多级本地化差分隐私算法推荐框架

王瀚仪[1,2]，李效光[3]，毕文卿[1,2]，陈亚虹[1,2]，李凤华[1,2]，牛犇[1]
（1. 中国科学院信息工程研究所，北京 100093；2. 中国科学院大学网络空间安全学院，北京 100049；
3. 西安电子科技大学网络与信息安全学院，陕西 西安 710071）

摘 要：本地化差分隐私（LDP）算法通常为不同用户分配相同的保护机制及参数，却忽视了不同用户终端设备资源与隐私需求的差异。为此，提出一种多级 LDP 算法推荐框架。该框架考虑服务商以及用户的需求，通过服务商和用户的多级管理实现多用户差异化隐私保护。将框架应用至频数统计场景形成 LDP 算法推荐方案，改进 LDP 算法以保证统计结果的可用性，设计协同机制保护用户的隐私偏好。实验结果证明了所提方案的可用性。

关键词：本地化差分隐私；资源自适应；个性化隐私预算
中图分类号：TP309
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022106

Multi-level local differential privacy algorithm recommendation framework

WANG Hanyi[1,2], LI Xiaoguang[3], BI Wenqing[1,2], CHEN Yahong[1,2], LI Fenghua[1,2], NIU Ben[1]
1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
3. School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract: Local differential privacy (LDP) algorithm usually assigned the same protection mechanism and parameters to different users. However, it ignored the differences among the device resources and the privacy requirements of different users. For this reason, a multi-level LDP algorithm recommendation framework was proposed. The server and the users' requirements were considered in the framework, and the multi-users' differential privacy protections were realized by the server and the users' multi-level management. The framework was applied to the frequency statistics scenario to form an LDP algorithm recommendation scheme. LDP algorithm was improved to ensure the availability of statistical results, and a collaborative mechanism was designed to protect users' privacy preferences. The experimental results demonstrate the availability of the proposed scheme.

Keywords: local differential privacy, resource adaptation, personalized privacy budget

0 引言

随着移动通信技术的发展，数据的重要性越来越突出，服务商试图通过分析用户的数据为用户提供更加个性化的优质服务。然而，随着互联网用户隐私保护意识的觉醒，以及多个国家相继出台隐私保护法案带来的双重压力，如何在保护用户隐私信息的前提下采集数据成为各大互联网服务商刻不容缓的任务。在学术界和工业界的共同推动下，本地化差分隐私（LDP, local differential privacy）技术逐渐成为保护用户隐私数据的重要技术手段之一。LDP 借助其严格的数学定义，在不需要可信第三方

收稿日期：2022−02−09；**修回日期**：2022−04−18
通信作者：牛犇，niuben@iie.ac.cn
基金项目：国家重点研发计划基金资助项目（No.2021YFB3100300）；国家自然科学基金资助项目（No.61872441, No.61932015）
Foundation Items: The National Key Research and Development Program of China (No.2021YFB3100300), The National Natural Science Foundation of China (No.61872441, No.61932015)

参与的情形下，抵御任意背景知识的攻击者，并且通过隐私预算参数，量化了算法隐私保护的强度，充分满足了被采集用户对隐私信息保护的需求；同时，LDP 还能够在用户数据量充足的条件下，保证统计分析结果的可用性。

LDP 技术的研究众多，各类统计分析任务下的 LDP 技术层出不穷，然而参与统计分析任务的用户均需要采用相同的 LDP 保护机制，选取相同的隐私预算参数，忽略了用户动态变化的资源环境与不同用户对资源、隐私的个性化需求。

面对现实中移动终端资源的动态变化与限制，以及用户对个性化的追求，单一的隐私保护算法不足以支撑用户的需求。例如，打车服务商想通过频繁采集用户位置打卡信息来选取合适的打车等候站点。一方面，用户不想泄露自己的位置信息，因此服务商在信息采集过程中需要为用户提供隐私保护；另一方面，服务商希望得到一个比较准确的统计结果，因此常用 LDP 算法解决这类问题。尽管 LDP 有很多不同的机制，但是现有方案中所有用户必须采用相同的 LDP 机制，服务商才能够计算出较为准确的结果，这种方案固化带来的问题是一旦用户设备资源受限，则会使结果不准确或者无法采集该用户的信息。例如，某用户手机电量不足，若和其他用户一样选择电量消耗较大的机制，手机很可能直接由于电量耗尽而关机，无法执行任务也无法正常通信；若所有用户都选择电量消耗较小的机制，考虑到消耗小的机制往往可用性较差，因此服务商计算结果的可用性就会急剧降低。

为了权衡用户对资源、隐私的偏好以及服务商对可用性的需求，本文设计了多级 LDP 算法推荐框架，考虑到服务商对统计结果可用性的要求以及用户对资源、隐私的个性化需求，通过多级管理为不同用户推荐不同的但适合当前自身资源环境的 LDP 算法，实现多用户差异化隐私保护。进一步，将框架应用至位置服务中的位置兴趣点（PoI, point of interest）频数统计任务中，改进 LDP 算法以保证统计结果的可用性，设计协同机制保护用户的隐私偏好。

本文主要的贡献如下。

1) 设计一个多级 LDP 算法推荐框架，解决 LDP 技术忽略用户之间差异的问题。从隐私信息全生命周期保护的角度出发，分 5 个步骤对用户信息进行保护。充分考虑用户对资源、隐私的个性化需求以及服务商对统计结果的可用性要求，通过服务商和用户对 LDP 算法进行选取，既保证了服务商对用户的管理，又不限制用户的个性化需求。

2) 将框架落地到 PoI 频数统计场景中，并将框架具体化。选择 4 种典型频数 LDP 算法，考虑电量、流量 2 个资源因素以及算法的可用性，为用户推荐当前环境下的最优 LDP 算法。通过对 LDP 机制进行改进，使不同用户即使采用不同的机制，仍能保证服务商计算出的统计结果为真实结果的无偏估计。在此基础上，设计开销较小的用户协同机制，保护用户的隐私需求。

3）实验结果表明，本文方案能够使资源自适应地为用户推荐个性化的 LDP 算法，并且能够保证统计结果的可用性。

1　相关工作

差分隐私（DP, differential privacy）[1]是一个具有严格数学定义的隐私保护概念，即任意一条记录的增加或删除都不会影响查询结果，也就避免了让攻击者了解更多的信息。基于此，学者们提出了多种差分隐私保护技术。Dwork 等[2]提出了 Laplace 机制，针对连续的数值型查询结果添加服从 Laplace 的噪声。Mcsherry 等[3]针对非数值型数据的查询结果提出了指数机制，以一定概率从结果集合中选择一个结果输出。

标准的 DP 通过可信的数据中心收集用户数据，并发布满足差分隐私的用户数据统计分析结果。然而在实际应用中，很难找到完全可信的数据中心，由此 LDP 应运而生，其能够在用户端执行满足差分隐私的保护机制，同时保证统计结果的准确性。按照服务商的数据统计分析的任务类型不同，LDP 机制主要可以划分为三类[4]：频数统计任务[5-10]、均值统计任务[11-12]和复杂任务[13-14]。还有一些工作对 DP 概念中的隐私预算进行研究，例如在 LDP 中为不同用户分配不同的隐私预算，或探讨如何为 DP 算法选择合适的隐私预算。

1.1　频数统计任务下的 LDP

有多种 LDP 机制被设计用来求取用户的频数结果，如统计某个年龄段的用户数。这些机制或不对数据进行编码处理[5-6]，或采用二进制向量[7]、hash 函数[7,15]、矩阵转换[8]等手段对用户数据进行编码，并对编码数据或原始数据扰动后发送给服务商，服务商再对数据进行校准、聚合，得到频数估

计结果。Google 团队提出了 Rappor 算法[16]，应用随机应答（RR, randomized response）扰动机制，使服务商能够在隐藏用户字段的前提下，从用户处获得字段频数的统计结果。

除此之外，还扩展出许多新型频数统计任务，例如频繁项集挖掘[9]、针对 key-value 类型数据的频数估计[10]等。Ye 等[10]提出 PrivKV 算法，能够在满足 LDP 的条件下对 key-value 数据进行保护，并且保留 key 和 value 之间的关联关系；为了提高结果的准确度，Ye 等在此基础上改进算法，对 PrivKV 进行多次迭代，为了减少网络时延，进一步将多次迭代优化为虚拟迭代，减少了用户的参与。

1.2 均值统计任务下的 LDP

均值统计任务是求取多个用户数据的均值，如求用户的平均年龄。针对均值统计任务，最基础的 LDP 机制是利用 Laplace 机制[2]分别对用户数据添加噪声。Duchi 等[11]提出了一种适用于多维数值型数据的 LDP 机制，其基本思想是利用随机响应技术，根据一定的概率分布扰动每个用户的数据，同时确保均值统计结果的无偏估计。然而，Nguyên 等[12]指出当数据维数为偶数时，Duchi 等的算法并不能满足 LDP 的定义，并对该算法进行了修正，同时提出了 Harmony 算法，满足与 Duchi 等的算法相同的隐私保护强度和可用性，但大大减小了算法的通信量。

Wang 等[17]提出了针对一维数值型数据的 PM 算法，相对 Duchi 等的算法而言，统计结果有着更小的方差，即更好的可用性，同时实现也更加简易；进一步地，Wang 等基于 Harmony 算法将 PM 算法扩展至高维数据，并设计 HM 算法以处理分类型的数据。

1.3 复杂任务下的 LDP

Yilmaz 等[13]提出利用 LDP 算法来训练朴素的贝叶斯分类器，首先将每个用户的标签和取值转化成一个新的值，然后对该值执行 LDP 机制下的扰动，在保留标签和取值之间关系的同时，保护用户的训练数据。Mahawage 等[14]改进已有频数统计任务下的 LDP 算法，用于控制卷积神经网络中的隐私泄露，同时提出效用增强随机化机制，进一步提高随机化二进制字符串的可用性。Shin 等[15]将 LDP 机制应用到推荐系统中，在用于矩阵分解的梯度下降算法中给梯度添加噪声，保护用户在交互过程中的真实梯度不被服务商获取，从而保护用户的真实属性以及对应评分。进一步地，为了减少开销，Shin 等[15]引入降维技术并提出基于采样的二元机制，减小算法的开销，同时在一定参数范围内也能保证较好的推荐准确性。Zhao 等[18]将联邦学习与 LDP 相结合，提出多种 LDP 机制，以在保护用户隐私、降低通信成本的前提下实现机器学习模型。

1.4 个性化的 LDP

Chen 等[19]首次提出了个性化本地化差分隐私（PLDP, personalized local differential privacy）的概念，并进一步提出了个性化计数估计协议，利用用户组聚类算法将该协议应用于不同隐私级别的用户。Nie 等[20]提出了一个框架来优化直方图估计，其中利用个性化隐私下的数据回收方案扩大估计的样本量，并证明该框架具有最优效用。Xia 等[21]利用本地化差分隐私技术执行 k-means 聚类任务，为了提高结果的可用性，给二进制数据的不同比特位分配了不同的隐私预算，并考虑到不同用户具有不同的隐私需求，因此设计不同用户参与扰动的比特位不同。Shen 等[22]提出了新的 PLDP 概念，改进频数统计任务下的 OUE 算法，设计了多维联合分布估计方案，为不同维度的数据分配不同的隐私预算，使其比传统 LDP 具有更好的效能。

1.5 DP 隐私预算参数选择

在差分隐私中，隐私预算参数 ε 表示算法的隐私保护程度，ε 越小，保护程度就越高。针对如何选取参数 ε 的问题，Naldi 等[23]提出了一种基于区间估算理论的选择方法，用置信区间和置信水平 2 个参数来衡量 ε。Shahani 等[24]在给定场景下通过权衡隐私保护和可用性对 ε 进行选择，并给出了 ε 的一个上界。

综上，LDP 算法可应用在多种任务场景下，种类众多，且不同 LDP 算法在资源开销、可用性上各有优势，但不同用户往往采用相同的 LDP 算法。并且，尽管隐私预算参数的大小决定了算法的隐私保护强度，但少有工作为不同用户选取不同的隐私预算。

2 预备知识

2.1 本地化差分隐私

定义 1 一个随机算法 M 满足 ε-LDP，当且仅当对于任意 2 条不同记录 t，$t' \in \mathrm{Domain}(M)$，以及任意 $y \in \mathrm{Range}(M)$，都有

$$\Pr(M(t)=y) \leqslant \mathrm{e}^{\varepsilon} \Pr(M(t')=y) \tag{1}$$

因此，2 个不同的数据经过随机算法 M 扰动

后，接收者无法通过收到的数据来分辨二者。

2.2　个性化本地化差分隐私

由于不同用户的隐私需求不同，Chen 等[19]提出 PLDP 的概念。PLDP 中包含 2 个参数，第一个参数是安全区域，即用户认为可以透露的最小区域 τ，例如，该用户不介意别人知道其所在位置为北京，可以将安全区域设置为北京，则用户希望算法能够使真实位置与安全区域内的其他位置无法区分开；第二个参数是隐私预算 ε，与 LDP 中的概念相同，ε 表示算法限制攻击者区分任意 2 个位置的能力大小，ε 越小，表示该能力越高，称 (τ,ε) 为某特定用户的隐私规范。基于此，(τ,ε)-PLDP 的基本概念如下。

定义 2　一个随机算法 M 对某用户满足 (τ,ε)-PLDP，当且仅当对于任意 2 个位置 $t,t'\in\tau$，以及任意 $O\subseteq \mathrm{Range}(M)$，都有

$$\Pr(M(t)\in O)\leqslant \mathrm{e}^{\varepsilon}\Pr(M(t')\in O) \tag{2}$$

2.3　问题与挑战

本文旨在解决本地化差分隐私机制各异，但不同用户往往采用相同的机制和参数，无法根据用户当前的资源环境和隐私需求灵活地为用户提供细粒度的隐私保护的问题。为了实现该目标，本文需解决以下 3 个挑战。

1) 现有 LDP 算法为了便于服务商对扰动后的用户数据进行无偏估计，所有用户均采用相同的机制和参数。然而不同 LDP 算法在资源开销、统计分析结果可用性上各有优劣，因此能否结合不同机制的不同特性，为资源开销、隐私需求各异且动态变化的用户推荐个性化的 LDP 算法，并且仍能保证结果的无偏，是本文面临的第一个挑战。2) 如何在有限且动态变化的资源环境下，为用户选择适配当前资源环境的 LDP 机制，是本文面临的第二个挑战。3) 不同用户对隐私的需求不同，用户选取的隐私预算参数能够反映用户对隐私的个性化偏好，若某个用户对隐私的需求较低，攻击者可以集中针对该用户进行攻击，因此，如何在不泄露该偏好的情况下执行方案，是本文面临的第三个挑战。

2.4　攻击模型

本文设定服务商和用户是诚实而好奇的（HBC, honest-but-curious），他们诚实地遵守设计的方案，但是会试图根据已知信息推测其他用户的更多信息。本文涉及的系统参数如表 1 所示。

表 1　系统参数

参数	含义
ALG	LDP 算法集合
alg_i	服务商为用户 i 选择的 LDP 算法集合
$r_1,r_2,\cdots,r_o$	LDP 算法的各类资源开销，其中，r_1 为电量，r_2 为流量
r_{var}	LDP 算法的可用性
$a_1,a_2,\cdots,a_h$	影响 LDP 算法资源开销与可用性的因素
$w_1,w_2,\cdots,w_o$	用户 i 对 o 类资源开销的权重
w_{var}	用户 i 对可用性的权重
$\mathrm{alg}_{\mathrm{opt}}$	用户选择的最优 LDP 算法
ε_i	用户 i 的个性化隐私预算
D	位置 PoI 集合
D_i	用户所选扰动范围，$D_i\subseteq D$
n	用户个数
g	后处理环节中的第 g 个群组
m_g	群组 g 内的用户数
G_g	群组 g 内用户统一的扰动结果集合
G'_g	群组 g 内用户统一的扰动范围集合
F^{d_l}	所有用户访问位置 d_l 的估计频数

3　多级 LDP 算法推荐

3.1　算法推荐框架

Li 等[25]提出隐私计算及其框架，该框架从隐私信息全生命周期保护的角度出发，分为隐私信息提取、场景抽象、隐私操作选取、隐私保护方案设计或选取、隐私保护效果评估 5 个步骤。本文基于该框架，设计了多级 LDP 算法推荐框架，分为 5 个步骤，图 1 简要描述了所提框架。

1) 隐私信息提取。用户对服务商想要采集的信息进行评估，并根据信息的敏感程度，设定自己的隐私预算大小。

2) 场景抽象。服务商通过自身需求确定 LDP 算法的计算场景，例如，是求取用户信息的平均值，还是执行更复杂的任务。

3) 隐私操作选取。确定场景后，服务商确定满足场景的算法集合 ALG；然后，服务商会根据用户的级别，依照策略从所有算法集合 ALG 中为每个用户 i 选择一个算法集合 $\mathrm{alg}_i\subseteq \mathrm{ALG}$，方便服务商管控。策略机制可以根据需求自定义，例如，作为普通用户 A，服务商为用户开放部分算法，即

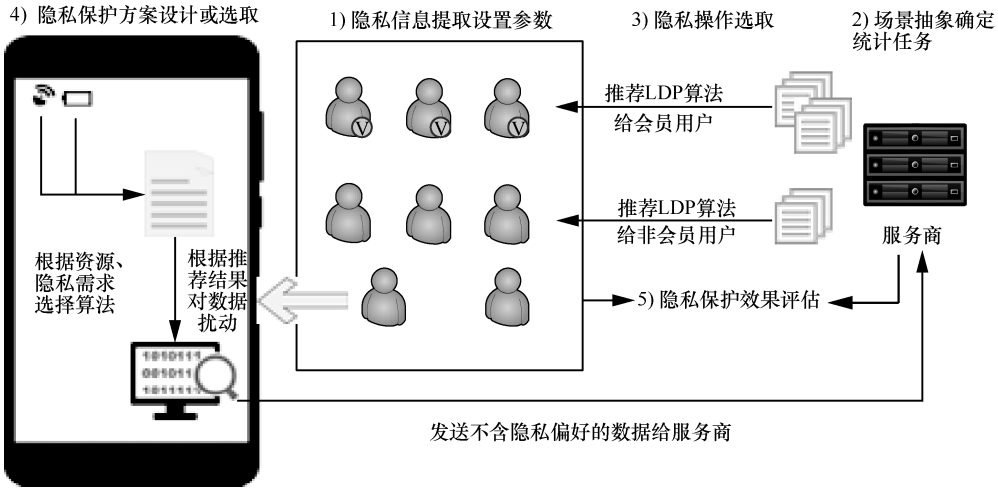

图 1　多级 LDP 算法推荐框架

$\text{alg}_\text{A} \subset \text{ALG}$，对资源灵活性要求不高的用户提供更简洁的选择；相反，若用户 B 对资源灵活性要求较高，成为服务商的会员，那么服务商会为用户 B 开放所有的算法供用户选择，即 $\text{alg}_\text{B}=\text{ALG}$。

4) 隐私保护方案设计或选取。具体细分为如下 4 个步骤。

① 服务商预处理。为了方便用户对算法进行选取，服务商需要先进行预处理，得到各项资源开销与影响因素之间的关系；同时从服务商自身考虑，也需要将算法的可用性纳入影响用户选取 LDP 算法的因素中，以便增加统计结果的可用性。考虑 o 类资源开销 $r_1,r_2,\cdots,r_o$，设定影响 LDP 算法资源开销因素有 h 个，分别为 $a_1,a_2,\cdots,a_h$。利用实验拟合集合 ALG 中每个算法 alg 的各类资源开销与影响因素之间的关系 $r_j=f_{\text{alg}}^{j}(a_1,a_2,\cdots,a_h),1\leqslant j\leqslant o,\text{alg}\in\text{ALG}$；同时，利用 LDP 算法结果的方差来衡量其可用性，因此可以通过理论推导得到算法 alg 的可用性 r_{var} 与相应的影响因素之间的关系 $r_{\text{var}}=f_{\text{alg}}^{\text{var}}(a_1,a_2,\cdots,a_h),\text{alg}\in\text{ALG}$。

② 用户具体算法选取。用户 i 定义自身对 o 类资源开销的权重 $0\leqslant w_1,w_2,\cdots,w_o\leqslant 1$，权重越高表示对该类资源的消耗越重视。服务商统一定义对算法可用性的权重 $0\leqslant w_{\text{var}}\leqslant 1$，同时根据用户在当前状态下影响因素与权重的取值，利用多目标决策方法选择最优算法 $\text{alg}_{\text{opt}}\in\text{alg}_i$。

③ LDP 算法扰动。用户选取本地化差分隐私个性化隐私预算 ε_i，利用所选算法对用户的隐私数据 data_i 进行扰动得到 $\text{result}_i=\text{alg}_{\text{opt}_i}(\text{data}_i,\varepsilon_i)$，再将结果 result_i 在隐藏自身隐私预算 ε_i 的前提下进行校准，并发给服务商，其中，用户对隐私保护强度要求越高，选择的隐私预算值就越小，具体选取方法可以参考 1.5 节中的相关工作。

④ 服务商结果整合。服务商收到用户的数据后对数据进行聚合，得到目标结果 R。

5) 隐私保护效果评估。由于本文对 LDP 算法进行推荐，因此可以用实验的形式来量化分析方案的偏差性和复杂性[25]。

3.2　基于位置 PoI 频数统计的 LDP 算法推荐

本节具体考虑位置 PoI 频数统计场景下对 LDP 算法的推荐，由于本文侧重点在于如何推荐 LDP 算法，因此图 1 中的步骤 1)、步骤 5)在此部分不作研究。本文框架不仅可以用于频数统计任务的 LDP 算法，还可以用于其他 LDP 算法中，但是，一是由于均值统计算法一般不需要校准，不涉及泄露用户隐私偏好的情况；二是复杂类型的 LDP 算法一般由均值或频数统计任务演化而来，因此本文以频数统计的场景为例进行讨论。

3.2.1　方案设置

设定位置 PoI 集合 $D=\{d_1,d_2,\cdots,d_L\}$，共 L 个 PoI。现有 n 个用户，服务商以商业选址为目的想要统计这 n 个用户中处在各个 PoI 的人数。用户 i 此时所在的 PoI 为 $d_s\in D$，即集合 D 中的第 s 个点，并且假设用户在该场景下关心隐私算法给用户移动终端带来的电量和流量的开销，即考虑不同算法带来的电量、流量资源开销 r_1,r_2。当用户场景发生变化，需要考虑更多因素（例如时间成本等）时，可以将资源因素进行扩展，并仍遵循

本节的方案步骤进行处理。

根据 PLDP 的定义，用户 i 根据自己的需求选定 PoI 集合 $D_i \subseteq D$，则用户扰动后的结果均在集合 D_i 内。为方便起见，本文给定 D_i 供用户选择，即 $D_i \in \{\bar{D}_1, \bar{D}_2, \cdots\}$，其中 $\bar{D}_j \subseteq D$，因此在传输 D_i 时，用户只需要传输一个下标即可；同时，用户根据隐私偏好选择隐私预算 ε_i。

本节选取了 4 种单值频数统计的 LDP 算法作为算法集合，分别是 DE（direct encoding）、OUE（optimized unary encoding）、OLH（optimized local hashing）和 THE（thresholding with histogram encoding）[7]，即 ALG = {DE,OUE,OLH,THE}。单值频数统计[26]是指每个用户只发送一个变量取值给数据收集者，数据收集者根据所有用户上传的取值统计每一个候选值的频数结果，并进行发布。本节选取的 4 种算法代表了最典型的 LDP 频数统计算法，4 种算法分别采用了不同的编码机制，故产生了有差别的资源消耗。下面，简单介绍这 4 种算法，其中假设用户选择的数据扰动范围集合为 D，本地化差分隐私预算为 ε。

1) DE 算法。该算法不对数据 d_s 进行编码，在用户端以概率对原始数据 d_s 进行扰动，其中，以概率 p 保持原始数据不变，以概率 q 将数据扰动至集合 D 内的任意除 d_s 的其他取值。用户将扰动后结果发送给服务商，服务商统计扰动后结果为 d_t 的用户数 k_t，并校准得到取值为 d_t 的用户的频数估计 $f_t = \dfrac{k_t - nq}{p - q}$。其中，$p = \dfrac{\mathrm{e}^{\varepsilon}}{\mathrm{e}^{\varepsilon} + |D| - 1}$，$q = \dfrac{1}{\mathrm{e}^{\varepsilon} + |D| - 1}$。

2) OUE 算法。该算法首先对 d_s 进行编码，得到一个长度为 $|D|$ 的 0-1 向量 $\boldsymbol{t}_i = (t_{i,d_1}, t_{i,d_2}, \cdots, t_{i,d_{|D|}})$，其中，$t_{i,d_s}$ 对应的分量为 1，其余分量均为 0。在用户端对 0-1 向量的每一个分量进行扰动，如果该分量为 1，则以概率 p 保持取值不变；如果该分量为 0，则以概率 q 对其进行翻转，由 0 变为 1。所有用户将扰动后的向量 $\boldsymbol{t}'_i = (t'_{i,d_1}, t'_{i,d_2}, \cdots, t'_{i,d_{|D|}})$ 发给服务商，服务商统计在 d_t 对应位置分量为 1 的用户数 k_t，校准扰动结果，得到取值为 d_t 的用户的频数估计为 $f_t = \dfrac{k_t - nq}{p - q}$。其中，$p = \dfrac{1}{2}$，$q = \dfrac{1}{\mathrm{e}^{\varepsilon} + 1}$。

3) OLH 算法。该算法需要对 d_s 进行哈希编码，哈希算法集 $\mathbb{H}$ 中的每个哈希算法都将集合 D 中的元素映射到大小为 $\lfloor \mathrm{e}^{\varepsilon} + 1 \rfloor$ 的集合 E 中，用户随机选取哈希算法集 $\mathbb{H}$ 中的一个哈希算法 H，对 $x = H(d_s)$ 进行扰动，以概率 p 保持 x 取值不变，以 $\dfrac{1}{2\mathrm{e}^{\varepsilon}}$ 的概率将数据扰动至集合 E 内的任意其他取值。用户将扰动结果 y 发给服务商，服务商对结果进行解码得到 x'，即服务商遍历 $x' \in D$，选取所有满足 $H(x') = y$ 的 x' 值作为该用户的解码结果。服务商统计所有解码后为 d_t 的用户数 k_t，对其校准得到取值为 d_t 的用户的频数估计为 $f_t = \dfrac{k_t - nq}{p - q}$。其中，$p = \dfrac{1}{2}$，$q = \dfrac{1}{\mathrm{e}^{\varepsilon} + 1}$。

4) THE 算法。该算法同样对 d_s 进行编码，得到长度为 $|D|$ 的 0-1 向量 $\boldsymbol{t}_i = (t_{i,d_1}, t_{i,d_2}, \cdots, t_{i,d_{|D|}})$，其中，$t_{i,d_s}$ 对应的分量为 1，其余分量均为 0。用户端对每一个分量添加一个满足分布为 $\mathrm{Laplace}\left(\dfrac{2}{\varepsilon}\right)$ 的噪声，得到 $\boldsymbol{t}'_i = (t'_{i,d_1}, t'_{i,d_2}, \cdots, t'_{i,d_{|D|}})$，同时设定阈值 θ，其中 $\dfrac{1}{2} < \theta < 1$，对每一个扰动后的分量 $t'_{i,j}$，若 $t'_{i,j} > \theta$，则令 $t'_{i,j} = 1$；若 $t'_{i,j} \leqslant \theta$，则令 $t'_{i,j} = 0$。用户将向量 $\boldsymbol{t}'_i$ 发送给服务商，服务商统计在 d_t 对应位置分量为 1 的用户数 k_t，校准扰动结果，得到取值为 d_t 的用户的频数估计为 $f_t = \dfrac{k_t - nq}{p - q}$。其中，$p = 1 - \dfrac{1}{2}\mathrm{e}^{\frac{\varepsilon}{2}(\theta - 1)}$，$q = \dfrac{1}{2}\mathrm{e}^{-\frac{\varepsilon}{2}\theta}$。

本文方案具体分为 5 个步骤，分别为服务商算法集管控、服务商预处理、用户具体算法选取、LDP 算法扰动、服务商结果整合，具体步骤在 3.2.2~3.2.6 节中分别详细介绍。

3.2.2　服务商算法集管控

服务商根据用户的级别为用户推荐不同的算法，在本节中设置策略如下：为普通用户推荐算法 DE 和 OLH，为会员用户推荐 4 种算法。这里选择 DE 和 OLH 是因为 DE 在 $|D_i|$ 较小时，开销很小，可用性也不大；OLH 开销较大，可用性却相对稳定，所以服务商为用户提供的算法集可供对灵活性要求不高的用户使用。记服务商给用户 i 推荐算法集合为 alg_i。

3.2.3　服务商预处理

服务商通过预处理分别建立资源开销、可用性与影响因素之间的关系，其中，根据在 3.2.5 节的

得到的 LDP 改进算法可以发现，算法流量和电量的开销主要与用户选取的扰动范围大小$|D_i|$以及隐私预算ε_i有关，而且当ε_i固定时，开销与$|D_i|$成正比例关系。因此，固定ε_i，测试在扰动范围$|D_i|$变化时，各个算法在用户终端上电量和流量的消耗，然后利用最小二乘法拟合得到关系式$r_1 = f_{\text{alg}}^1(|D_i|)$以及$r_2 = f_{\text{alg}}^2(|D_i|)$，并且调节$\varepsilon_i$，得到在不同$\varepsilon_i$下的关系式。此外，本节利用不同 LDP 算法理论上的方差来描述可用性与各个影响因素之间的关系，由于算法的方差同样与$|D_i|$和ε_i有关，因此固定ε_i时，可以得到算法的可用性关系式$r_{\text{var}} = f_{\text{alg}}^{\text{var}}(|D_i|)$。

3.2.4 *用户具体算法选取*

首先建立用户对各项资源的权重，可以根据用户偏好主观地进行设置，也可以由服务商建立指导设置，后期可以根据用户需求进行调整。本节给出了一种客观判断权重的方案，为用户提供指导：根据用户当前的剩余电量b_r（满电量是 100），使用的流量套餐中的剩余流量f_r（MB），以及每超出套餐 1 MB 需要的价格f_p（元），建立用户本身对电量和流量重视程度的权重值w_1和w_2。其中，$w_1 = 1 - \dfrac{b_r}{100}$，当手机充电时，用户不会担心电量消耗，因此$w_1 = 0$；当手机连接 Wi-Fi 或者剩余流量充足时，用户不会担心流量消耗，因此$w_2 = 0$；当使用流量且剩余流量不足时，令$w_2 = 3.33 f_p$（现有流量套餐超出的最高单价为 0.3 元/MB）。因此，权重设置为

$$w_1 = \begin{cases} 1 - \dfrac{b_r}{100}, & \text{未充电} \\ 0, & \text{充电} \end{cases} \tag{3}$$

$$w_2 = \begin{cases} 0, & \text{连接Wi-Fi 或} f_r > 1\ 024 \\ 3.33 f_p, & f_r \leqslant 1\ 024 \end{cases} \tag{4}$$

服务商对可用性的权重设置为$w_3 = 1$。

用户根据所选择的扰动范围的大小$|D_i|$得到每个算法的具体电量和流量消耗$r_1^{\text{alg}} = f_{\text{alg}}^1(|D_i|)$，$r_2^{\text{alg}} = f_{\text{alg}}^2(|D_i|)$，以及算法可用性$r_{\text{var}}^{\text{alg}} = f_{\text{alg}}^{\text{var}}(|D_i|)$，本节采用算法的方差来衡量可用性。利用带权重的优劣解距离法（TOPSIS, technique for order preference by similarity to an ideal solution）找到服务商推荐的算法中的最优方案。具体操作如下。

1) 数据正向化。由于算法的电量和流量消耗以及可用性都是越小越好，因此需要将这 3 个指标值进行正向化处理，使之越大越好，故计算正向化后的结果为$\overline{r}_1^{\text{alg}} = \dfrac{1}{r_1^{\text{alg}}} = \dfrac{1}{f_{\text{alg}}^1(|D_i|)}$，$\overline{r}_2^{\text{alg}} = \dfrac{1}{r_2^{\text{alg}}} = \dfrac{1}{f_{\text{alg}}^2(|D_i|)}$，$\overline{r}_{\text{var}}^{\text{alg}} = \dfrac{1}{r_{\text{var}}^{\text{alg}}} = \dfrac{1}{f_{\text{alg}}^{\text{var}}(|D_i|)}$。

2) 归一化。正向化后，由于电量、流量、可用性的量纲不同，因此要对$\overline{r}_1$、$\overline{r}_2$、$\overline{r}_{\text{var}}$进行归一化处理。采用最大最小归一化方法，参考文献[27]，避免归一化区间过小导致数据区分度不大，以及避免后续步骤中将数据放至分母处导致不可计算，因此选取归一化区间[0.01,0.99]，对所有$\text{alg} \in \text{ALG}$计算归一化结果$\tilde{r}_1^{\text{alg}} = 0.01 + 0.98\dfrac{\overline{r}_1^{\text{alg}} - \min_1}{\max_1 - \min_1}$，$\tilde{r}_2^{\text{alg}} = 0.01 + 0.98\dfrac{\overline{r}_2^{\text{alg}} - \min_2}{\max_2 - \min_2}$，$\tilde{r}_{\text{var}}^{\text{alg}} = 0.01 + 0.98\dfrac{\overline{r}^{\text{alg}} - \min_{\text{var}}}{\max_{\text{var}} - \min_{\text{var}}}$，其中，$\max_x = \max\{\overline{r}_x^{\text{alg}} \mid \text{alg} \in \text{ALG}\}, x = \{1, 2, \text{var}\}$，$\min_x = \min\{\overline{r}_x^{\text{alg}} \mid \text{alg} \in \text{ALG}\}, x = \{1, 2, \text{var}\}$。

3) 评分构建。建立最优方案$\text{V}^+ = \{\max_1', \max_2', \max_{\text{var}}'\}$和最差方案$\text{V}^- = \{\min_1', \min_2', \min_{\text{var}}'\}$，分别表示各项指标全部最优或最差情况下的方案。其中，$\max_x' = \max\{\tilde{r}_x^{\text{alg}} \mid \text{alg} \in \text{ALG}\}, x = \{1, 2, \text{var}\}$，$\min_x' = \{\tilde{r}_x^{\text{alg}} \mid \text{alg} \in \text{ALG}\}, x = \{1, 2, \text{var}\}$。

最终计算算法的评分数为$C_{\text{alg}} = \dfrac{D_{\text{alg}}^-}{D_{\text{alg}}^- + D_{\text{alg}}^+}$，其中，$D_{\text{alg}}^+$和$D_{\text{alg}}^-$分别是算法 alg 到最优方案和最差方案的加权距离，即

$$D_{\text{alg}}^+ = \sqrt{\sum_{x=1,2,\text{var}} w_x(\tilde{r}_x^{\text{alg}} - \max_x')^2}, x = \{1, 2, \text{var}\}$$

$$D_{\text{alg}}^- = \sqrt{\sum_{x=1,2,\text{var}} w_x(\tilde{r}_x^{\text{alg}} - \min_x')^2}, x = \{1, 2, \text{var}\}$$

4) 推荐算法。评分最高的算法alg_{opt}将作为用户的推荐结果。

3.2.5 *LDP 算法扰动*

用户使用算法alg_{opt}对原始数据d_l进行处理。在前述介绍的 4 种频数统计的 LDP 算法中，分别经过了在用户端编码、扰动和在服务商端校准的过程。然而为了保护用户的隐私预算ε_i，需要对原算法进行改进，避免在服务商端进行校准，为此，本文将

校准过程迁移至用户端，并设计后处理算法，防止扰动结果泄露用户的隐私偏好。具体操作如下。

1) 编码。用户数据 d_l 可转化为向量 $\boldsymbol{v}_i=(v_{i,d_1},v_{i,d_2},\cdots,v_{i,d_L})$ 的形式，其中 $v_{i,d_l}=1$，其余分量为 0。用户在编码环节，采用最优算法 $\mathrm{alg}_{\mathrm{opt}}$ 的编码方式，将数据 d_l 编码至所选扰动范围 D_i 内的结果，该结果可以转化为 0-1 向量的形式，即 $\boldsymbol{d}_l^{\mathrm{code}}=(v_{i,D_{i,1}},v_{i,D_{i,2}},\cdots,v_{i,D_{i,|D_i|}})$，$D_{i,j}$ 表示集合 D_i 中的第 j 个元素，$1\leqslant j\leqslant|D_i|$。其中，对于 DE 和 OUE，$v_{i,d_l}=1$；对于 OLH，由于使用哈希编码，结果为 $d_l^{\mathrm{code}}=H(d_l)<|D_i|$，因此仍可以将其视作 $\boldsymbol{d}_l^{\mathrm{code}}=(v_{i,D_{i,1}},v_{i,D_{i,2}},\cdots,v_{i,D_{i,|D_i|}})$ 的形式，其中 $v_{i,d_l^{\mathrm{code}}}=1$；THE 编码后的结果为 $\boldsymbol{d}_l^{\mathrm{code}}=(v_{i,D_{i,1}},v_{i,D_{i,2}},\cdots,v_{i,D_{i,|D_i|}})$，其中 $v_{i,d_l}=1$。

2) 扰动。利用算法 $\mathrm{alg}_{\mathrm{opt}}$ 的扰动机制对 $\boldsymbol{d}_l^{\mathrm{code}}$ 进行扰动得到 x_i。OUE 和 THE 这 2 个算法的扰动结果可转化为向量 $\boldsymbol{x}_i=(x_{i,D_{i,1}},x_{i,D_{i,2}},\cdots,x_{i,D_{i,|D_i|}})$ 的形式，其中 $\boldsymbol{x}_i$ 每个分量取值为 0 或 1；对于算法 DE 而言，扰动结果为 $d_{l'}\in D_i$，转化为 $\boldsymbol{x}_i=(x_{i,D_{i,1}},x_{i,D_{i,2}},\cdots,x_{i,D_{i,|D_i|}})$ 的形式后，仅有 $x_{i,d_{l'}}=1$，其余分量为 0；对于 OLH 算法，扰动结果 $d_{l'}\in\mathrm{Range}(H)$，这里用户需要遍历一遍 D_i，记录所有满足 $d_{l'}=H(d_l^{\mathrm{code}})$ 的 $d_{l'}$ 值，同样可以转化为相同的形式 $\boldsymbol{x}_i=(x_{i,D_{i,1}},x_{i,D_{i,2}},\cdots,x_{i,D_{i,|D_i|}})$，其中 $x_{i,d_{l'}}=1$，其余分量为 0（OLH 中的 $d_{l'}$ 可能不止一个）。

3) 校准。将原本在服务商端做的校准工作前移到用户端。校准后的扰动结果可转化为向量 $\boldsymbol{x}_i'=(x_{i,D_{i,1}}',x_{i,D_{i,2}}',\cdots,x_{i,D_{i,|D_i|}}')$ 的形式，对 4 种算法都有 $x_{i,j}'=\dfrac{x_{i,j}-q_i}{p_i-q_i}$，其中 p_i 和 q_i 与算法以及 ε_i 有关，具体取值见 3.2.1 节的 p 和 q。

4) 后处理。为了防止服务商根据 $x_{i,j}'$ 值推算出 p_i、q_i 的取值，从而判断出用户的隐私预算 ε_i，使用户的敏感信息泄露，需要隐藏 $x_{i,j}'=\dfrac{x_{i,j}-q_i}{p_i-q_i}$ 的取值，同时为了防止算法的开销过度增大，本文设计了一个协同机制，如图 2 所示，具体操作如下。

① 用户随机生成 2 个参数 $\boldsymbol{\delta}_i=(\delta_{i,1},\delta_{i,2})$ 作为扰动因子，并将扰动因子添加到 2 个不同的 $x_{i,j}'$ 值上得到 $\boldsymbol{x}_i''=\left(\dfrac{1-q_i}{p_i-q_i}+\delta_{i,1},\dfrac{0-q_i}{p_i-q_i}+\delta_{i,2}\right)$，以隐藏 p_i 和 q_i 的取值；同时构建集合 $M_i=\{j\mid x_{i,j}=1\}$，显然，$D_i-M_i=\{j\mid x_{i,j}=0\}$，并将 M_i、$\boldsymbol{x}_i''$ 和 D_i 发给服务商。

② 服务商收到信息后，将相同 M_i 和 D_i 的用户整合形成一个群组，记该群组的 M_i 和 D_i 分别为 G_g 和 G_g'。设共有 k 个群组，每个群组 $g(1\leqslant g\leqslant k)$ 的群成员有 m_g 个。服务商给第一个群成员发送长度为 2 的随机初始化噪声向量 $\boldsymbol{S}_0=(S_{0,1},S_{0,2})$。第一个群成员接收向量后，计算 $\boldsymbol{S}_1=\boldsymbol{S}_0-\boldsymbol{\delta}_i=(S_{0,1}-\delta_{i,1},S_{0,2}-\delta_{i,2})$，再将 $\boldsymbol{S}_1$ 发送给下一个群成员。每个群成员都进行上述操作，直到最后一个群成员计算 $\boldsymbol{S}_{m_g}$，并把 $\boldsymbol{S}_{m_g}$ 发回给服务商。服务商分别计算每个群组 g 的求和向量 $\boldsymbol{R}_g=(R_g^1,R_g^2)=\sum\limits_{\substack{i:M_i=G_g,\\ D_i=G_g'}}\boldsymbol{x}_i''+\boldsymbol{S}_0-\boldsymbol{S}_{m_g}$。显然 $R_g^1=\sum\limits_{\substack{i:M_i=G_g,\\ D_i=G_g'}}\dfrac{1-q_i}{p_i-q_i}$，$R_g^2=\sum\limits_{\substack{i:M_i=G_g,\\ D_i=G_g'}}\dfrac{-q_i}{p_i-q_i}$。同时服务商也无法得知每个用户 p_i

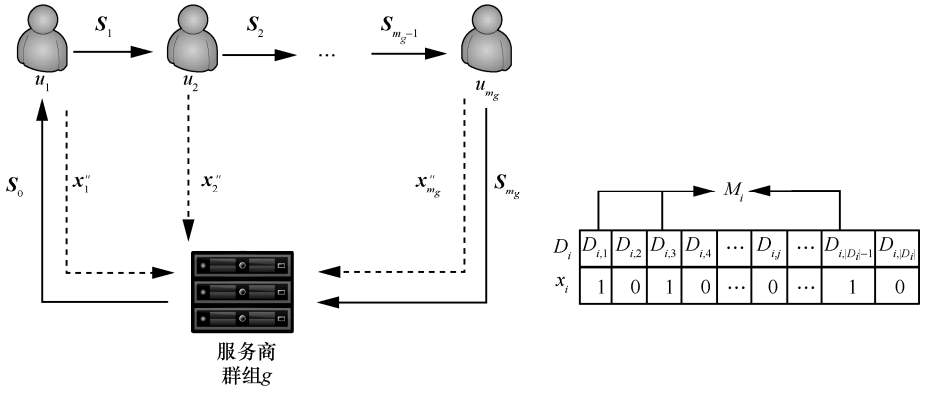

图 2　协同机制

和 q_i 的具体取值。

3.2.6 服务商结果整合

服务商为群组 g 构建向量 $\boldsymbol{V}_g=(R_g^{b_1},R_g^{b_2},\cdots,R_g^{b_{|G_g|}})$，其中当 $j\in G_g$ 时，$b_j=1$；当 $j\in G_g'-G_g$ 时，$b_j=2$。随后服务商整合所有群组的向量 $\boldsymbol{V}_g$，并估计每个 PoI d_l 的访问频数 $F^{d_l}=\sum\limits_{g:d_l\in G_g}R_g^1+\sum\limits_{g:d_l\in G_g'-G_g}R_g^2$。

3.3 方案证明

定理 1 本文方案为每个用户提供了 (D_i,ε_i) 的个性化本地化差分隐私。

证明 由于本地化差分隐私具有后处理的性质，因此只需证明在 3.2.5 节步骤 2)的扰动后，对于 $\forall j_1,j_2\in D_i$，有 $\Pr(M(j_1)=\boldsymbol{x}_i)\leqslant \mathrm{e}^{\varepsilon}\Pr(M(j_2)=\boldsymbol{x}_i)$。

假设方案 M 包含 K 个 LDP 算法，其中推荐算法 M_k 的概率为 p_k，所以对于 $\forall j_1,j_2\in D_i$，$\Pr(M(j_1)=\boldsymbol{x}_i)=\sum\limits_{k=1}^{K}p_k\Pr(M_k(j_1)=\boldsymbol{x}_i)$。由于任意 M_k 算法都满足 (D_i,ε_i) 本地化差分隐私，因此有

$$\sum_{k=1}^{K}p_k\Pr(M_k(j_1)=\boldsymbol{x}_i)\leqslant\sum_{k=1}^{K}p_k\mathrm{e}^{\varepsilon}\Pr(M_k(j_2)=\boldsymbol{x}_i)=\mathrm{e}^{\varepsilon}\sum_{k=1}^{K}p_k\Pr(M_k(j_2)=\boldsymbol{x}_i)=\mathrm{e}^{\varepsilon}\Pr(M(j_2)=\boldsymbol{x}_i)\tag{5}$$

综上，有 $\Pr(M(j_1)=\boldsymbol{x}_i)\leqslant\mathrm{e}^{\varepsilon}\Pr(M(j_2)=\boldsymbol{x}_i)$ 不等式成立。证毕。

定理 2 服务商对 PoI d_l 的访问频数的统计结果 F^{d_l} 是真实频数统计结果的无偏估计。

证明 真实频数统计结果 $F_{\text{real}}^{d_l}$ 为

$$F_{\text{real}}^{d_l}=\sum_i v_{i,d_l}=\sum_{i:d_l\in D_i}v_{i,d_l}=\sum_{g:d_l\in G_g'}\sum_{\substack{i:D_i=G_g',\\M_i=G_g}}v_{i,d_l}\tag{6}$$

F^{d_l} 的期望为

$$\mathbb{E}(F^{d_l})=\mathbb{E}\left(\sum_{g:d_l\in G_g}R_g^1+\sum_{g:d_l\in G_g'-G_g}R_g^2\right)=\mathbb{E}\left(\sum_{g:d_l\in G_g}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\frac{1-q_i}{p_i-q_i}+\sum_{g:d_l\in G_g'-G_g}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\frac{-q_i}{p_i-q_i}\right)=\mathbb{E}\left(\sum_{g:d_l\in G_g'}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\frac{x_{i,d_l}-q_i}{p_i-q_i}\right)=\sum_{g:d_l\in G_g'}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\mathbb{E}\left(\frac{x_{i,d_l}-q_i}{p_i-q_i}\right)\tag{7}$$

根据本文算法定义，有

$$\mathbb{E}\left(\frac{x_{i,d_l}-q_i}{p_i-q_i}\right)=\frac{1}{p_i-q_i}\mathbb{E}(x_{i,d_l}-q_i)=\frac{1}{p_i-q_i}\left(p_iv_{i,d_l}+q_i(1-v_{i,d_l})-q_i\right)=v_{i,d_l}\tag{8}$$

故式(6)与式(7)相等，因此有 $\mathbb{E}(F^{d_l})=F_{\text{real}}^{d_l}$，即服务商通过本文方案对 PoI d_l 的访问频数的统计结果 F^{d_l} 是真实频数统计结果的无偏估计。证毕。

定理 3 本文方案得到的频数估计的方差约为 $\sum\limits_{i:d_l\in D_i}\frac{q_i(1-q_i)}{(p_i-q_i)^2}$。

证明 F^{d_l} 的方差为

$$\operatorname{var}(F^{d_l})=\operatorname{var}\left(\sum_{g:d_l\in G_g}R_g^1+\sum_{g:d_l\in G_g'-G_g}R_g^2\right)=\operatorname{var}\left(\sum_{g:d_l\in G_g}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\frac{1-q_i}{p_i-q_i}+\sum_{g:d_l\in G_g'-G_g}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\frac{-q_i}{p_i-q_i}\right)=\operatorname{var}\left(\sum_{g:d_l\in G_g'}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\frac{x_{i,d_l}-q_i}{p_i-q_i}\right)=\sum_{g:d_l\in G_g'}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\operatorname{var}\left(\frac{x_{i,d_l}-q_i}{p_i-q_i}\right)\tag{9}$$

根据本文算法定义，有

$$\operatorname{var}\left(\frac{x_{i,d_l}-q_i}{p_i-q_i}\right)=\frac{1}{(p_i-q_i)^2}\operatorname{var}(x_{i,d_l})=\frac{1}{(p_i-q_i)^2}\left(\mathbb{E}\left((F^{d_l})^2\right)-\mathbb{E}^2(F^{d_l})\right)=\frac{1}{(p_i-q_i)^2}\left(p_iv_{i,d_l}+q_i(1-v_{i,d_l})-\left(p_iv_{i,d_l}+q_i(1-v_{i,d_l})\right)^2\right)=\frac{v_{i,d_l}(1-p_i-q_i)}{p_i-q_i}+\frac{q_i(1-q_i)}{(p_i-q_i)^2}\tag{10}$$

因此有

$$\operatorname{var}(F^{d_l})=\sum_{g:d_l\in G_g'}\sum_{\substack{i:M_i=G_g,\\D_i=G_g'}}\left(\frac{v_{i,d_l}(1-p_i-q_i)}{p_i-q_i}+\frac{q_i(1-q_i)}{(p_i-q_i)^2}\right)=$$

$$\sum_{i:d_l\in D_i}\left(\frac{v_{i,d_l}(1-p_i-q_i)}{p_i-q_i}+\frac{q_i(1-q_i)}{(p_i-q_i)^2}\right) \tag{11}$$

由于当用户数足够多时，$\sum_{i:d_l\in D_i}\frac{v_{i,d_l}(1-p_i-q_i)}{p_i-q_i}$ 这一项可以忽略，因此有 $\mathrm{var}(F^{d_l})=\sum_{i:d_l\in D_i}\frac{q_i(1-q_i)}{(p_i-q_i)^2}$。证毕。

根据方差公式可以看到，在每个用户选择的 LDP 算法固定的情况下，扰动范围 D_i 和隐私预算 ε_i 影响了本文方案的可用性。

4　实验分析

4.1　实验设置

本节参考文献[25]提出的隐私保护效果评估，从可用性、复杂度分析的角度对所设计方案进行衡量。首先，利用实验拟合出 4 种 LDP 算法的资源开销与影响因素的关系式，并利用 LDP 算法理论上的方差来衡量可用性；其次，将表达式代入方案，并观察用户处于不同资源场景时的最优算法结果，证明本文方案的可行性；再次，通过调节参数，比较本文方案与常见 LDP 算法可用性的差异；最后，从理论和实验上分析本文方案的复杂度。

为了简便起见，假设服务商给所有用户都推荐了 LDP 的算法全集 ALG，且所有用户扰动范围集合 D_i 与隐私预算 ε_i 均相同，简记为 D 与 ε，令集合 D 的大小为 $L=|D|$。固定参与统计的用户人数为 $n=10\,000$，固定 THE 中的参数 $\theta=1$。本文执行多次实验，并对实验结果求取平均值。

4.2　LDP 算法的资源拟合关系

本节中的测试环境为：硬件设备为小米 MIX 终端，软件系统为 MIUI 10 9.3.28，运行内存为 6 GB。

为了拟合电量、流量与影响因子之间的关系，本节以 $\varepsilon=4$ 的情况为例，测试 L 变化时，执行 1 000 次改进后的 DE、OUE、OLH、THE 在电量和流量上的开销。

从图 3 和图 4 中可以发现，4 种算法的流量和电量均随着 L 的增大而增大，趋近于一条直线，利用最小二乘法拟合出 4 种算法的流量和电量与 L 的关系式，得到

$$\begin{aligned}
r_1^{\mathrm{DE}}&=1.750\,6\times10^{-8}L+1.728\,0\times10^{-4}\\
r_1^{\mathrm{OUE}}&=7.473\,0\times10^{-8}L+2.776\,6\times10^{-4}\\
r_1^{\mathrm{OLH}}&=5.4111\times10^{-8}L+2.322\,9\times10^{-4}\\
r_1^{\mathrm{THE}}&=4.008\,9\times10^{-8}L+7.568\,7\times10^{-4}\\
r_2^{\mathrm{DE}}&=2.366\,3\times10^{-9}L+0.010\,2\\
r_2^{\mathrm{OUE}}&=1.646\,3\times10^{-5}L+0.008\,6\\
r_2^{\mathrm{OLH}}&=1.603\,1\times10^{-5}L+0.008\,7\\
r_2^{\mathrm{THE}}&=7.131\,5\times10^{-5}L+0.568\,5
\end{aligned} \tag{12}$$

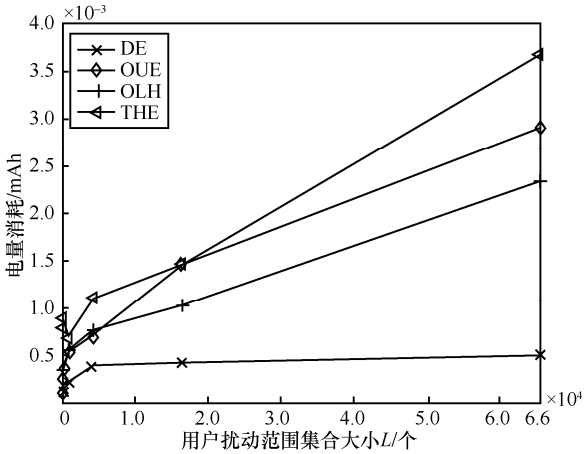

图 3　不同算法电量消耗情况随 L 的变化

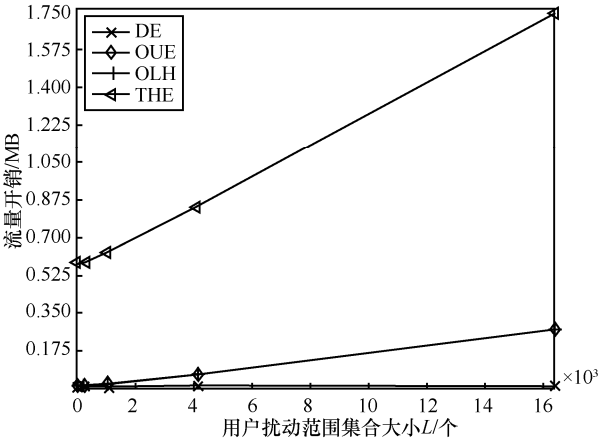

图 4　不同算法流量开销情况随 L 的变化

对于各种算法的可用性，类似定理 3 的证明，当 $\varepsilon=4$ 时，可以得到 DE、THE、OUE 和 OLH 算法的方差与影响因素的关系式分别为

$$r_{\mathrm{var}}^{\mathrm{DE}}=\frac{L-2+\mathrm{e}^4}{(\mathrm{e}^4-1)^2}$$

$$r_{\mathrm{var}}^{\mathrm{THE}}=\frac{2\mathrm{e}^2-1}{(1-\mathrm{e}^2)^2}\quad r_{\mathrm{var}}^{\mathrm{OLH}}=r_{\mathrm{var}}^{\mathrm{OUE}}=\frac{4\mathrm{e}^4}{(\mathrm{e}^4-1)^2} \tag{13}$$

类似地，能得到 ε 在其他取值下的关系式，再利用关系式代入 3.2.4 节的算法选取方案，并观察用户在不同状态时推荐的最优算法，具体结果如表 2 所示。

需要说明的是，改进后的 THE 算法尽管其可用性比 DE 好，但是与其他算法相比开销较大，不足以成为最优算法。因此本文实验对改进后的 THE 算法采用了更高效的传输方式，增大其流量开销，以换

表 2 用户在不同状态下的算法推荐

电量状态	流量状态	D_i	ε_i	最优算法
未充电，剩余电量 5	未连接 Wi-Fi，剩余流量 0 MB，套餐价格 0.3 元/MB	10	4.0	DE
充电中，电量充足	未连接 Wi-Fi，剩余流量 0 MB，套餐价格 0.3 元/MB	128	4.0	OUE
未充电，剩余电量 5	连接 Wi-Fi，流量充足	4 096	4.0	OLH
未充电，剩余电量 5	连接 Wi-Fi，流量充足	1 024	0.5	THE

取电量开销的减小，具体传输方式为在 3.2.5 节的步骤 4)中，用户要将 M_i 发送给服务商，其他 3 种算法会将 M_i 集合中的多个元素整合成一个字符发出，而 THE 的 M_i 会以 list 的形式发出，因此 THE 的传输开销会更高，但是同时少了整合步骤，因此计算量的消耗会降低，即电量消耗会减小。更新后的 THE 算法可以视作另一种新型的 LDP 算法。由此可以看到，当用户的状态发生变化时，本文方案可以为用户提供自适应的、个性化的算法推荐服务。

4.3 方案的可用性

本节采用改进后的 DE、OUE、OLH 和 THE 这 4 种算法作为对比方案。考虑在用户所选扰动范围集合大小 L 和隐私预算 ε 发生变化时本文方案与对比方案的可用性。本文利用每个 d_l 的真实频数 $f_{\text{real}}^{d_l}$ 与估计频数 $f_{\text{estimate}}^{d_l}$ 之间的均方根误差（RMSE, root mean square error）来衡量方案的可用性。具体计算式为

$$\text{RMSE}=\sqrt{\frac{1}{L}\sum_{d_l\in D}(f_{\text{estimate}}^{d_l}-f_{\text{real}}^{d_l})^2}=\sum_{i:d_l\in D_i}v_{i,d_l} \quad (14)$$

从式(14)可以看到，可用性越高， RMSE 的值就越小。

实验假设4种算法被用户选择作为最终方案的概率是相等的，在该前提下，首先固定 $\varepsilon=4$，研究 L 对方案可用性的影响。图 5 展示了本文方案与其他 4 种方案可用性区别与变化趋势。为了表现 L 在不同数量级下方案的可用性，参考文献[10]，采取指数型的横坐标进行实验。从图 5 可以看到，当 $L<2^6$ 时，DE 的可用性优势明显，但当 L 继续增大时，DE 可用性变差。而其他3个对比方案的可用性受 L 的影响不大，OLH 和 OUE 的可用性很接近，维持在 30 左右，THE 的可用性略微差一些，维持在 50 左右。本文方案由于受到 DE 的影响，可用性随着 L 的增大有所减小，但是好于 DE 本身，并且在 $L<2^{10}$ 时可用性甚至优于 THE。这说明本文方案的可用性中和了 LDP 算法集中不同算法的可用性，这是因为独立的多个随机变量的和的方差等于变量的方差和。

固定 $L=2^{10}$，观察 ε 对方案可用性的影响，如图 6 所示。从图 6 可以看到，所有方案的可用性都随着 ε 的增大而增加。当 ε 较小时，本文方案的可用性要优于 DE 和 OLH 这 2 个方案，且随着 ε 的增大，本文方案与 OLH 和 OUE 的可用性逐渐接近，甚至优于 THE。另外，尽管理论上 OUE 与 OLH 的可用性是一致的，但是由于在 ε 较小时，OLH 中参数 $\lfloor e^{\varepsilon}+1\rfloor$ 很小，即哈希函数会将 1 024 个数字映射到 $\lfloor e^{\varepsilon}+1\rfloor$ 个数上，使碰撞较大，从而影响了 OLH 的可用性；而当 ε 增大时，这二者的可用性又恢复了一致。

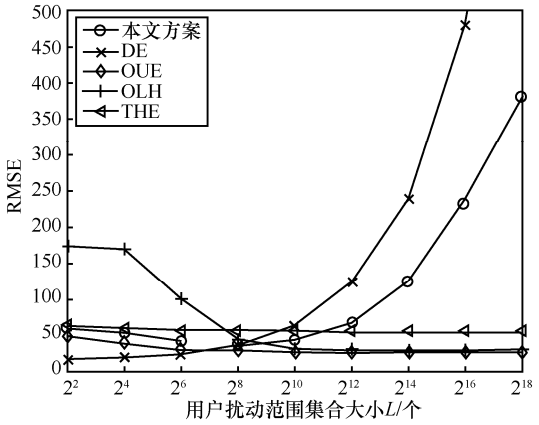

图 5 不同方案在固定 ε 下 RMSE 的变化

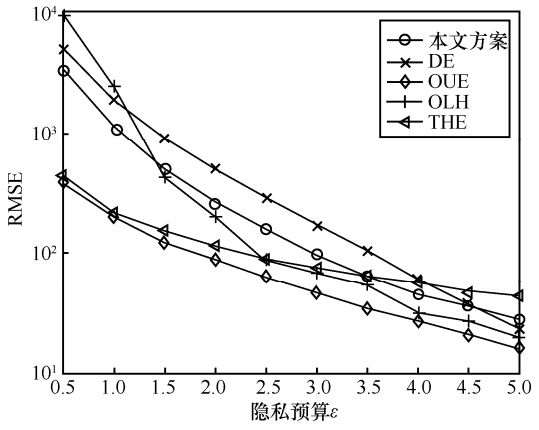

图 6 不同方案在固定 L 下 RMSE 的变化

综上所述，本文方案与其他方案相比，在 L 和 ε 的变动下，其可用性始终能够保持较好的状态。

4.4 方案的复杂度

本文在 3.2 节中提出的方案的算法复杂度与所

选择的算法集合大小有关，当算法集合大小固定时，算法复杂度与用户所选扰动范围$|D_i|=L$有关。

具体而言，设服务为用户 i 推荐了$|\text{alg}_i|$个算法，并且算法的拟合关系式在用户资源充足时已由服务商更新，权重也已预计算。因此在本文方案中，每个用户均执行了$20|\text{alg}_i|+8$次加法运算、$15|\text{alg}_i|$次乘法运算、$7|\text{alg}_i|+2$次除法运算、$2|\text{alg}_i|$次开根运算和 2 次随机运算。用户选择不同的扰动算法会有不同的计算开销：采用 DE 算法会进行一次随机；采用 OUE 算法会进行 L 次随机；采用 OLH 算法会进行 $L+1$次哈希和一次随机；采用 THE 算法会进行 L 次随机和 L 次加法运算。

在 4.1 节的设置下，$|\text{alg}_i|=4$，且用户均匀选择 4 种算法时，通过实验得出本文方案每次执行时间与 L 的关系如图 7 所示，从图 7 中可以看到，随着 L 的变化，本文方案的时间开销会增加。

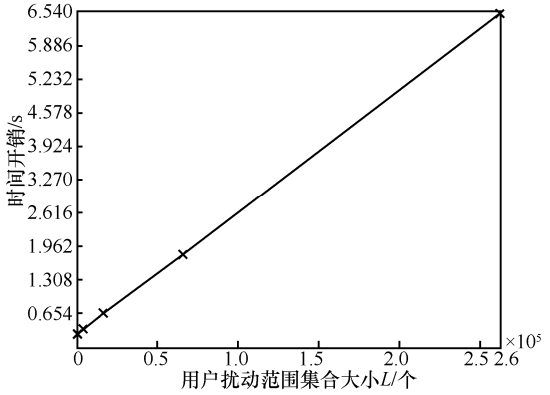

图 7　本文方案的时间开销随 L 的变化

本文方案共传输数据 M_i、$\boldsymbol{x}_i''$、D_i、$\boldsymbol{S}_i$ 至服务商，通过实验测试得出的本文算法每次通信量大小与 L 的关系如图 8 所示。从图 8 可以发现，随着 L 增大，本文方案的通信量也随之增加。

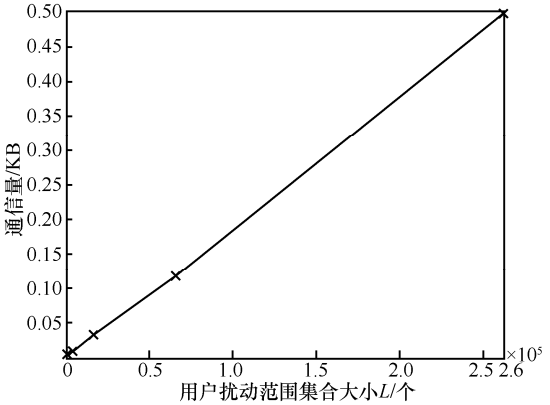

图 8　本文方案的通信量随 L 的变化

5　结束语

本文设计了基于 LDP 的多级资源自适应的算法推荐框架，能够灵活地为用户提供资源、隐私个性化的算法推荐。同时，将框架应用在位置 PoI 频数统计任务的场景下，并在此基础上对 LDP 算法进行改进，使不同用户能够选取不同的 LDP 机制，且可以设定不同的隐私需求；考虑到用户的隐私偏好包含了用户的敏感信息，因此设计了后处理机制，防止服务商推测出用户的隐私偏好。最后，通过实验证明了本文方案的可行性、可用性和效率。

参考文献：

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

[2] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography. Berlin: Springer, 2006: 265-284.

[3] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 2007: 94-103.

[4] WANG T, ZHANG X F, FENG J Y, et al. A comprehensive survey on local differential privacy toward data statistics and analysis[J]. Sensors, 2020, 20(24): 7030.

[5] KAIROUZ P, OH S, VISWANATH P. Extremal mechanisms for local differential privacy[J]. Journal of Machine Learning Research, 2016, 17(17): 1-51.

[6] KAIROUZ P, BONAWITZ K, RAMAGE D. Discrete distribution estimation under local privacy[C]//Proceedings of 2016 International Conference on Machine Learning (ICML). New York: ACM Press, 2016: 2436-2444.

[7] WANG T H, BLOCKI J, LI N H, et al. Locally differentially private protocols for frequency estimation[C]//Proceedings of 2017 USENIX Security Symposium (USENIX Security). Berkeley: USENIX Association, 2017: 729-745.

[8] BASSILY R, SMITH A. Local, private, efficient protocols for succinct histograms[C]//Proceedings of the 47th Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2015: 127-135.

[9] WANG T H, LI N H, JHA S. Locally differentially private frequent itemset mining[C]//Proceedings of 2018 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2018: 127-143.

[10] YE Q Q, HU H B, MENG X F, et al. PrivKV: key-value data collection with local differential privacy[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 317-331.

[11] DUCHI J C, JORDAN M I, WAINWRIGHT M J. Privacy aware learning[J]. Journal of the ACM, 2014, 61(6): 1-57.

[12] NGUYÊN T T, XIAO X, YANG Y, et al. Collecting and analyzing data from smart device users with local differential privacy[J]. arXiv Pre-

print, arXiv:1606.05053, 2016.

[13] YILMAZ E, AL-RUBAIE M, CHANG J M. Locally differentially private naive bayes classification[J]. arXiv Preprint, arXiv: 1905.01039, 2019.

[14] MAHAWAGA A P C, BERTOK P, KHALIL I, et al. Local differential privacy for deep learning[J]. IEEE Internet of Things Journal, 2020, 7(7): 5827-5842.

[15] SHIN H, KIM S, SHIN J, et al. Privacy enhanced matrix factorization for recommendation with local differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1770-1782.

[16] ERLINGSSON Ú, PIHUR V, KOROLOVA A. RAPPOR: randomized aggregatable privacy-preserving ordinal response[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2014: 1054-1067.

[17] WANG N, XIAO X K, YANG Y, et al. Collecting and analyzing multidimensional data with local differential privacy[C]//Proceedings of 2019 IEEE 35th International Conference on Data Engineering. Piscataway: IEEE Press, 2019: 638-649.

[18] ZHAO Y, ZHAO J, YANG M, et al. Local differential privacy-based federated learning for Internet of things[J]. IEEE Internet of Things Journal, 2021, 8(11): 8836-8853.

[19] CHEN R, LI H R, QIN A K, et al. Private spatial data aggregation in the local setting[C]//Proceedings of 2016 IEEE 32nd International Conference on Data Engineering. Piscataway: IEEE Press, 2016: 289-300.

[20] NIE Y W, YANG W, HUANG L S, et al. A utility-optimized framework for personalized private histogram estimation[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(4): 655-669.

[21] XIA C, HUA J Y, TONG W, et al. Distributed k-means clustering guaranteeing local differential privacy[J]. Computers & Security, 2020, 90: 101699.

[22] SHEN Z X, XIA Z H, YU P P. PLDP: personalized local differential privacy for multidimensional data aggregation[J]. Security and Communication Networks, 2021, 2021: 6684179.

[23] NALDI M, D'ACQUISTO G. Differential privacy: an estimation theory-based method for choosing epsilon[J]. arXiv Preprint, arXiv: 1510.00917, 2015.

[24] SHAHANI S, ABRAHAM J, VENKATESWARAN R. Selection and verification of privacy parameters for local differentially private data aggregation[C]//Proceedings of the 5th International Conference on Information System and Data Mining. New York: ACM Press, 2021: 84-89.

[25] LI F H, LI H, NIU B, et al. Privacy computing: concept, computing framework, and future development trends[J]. Engineering, 2019, 5(6): 1179-1192.

[26] 叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述[J]. 软件学报, 2018, 29(7): 1981-2005.
YE Q Q, MENG X F, ZHU M J, et al. Survey on local differential privacy[J]. Journal of Software, 2018, 29(7): 1981-2005.

[27] NIU B, LI Q H, WANG H Y, et al. A framework for personalized location privacy[J]. IEEE Transactions on Mobile Computing, 2021: doi.org/10.1109/TMC.2021.3055865.

[作者简介]

王瀚仪（1994− ），女，吉林省吉林市人，中国科学院信息工程研究所博士生，主要研究方向为隐私计算。

李效光（1995− ），男，陕西西安人，西安电子科技大学博士生，主要研究方向为差分隐私。

毕文卿（1997− ），女，山东菏泽人，中国科学院信息工程研究所硕士生，主要研究方向为隐私计算。

陈亚虹（1995− ），女，福建泉州人，中国科学院信息工程研究所博士生，主要研究方向为隐私计算。

李凤华（1966− ），男，湖北浠水人，博士，中国科学院信息工程研究所研究员、博士生导师，主要研究方向为网络与系统安全、信息保护、隐私计算。

牛犇（1984− ），男，陕西西安人，博士，中国科学院信息工程研究所副研究员、博士生导师，主要研究方向为隐私计算、网络安全防护。