值,该值与位置频繁次数成正比,通过指数机制选取前 k 个位置数据可有效保护位置的频繁次数,不被数据分析人员获取。

1.2.2　基于拉普拉斯机制的加噪处理

运用差分隐私下的指数机制思想选取前 k 个频繁位置数据,有效保护了频繁位置选取过程中的频繁次数隐私安全。然而,指数机制只用于频繁项选取,并未对位置数据进行扰动,若作为最终结果输出会造成严重的用户隐私泄露问题。因此,在频繁位置选取后,仍需采用拉普拉斯机制对选出的前 k 个频繁位置加噪,扰动其真实频繁次数,从而保护用户的隐私安全。拉普拉斯函数的概率密度如式(4)所示:

$$Pr(x,b)=\frac{1}{2b}\exp(-\frac{|x|}{b})。\tag{4}$$

令拉普拉斯机制加噪所分配的隐私预算为 ε_2,则对选取的 k 个频繁位置次数加噪,每次所分配的隐私预算为 ε_2/k。查询函数 $Q(F_L)$ 的敏感度为 ΔQ,当用 $Q(F_L)$ 表示位置频繁次数时,ΔQ 的值为 1。定义 $b=\Delta Q/(\varepsilon_2/k)=k/\varepsilon_2$。则在式(4)的基础上,拉普拉斯噪声的定义如式(5)所示:

$$\mathrm{Lap}(\frac{k}{\varepsilon_2})=\frac{\varepsilon_2}{2k}\exp(-\frac{\varepsilon_2|x|}{k})。\tag{5}$$

每次加噪扰动后的查询函数 $Q(F^{*}\ L_i)$ 如式(6)所示:

$$Q(F^{*}_{L_i})=Q(F_{L_i})+\mathrm{Lap}(\frac{k}{\varepsilon_2})。\tag{6}$$

通过拉普拉斯机制对 k 个频繁位置的次数进行扰动,泛化了用户敏感位置,从而保护了用户的隐私安全。然而,加噪后的频繁位置次数可能不为整数,若直接覆盖原位置数据会降低数据的可用性。因此,需要对其进行后置处理,以提高数据的可用性。

1.2.3　后置处理

采用一致性对数据进行后置处理,以提高数据的可用性。一致性约束将后置处理转化为一个保序回归问题,不妨令 $C_L(F^{*}_L)=\{CL_1(F^{*}_{L_1}),C_{L_2}(F^{*}\ L_2),\cdots,C_{L_k}(F^{*}_{L_k})\}$ 表示扰动后的位置频繁次数集合,同时令 $C_L(F^{\Delta}_L)=\{C_{L_1}(F^{\Delta}_{L_1}),C_{L_2}(F^{\Delta}_{L_2}),\cdots,C_{L_k}(F^{\Delta}_{L_k})\}$ 表示一致性约束处理后位置频繁次数集合。即在 $C_{L_i}(F^{\Delta}_{L_i})\geqslant C_{L_{i+1}}(F^{\Delta}_{L_{i+1}})$ 条件下,对式(7)进行求解

$$\min\sum_{i=1}^{k}(C_L(F^{*}_L)-C_L(F^{\Delta}_L))^2。\tag{7}$$

定理 1　令 $L_k=\min_{j\in[t,k]}\max_{i\in[1,j]}M[i,j]$,$U_t=\max_{i\in[1,t]}\min_{i\in[i,k]}M[i,j]$,则在式(7)和其一致性条件约束下的解为 $C_L(F^{D}_L)=\{L_1,L_2,\cdots,L_k\}=\{U_1,U_2,\cdots,U_k\}$。其中,$M[i,j]=\sum_{t=i}^{j}C_{L_t}(F^{\Delta}_{L_t})/(j-i+1)$。

基于定理 1[20]求解出一致性约束处理后的位置频繁次数集合,再对其进行向上取整处理,以满足频繁位置次数为整数的性质。例如,扰动后的位置频繁次数集合为 $C_L(F^{*}_L)=\{14.8,12.5,13.3\}$,经过一致性约束处理后位置频繁次数集合为 $C_L(F^{\Delta}_L)=\{14.8,12.9,12.9\}$,采用向上取整处理后位置频繁次数集合为 $\{15,13,13\}$。将后置处理的结果更新到共享位置数据集中进行发布,保护敏感位置的同时具有较好的位置

数据可用性。

2 安全性分析

对方案的安全可行性进行分析，证明其满足 ε-差分隐私[11]。首先对指数机制进行差分隐私分析，证明其满足 ε_1-差分隐私；其次对拉普拉斯机制进行差分隐私分析，证明其满足 ε_2-差分隐私。综合以上 2 点，本方案满足 ε-差分隐私。

定理 2 令 $S=\{F_{L_1},F_{L_2},\cdots,F_{L_N}\}$ 为频繁位置集合，差分隐私下的指数机制隐私预算为 ε_1，采用均匀分配每次选取所分配的隐私预算为 ε_1/k 。指数机制的打分函数为 $u(S,F_{L_i})$，打分函数 $u(S,F_{L_i})$ 的敏感度 Δu 为 1。则指数机制满足 ε_1-差分隐私。

证明 令 $M(S,i)(1\leqslant i\leqslant k)$ 表示第 i 次从频繁位置集合 S 中选取频繁位置 F_{L_i}；S^* 表示集合 S 的相邻数据集。则指数机制的概率比值如式(8)所示：

$$\frac{Pr[M(S,i)=F_{L_i}]}{Pr[M(S^*,i)=F_{L_i}]}=\frac{\dfrac{\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}}{\dfrac{\exp(\frac{\varepsilon_1 u(S^*,F_{L_i})}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S^*,F_{L_i})}{2k})}}=$$

$$\frac{\exp(\frac{\varepsilon_1(u(S,F_{L_i})-u(S^*,F_{L_i}))}{2k})\times\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S^*,F_{L_i})}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}\leqslant\exp(\frac{\varepsilon_1}{2k})*\frac{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S^*,F_{L_i})}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}。\tag{8}$$

由敏感度定义对式(8)化简，结果如式(9)所示：

$$\frac{Pr[M(S,i)=F_{L_i}]}{Pr[M(S^*,i)=F_{L_i}]}\leqslant\exp(\frac{\varepsilon_1}{2k})\times\frac{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S^*,F_{L_i})}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}\leqslant$$

$$\exp(\frac{\varepsilon_1}{2k})\times\frac{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1(u(S,F_{L_i})+1)}{2k})}{\sum_{i=1}^{N}\exp(\frac{\varepsilon_1 u(S,F_{L_i})}{2k})}=\exp(\frac{\varepsilon_1}{2k})\times\exp(\frac{\varepsilon_1}{2k})=\exp(\frac{\varepsilon_1}{k})。\tag{9}$$

即每一次频繁位置选取均满足 ε_1/k -差分隐私。由差分隐私的串行性质可知，本指数机制满足 ε_1-差分隐私。

定理 3 令 $S=\{F_{L_1},F_{L_2},\cdots,F_{L_N}\}$ 表示频繁位置集合，差分隐私下的拉普拉斯机制隐私预算为 ε_2，采用均匀分配每次添加拉普拉斯噪声所分配的隐私预算为 ε_2/k 。则添加噪声的过程满足 ε_2-差分隐私。

证明：不妨令扰动后的位置频繁次数集合为 $C_L(F_L^*)=\{C_{L_1}(F_{L_1}^*),C_{L_2}(F_{L_2}^*),\cdots,C_{L_k}(F_{L_k}^*)\}$。同时，不妨令扰动前的频繁次数集合为 $C_L(F_L)=\{C_{L_1}(F_{L_1}),C_{L_2}(F_{L_2}),\cdots,C_{L_k}(F_{L_k})\}$。$S^*$ 表示集合 S 的相邻数据集，$Q(F_L)$ 表示查询函数。对于每一次向 F_{L_i} 的频繁次数 $C_{L_1}(F_{L_i})$ 添加噪声扰动而言，拉普拉斯机制的概率比值如式(10)所示：

$$\frac{Pr[Q(S,F_{L_i})=C_L(F_{L_i}^*)]}{Pr[Q(S^*,F_{L_i})=C_L(F_{L_i}^*)]}=\frac{\exp(-\dfrac{\varepsilon_2\mid C_L(F_{L_i}^*)_S-C_L(F_{L_i})\mid}{k})}{\exp(-\dfrac{\varepsilon_2\mid C_L(F_{L_i}^*)_{S^*}-C_L(F_{L_i})\mid}{k})}=$$

$$\exp(\frac{\varepsilon_2}{k}\times(\mid C_L(F_{L_i}^*)_{S^*}-C_L(F_{L_i})\mid-\mid C_L(F_{L_i}^*)_S-C_L(F_{L_i})\mid))\leqslant$$

$$\exp(\frac{\varepsilon_2}{k}\times\mid C_L(F_{L_i}^*)_{S^*}-C_L(F_{L_i}^*)_S\mid)\leqslant\exp(\frac{\varepsilon_2}{k})。\tag{10}$$

综上所述，每一次加噪均满足 ε_2/k -差分隐私。由差分隐私的串行性质可知，拉普拉斯加噪过程满足 ε_2-差分隐私。故从整体上而言，满足 ε -差分隐私。

3　仿真实验及结果分析

在 Intel(R) Core(TM) i5-4210U CPU @ 1.70 GHz2.40 GHz 处理器、4 GB 内存、Windows 10 操作系统下进行仿真实验。实验代码采用 Python 3.7 编写，并于 PyCharm 上运行，所用的数据集为 Foursquare dataset[22]，该数据集经过处理后，包含 24 941 个用户，28 593 个位置兴趣点和 1 196 248 个用户的位置签到数据。以下从时间和效用性 2 个方面分析讨论方案的敏感隐私保护和数据可用性。

3.1　时间分析

时间是衡量算法效率的因素之一，从建立 Trie 树的时间和挑选出前 k 个频繁位置的时间 2 个角度进行计算分析，并与 LPT-DP-K[20] 的算法进行对比分析。先对原数据集进行预处理，使预处理后的数据集仅包含位置编号和位置信息。从预处理后的数据集中选取 100，200，500，1 000行数据，本方案(Trie)和 LPT-DP-K 方案(ergod)建立 Trie 树的时间如图 3 所示。从图 3 可看出，本方案和 LPT-DP-K 方案建立 Trie 树的时间效率都很高，且本方案优于 LPT-DP-K 方案。通过新建字典的方式构建位置数据库，当数据集位置在字典中，则位置频繁次数加 1，否则添加位置进字典中，位置频繁次数值取 1。通过查询字典，只需遍历一次数据集。而 LPT-DP-K 方案通过先选出数据集中不同的项集 N，再对每一个项集遍历数据集统计每一项的频率次数。

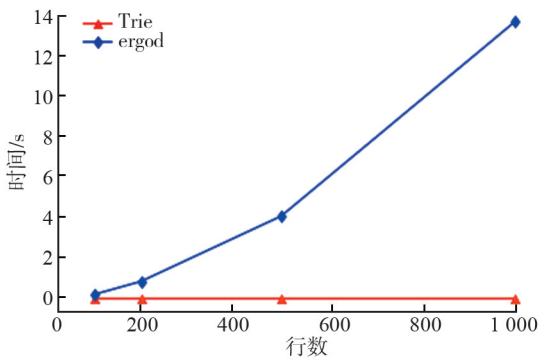


图 3　建立 Trie 树的时间对比

Fig. 3　Time comparison of establishing Trie tree

从预处理后的数据集中选取前 1 000 行位置数据建立 Trie 树，使用无放回取最大值的思想和指数机制频繁位置选取 2 种方法分别从 Trie 树上选取前 $k=[10,20,50,100]$ 个频繁位置。2 种方法选择前 k 个频繁位置的时间如图 4 所示。从图 4 可看出，使用无放回取最大值的时间效率优于指数机制频繁位置选取。第 1 种是基于无放回取最大值的思想，取 k 次后即可获取前 k 个频繁位置数据；第 2 种是基于差分隐私下的指数机制思想，先按层遍历选出频繁次数不小于 m 的位置数据，再运用指数机制进行筛选获取前 k 个频繁位置数据。第 1 种方法获取前 k 个频繁位置数据时间效率高，第 2 种方法可在选取前 k 个频繁位置数据时保护位置的频繁次数不被数据分析人员获取。

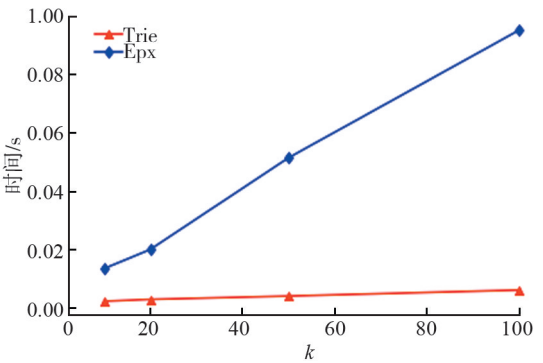


图 4　2 种方法的时间对比

Fig. 4　Time comparison of two methods

3.2　效用性分析

效用性可用来衡量算法的隐私保护有效性和位置数据的可用性。

3.2.1　有效性分析

对有效性而言，对隐私保护前后的敏感位置进行分析。从预处理后的数据集中选取前 1 000 行位置数据，取 k 频繁位置分别为 10，20，50 和 100 时对隐私保护前后的敏感位置进行对比分析，从而验证位置隐私保护方案的有效性。算法中定义访问频率 h 划分敏感位置和非敏感位置。当位置的频繁次数值不小于 h 时，则将此位置划分为敏感位置；否则，则将此位置划分为非敏感位置。通过实验可得无差分隐私[9] 和差分隐私后的敏感位置对比如表 2 所示。

表 2　敏感位置对比

Tab.2　Comparison of sensitive locations

模式	无差分隐私		差分隐私	
个数	敏感	非敏感	敏感	非敏感
10	2	8	4	6
20	4	16	8	12
50	11	39	23	27
100	21	79	51	49

为了更直观地对比分析，基于表2的数据绘制隐私保护前后的敏感位置对比图，如图5所示。由图5可知，采用差分隐私对频繁位置进行保护后敏感位置变多了，从而对原敏感位置进行泛化。当 k 为10时，敏感位置从2个变为4个；当 k 为20时，敏感位置从4个变为8个；当 k 为50时，敏感位置从11个变为23个；当 k 为100时，敏感位置从21个变为51个。

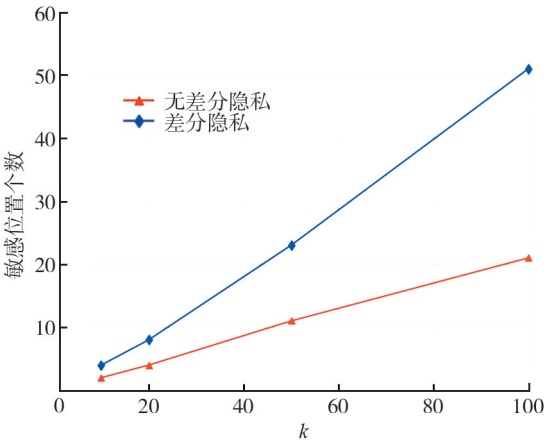
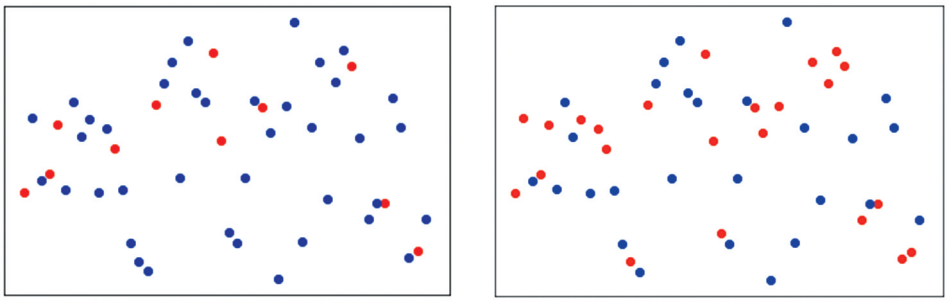


图5 隐私保护前后的敏感位置数量对比

Fig. 5 Comparison of the number of sensitive locations before and after privacy protection

图6更为直观地显示了50个模式隐私保护前后的敏感位置数量，其中蓝点表示非敏感位置，红点表示敏感位置。可见，采用拉普拉斯机制对频繁位置的次数加噪后，使得部分非敏感位置转变成敏感位置，泛化了最初的敏感位置，从而保护了用户的位置隐私。此外，经过泛化后的敏感位置集合仍包含用户敏感位置，使得第三方对位置数据集的分析依然有着较高的可用性。



a）50个模式隐私保护前的敏感位置数量图　　b）50个模式隐私保护后的敏感位置数量图

图6 50个模式隐私保护前后的敏感位置数量图

Fig. 6 Number of sensitive locations before and after privacy protection of 50 modes

3.2.2 可用性分析

对于可用性而言，采用精确率和拒真率作为本方案的评价标准。设 A 为选取的原前 k 个频繁位置的集合，B 为拉普拉斯加噪后的前 k 个频繁位置的集合。精确率定义为算法输出的频繁位置出现在集合 A 和集合 B 中的数量与集合 B 中的数量的比值，如式(11)所示。拒真率定义为算法输出的频繁位置出现在集合 A 中却未在集合 B 中的比值，如式(12)所示：

$$\text{精确率}=\frac{|A\cap B|}{|B|}, \tag{11}$$

$$\text{拒真率}=\frac{|A\cup B-B|}{K}。 \tag{12}$$

采用向上后置处理和一致性约束后置处理2种手段对加噪的位置频率进行处理。将从精确率和拒真率2个评价指标对无后置处理[19-20]、向上后置处理[21]和一致性约束后置处理3种方法进行对比分析。通过给定确切 ε 的值和变化 k 的取值计算3种方法的精确率和拒真率。实验中设 $\varepsilon=1$，参数 k 的取值分别为20，40，60，80，100。则参数 k 值变化时，3种方法的精确率和拒真率分别如图7和图8所示。

从图7可看出，随着 k 值的增加，3种后置处理方案的准确率均有所下降。使用向上后置处理和一致性约束后置处理后的准确率明显优于无后置处理的准确率。此外，随着 k 值的增加，经过一致性约束后置处理后的数据准确率的变化曲线较为平缓。当 k 取值为100时，一致性约束后置处理后的准确率依然在80%左右，从另一方面验证了对加噪数据采用后置处理，可提高数据的可用性。从图8可看出，随着 k 值的增加，3种后置处理方案的拒真率均有所提升。拒真率用来衡量算法的效用性，拒真率越小，效用性越高。使用向上后置处理和一致性约束后置处理的数据拒真率较低，具有很好的效用性。

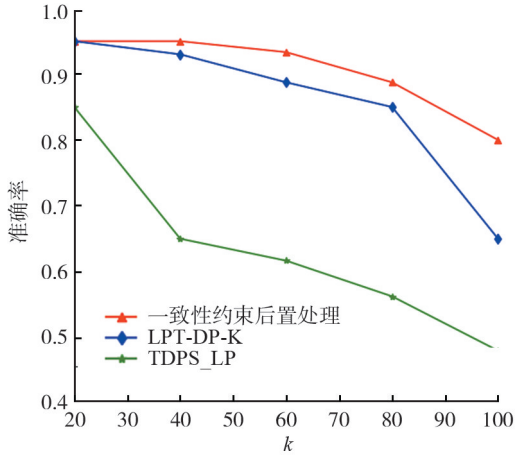


图 7　k 值变化时的精确率

Fig. 7　Accuracy rate when k value changes

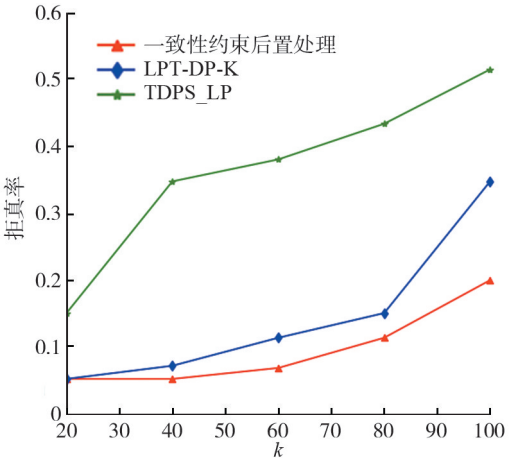


图 8　k 值变化时的拒真率

Fig. 8　Rejection rate when k value changes

通过给定确切的 k 值和变化 ε 取值范围计算 3 种方法的精确率和拒真率。实验中设 $k=200$，参数 ε 的取值为 0.05，0.10，0.50，0.75，1.00，1.50。则参数 ε 值变化时 3 种方法的精确率和拒真率分别如图 9 和图 10 所示。从图 9 可看出，随着 ε 值的增加，3 种后置处理方案的精确率均有所上升。使用向上后置处理和一致性约束后置处理后的精确率明显优于无后置处理的精确率。随着隐私预算的增加，数据可用性越高，但隐私保护力度降低。此外，当 ε 取值为 1 时，一致性约束后置处理后的准确率在 85%左右，从另一方面验证了对加噪数据采用后置处理，可提高数据的可用性。从图 10 可看出，随着 ε 值的增加，3 种后置处理方案的拒真率有所下降。拒真率用来衡量算法的效用性，拒真率越小，效用性越高。从图 10 可发现，固定 k 值和改变 ε 取值时，使用向上后置处理和一致性约束后置处理的数据拒真率较低，具有很好的效用性。因此，本方案能有效保护用户的隐私安全，并具有较好的数据可用性。

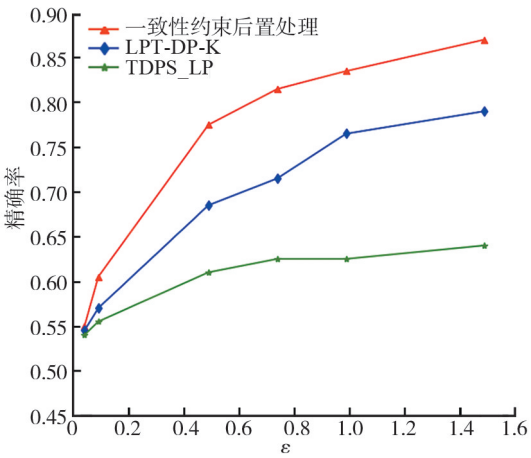


图 9　ε 值变化时的精确率

Fig. 9　Accuracy rate when ε value changes

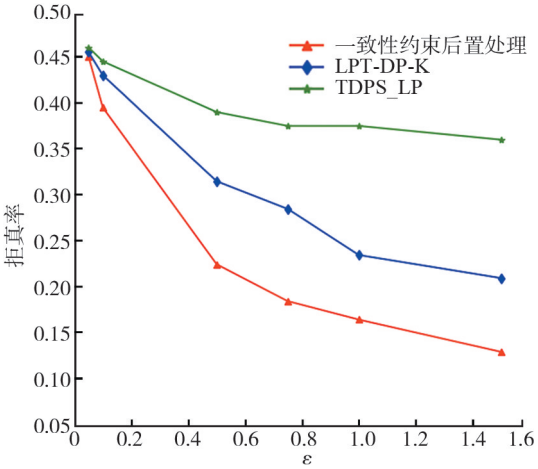


图 10　ε 值变化时的拒真率

Fig. 10　Rejection rate when ε value changes

4　结　语

为了解决共享位置数据的隐私保护和可用性不足的问题，提出了一种基于差分隐私的位置隐私保护方案。首先，采用 Trie 数据结构存储位置数据和其支持度，保持了频繁位置数据特性，提高了位置数据处理效率。其次，使用差分隐私下的拉普拉斯机制扰动频繁位置支持度，保护了用户的敏感位置。最后，对扰动后的数据进行后置处理，并通过不同的后置处理方法进行对比，提高了数据的可用性。

未来工作将围绕快速建立 Trie 树提高算法效率和考虑多因素筛选敏感位置 2 个方面开展研究，首先使用 hash 表完成位置事务数据库的统计构建，基于堆结构进行查找并完成频繁位置指数机制选取；其次综合考虑频繁次数和时间等多个位置因素，通过不同的权重，从可信计算的角度更为准确地筛选敏感位置，设计出更适用于 LBS 用户的位置隐私保护方案。

参考文献/References:

[1] WANG Jinbao,CAI Zhipeng,LI Yingshu,et al.Protecting query privacy with differentially private k-anonymity in location-based services[J].Personal and Ubiquitous Computing,2018,22(3):453-469.

[2] DARGAHI T,AMBROSIN M,CONTI M,et al.ABAKA:A novel attribute-based k-anonymous collaborative solution for LBSs[J].Computer Communications,2016,85(1):1-13.

[3] 刘海,李兴华,雒彬,等.基于区块链的分布式 K 匿名位置隐私保护方案[J].计算机学报,2019,42(5):942-960.
LIU Hai,LI Xinghua,LUO Bin,et al.Distributed K-anonymous location privacy protection scheme based on blockchain[J].Chinese Journal of Computers,2019,42(5):942-960.

[4] WANG Shengling, HU Qin, SUN Yunchuan, et al. Privacy preservation in location-based services[J]. IEEE Communications Magazine, 2018,56(3):134-140.

[5] 万盛,李凤华,牛犇,等.位置隐私保护技术研究进展[J].通信学报,2016,37(12):124-141
WAN Sheng,LI Fenghua,NIU Ben,et al.Research progress on location privacy-preserving techniques[J].Journal on Communications,2016,37(12):124-141.

[6] SWEENEY L.K-anonymity:A model for protecting privacy[J].International Journal of Uncertainty Fuzziness & Knowledge Based Systems,2002,10 (5):557-570.

[7] 张啸剑,孟小峰.面向数据发布和分析的差分隐私保护[J].计算机学报,2014,37(4):927-949.
ZHANG Xiaojian,MENG Xiaofeng.Differential privacy in data publication and analysis[J].Chinese Journal of Computers,2014,37(4):927-949.

[8] WANG Yue,YANG Lin,CHEN Xiaoyun,et al.Enhancing social network privacy with accumulated non-zero prior knowledge[J].Information Sciences,2018,445/446:6-21.

[9] ZHANG P, DURRESI M, DURRESI A. Internet network location privacy protection with multi-access edge computing[J]. Computing, 2021,103:473-490.

[10] 付钰,俞艺涵,吴晓平.大数据环境下差分隐私保护技术及应用[J].通信学报,2019,40(10):157-168.
FU Yu,YU Yihan,WU Xiaoping.Differential privacy protection technology and its application in big data environment[J].Journal on Communications,2019,40(10):157-168.

[11] DWORK C.Differential privacy[C]// Proceedings of the 33rd International Conference on Automata, Languages and Programming (ICALP 2006).Berlin:Springer,2006:1-12.

[12] GENG Q,VISWANATH P.The optimal noise-adding mechanism in differential privacy[J].IEEE Transactions on Information Theory, 2016,62(2):925-951.

[13] 丁丽萍,卢国庆.面向频繁模式挖掘的差分隐私保护研究综述[J].通信学报,2014,35(10):200-209.
DING Liping,LU Guoqing.Survey of differential privacy in frequent pattern mining[J].Journal on Communications,2014,35(10):200-209.

[14] 王斌,张磊,张国印.敏感渐进不可区分的位置隐私保护[J].计算机研究与发展,2020,57(3):616-630.
WANG Bin,ZHANG Lei,ZHANG Guoyin.A gradual sensitive indistinguishable based location privacy protection scheme[J].Journal of Computer Research and Development,2020,57(3):616-630.

[15] 卢国庆,张啸剑,丁丽萍,等.差分隐私下的一种频繁序列模式 挖掘方法[J].计算机研究与发展,2015,52(12):2789-2801.
LU Guoqing,ZHANG Xiaojian,DING Liping,et al.Frequent sequential pattern mining under differential privacy[J].Journal of Computer Research and Development,2015,52(12):2789-2801.

[16] 张啸剑,王淼,孟小峰.差分隐私保护下一种精确挖掘 top-k 频繁模式方法[J].计算机研究与发展,2014,51(1):104-114.
ZHANG Xiaojian,WANG Miao,MENG Xiaofeng.An accurate method for mining top-k frequent pattern under differential privacy[J].Journal of Computer Research and Development,2014,51(1):104-114.

[17] BODON F.A fast apriori implementation[J].Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, 2003,90:654-667.

[18] LE P,VO B,HONG P,et al.Incremental mining frequent itemsets based on the trie structure and the pre-large itemsets[C]// IEEE International Conference on Granular Computing.Taiwan:IEEE,2012:369-373.

[19] 欧阳佳,印鉴,刘少鹏,等.一种有效的差分隐私事务数据发布策略[J].计算机研究与发展,2014,51(10):2195-2205.
OUYANG Jia,YIN Jian,LIU Shaopeng,et al.An effective differential privacy transaction data publication strategy[J].Journal of Computer Research and Development,2014,51(10):2195-2205.

[20] LI C,PALANISAMY B,JOSHI J.Differentially private trajectory analysis for points-of-interest recommendation[C]// 2017 IEEE International Congress on Big Data (BigData Congress).Honolulu:IEEE,2017:49-56.

[21] YIN Chunyong,XI Jinwen,SUN Ruxia,et al.Location privacy protection based on differential privacy strategy for big data in industrial internet of things[J].IEEE Transactions on Industrial Informatics,2018,14(8):3628-3636.

[22] LI Huayu,GE Yong,HONG R,et al.Point-of-interest recommendations:Learning potential check-ins from friends[C]//The 22nd ACM SIGKDD International Conference.[S.l]:ACM,2016:975-984.