四川农业大学

本科毕业论文（设计）

（2022届）

|  |  |
| --- | --- |
| 题 目： | 基于差分隐私的位置轨迹隐私数据保护与应用 |
| 学 院： | 信息工程学院 |
|  |  |
| 专 业： | 信息管理与信息系统 |
| 学生姓名： | 陈洪海 学号： 201803897 |
| 导师： | 段旭良 职称： 副教授 |
|  |  |
|  |  |
| 完成日期： | 2022年 5 月 日 |

目 录

[1 绪论 2](#_Toc102313778)

[1.1 论文研究的背景及意义 2](#_Toc102313779)

[1.2 国内外研究现状 3](#_Toc102313780)

[1.2.1 LBS发展概述 3](#_Toc102313781)

[1.2.2 基于LBS的隐私保护研究进展 3](#_Toc102313782)

[2 基于地理位置服务的隐私数据保护常见技术与分析 3](#_Toc102313783)

[2.1 扭曲技术 3](#_Toc102313784)

[2.1.1 假名法 3](#_Toc102313785)

[2.1.2 随机化 4](#_Toc102313786)

[2.1.3 假轨迹法 5](#_Toc102313787)

[2.2 加密技术 5](#_Toc102313788)

[2.2.1 PIR方案 6](#_Toc102313789)

[2.2.2 空间转换技术 6](#_Toc102313790)

[2.3 匿名技术 7](#_Toc102313791)

[2.4 三种隐私保护技术分析 7](#_Toc102313792)

[3 差分隐私原理 8](#_Toc102313793)

[3.1 基本思想 8](#_Toc102313794)

[3.1.1 差分隐私机制 9](#_Toc102313795)

[3.1.2 地理不可区分性 9](#_Toc102313796)

[3.1.3 差分隐私实现机制 10](#_Toc102313797)

[3.3差分隐私保护架构 10](#_Toc102313798)

[3.3.1 集中式架构 10](#_Toc102313799)

[3.3.2 分布式架构 11](#_Toc102313800)

[3.3.3 混合式架构 11](#_Toc102313801)

[3.3.4 系统架构对比 12](#_Toc102313802)

[3.4 代表算法浅析 13](#_Toc102313803)

[3.4.1 集中式差分隐私 13](#_Toc102313804)

[3.4.2 本地化差分隐私 13](#_Toc102313805)

[3.4.3 缺陷与不足 14](#_Toc102313806)

[4 保护架构及算法改进 14](#_Toc102313807)

[4.1 瓶颈问题与改进挑战 14](#_Toc102313808)

[4.2 隐私保护框架改进 16](#_Toc102313809)

[4.3 轮询算法改进 19](#_Toc102313810)

[4.4 核心代码浅析 22](#_Toc102313811)

[5 改进框架及算法分析 26](#_Toc102313812)

[5.1 改进框架分析 26](#_Toc102313813)

[5.2 算法复杂性分析 27](#_Toc102313814)

[5.3 算法隐私保护效果分析 27](#_Toc102313815)

[5.4 横向对比分析 27](#_Toc102313816)

[6 仿真实验应用及结果分析 29](#_Toc102313817)

[6.1 移动对象生成器 29](#_Toc102313818)

[6.2 数据获取 31](#_Toc102313819)

[6.3 仿真过程 32](#_Toc102313820)

[6.4 仿真结果分析 32](#_Toc102313821)

[6.4.1 隐私预算ε因素分析 32](#_Toc102313822)

[6.4.2 敏感位置个数因素分析 33](#_Toc102313823)

[6.4.3 LBS服务器个数因素分析 33](#_Toc102313824)

[6.5算法应用性分析 34](#_Toc102313825)

[7 总结 38](#_Toc102313826)

[参考文献 39](#_Toc102313827)

[致谢 42](#_Toc102313828)

[附录 43](#_Toc102313829)

基于差分隐私的位置轨迹隐私数据保护

信息管理与信息系统 陈洪海

导师：段旭良

**摘要**：随着通信技术与互联网科技地迅猛发展，基于地理位置服务（location-based services，LBS）的技术逐渐地普及世界。与此同时，地理位置、移动轨迹等个人隐私数据的泄露风险也日益增加，从而引起了国内外的热烈地讨论与积极地研究。本文首先详细地对比分析了目前国内外研究时间最久的三种主流的地理位置隐私数据保护技术——位置扭曲技术、匿名技术、位置数据加密技术，阐明各自优缺点以及其适用场景。进而介绍差分隐私数据保护技术，以及相对于前文所提主流三大技术，其在地理位置数据保护的高效性与低耗性。在前人的基础上，进一步分析、改进，着重从混合式架构以及本地化两个要点进行深入研究，改进出一种适用于本地化差分隐私地理位置数据保护算法。本文同时还将实现的移动对象数据生成器，能够利用下载的路网数据集进行采点生成地理位置及轨迹数据，以此对改进后的把本地化差分隐私算法进行仿真实验。最终得出该改进后的算法在隐私保护、效率、提供服务的准确性以及空间占用上均达到预期目标的结论。

**关键词**：基于位置服务；本地化差分隐私；移动对象生成器

**Location track privacy data protection**

**based on differential privacy**

Information Management and Information System CHEN Hong-hai

Supervisor: DUAN Xu-liang

**Abstract**: The rapid development of communication technology has led to location-based services (LBS Popularizing the public has brought great convenience to people's life.But on the other hand, the risk of leakage of personal privacy data such as geographical location and mobile trajectory is also increasing, which has aroused heated discussion and active research at home and abroad. This paper first compares and analyzes in detail the three mainstream geolocation privacy data protection technologies that have been studied for the longest time at home and abroad - location distortion technology, anonymity technology, and location data encryption technology, and clarifies their advantages and disadvantages and their applicable scenarios. In addition, it introduces the differential privacy data protection technology, and the efficiency and low consumption of geographical data protection compared with the three mainstream technologies mentioned above. On the basis of the predecessors, further analysis and improvement, focusing on the two key points of hybrid architecture and localization, in-depth research, improve a data protection algorithm suitable for localization of differential privacy geographical location data protection. At the same time, the mobile object data generator that will be implemented in this paper can use the downloaded road network dataset to generate geolocation and trajectory data, so as to simulate the improved localization differential privacy algorithm. In the end, it was concluded that the improved algorithm achieved the expected goals in terms of privacy protection, efficiency and space occupation.

**Key words**: LBS (location-based services); Localization differential privacy; Mobile Object Generator

1 绪论

1.1 论文研究的背景及意义

近年来，无线通信技术取得了突飞猛进的发展，随之而来的是通讯成本的快速降低，智能终端设备的广泛普及[1]。位置服务使人们的生活更加舒适和方便，基于 LBS研发的各类软件已经成为人们生活中的热门应用[2]。LBS是依赖于用户位置信息的服务，如路径导航，查找酒店、餐馆或加油站等目的地位置[1]。然而，位置信息作为一种特殊的个人信息，可以被攻击者所利用从而获得诸如用户的偏好、行动轨迹甚至身体健康状况等隐私信息。因此，位置信息的滥用可能会造成个人隐私信息的严重泄露[3]。

差分隐私( differential privacy，DP) 于2006年被DWORK提出，基于严格的数学理论的支架[4]，以数学定理的方式对隐私保护进行了严格地定义，可以有效地防范攻击者的基于用户的背景知识的攻击[5]。此外，差分隐私并不受到用户所提交的查询信息中的某条具体数据变化的影响[3]。

基于差分隐私技术的隐私保护方案以严格的数学理论为支撑，适用于攻击者具有一定的背景知识，该技术可以与其他多种隐私保护方案相结合使用，应用前景广阔。是目前 LBS 隐私保护的主流技术[6]。

1.2 国内外研究现状

1.2.1 LBS发展概述

随着5G时代的到来，新型智能手机在世界范围内的普及，同时在数据存储与数据的运算方面取得了极大的进展。为了满足信息时代人们的美好生活的需要，世界各国学者掀起了对于LBS的讨论研究。如今，世界范围内已经实现了5G的商用与民用，地理位置服务已经全面地融入了我们的生活。同时，随之出现的LBS服务的隐私安全问题，也已经成为了网络安全领域的一个新的热门研究方向[7]。

1.2.2 基于LBS的隐私保护研究进展

当前，使用最为广泛是K-匿名技术，具体是通过将用户的真实位置与算法生成的虚拟位置映射关系为1:K实现匿名目的，从而达到隐私保护的效果。而针对于连续性的轨迹隐私数据保护，一般是通过构建匿名区域来将用户的真实出行轨迹进行混淆。如张璐等人[3]提出基于实际道路条件构建匿名区域的方法。Shuhei等人[9]使用基于估计的虚拟轨迹生成方法，其原理是通过用户之前的行进轨迹来估测用户的移动趋势从而生成多条虚拟轨迹来达到混淆的目的。而针对于匿名技术存在的时间以及空间占用较大的问题，研究者进一步提出了轨迹抑制技术——历史轨迹发布抑制和实时发布位置抑制，原理是根据用户的真实位置的敏感度选择性抑制发布[10]，从而达到对用户隐私的保护。

差分隐私( differential privacy，DP)虽问世较晚，但因其高效低耗性和拥有数学方法严格定义的攻击模型，同时量化了隐私泄露风险等优点而让其成为当前研究的主流地理位置服务隐私数据保护技术[10]。

2 基于地理位置服务的隐私数据保护常见技术与分析

2.1 扭曲技术

基于扭曲技术是的 LBS 查询保护，是指在获取用户的真实地理位置时，对于获得的信息进行必要的扰动操作以此避免非法入侵者直接获取用户的真实信息。包括假数据法（又称假名法）、随机化方法和假轨迹几类[10]。

2.1.1 假名法

假名法使用集中式架构。用户在发起了基于当前位置的服务请求时，数据的查询以及发布均在中心服务器上完成[11]。中心服务器由于被假定绝对可靠信任，又被称之为可信第三方平台（trusted third platform，TTP）。这种情况下，如果攻击者成功入侵TTP，则很容易搜集到客户的全部真实信息。为了解决客户身份和TTP查询服务之间的关联，Jong 等人[13]提出了假名法。

假名法的原理是：当用户在基于地理位置服务提交查询时，将用户的真实姓名用系统生成的假名进行替换。因此，用户所提交的查询请求不会被关联到任何可能暴露用户的真实身份以及其敏感数据的信息，通过切断用户查询与其真实身份之间的联系，达到保护用户隐私的要求。

但当攻击者掌握一定的背景知识，同时结合用户的历史消息，则存在着攻击者能够推测出用户的标识符的风险[12]。即使服务器多次修改用户的标识，依然存在着用户的偏好记录被窃取并利用到针对于用户真实身份攻击的风险。

2.1.2 随机化

随机化使用分布式架构。是指在用户基于地理位置发起服务请求时，随机化的在用户提交的信息中加入虚拟信息，然后将混合后的查询请求提交给TTP进行后续服务提供，以此达到保护用户隐私的目的[14]。而添加进入用户提交查询服务的虚假信息称为哑元，哑元主要包括哑元位置和哑元查询两类。

哑元位置是随机生成的与用户真实位置格式一致的假位置，用以在查询过程中混淆用户真实地理位置。关于哑元位置的生成，Shin 等人[15]提出的SpaceTwist方案，是当用户在发起请求时，由用户的客户端随机选定一点作为锚点，将该锚点作为用户的位置数据发送给LSP，LSP以该锚点为中心，不断扩大搜索周围的服务发起点，直至包含了客户的真实地理位置为止。具体流程如图1.

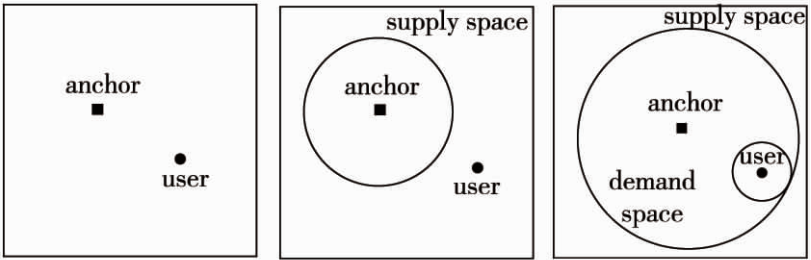


图1 SpaceTwist方案示例图

哑元查询即使添加了哑=哑元位置后发起的查询，又称为假查询。生成的哑元查询与用户的真实查询信息相似，其与用户的真实位置对应，并且添加时充分考虑到时间、空间的相关性，以此起到混淆攻击者的作用[16]。

图2为随机化示例。在图上，点A代表着用户的真实地理位置，“电影院”则代表用户提交的真实查询请求。为了达到防止用户的隐私数据泄露的目的，当用户在查询A地附近的“电影院”时，系统会就近生成两个哑元位置点B、点C，并且同时生成两个哑元查询，分别为“饭店”、“网吧”，然后将所有的信息一并发送给位置服务平台（location services platform，LSP）进行查询，而用户随后通过自己真实的LBS请求过滤LSP返回的查询结果列表，从而得到距离自己A点最近的“电影院”。

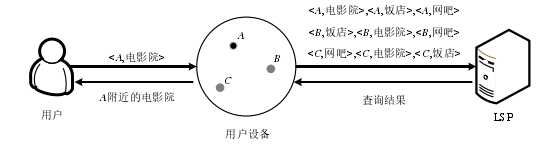


图2 随机化查询示意图

2.1.3 假轨迹法

假轨迹法采用的是分布式架构。假轨迹隐私保护技术核心依然是基于假位置法[17]。与前者仅仅保护用户单点位置隐私不同的是，假轨迹法侧重于对于用户的整个行进轨迹进行隐私保护处理。基本思想是基于用户的真实行进轨迹构造出多条类似的虚假轨迹欺骗攻击者，以此降低隐私泄露的风险[18]。最早提出来的建立在假位置的基础上的虚假轨迹法，其是根据用户前两次位置请求得到的假位置来构建虚假轨迹[20]。Wu 等人针对于用户轨迹数据之间的距离指标进行改进，对生成的虚假轨迹进行适当的扰动，使其更加接近用户真实轨迹，满足用户的需求[22]。

同时由假轨迹法衍生出来的轨迹抑制法。基本思想是删除用户提交请求中的高频、敏感位置后再进行服务发布[19]。Gruteser 等人最早提出将轨迹区域的敏感度划分。将用户的轨迹区域标记为敏感区域和非敏感区域，当户处于敏感区域时，轨迹数据将会进行相应的抑制、扰动操作[23]。轨迹抑制法存在的问题是如何在兼顾服务效率与隐私性的前提下进行合理的地理位置抑制。若是敏感区域过于繁密，抑制的敏感点过多，则会降低用户的服务体验；反之，抑制位置点没达到相当标准，则隐私保护效果较差。

2.2 加密技术

基于加密技术的LBS隐私保护操作采用分布式架构。其基本原理是：对于用户基于地理位置服务而提交的LBS查询请求进行加密保护，使得请求数据在传输过程中以及在LSP的服务器上均保持一种不可见状态。即使攻击者拦截、获取了用户所发送的请求数据，依然无法从中破解出能够表明用户的真实身份信息以及其他隐私数据[24]。基于加密的技术是将密码学与地理位置隐私被保护的融合[25]，从而更好的保证了服务的准确性以及数据的可用性。目前基于加密技术的LBS隐私数据保护技术主要有PIR方案以及空间转换技术。

2.2.1 PIR方案

PIR（private information retrieval）——隐私信息检索方案。PIR方案基于数据加密技术，通过对用户的数据进行加密处理从而达到隐私保护的目的。该方案采用的是分布式架构。方案对于隐私的保护主要基于计算能力，不需要经过中间服务器TTP，从而进一步降低数据在传输过程的中间环节泄露风险[26]。

PIR协议表示为：在服务器的数据库中，各项数据表示为二进制字符串的数据集合{Xi}。当用户向服务器请求数据Xi时，便向服务器发送加密请求q(i)，服务器在对查询请求进行分析配对后，向客户端返回加密响应r(X,q(i))，然后由用户对返回结果进行解密操作，从而得到所需X。图3表示PIR实现流程。

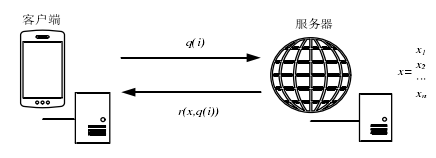


图3 PIR协议时序图

PIR方案代表算有Khoshg 等人提出了一种基于PIR方案，并且融合模糊位置查询获取服务的方法，这种方法可以明显地降低了计算和通信复杂度[27]。同时，为了解决PIR方案存在的预处理时间耗费长以及查询效率低下等缺点，Chen Xi hui等人提出了基于PIR方案的近邻查询方法[28]，利用空间加密算法使得用户能够快速实现本地化近邻查询，有效地降低了算法的预处理时间。

2.2.2 空间转换技术

基于加密技术的空间转换方案使用分布式架构，不需要中间服务器TTP[29]。其基本思想是：利用加密技术，将LBS查询信息从空间三维数据映射到一维加密空间并对转换后的数据进行隐私泄露的风险评估，确保转换后的数据与原数据的对应关系[30]。

空间转换技术代表性算法有Khoshgoz等人[31]提出的基于空间曲线变换的位置匿名算法。其将旋转后的空间利用加密密钥K创建曲线，将三维的地理位置坐标映射到曲线上。为了提高针对于轨迹的隐私保护，他们在此基础上提出了组合双曲线解析查询技术。梁慧超等人将空间曲线转换结合匿名技术[32]，提出了一种基于路网环境下的查询的隐私保护方法，降低了针对位置信息的推断攻击风险。袁健等人基于空间变换技术[33]，在优化用户轨迹划分定位算法( UTPP ) 时，将客户的地理位置进行重排序，保证了每个用户标识符的唯一性。

2.3 匿名技术

匿名技术的原理是对于用户查询请求的位置与时间利用空间泛化技术加以扰动，将单个用户之间构建成一个匿名区域，采集模糊的数据来提供服务[34]。

代表算法为K-匿名算法。该算法采用集中式机构，是由Sweeney提出，其原理是对原数据进行脱敏泛化的操作，其结果是使得该处理后的数据集中，某一任意数据与其他的剩下的所有数据区分开来。当用户的某一时刻所处的地理位置无法区分于其他K-1个移动对象时。我们可以说此位置满足K-匿名。

匿名技术的核心在于构建匿名区域[39]。若区域太大，则会极大地负荷用户设备与服务器的存储与计算能力；若构建区域过小，则难以抵抗背景攻击，难以发挥隐私保护作用[35]。

2.4 三种隐私保护技术分析

基于LBS的隐私保护技术侧重点不同，各有优缺点。选取时要结合具体应用场景及用户需要。扭曲技术一般用于以较低的开销获取很高的隐私保护效果的场景，但是存在原始数据失真的问题；当用户设备与服务器的存储与计算能力较高的情况下则采用加密的方案获得隐私保护与系统服务效果；匿名技术一般用于人流量较大时，方便构建匿名区域[36]。三种技术方案详细对比情况见下表1：

表1 三种基于LBS隐私保护方法的比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 技术分类 | 主要优点 | 主要缺点 | 代表方法 | 服务质量 |
| 基于扭曲技术 | 实现简单，平衡服务质量与隐私性保护，开销小 | 容易遭受背景攻击，数据存在失真情况 | 随机化、隐蔽化、假名技术、轨迹抑制技术 | 由生成的哑元位置所决定，同时受到用户所处环境影响 |
| 基于加密的技术 | 隐私保护效果与服务质量均较高 | 耗费大量的存储空间，同时对于设备的运算能力有很高要求 | PIR方案、空间转换技术 | 取决于终端的运算能力以及曲线查询分析技术的准确与安全性 |
| 基于匿名技术 | 构建的匿名区域可有效降低用户的隐私泄露风险，用户针对性强，算法移植性强 | 匿名区域尺度把握困难，且容易遭受重放攻击 | K-匿名技术 | 取决于匿名位置的构建以及区域内用户的数量，同时要求区域内的用户的真实性 |

3 差分隐私原理

3.1 基本思想

差分隐私（differential privacy，DP）是一种以数学理论为基础的隐私保护技术。该技术可以有效地防范攻击者的基于用户的背景知识的攻击，同时差分隐私还具备不受用户查询信息中某条具体数据的变化的影响的特点。

针对于前文提到的三种隐私保护技术，目前最具有攻击性的技术就是关于用户位置的背景知识攻击[37]。攻击者可基于一定的背景知识，构造用户位置数据集，从而表示出用户分布在数据集上每个位置的概率[38]。

差分隐私与上述所列举的包括但不限于加密技术、扭曲技术、匿名技术等方案所不同之处就在于——前者对于用户的地理位置轨迹的保护的结果仅仅是局限在一个可观察到的范围，并没有一定量的标准来对隐私保护效果加以衡量；与其不同的是，差分隐私不仅可以防范一般攻击者常常针对用户地理数据隐私发起的背景攻击，更是以数学的方法量化了隐私保护的效果，使得隐私保护具有了定量的指标来进行量化分析。

差分隐私技术虽然出现时间不长，但因其高效低耗性和拥有数学方法严格定义的攻击模型，同时可以有效防范上述所列攻击者基于背景知识的发起攻击等优点而让其成为当前研究的主流地理位置服务隐私数据保护技术[40]。

3.1.1 差分隐私机制

对于机制M，设k表示机制M的任意可能输出结果，D与D'表示相邻近的数据集（在基于LBS隐私保护场景下为用户的地理位置隐私数据集），若M满足下列公式：

则称机制M满足ε-差分隐私。其中ε称之为隐私预算。差分隐私机制的保护性跟差分隐私的隐私参数ε相关，ε越小，安全性则越高；反之，ε越大，安全性就越低。同时数据集D与D'最多只能有一条数据的差别。若是两数据集在数量上相差为一条记录，则机制M满足无界差分隐私；若两数据集所包含数据数量一致，其中仅一条数据记录不同，则机制M满足有界差分隐私[41]。如下图4：

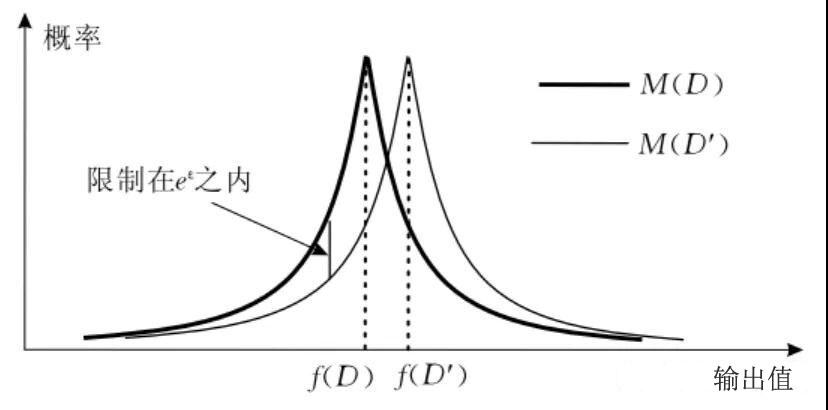


图4 差分隐私机制

3.1.2 地理不可区分性

地理不可区分性的基本思想是：对于攻击者来说，地理位置相近的两个位置x、y，如果位置扰动机制发布扰动位置的概率密度函数满足下列公式：

则称其满足参数为-的地理位置不可区分。

地理位置相近的两个位置若满足地理不可区分性机制，则其具备相似的发布概率，从而实现对于攻击者的混淆效果。位置发布概率是决定于扰动位置与用户的真实位置之间的距离。两者位置越近，则发布概率越大；扰动位置与用户真实位置距离越小，则越大[42]。

3.1.3 差分隐私实现机制

关于差分隐私安全机制有以下两种技术，分别是拉普拉斯机制以及指数机制。前者一般用于对于返回数值型查询结果的保护，而后者偏重于对于非数值型的保护[41]。前者是通过向源数据添加拉普拉斯噪声进行扰动从而达到隐私保护效果，而后者则是通过添加服指数分布的噪声。

3.3差分隐私保护架构

3.3.1 集中式架构

基于集中式架构的地理位置隐私保护系统又被称之为中心服务器结构，集中式架构是由LBS服务平台LSP、客户端设备以及中心服务器（位置匿名服务器）三部分组成[42]。图5所示：

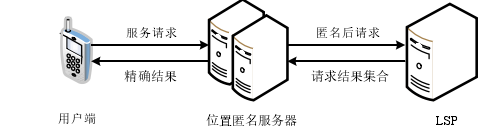


图5 集中式保护架构

匿名服务器在整个基于LBS发起的查询中，主要的实现功能包括用户位置的匿名化处理以及对于查询结果的精炼。在用户提起了基于其真实地理位置信息的服务请求后，位置匿名服务器的匿名化处理功能则将用户的请求信息进行匿名化处理，然后发送给地理位置服务平台LSP。地理位置服务平台则开始处理这一匿名请求，随后返回处理之后的服务请求结果数据集给位置匿名服务器，接着再由位置匿名服务器将返回结果精炼之后发回用户端[43]。至此，一个查询查询流程结束[43]。

3.3.2 分布式架构

分布式架存在的前提是能够保证参与在这个架构中的每一个设备，必须同时具备一定的计算与存储能力。当采用分布式架构时：用户需要经由基站或其他通信设备与周围的有相同需求的用户自组织通信。通过构建一个具有相当数量需求的用户的匿名区域，从而达到保护自己隐私的目的。当处在这一匿名区域的某一用户发起了基于LBS的服务请求时，将其自身的真实地理位置模糊为匿名区域的位置信息，然后直接发送到LSP请求服务。LSP将处理后的请求消息以集合的方式直接返回给用户端，用户则根据自己的查询需求，从集合中筛选出请求结果。具体流程如下图6所示：

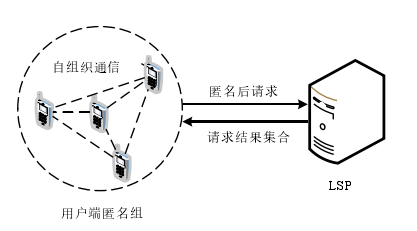


图6 分布式架构通信示意图

分布式架构不存在像集中式架构那样的第三方无法完全可信的缺陷。其所构建的匿名区域信息能够相较于集中式架构更有效地保护用户的隐私。但是，分布式架构要求用户设备有一定的计算存储能力，增加了用户设备的开销[44]。

3.3.3 混合式架构

混合式架构优化、改进集中式与分布式架构而来，性。混合式与集中式不同之处在于，单个的用户被构建的匿名用户组代替——表现为分布式架构的特点。

基于混合式架构的LBS服务请求时：首先是根据发起的要求构建匿名用户组的广播消息，将满足隐私要求的对等用户构建成一个匿名组[44]。当匿名组成员数量满足匿名要求时，则组中的所有用户都可以通过该匿名组发起匿名请求，即是实现分布式架构的处理方案；若组内用户数量不满足要求时，则使用匿名服务器直接接受用户发起的服务请求，即是实现集中示架构方案。具体处理流程见下图7：

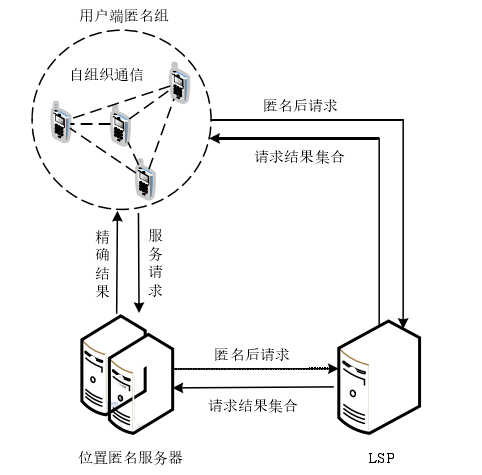


图7 基于混合式架构位置服务流程图

混合式架构保护方案仍有一定缺陷使其适用性受到一定限制：结构比较复杂、参数过多、布置调整繁琐[44]。

3.3.4 系统架构对比

上文所列举三种基于地理位置隐私保护服务的架构各有优缺点。详细对比见下表2：

表2 关于三种架构的分析、对比

|  |  |  |
| --- | --- | --- |
| 系统架构 | 优点 | 缺点 |
| 集中式架构 | 对用户端要求较低、服务质量高、隐私保护效果好 | 匿名服务器无法保证绝对可信、存在隐私泄露风险、容易遭受攻击、TTP成为系统性能提升瓶颈 |
| 分布式架构 | 不存在第三方不可信造成隐私泄露风险 | 增加用户端设备的压力、效率相对较低 |
| 混合式架构 | 适用环境广泛、安全性高 | 参数调整困难、部署繁杂、结构较为复杂 |

3.4 代表算法浅析

截至目前，差分隐私安全机制主要包括两大类技术[42]——集中式差分隐私技术（centralized differential privacy，CDP）以及本地化差分隐私（local differential privacy，LDP）。前者基于集中式架构，而本地差分隐私则基于分布式保护架构。

3.4.1 集中式差分隐私

集中式差分隐私（CDP）的基本原理是对数据添加干扰噪声，以此达到保护用户隐私的目的。

处理地理位置数据时，最为常用的方案是K-means算法，然而K-means聚类算法处理位置数据时，聚类中心很容易遭受到攻击。因此，K-means聚类算法通常结合差分隐私技术——通过在聚类中心添加服从Laplace分布的噪声数据对用户隐私进行保护[46]。基于以上处理方法，有人提出了一种基于差分隐私技术并结合K-means聚类算法的混合地理位置信息保护方案[46]：方案划分用户位置为离散位置与非离散位置点，对于前者则直接使用差分隐私技术进行噪声处理；而对于非离散的数据则使用K-means算法进行泛化处理。张琳等人将算法改进，他们将位置信息中的包含位置数据与未包含位置数据区别处理，提出了一种新的地理位置发布算法Diff\_Anonmity。

而在用户的轨迹数据保护方面，Chen等人最早将差分隐私技术应用到轨迹数据保护中——基于前缀树的数据结构的隐私轨迹发布算法STM-Full。该算法先将高斯噪声进行滤波处理，使其轨迹序列的自相关函数相同，再将其叠加到用户的轨迹数据进行发布。同时，考虑到轨迹数据信息的多维度性，提出了满足差分隐私的二维拉普拉斯噪声，将其以极坐标的形式加入到用户轨迹中去，提高了保护效果。

3.4.2 本地化差分隐私

本地化差分隐私保护技术是以集中式差分隐私技术为基础的一种新的隐私保护模型。该方案不仅消除了TTP不可完全信任的安全风险，同时能够高效地抵御攻击者发起的基于背景知识的攻击行为。其从数据获取到发布整个流程，均能得到全面高效地保护。

本地化差分隐私（LDP）的基本原理是：将原本由中心服务器——可信第三方（TTP）实现的数据隐私化处理功能，均摊到每一个用户的本地设备中去，从而避免了用户的隐私信息经由第三方泄露的风险。

LDP的主流扰动机制为随机响应技术。随机响应技术的基本理念为——对于存在有用户敏感数据的服务响应是完全不能够确定的。这种不确定性能够很好的对于用户的服务请求数据起到保护效果。当前，基于随即应答机制的本地化差分隐私技术因其良好的性能吸引了大量学者的注意。针对于数据采集过程中的敏感数据的保护，高志强等人提出了一种基于LDP技术的位置数据采集方案，利用期望最大法和直接统计法分析数据。

3.4.3 缺陷与不足

传统的差分隐私通常依赖于在用户的原始数据中添加服从拉普拉斯或者是指数分布的噪声，以此来达到隐私保护的效果。但是往往很难把控噪声的添加的量：噪声添加过少则难以起到保护效果，但若是添加过多，则会大大降低原始数据的可用性，进而影响服务的质量。同时，针对于本地化差分隐私的随机响应机制，因为其对于存在着用户敏感数据的服务响应是一种不确定的状态，服务的质量也很难保证。

同时，集中化差分隐私还存在第三方信任等问题，而本地化差分隐私的也存在着匿名区域用户的身份难以验证等问题。具体对照见下表3。

表3 两种差分隐私比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 名称 | 数据的处理时间 | 数据处理位置 | 数据类型 | 架构 | 扰动机制 | 风险 |
| 集中式差分隐私机制 | 数据发送之后 | TTP | 初始数据 | 集中式架构 | 拉普拉斯机制、指数机制 | 难以保证TTP的完全可信任、用户个人对信息不可控 |
| 本地化差分隐私机制 | 数据发送前 | 本地 | 经过扰动后的数据 | 分布式架构 | 随即响应机制 | 初始数据存储在本地、信息用户可控 |

4 保护架构及算法改进

4.1 瓶颈问题与改进挑战

LDP在相较之于CDP在计算用于的查询结果方式上完全不同。基于LDP的地理隐私保护技术从用户设备中获取而来的数据，几乎全是噪声信息。LSP服务器的作用只是：将用户设备上提交来的独立分布的噪声抵消，从而获得唯一准确的结果。LDP的随即响应机制不是依赖TTP服务器，而是基于客户端的一定计算、存储能力的隐私解决机制。对于数据的控制权力完全在于客户端设备。客户端在本地将本人的地理隐私数据执行了随机扰动算法之后，向服务器发送经过扰动之后的值。

差分隐私的根本在于：攻击者在面对即是数据信息仅仅差别一两条的数据集发起的服务请求时，对于查询结果的之间概率区别微乎其微。这样就保证了假设攻击者已经知道绝大部分记录的敏感信息，也仍然无法推断出该数据集的任何重要性息。

基于本地化差分隐私机制的地理位置隐私保护不单单是在理论上有着数量众多的研究成果，同时也被广大学者加以应用于实际生活当中。但同时，面对越来越丰富的地理数据类型以及存在于客户端的源数据处理任务，当前的一些基于差分隐私的隐私保护算法并不能做到很好的处理。本地化差分隐私算法的改进任务，仍然存在着以下的挑战。

（1）地理位置数据的复杂性。当前智能终端服务功能越来越丰富，而随着在线服务的兴起，用户所产生的位置数据也愈来愈趋向于多元化，而不再仅仅停留在一维或者是二维数据，例如说键值、图表等。另一方面，这类数据所需要的处理空间与时间往往十分庞大。同时，在这些数据内部，还存在有一定的数据结构，在对其进行隐私保护时，需得要在满足约定的隐私保护条件下尽可能地保留源数据的结构。所以在面对这类复杂数据时，基于本地化差分隐私的算法改进将更加具有挑战。

（2）用户位置点数据之间相互关联。地理位置数据不同于其他数据最为明显的一点即是：用户的轨迹位置是数据是属于在时间和空间上连续产生的数据类型。其中包括空间相关性、行进相关性以及历史相关性。算法在对其进行隐私性保护时往往会引起“牵一发而动全身”的状况。所以针对于用户位置进行单独地保护会产生噪声叠加现象，而对于数据可用性的要求又使其不能采用直接保护整个数据集。

（3）数据隐私保护的可用性。用户所产生的地理位置数据，归根结底都是为了基于地理位置所产生的服务的需要。仅仅是在这个过程中，为了避免隐私的泄露才迫不得已进行数据保护这一操作。在基于本地化的差分隐私算法改进中，在要求的隐私保护限制前提下尽可能地保持用户数据的准确于可用性是关键。在算法改进任务中，满足隐私保护是前提，而能够在隐私保护之后依然能够提供较高准确性的服务才是首要的目标。因此，针对地理位置数据的隐私保护任务怎样在保证源数据的可用性前提下从而改进基于本地化的差分隐私保护算法是一项巨大的挑战。

4.2 隐私保护框架改进

差分隐私的根本在于：攻击者在面对即是数据信息仅仅差别一两条的数据集发起的服务请求时，对于查询结果的之间概率区别微乎其微。这样就保证了假设攻击者已经知道绝大部分记录的敏感信息，也仍然无法推断出该数据集的任何重要性息。

集中式架构中引进的可信第三方TTP为匿名服务器，可以达到很好的隐私保护效果，同时减轻了客户端的隐私保护和数据处理的负担。但其存在以下三点风险：（1）难以保证可信第三方的绝对信任，即使匿名服务器声称完全不会泄露用户的个人隐私数据。在大多数的实际应用中，很少存在将数据完全的依赖于某一处理中心进行存储、处理等多种功能为一身的情况，一旦数据丢失或是服务器被入侵，则会导致整个服务崩溃（2）匿名服务器或成为整个系统的瓶颈和攻击点。随着日渐的智能设备接入互联网中，越来越多的用户急需基于地理位置的服务，在这种呈现出指数级别增长的服务需求的增长中，如果采用这种集中式的架构，会对之后的服务扩展以及处理数据的能力的扩展带来很大的不便（3）大量的部署第三方服务器在现实生活中具有极大的困难。服务器尤其是大型的服务器，不仅仅时价格昂贵，难以承担，且后期的维护以及更新所耗资源是极其庞大的，在现实生活中往往是不现实的。

分布式架存在的前提是能够保证参与在这个架构中的每一个设备，必须同时具备一定的计算与存储能力。当采用分布式架构时：用户需要经由基站或其他通信设备与周围的有相同需求的用户自组织通信。通过构建一个具有相当数量需求的用户的匿名区域，从而达到保护自己隐私的目的。但是，分布式架构更多的还是考虑到用户之间的自组织通信，这也是其区别于混合式架构的唯一不同之处。虽然避免了中心服务器的存在带来的诸如较高的攻击风险、性能瓶颈以及耗费等问题。但是其仍然具备相当大的问题：如匿名用户并不能满足当前要隐私保护要求所需要的用户数量，更或是其难以保证所加入匿名组的用户的可靠性。

针对于差分隐私保护机制存在的地理位置数据及轨迹数据天然存在的相关性，以及考虑到过量的添加噪声会降低位置服务查询的效率的情况。本文在综合分析前文所提到的基于差分隐私的地理位置数据保护方法的基础上，改进一种轨迹拆分保护算法。

该保护机制的基本思想是：通过引进多个LBS服务器，构建一个满足本地化ε-差分隐私的服务框架。同时，基于该多LBS服务器地理位置隐私保护框架，改进出一种轮询算法——用户端设备在发起基于地理位置的服务请求时，对当前的位置是否满足ε-差分隐私进行计算，且根据计算结果，选择进行位置发布以获得服务或者是进入轮询依次判断下一个LBS服务器是否能够满足隐私保护的要求。在经过一系列定义严格的计算之后，用户最终可以选择最能够起到隐私保护的的LBS服务器进行服务查询，亦或是选择抑制当前敏感位置的发布。

通过引入这种基于多LBS服务器的隐私保护架构，首先，本地设备可以预先定义每个用户自身的敏感数据集，可以很好的实现个性化的服务。同时，数据在本地会进行事先的预处理（剔除用户基于位置发起的服务数据中存在于自定义敏感数据集中的暴露用户个人隐私的数据点）。同时LBS服务器的部署难度与所耗费的人力物力资源是远远小于中心服务的，因为其采用的混合式架构，可以预先在缓冲区中缓存一部分先前用户所发起的服务请求数据，当一个新用户通过信号基站接入当前LBS服务器组中并发起服务请求时，实际上所需要处理的数据大概率已经存在与LBS服务器的缓冲区中，这是由于地理位置数据的相关性所引起的。

如下图8所示：用户端设备在发起服务查询请求时，第一步是判断自身当前所处位置的敏感程度，即是当前位置，是否是处于用户根据自身实际情况定义的敏感位置数据集合当中，若回答为肯定，表明当前所查询位置包含有敏感信息，抑制发布；若回答为否定，则计算当前选择哪一个LBS服务器进行查询服务不会造成用户本人的隐私信息地泄露。通过这种多LBS服务器框架结合对于用户的轨迹进行拆分的地理位置信息保护机制，能够很好的消除用户位置数据之间的相关性，提高查询效率。

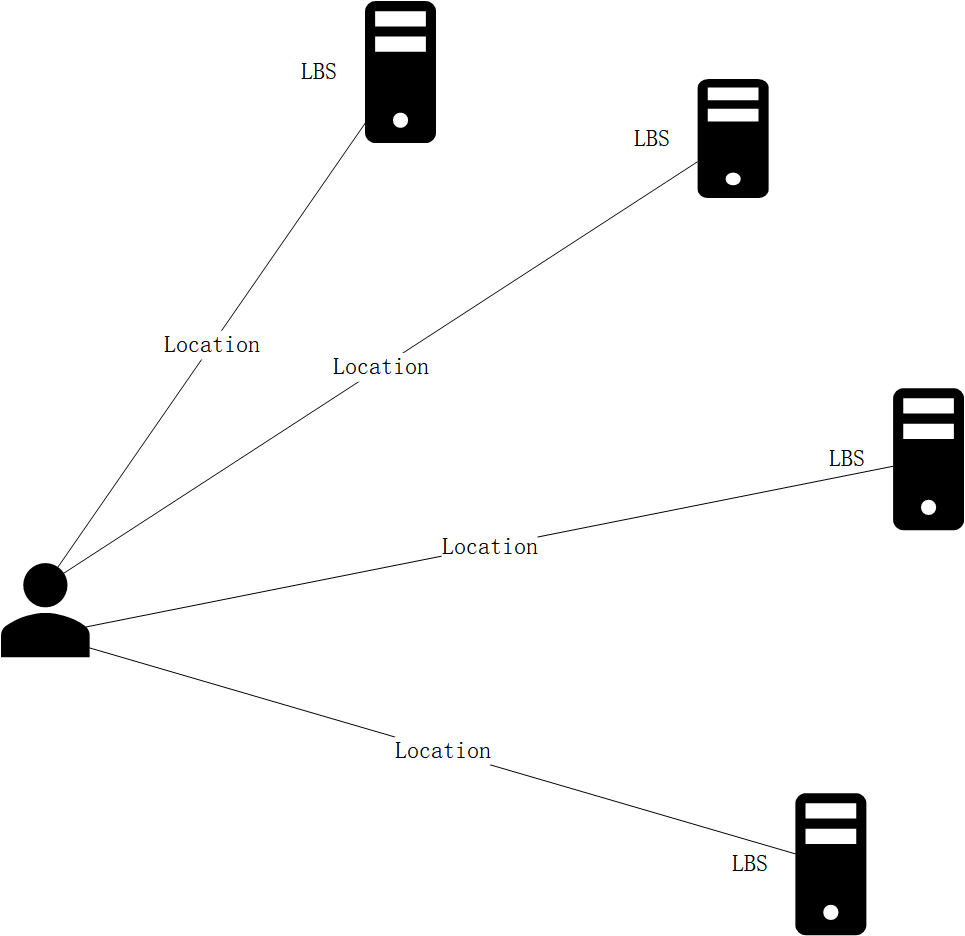


图8 基于多LBS的地理位置数据隐私保护架构

4.3 轮询算法改进

针对于将噪声加入到用户真实地理位置数据进而保护用户的敏感信息的差分隐方法，其局限性在于该算法仅仅针对于用户的地理位置数据，而无法对于用户的移动轨迹进行良好的保护，除此之外，噪声在初始数据中的的大量叠加会引起原数据的失真，降低了数据的可用性，使得查询返回的结果的有效性降低[46]；而对于将K-means聚类算法结合差分隐私的保护方案来讲，虽然其通过对于K-means的聚类中心添加噪声来进行扰动从得到泛化保护效果，但是这种保护方案所针对用户构建匿名区域所包含的位置信息点范围比较大时，用户位置不集中时难以适用。而且针对于集中的数据点运用K-means聚类算法时，对于不同的数据聚类中心的选取也对数据最终的隐私保护效果影响偏差很大[47]。

总的来讲，现有的基于差分隐私机制的地理位置保护的算法有着以下一些不足：（1）如何在确保用户隐私信息得到有效地保护的前提下，减少添加噪音带来的数据失真、可用性降低以及一些其他的负面影响；（2）位置数据集过大，在结合聚类算法时存在的异常点问题、离散点敏感、聚类中心的选择等问题，都会导致聚类效果不理想，影响最终的服务效果；（3）用户位置点数据之间相互关联[47],其中包括空间相关性、行进相关性以及历史相关性，所以针对于用户位置进行单独地保护会产生噪声叠加现象，而对于数据可用性的要求又使其不能采用直接保护整个数据集。

在本文提出的多LBS服务器的地理位置数据隐私保护架构中，将这n个LBS服务器依次表示为L1、L2、L3……Ln；而用户所处的位置数据点以X1、X2、X3……Xi表示；同时将用户的所提交到各个LBS服务器上的轨迹数据表示为trl1{X11、X12……X1i}；trl2{X21、X22……X2i}，…，trln{Xn1、Xn2……Xni}。

为了保证各个LBS服务器的均衡负载要求，考虑到提供服务的机会均等的公平性原则，算法首先采用调度方案为时间片轮转算法，对初始的LBS服务器Lw进行选择。

算法其次计算当前位置Xi ，若使用初始设置LBS服务器Lw 时的，其值是否满足上文定义的先验概率以及后验概率对于ε-的差分隐私要求，以此判断若发布当前位置Xi是否会造成用户所处轨迹前后敏感位置的泄露。如若不会，则直接向该LBS服务器发起查询；若会造成敏感位置信息的泄露，则进入算法的轮询阶段，选择下一个LBS服务器Lw+1 ，再次进行其隐私泄露风险计算。直到找到可以发布满足当前位置Xi 的隐私保护的LBS服务器Ln为止。若轮询完所有的LBS服务器均无法满足隐私保护要求，则禁止该位置的发布。

通过对该改进后的轮询算法步骤进行分析可知：该算法是以当前单个位置点来针对某一特定的LBS位置服务器进行隐私保护计算，并不存在如传统的聚类算法存在某一聚类中心，并且规避了用户位置的异常点与离散数据值的问题。这不仅确保了数据集的大小之于用户设备可忍受的范围，同时也极大地降低了由于数据聚类而导致的背景攻击风险；同时，本算法所提供的隐私安全保护是基于在多个LBS服务器框架内的每个用户提交的位置点的风险计算，并没有往源数据中添加任何量的噪声。即是该算法并不会造成源数据的失真，不会降低数据的使用性；另外，本文所改进算法在每次进行隐私风险计算时均以当前单独用户位置为单位，所对应的LBS服务器在当前一轮循环时数量也仅仅为1。所以，用户所提交的地理位置数据在隐私保护领域内的固有缺点——空间相关性、行进相关性以及历史相关性将均被规避。

改进后算法核心代码示意如下：

|  |
| --- |
| 基于n个LBS服务器的差分隐私保护机制的轮询算法 |
| 输入：用户地理位置Xi ；用户提交的带查询的轨迹数据集{trl1、trl2、trl3……trln}；隐私保护预算ε；用户自定义敏感数据集S  输出：LBS服务器序号w |
| 1. w，sum=1 2. if Xi  in S: 3. 禁止查询发布 4. elif: 5. while (sum<n or w<n): 6. if (w<n and sum<n): 7. 计算Xi在该LBS服务器发布时对于之后最近的敏感位置的safter隐私暴露风险的先验概率Pbefore与后验概率Pafter 8. 轨迹位置前移：trlw=trlw+Xi 9. 计算Xi在该LBS服务器发布，对时间序列上前一敏感位置的safter隐私暴露风险的[48]先验概率Pbefore-t与后验概率Pafter-t 10. if (Pbefore /Pbefore-t≤eε and Pafter/Pafter-t≤eε): #对前后位置发布是否满足ε-差分隐私判断 11. 返回当前LBS服务器序号w 12. else： 13. w+=1 and sum+=1 14. #轮询完所有LBS服务器均不能达到隐私保护要求 15. else: 16. return None |

4.4 核心代码浅析

为了测试改进后算法对于用户地理数据隐私性的保护，实验将使用由移动对象生成器生成的数据集测试示例。由于数据生成时便自动提供了标签，因此我们可以通过要求模型对每个样本进行分类，然后将预测的类与数据集中的实际标签进行比较来测试模型的准确性。我们将数据拆分为包含80%示例的训练集和包含20%示例的测试集。对于数据的具体处理代码如下：

#引入python的matplotlib库进行函数图像绘制

import matplotlib pyplot as plt

#引入python的pandas库与numpy库以便后续对于源数据的不同维度数据进行函数处理

import pandas as pd

import numpy as np

#定义拉普拉斯噪声函数

def laplace\_ mech(v, sensitivity, epsilon):

return v + np.random. laplace(loc =0, scale -sensitivity 1 epsilon)

def gaussian\_ mech(v, sensitivity, epsilon, delta):

return v+np.random.normal(loc-0,scale=sensitivity\*np.sqrt(2\*nplog(1.25/delta))/epsilon)

#定义高斯噪声函数

def gaussian\_ mech. ved(v, sensitivity, epsilon, delta):

return v+np.random.normal(loc=0,scale=sensitivity\* np sqrt(2 np.log(1.25/delta))epsilon, size =len(v))

def pct\_ error(orig, priv):

return np .abs(orig - priv)/orig \* 100.0

def z\_ clip(xs, b):

return [min(x, b) for xin xs]

def g\_ clip(v):

n= np.linalg. norm(v, ord=2)

if n> 1:

return v/n

else:

return V

#定义数据测试集大小

training\_ size = int(X .shape[0] \* 0.8)

#切片划分数据集

X\_train= X[training\_ size]

X\_ test = X[training\_ size:]

y\_train =y[:training.size]

y\_ test = y[training. size:]

y\_ test.shape

当由移动对象生成器生成的数据集经过了训练集与测试集的拆分之后，将会进行隐私处理算法的迭代。此项工作的最主要目的即是通过往生成的用户数据集中添加规定量值的噪声，并同时轮询LBS服务器，最终达到在确保用户隐私得到保护的前提下能够成功得到基于地理位置的服务。代码如下：

#设置数据集batch size

gaussian\_mech = DPOptimizerClassFactory(args.micro\_batches)

#实例化差分隐私

gaussian\_mech.set\_mechanisms('Gaussian',

norm\_bound=args.l2\_norm\_bound,

initial\_noise\_multiplier=args.initial\_noise\_multiplier)

#设置差分隐私的噪声机制——自适应高斯噪声机制

net\_opt = gaussian\_mech.create('Momentum')(params=network.trainable\_params(),

learning\_rate=cfg.lr,

momentum=cfg.momentum)

```设置差分隐私预算监测器，用于设定每次轮询中的差分隐私预算,

监控隐私预算的原因是：如果隐私预算ε处于较低的水平，虽然

可以保证较高的安全性，但是会导致源数据的可用性下降，进而导

致服务的准确性降低

```

rdp\_monitor = PrivacyMonitorFactory.create('rdp',

num\_samples=60000,

batch\_size=16,

initial\_noise\_multiplier=5,

target\_delta=0.5,

per\_print\_times=10)

#trl是生成的用户的轨迹数据，ε是隐私保护预算，S是敏感数据集合，n为LBS服务器个数  
def iteration(trl**,**ε**,**n**,**S):  
 #w是LBS服务器序号，默认从1开始  
 w**,**sum=1  
 for i in trl:  
 #如果当前用户位置在敏感数据集中，退出本次循环  
 if i in S:  
 continue  
 elif:  
 while(sum<n or w<n):  
 if (w<n and sum<n):  
 #计算当前位置前后是否满足差分隐私隐私  
 Pa**,**Pb=CountDP(i)  
 trl[i]=trl[i+1]  
 #计算当前位置下一位置前后是否满足差分隐私隐私  
 Paa**,**Pbb=CountDP(i+1)  
 #如果满足差分隐私，则返回LBS服务器序号，查询成功  
 if(Pa/Paa<=e^ε and Pb/Pbb<=e^ε)  
 return w  
 #否则，对下一LBS服务器发起查询请求  
 else:  
 w+=1  
 sum+=1  
 #如果轮询完是sum=n时，则没有找到符合差分隐私的LBS服务器，返回空  
 else:  
 return None

当数据集经过了训练集与测试集的拆分之后，又经过轮询算法的迭代。还需要测试经过隐私保护运算之后的数据集的提供服务查询准确性的校验，并进行可视化处理见图16。定义函数代码如下：

def accuracy(theta):

#利用Numpy函数将查询成功次数与总发起查询次数进行模运算以求得准确率

return np. sum(predict(X\_test, theta) == y\_ test)/x\_ test. shape[0]

accuracy(theta)

#根据求得查询准确率利用Matplotlib库进行可视化展示

delta= 1e-5

epsilons = [0.001, 0.003, 0.005, 0.008, 0.01, 0.03, 0.05, 0.08, 0.1]

thetas = [noisy\_ gradient descent(10, epsilon, delta) for epsilon in epsilons]

accs = [accuracy(theta) for theta in thetas]

#定义横纵坐标

plt xlabel'Epoch')

plt. ylabel('Accuracy')

#调用绘图函数

plt plot(epsilons, accs)

5 改进框架及算法分析

5.1 改进框架分析

本文在综合前人的研究成果的基础上，该进出基于多LBS服务器的差分隐私保护框架。该保护机制的基本思想是：通过引进多个LBS服务器，构建一个满足本地化ε-差分隐私的服务框架。同时，结合改进后的算法，用户端设备在发起基于地理位置的服务请求时，对当前的位置是否满足ε-差分隐私进行计算，且根据计算结果，选择进行位置发布以获得服务或者是进入轮询依次判断下一个LBS服务器是否能够满足隐私保护的要求。以此来达到保护的效果。

框架采用混合式架构。本地设备能够预先定义每个用户自身的敏感数据集，可以很好的实现个性化的服务。同时，用户所输入的数据可以预先在本地根据用户的自定义敏感数据集进行敏感位置点地剔除，能够对用户所提交的原始数据进行一定的预处理，剔除用户自定义的敏感点后，可以大大提高数据在提交到LBS服务器过程中的隐私程度，同时也可以减轻LBS的后续的运算压力。

采取混合式架构的另一个好处是：能够通过预先在缓冲区中缓存一部分先前用户所发起的服务请求数据，当一个新用户通过信号基站接入当前LBS服务器组中并发起相类似服务请求时，能够直接进行服务提供。这可以很好的提升服务的效率与品质。

5.2 算法复杂性分析

针对于上文所提出的基于n个LBS服务器的差分隐私保护机制的轮询算法的时间耗费进行分析如下:该算法耗时最多的步骤在于第7、9、10步，其主要实现的功能是——用户在当前位置Xi发起基于LBS的查询服务时，本文所提出算法仅仅只需要计算对于最近前后的两个敏感位置的隐私泄露风险，时间复杂度为O(nr),n代表着LBS服务器的个数，r为当时刻t时，查询数据序列长度，因为轨迹数据是有限长度，所以时间复杂度在于可接受范围内。本文虽然是基于多个LBS服务器进行隐私保护，但是算法在实际运算时仅仅需要计算前后两个敏感位置的隐私风险暴露，计算开销为中等，在计算中并不会产生额外的运行数据，且由于并没有源数添加噪声进行扰动，使得地理位置元数据可用性大大增加，在算法处理中并不需要额外添加其他数据，空间复杂度亦在可接受范围之内。

5.3 算法隐私保护效果分析

对于任意一条轨迹数据trl中的位置Xi，令离Xi最近以及次近的地理敏感位置s1、s2，则根据轨迹位置的位置相关性可知，s2相较于s1与位置Xi之间关联性更低——关联性与两点之间距离成反比关系[49]。则我们可得出，当Xi对于s1满足差分隐私时，也对于s2满足差分隐私。由此可以确保所有LBS服务器上的隐私数据集中的敏感位置均能保证其安全性。

5.4 横向对比分析

改进后的算法是以当前单个位置点来针对某一特定的LBS位置服务器进行隐私保护计算，并不存在如传统的聚类算法存在某一聚类中心，并且规避了用户位置的异常点与离散数据值的问题。这不仅确保了数据集的大小之于用户设备可忍受的范围，同时也极大地降低了由于数据聚类而导致的背景攻击风险；同时，本算法所提供的隐私安全保护是基于在多个LBS服务器框架内的每个用户提交的位置点的风险计算，并没有往源数据中添加任何量的噪声。即是该算法并不会造成源数据的失真，不会降低数据的使用性；另外，本文所改进算法在每次进行隐私风险计算时均以当前单独用户位置为单位，所对应的LBS服务器在当前一轮循环时数量也仅仅为1。所以，用户所提交的地理位置数据在隐私保护领域内的固有缺点——空间相关性、行进相关性以及历史相关性将均被规避。

同时，本文选取吴云乘等人提出的一种基于时空相关性的差分隐私的轨迹保护隐私机制以及赵婧等人提出的基于轨迹频率抑制差分隐私保护机制进行横向对比分析。本文提出的基于差分隐私保护机制的多LBS服务器的轮询算法与吴云乘等人提出的保护机制均为基于用户实时地理位置进行发布，而在位置数据发布之前，便已经在用户设备上进行满足差分隐私的检测，能够防止服务器的内部攻击，而由赵婧等人提出的保护机制是基于离线的处理方式，无法抵御服务器内部发起的攻击；其次，本文所提出的算法并未对原始数据进行添加噪声扰动，查询结果准确性更高；最后，本文虽然是基于多个LBS服务器进行隐私保护，但是算法在实际运算时仅仅需要计算前后两个敏感位置的隐私风险暴露，计算开销为中等。

表4 三种方案的横向对比分析

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方案 | 是否为在线模式 | 开销 | 查询成功率 | 是否添加噪声扰动 |
| 本文方案 | 是 | 中等 | 高 | 否 |
| 吴云乘等人[53]方案 | 是 | 较高 | 中等 | 是 |
| 赵婧等人[54]方案 | 否 | 中等 | 较高 | 否 |

6 仿真实验应用及结果分析

本文采用软件实验环境为：PyCharm Community（算法实现语言为Python）；Anaconda（数据分析语言为Python，使用到的包括Numpy、Matplotlib、Pandas等数据分析常用库及专业绘图库）；Visual Studio 2022 professional（移动对象生成器实现语言为java）。硬件环境为：处理器为Intel(R) Core(TM) i7-7700HQ CPU，2.80GHz，8.00 GB内存。64位Windows11专业版系统。

6.1 移动对象生成器

基于LBS服务所提交的查询数据是一种很典型的同时包含时间和空间的数据[45]，由此在处理此类数据时需要定义模拟移动对象典型行为的合适基准。以前的方法不考虑移动对象的一个重要属性：在现实生活中，目前许多获取并且处理用户的带有时空属性的地理位置及轨迹数据的应用程序中，用户的行进总是遵循给定的网络（即是用户的运动或者是位置数据，总是在一个特定的地理区域中的，为了更好的对数据进行后期加工处理，我们可以将该地理区域以网格的方式进行划分）。因此，改进后的算法的测试需要具有这种"基于网络"的移动对象的轨迹位置数据集。

生成此类数据集的重要方面是连接的最大速度和最大容量，以及其他移动对象对用户的速度和路线的影响，用户的开始和目的地的充分确定，外部对象和事件的影响以及预定流量。

Thomas Brinkhoff的这种基于道路网格的时空数据生成器[50]。它能够将用户当前的真实位置、轨迹数据集（以网格形式进行切分）与用户定义的数据集属性规范相结合。在本文背景条件下，即是该基于路网的移动对象生成器能够结合用户所提交的基于LBS服务所发起的查询请求以及用户自定义的敏感信息数据集，生成新的符合时空特性的、便于改进后的基于差分隐私机制的轮询保护算法的后续处理的数据合。其基本原理是：在输入的地图节点数据数据读入后，随机生成多个对象来模拟真实用户在地图网格上的运动，然后将运动数据经空间转换和数学工具处理后输出规范的、时空特性的轨迹数据文件[52]，如下图9所示：

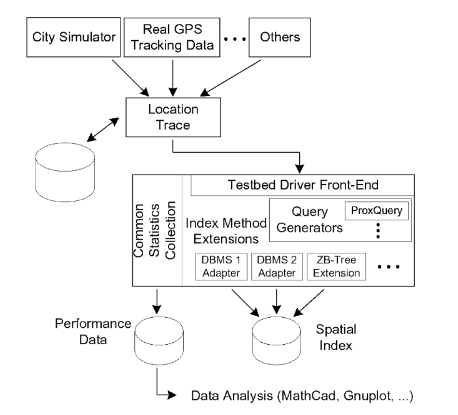
****

图9 基于路网的移动生成器原理示意

下图10为移动对象生成器工作截图。

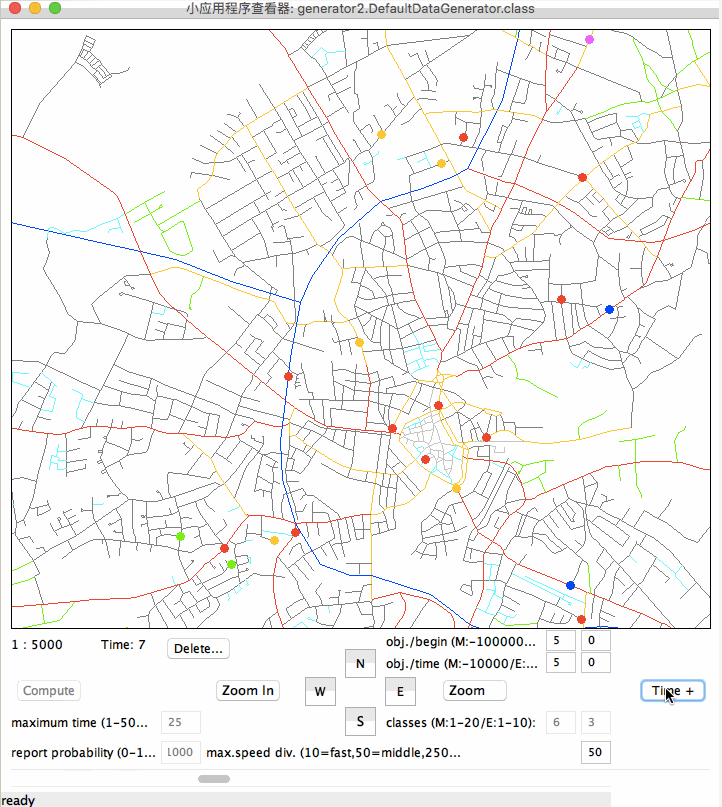


图10 移动对象生成器运行截图

6.2 数据获取

本文使用奥尔登堡城市路网数据集和新加坡城市路网数据生成运动轨迹。同时将两地图数据集分划分网格大小为0. 01 km与0. 06 km。以此来测试改进后算法在稀疏以及稠密路网中的表现。

在使用基于路网的移动生成器处理下载的数据集时，将会生成多个对象、随机的、带有时间戳的坐标点组成的移动轨迹。不同移动对象会形成不同的轨迹，通过移动对象的id号来区分不同的移动物体[52]。生成数据具体情况下图11所示：

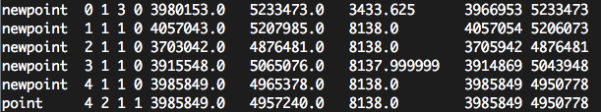


图11 生成数据示例

6.3 仿真过程

仿真实验主要分为以下几个步骤：（1）由基于路网的移动生成器随机生成的“用户”在所下载的地图包上进行运动，采集其地理位置集轨迹数据；（2）改进后的算法读入采集到的数据集合，并且添加进在轨迹上根据个人偏好选择的敏感位置构成的隐私数据集；（3）将能够找到满足隐私保护需求的LBS服务器的位置个数，以及总的轨迹上的位置点输出并进行处理；（4）利用Python的Numpy、Pandas库进行数据处理，并且用Matplotlib画出折线图进行结果分析。

本文在前人基础上拟实现的基于差分隐私保护机制的轮询算法改进，其主要目的是：用户在基于LBS服务发起查询请求时，能够在确保用户隐私得到保护的前提下，通过降低查询请求数据中，上文所提到的地理数据存在的相关性，来提高查询的效率性以及查询的准确性。因之下文所作针对实验结果分析指标为：运行时间、查询成功率（query success probability,QSP），服务准确性（在符合差分隐私的数据保护前提下，是否能够很好的满足用户的基于位置服务的请求）。其中，QSP定义如下：

其中，TOS代表查询中成功的次数，TOA代表对于轨迹数据集中非敏感位置发起的查询总次数。同时，在本文最后，将本文算法与吴云乘等人提出的一种基于时空相关性的差分隐私的轨迹保护隐私机制以及赵婧等人提出的基于轨迹频率抑制差分隐私保护机制进行横向对比分析。

6.4 仿真结果分析

6.4.1 隐私预算ε因素分析

在探究隐私预算ε的大小对于查询成功率的影响时，将敏感位置个数m确定为5，n为LBS服务器的个数。对于隐私预算ε来讲，其决定了轨迹数据中非敏感的位置点的发布对于那些敏感位置暴露的影响。当ε越小时，所能够查询到的敏感位置越少，因此QSP越低。可以得出结论：隐私保护预算ε与QSP呈正比趋势。而针对两个地图数据的分析证明这一结论。如下图12：

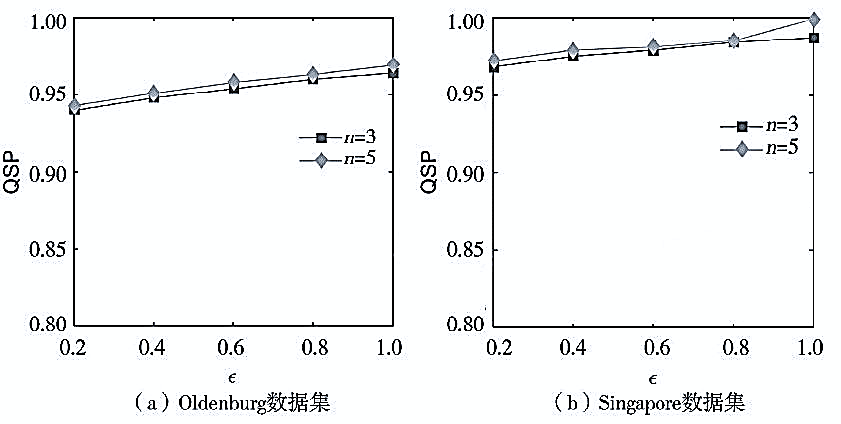


图12 ε指标分析

6.4.2 敏感位置个数因素分析

当确定基于差分隐私保护机制的查询算法的ε为0.6时，讨论敏感位置的个数m对于查询成功率QSP的影响情况。当随着敏感位置个数m增加时，会使得其他非敏感位置的发布对于敏感位置的暴露的风险升高，随之保护机制则会抑制该位置的发布，造成查询成功率QSP的下降。数据分析得出结论类似，详情如图13：

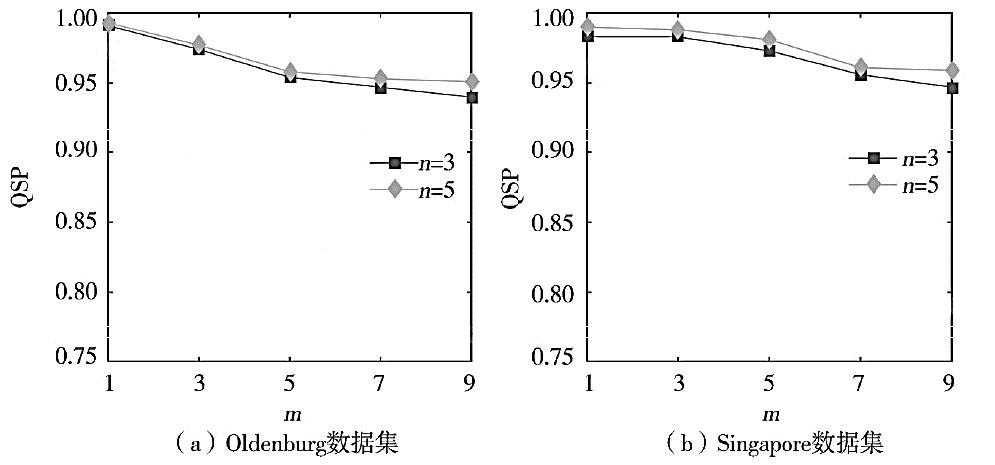


图13 敏感位置个数指标分析

6.4.3 LBS服务器个数因素分析

如上图11可见，当LBS服务器个数n增大时，地理位置之间的关联性随之减小，而查询成功率QSP有着极大的提高。同时当服务器个数增加时，运算时间相应增加，因为在每次查询时都要对隐私风险进行更多的计算。服务器个数与所耗时间大致比例如下图14：

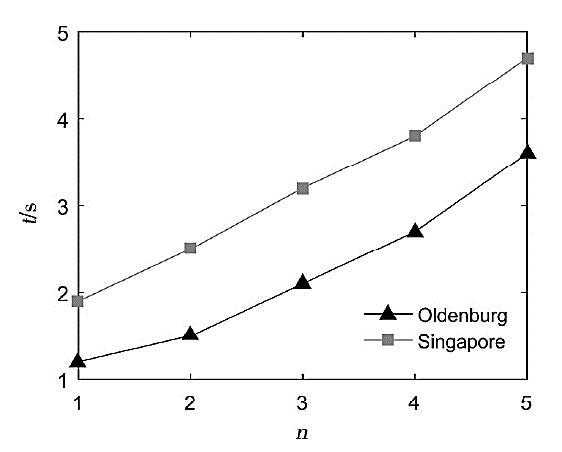


图14 服务器个数指标分析

6.5算法应用性分析

在基于本地化的差分隐私算法改进中，在要求的隐私保护限制前提下尽可能地保持用户数据的准确于可用性是关键。在算法改进任务中，满足隐私保护是前提，而能够在隐私保护之后依然能够提供较高准确性的服务才是首要的目标。因此，针对地理位置数据的隐私保护任务怎样在保证源数据的可用性前提下从而改进基于本地化的差分隐私保护算法是一项巨大的挑战。

在经过以上仿真实验证明了本文改进算法在时间复杂度与空间复杂度均处于可以接受的水平之外，同时在针对于用户所产生地理位置数据的敏感位置个数以及存在着的LBS服务器个数等影响因素的探讨，确保了在一定条件之下本文所改进算法在能够接受的时间与空间范围内，在存在着多个LBS服务器的保护框架下，能够很好的满足隐私保护的要求。

在对算法的应用进行分析时，首先将已经满足隐私保护的数据进行归一化的模型处理：

#引用sklearn库进行数据处理

from sklearn import datasets

from sklearn.model\_selection import train\_test\_split

#数据集使用已经进行隐私处理后的地理数据集

dataset = datasets.load\_lbs()

X\_train,X\_test,y\_train,y\_test=train\_test\_split(dataset.data,dataset.target, test\_size=0.2)

如图15为数据处理运行截图示例：

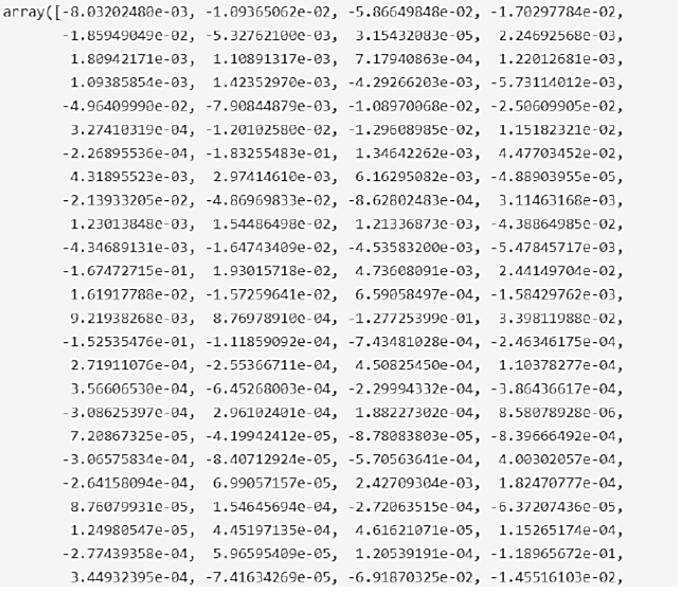


图15 数据处理截图

```

diffprivlib.models.GaussianNB可以在不带任何参数的情况下运行，但这会引发警告（我们需要指定bounds参数来避免这种情况）。隐私级别由参数epsilon控制，该参数在初始化时传递给分类器（例如GaussianNB(epsilon=0.1)）。默认值是epsilon = 1.0

```

from diffprivlib.models import GaussianNB

#我们现在可以对看不见的例子进行分类

clf = GaussianNB()

clf.fit(X\_train, y\_train)

clf.predict(X\_test)

#模型训练与测试

LOGGER.info(TAG, "============== Starting Training ==============")

model.train(cfg['epoch\_size'], ds\_train, callbacks=[ckpoint\_cb, LossMonitor(), rdp\_monitor],

dataset\_sink\_mode=args.dataset\_sink\_mode)

LOGGER.info(TAG, "============== Starting Testing ==============")

ckpt\_file\_name = 'trained\_ckpt\_file/checkpoint\_lenet-10\_1875.ckpt'

#对于其准确性的计算（抑制发布位置点与敏感位置点比较得来）

param\_dict = load\_checkpoint(ckpt\_file\_name)

load\_param\_into\_net(network, param\_dict)

ds\_eval=generate\_mnist\_dataset(os.path.join(args.data\_path,'test'),batch\_size=cfg.batch\_size)

acc = model.eval(ds\_eval, dataset\_sink\_mode=False)

LOGGER.info(TAG, "============== Accuracy: %s ==============", acc)

如下图16是改进后算法在运行完所有轮询之后呈现出的准确率，可见是符合一般用户的要求的。

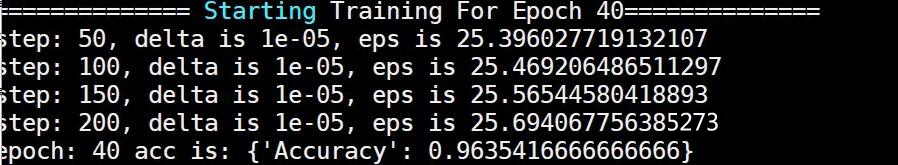


图16 应用服务服务准确性

在将源数据在基于多LBS服务器地理隐私保护架构中、利用轮询算法将用户提交的地理位置数据中的位置点进行符合ε-的差分隐私计算并且做出了该位置点是否进行发布的决定。本文将对于经过处理已经满足差分隐私的数据在提供服务时的准确性指标进行分析，并于未进行基于差分隐私保护的源数据进行比较。处理过程如下：

import numpy as np

import matplotlib.pyplot as plt

epsilons = np.logspace(-2, 2, 50)

bounds = ([4.3, 2.0, 1.1, 0.1], [7.9, 4.4, 6.9, 2.5])

accuracy = list()

for epsilon in epsilons:

clf = GaussianNB(bounds=bounds, epsilon=epsilon)

clf.fit(X\_train, y\_train)

accuracy.append(clf.score(X\_test, y\_test))

#使用matplotlib函数库进行绘制

plt.semilogx(epsilons, accuracy)

plt.title("Differentially Privacy")

plt.xlabel("Epoch")

plt.ylabel("Accuracy")

plt.show()

如下图17是经过算法处理之后满足差分隐私的数据以及未经差分隐私保护的源数据，在模拟用户发起的基于地理位置发起的查询准确性的比较。可以看出，虽然经过差分隐私的轮询计算从而丢失相关敏感地理位置数据，但是相较于源数据来讲，准确性并未有较大滑落——应用服务准确性与质量均在较高水平。

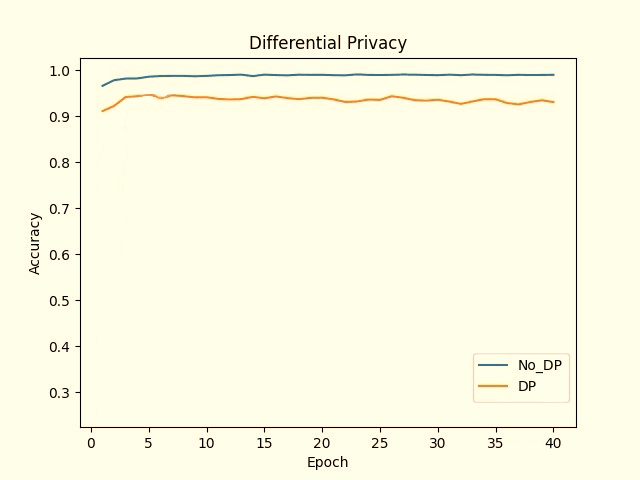


图17 应用算法前后准确性比较分析

本文所改进的基于本地化的隐私改进算法与其他差分隐私所不同的一点是：并不是通过添加高斯噪声或者是指数噪声，从而达到服从差分隐私的保护效果。而是通过算法的轮询机制，在设定好隐私预算ε后，逐个计算当前位置点若选择当前LBS服务器进行发布时之于前与后的地理位置点的隐私暴露风险，以此来决定发布或是抑制当前位置的发布，从而保护用户的隐私。因为并没有对用户所提交进行服务请求的原始数据进行过多的扰动——仅仅由框架在初始时剔除用户预先定义的敏感位置点，所以并没有过多的降低用户源数据的可用性。

我们可以从图17得出：经过隐私保护后的数据与原始数据相比，所提供的应用服务准确性依然在可以接受的范围内，实际上相比于用户初始数据并没有多少下降。即是本文算法在满足其对于隐私保护得前提下在基于地理位置的服务得应用性能够得到保证。

7 总结

本文结合论文的研究背景及意义，讨论了当今基于LBS服务隐私保护的国内外研究现状。同时在广泛查阅文献以及学习的基础上，综合对比分析了地理位置轨迹隐私保护领域的三大主流技术，指出其适用范围以及在针对位置轨迹保护的不足之处。由此引出了对于差分隐私技术在该领域的重要地位以及其优势所在。在介绍了差分隐私的基本原理以及数学定义之后，本文就关于地理位置及轨迹数据的保护架构进行了讨论，由此确定了采用本地化差分隐私机制进行算法的分析、改进以及对比。其中，重点是对基于差分隐私保护机制的多LBS服务器轮询算法，就其在效率性以及查询的准确度方面进行了改进。仿真实验的数据分析能够明确地表明，改进过后的算法能够解决地理位置轨迹数据存在的三种相关性导致的查询成功率较低的情况，同时，本文虽然是基于多个LBS服务器进行隐私保护，但是算法在实际运算时仅仅需要计算前后两个敏感位置的隐私风险暴露，降低了计算带来的开销。

同时，本文尚存在许多不足之处——对于仿真实验设计不够合理，不能够较为直观的凸显出算法改进的工作量以及难度；同时，仿真实验设计没有完美的展示出数据经过改进算法处理之后的数据情况，而是直接将处理后的数据进行了就针对QSP与时间指标的可视化处理并呈现，忽略了如何将隐私处理中的数据变换情况进行展示。

参考文献

1. 郑磊, 张俊星. 面向连续位置服务的位置感知隐私保护方案[J]. 浙江大学学报(工学版), 2020, 54(12): 2437-2444.
2. 程保容, 叶阿勇, 张强, 等. 一种基于隐私拆分的轨迹隐私保护方法[J]. 福建师范大学学报(自然科学版), 2020, 36(06): 28-35.
3. 张璐. 基于移动趋势的轨迹隐私保护方法研究[D]. 曲阜: 曲阜师范大学, 2020.
4. 张青云, 张兴, 李万杰, 等. 基于LBS系统的位置轨迹隐私保护技术综述[J]. 计算机应用研究, 2020, 37(12): 3534-3544.
5. 张青云, 张兴, 李万杰, 等. 位置轨迹隐私保护技术综述[J/OL]. 计算机应用研究: 1-12[2021-03-20]. https://doi.org/10.19734/j.issn.1001-3695.2019.10.0582.
6. 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.
7. 李杨, 温雯, 谢光强. 差分隐私保护研究综述[J]. 计算机应用研究, 2012, 29(09): 3201-3205.
8. 马银方. 社交网络中基于差分隐私保护的数据发布与挖掘研究[D]. 南京: 南京邮电大学, 2017.
9. 马银方, 张琳. 基于差分隐私的LBS群组最近邻查询[J]. 计算机科学, 2017, 44(S1): 336-341.
10. 张学军, 桂小林, 伍忠东. 位置服务隐私保护研究综述[J]. 软件学报, 2015, 26(09): 2373-2395.
11. 周长利, 马春光, 杨松涛. 基于敏感位置多样性的LBS位置隐私保护方法研究[J]. 通信学报, 2015, 36(04): 129-140.
12. Yi X, Paulet R, Bertino E, et al. Practical approximate K nearest neighbor queries with location and query privacy[J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(06): 1546-1559.
13. Jong S P, Philip S Y. An effective hash based algorithm for mining association rules[J]. ACM Sigmod record, 1995, 24(02): 175-186.
14. ]赵萍. 位置服务中的隐私保护关键技术研究[D].武汉: 华中科技大学, 2018.
15. Shin K G, Ju X E, Chen Z G, et al. Privacy protection for users of location-based services[J]. IEEE Wireless Communications, 2012, 19(01): 30-39.
16. 周长利, 马春光, 杨松涛. 基于敏感位置多样性的LBS位置隐私保护方法研究[J]. 通信学报, 2015, 36(04): 129-140.
17. Zhou C L, Ma C G, Yang S T, et al. A location privacy preserving method based on sensitive diversity for LBS[M]/ / Network and Parallel Computing. Berlin: Springer, 2014: 409-422.
18. 李敏, 秦志光. 路网环境下位置隐私保护技术研究进展[J]. 计算机应用研究, 2014, 31(09): 2576-2580.
19. 商宗民. 数据发布中匿名技术研究[D]. 哈尔滨: 哈尔滨工程大学, 2013.
20. Kido H, Yanagisawa Y, Satoh T. Protection of location privacy using dummies for location-based services[C]/ / Proc of International Conference on Data Engineering Workshops. Washington DC: IEEE Computer Society, 2005: 1248-1248.
21. Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location based services[C]/ / Proc of the 2nd International Conference on Pervasive Services. Piscataway, NJ: IEEE Press, 2005: 88-97.
22. Wu X C, Sun G Z. A novel dummy-based mechanism to protect privacy on trajectories[C]/ / Proc of IEEE International Conference on Data Mining Workshop. Piscataway, NJ: IEEE Press, 2014: 1120-1125.
23. Gruteser M, Liu X. Protecting privacy, in continuous location-tracking applications[J]. IEEE Security & Privacy, 2004, 2(02): 28-34.
24. 万盛. 位置服务的隐私保护关键技术研究[D]. 西安: 西安电子科技大学, 2018.
25. 张莹, 郑岩, 李华. 位置轨迹隐私保护专利技术综述[J]. 中国科技信息, 2021(22): 15-17+8.
26. 张峰. 基于PIR的保护位置隐私近邻查询技术研究[D]. 南京: 东南大学, 2017.
27. Khoshgozaran A, Shahabi C, Shirani-Mehr H. Location privacy: going beyond k-anonymity, cloaking and anonymizers[J]. Knowledge and Information Systems, 2011, 26(03): 435-465.
28. Chen X H, Pang J. Measuring query privacy in location-based services[C]/ / Proc of the 2nd ACM Conference on Data and Application Security and Privacy. New York: ACM Press, 2012: 49-60.
29. 武健. 位置服务中轨迹隐私保护方法研究[D]. 秦皇岛: 燕山大学, 2018.
30. 杨钊. 网络信息安全中的数据加密技术分析[J]. 集成电路应用, 2021, 38(12): 156-157. DOI:10.19339/j.issn.1674-2583.2021.12.069.
31. Khoshgozaran A, Shahabi C. Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy[C]/ / Proc of the 10th International Conference on Advances in Spatial and Temporal Databases. Berlin: Springer-Verlag, 2007: 239-257.
32. 梁慧超, 王斌, 崔宁宁, 等. 路网环境下兴趣点查询的隐私保护方法[J]. 软件学报, 2018, 29(03): 703-720.
33. 袁健, 王迪, 高喜龙, 等. 基于差分隐私的匿名组LBS轨迹隐私保护模型[J]. 小型微型计算机系统, 2019, 40(02): 341-347.
34. 安书山. 基于匿名化的地理社交网络隐私保护方法研究[D]. 北京: 北京交通大学, 2019.
35. 裴卓雄. 位置隐私保护中k-匿名技术研究[D]. 西安: 西安电子科技大学, 2017.
36. 张莹, 郑岩, 李华. 位置轨迹隐私保护专利技术综述[J]. 中国科技信息, 2021(22): 15-17+8.
37. 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(03): 222-230.
38. 王明. 面向地理位置服务的隐私保护机制研究[D]. 重庆: 重庆邮电大学, 2021. DOI:10.27675/d.cnki.gcydx.2021.001193.
39. 宋国超. 基于匿名技术的位置及轨迹隐私保护方法研究[D].青岛: 山东科技大学, 2020. DOI:10.27275/d.cnki.gsdku.2020.001559.
40. 王乐业. 群智感知中的地理位置本地化差分隐私机制: 现状与机遇[J]. 计算机科学, 2021, 48(06): 301-305.
41. 杨文娟. 基于差分隐私的网络社交中的隐私保护[J]. 无线互联科技, 2021, 18(22): 28-30.
42. 聂熠文. 基于差分隐私的数据发布技术研究[D].合肥: 中国科学技术大学, 2019.
43. 金媛媛, 倪志伟, 朱旭辉, 等. 基于本地差分隐私的空间数据自适应划分算法[J/OL]. 计算机工程: 1-13[2022-02-27]. DOI:10.19678/j.issn.1000-3428.0062382.
44. 朱勇. 基于差分隐私的聚类算法相关研究[D].南京: 南京邮电大学, 2021. DOI:10.27251/d.cnki.gnjdc.2021.000659.
45. 柳杰, 孔令山, 郭嵩, 等. 一种基于差分隐私的移动边缘计算算法[J]. 电脑与信息技术, 2021, 29(06): 5-7. DOI:10.19414/j.cnki.1005-1228.2021.06.002.
46. 张国鹏, 陈学斌, 王豪石, 等. 面向本地差分隐私的K-Prototypes聚类方法[J/OL]. 计算机应用: 1-10[2022-02-27].
47. 李晓会, 陈潮阳, 张兴, 等. 一种基于差分隐私的个性化服务推荐算法[J]. 现代电子技术, 2022, 45(04): 83-88. DOI:10.16652/j.issn.1004-373x.2022.04.016.
48. 张啸剑, 徐雅鑫, 孟小峰. 基于本地化差分隐私的空间数据近似k-近邻查询[J/OL]. 计算机研究与发展: 1-15[2022-02-27].
49. 彭春春, 陈燕俐, 荀艳梅. 支持本地化差分隐私保护的k-modes聚类方法[J]. 计算机科学, 2021, 48(02): 105-113.
50. 林静, 胡德敏, 王揆豪. 一种改进的差分隐私聚类位置保护方法[J]. 软件导刊, 2021, 20(12): 133-137.
51. 胡郁葱, 司徒炳强, 靳文舟, 等. 基于集散点选择模型的地下快速路网规划[J]. 华南理工大学学报(自然科学版), 2009, 37(02): 65-69.
52. 刘海燕. 移动数据生成与查询测试平台的设计与实现[D]. 沈阳: 东北大学, 2012.
53. 吴云乘, 陈红, 赵素云, 等. 一种基于时空相关性的差分隐私轨迹保护机制[J]. 计算机学报, 2018, 41(02): 309-322.
54. 赵婧, 张渊, 李兴华, 等. 基于轨迹频率抑制的轨迹隐私保护方法[J]. 计算机学报, 2014, 37(10): 2096-2106.

致谢

谨向敬爱的导师段旭良副教授致以诚挚的感谢！从论文的构思到最终定稿，都凝结着老师的汗水。感谢就该论文论述内容做出了巨大贡献的前辈们。谢谢我的婆婆爷爷、爸爸妈妈对我学习上的鼓励与支持！还有就是感谢室友们积极上进，为我营造了良好的学习氛围。对完成本论文起到任何作用的人或事，感谢！

附录

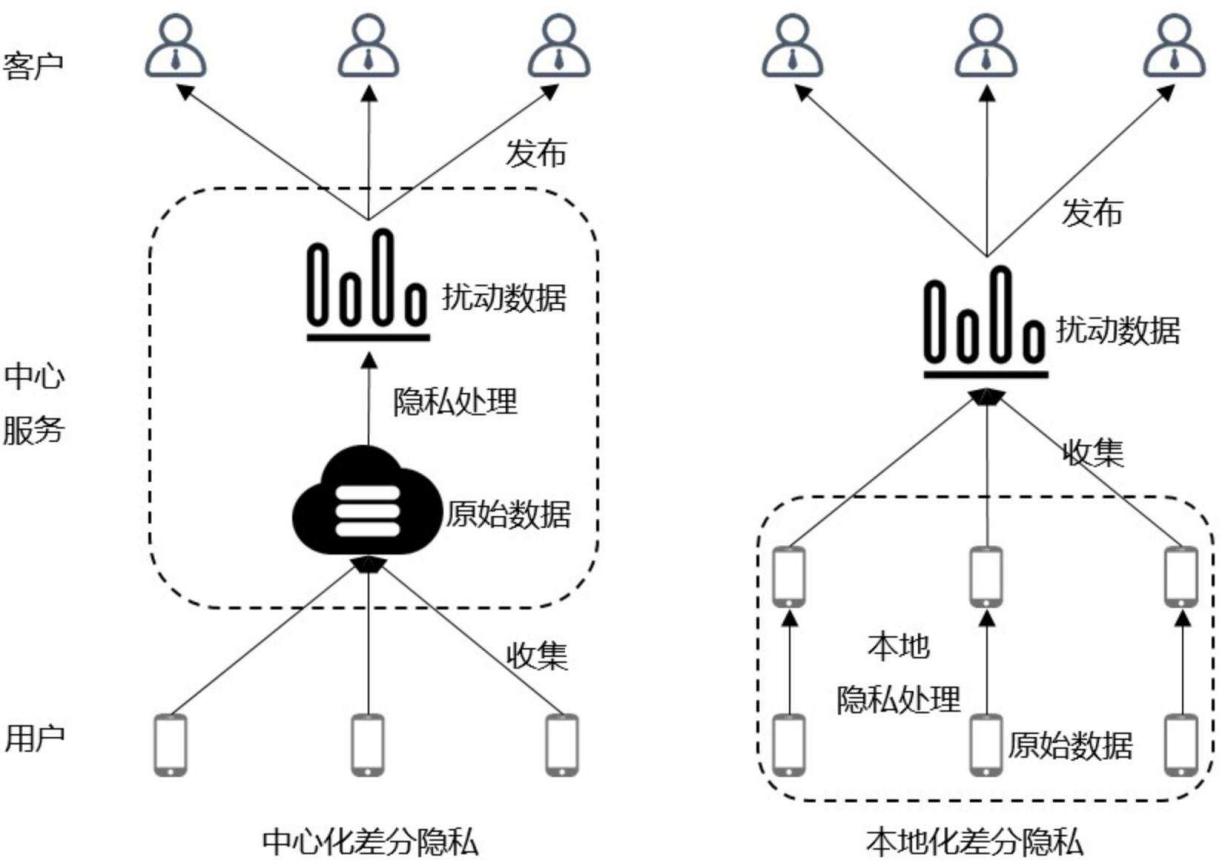


图17 集中式与本地化差分隐私比较