软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006044]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向频繁项集挖掘的本地差分隐私事务数据收集方法*

欧阳佳[1], 印　鉴[2], 肖政宏[1], 赵慧民[1], 刘少鹏[1], 梁　鹏[1], 肖茵茵[1]

[1](广东技术师范大学 计算机科学学院, 广东 广州 510665)

[2](中山大学 数据科学与计算机学院, 广东 广州 510275)

通讯作者:肖政宏, E-mail: huasxzh@126.com

摘　要: 事务数据常见于各种应用场景中,如购物记录、页面浏览历史等.为提供更好的服务,服务提供商收集用户数据并进行分析,但收集事务数据会泄露用户的隐私信息.为解决上述问题,本文基于压缩的本地差分隐私模型,提出一种事务数据收集方法.首先,定义一种新的候选项集分值函数;其次,基于该函数将候选项集的样本空间划分为多个子空间;第三,随机选择其中一个子空间,基于该子空间随机生成事务数据并发送给不可信的数据收集者;最后考虑到隐私参数的设置问题,基于最大后验置信度攻击模型设计启发式隐私参数设置策略.理论分析表明该方法能同时保护事务数据的长度与内容,满足压缩的本地差分隐私要求.实验表明,与目前最优的工作相比,本文收集的数据具有更高的效用性,隐私参数设置更具有语义性.

关键词: 隐私保护;数据收集;事务数据;本地差分隐私;隐私参数

中图法分类号: TP311

中文引用格式: 欧阳佳,印鉴,肖政宏,赵慧民,刘少鹏,梁鹏,肖茵茵等.面向频繁项集挖掘的本地差分隐私事务数据收集方法.软件学报. http://www.jos.org.cn/1000-9825/6044.htm

英文引用格式: Ouyang J, Yin J, Xiao ZH, Zhao HM, Liu SP, LIANG P, XIAO YY. Transaction Data Collection for Itemset Mining under Local Differential Privacy. Ruan Jian Xue Bao/Journal of Software, (in Chinese). http://www.jos.org.cn/1000-9825/6044.htm

Transaction Data Collection for Itemset Mining under Local Differential Privacy

OUYANG Jia[1], YIN Jian [1], XIAO Zheng-Hong[1], ZHAO Hui-Min[1], LIU Shao-Peng[1], LIANG Peng[1], XIAO Yin-Yin[1]

[1](College of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China)

[2](School of Data and Computer Science, SUN YAT-SEN University, Guangzhou 510275, China)

Abstract: Transaction data is commonly in various application scenarios, such as shopping records, page browsing history, etc., service providers collect and analyze transaction data for providing better services. However, collecting transaction data will disclose privacy information. To solve the problem, this paper proposes a transaction data collection mechanism based on Condensed Local Differential Privacy (CLDP). Firstly, we define a new score function of the candidate set. Secondly, we separate the output domain of the candidate set into several subspaces according to the function. Thirdly, the client select one subspace randomly, and generate transaction data randomly based on the subspace, then, send it to the untrusted data collector. Finally, considering the difficulty for setting the privacy parameter, we design the heuristic privacy parameter setting strategy, based on the maximum posterior confidence threat model (MPC). The theoretical analysis shows that this method can protect the length and content of transaction data at the same time and satisfies α-CLDP. The

* 基金项目: 国家自然科学基金(61702119,U1711262,U1501252,U1711261);广州市科技计划项目(201804010236,201607010152);广东省基础与应用基础研究基金(2019A1515012048).广东省教育厅创新团队项目(2017KCXTD021)

Foundation item: National Natural Science Foundation of China (61702119, U1711262,U1501252, U1711261); Science and Technology Program of Guangzhou (201804010236, 201607010152); Guangdong Basic and Applied Basic Research Foundation (2019A1515012048). The Innovation Team Project of the Education Department of Guangdong Province(2017KCXTD021)

收稿时间: 2019-11-06; 修改时间: 2020-01-30, 2020-03-09; 采用时间: 2020-03-20; jos 在线出版时间: 2021-05-20

experiments demonstrate that the transaction data collected in this paper has higher utility than the state-of-the-art approaches, and the privacy parameter setting is semantic.

Key words: privacy preserving; data collecting; transaction data; local differential privacy; privacy parameter

事务数据是项的集合,其中包含丰富的信息并可应用于不同的场景(如:购买的商品,看过的电影记录,搜索日志,网页浏览历史等).随着大数据技术的发展,海量的事务数据被收集,其中蕴含丰富的知识,数据收集者通过数据分析(如:协同过滤、关联规则等),基于得到的模型为用户提供更好的服务.

然而,事务数据中往往包含个人隐私信息,如搜索网页时产生的搜索日志会泄露自己的健康状态、居住地点等信息;网上购物时产生的购物记录将泄露自己所购买的隐私物品,甚至会泄露自己的购物习惯以及行为模式;浏览网页时产生的WEB点击流将泄露自己的上网习惯等等.如果不采取任何保护措施直接收集并分析用户的数据,将会导致个人隐私信息的泄露,造成严重危害.震惊世界的 AOL 日志隐私泄露事件已敲响警钟,因隐私泄露所带来的困扰将严重影响着人们的合法权益和生活质量.

目前,事务数据隐私保护发布是研究热点,大多数研究是将真实完整的事务数据发送到数据中心,并假设数据中心或数据收集者是可信的.数据收集者对数据进行扰乱处理后,发布满足差分隐私约束的数据集或相关统计信息,该方法统称为中心化差分隐私技术.尽管数据收集者宣称不会泄露或窃取用户的敏感信息,但在商业或利益的驱使下,用户隐私很难得到保证,因此假设数据中心或数据收集者是可信的这一点不切实际.

本地差分隐私(Local Differential Privacy, LDP)技术是一种本地化的数据收集方法,与中心化差分隐私不同的是,其针对的是不可信的第三方数据收集者.客户端基于LDP在本地独立对数据进行随机响应,然后再将扰乱后的数据发送给数据收集者,即数据收集者得到的数据是不完整的用户数据,但又保留了一定数据统计信息,具有较好的数据效用性.另外,LDP 避免了大规模计算以及与数据中心频繁交互的通讯代价,非常适用于资源受限的客户端,如:移动设备、无线传感器等.目前已在工业界得到推广应用,包括微软、Google、Apple 等公司均已将 LDP 嵌入到应用中.[1]

基于 LDP 的数据隐私保护已有研究中,主要集中于类型数据、数据数值、离散数据等.事务数据由于其应用非常广泛,一直以来都是研究的难点与重点.然而由于事务数据高维、稀疏以及长度不等的特性,导致事务数据的研究往往比常规数据要复杂得多.目前有许多基于 LDP 的工作对事务数据的内容(项)以及长度进行了有效的保护,这些工作首先对事务数据作等长处理,然后对事务数据进行子集抽样.文献[2]提出的 PrivSet 方法就是其中的典型代表,该方法设计了一种效用性函数,然后基于一种高效的随机化方法得到扰乱后的事务数据,达到了非常好的效果与效率.受 PrivSet 方法启发,本文基于压缩的本地差分隐私(Condensed Local Differential Privacy, CLDP)提出一种新的事务数据收集方法 TDC_LDP.首先设计一种新的候选项集的分值函数,然后基于指数机制从候选项集中随机抽取一个项集,由于本文提出的分值函数与 PrivSet 的效用性函数相比保留了更多的信息,因此整体效果比 PrivSet 方法要好.

本文的主要贡献如下:

(1) 提出一种新的候选项集的分值函数,基于该函数为每个候选项集打分,并将候选项集的样本空间划分为多个子空间,其中相同分值的候选项集位于同一个子空间;

(2) 抽取其中一个子空间,基于该子空间随机生成事务数据并发送给不可信的数据收集者,同时保证了项支持度计数与频繁项集挖掘的效用性;

(3) 考虑到隐私参数的设置困难,为直观的设置隐私参数,基于 MPC 攻击模型,提出一种启发式隐私参数设置策略.

本文其他内容组织如下.第 1 章介绍相关工作;第 2 章介绍本文相关的预备知识;第 3 章对 TDC_LDP 进行详细描述;第 4 章进行理论分析以及介绍隐私参数设置策略;第 5 章进行实验并分析实验结果;第 6 章是本文的总结.

1 相关工作

事务数据隐私保护收集涉及到事务数据隐私保护、隐私参数设置策略等,本节简要介绍并总结相关工作.

1.1 事务数据隐私保护

关于事务数据的隐私保护方法得到了广泛的研究,按收集者是否可信,可分为两大类型:中心化方法与本地化方法.其中中心化方法又可以分为基于传统的 k -匿名分组隐私模型与差分隐私模型;本地化方法可划分为本地差分隐私模型与压缩的本地差分隐私模型.如图 1-1 所示.差分隐私模型的隐私参数设置策略的研究也是热门领域,本文有涉及该工作,这部分将在 1.2 节进行综述.

1.1.1 中心化

在中心化环境中,假设数据中心(也称数据收集者)是可信的,用户会将自己的真实事务数据发送到数据中心,数据中心对收集到的数据进行随机化处理后,将数据发布出去用于统计与大数据分析.传统的基于分组的隐私模型中,k-匿名[3]与 l-多样化[4]是其中的典型代表,能保证用户的事务数据在组中不可识别.常用的分组技术主要是泛化[5-7]与消除[8, 9].

Ghinita 等人[10]基于 k-匿名模型与 l-多样化模型,基于行列重排列提出一种新的匿名分组技术,解决了数据的高维问题.Cao 等人[11]首次提出一种更加新颖的、更符合现实的隐私模型,假设攻击者的背景知识同时包含敏感项和非敏感项,提出 ρ -uncertainty 隐私模型,要求包含敏感项的关联规则的置信度不能超过 ρ .Terrovitis 等人[12]不区分项的敏感性,假设攻击者的背景知识最多包含 m 个项,提出一种新的隐私模型 k^m -anonymity,不采用项消除的方式,而是采用全局泛化方式对数据进行匿名.He 等人[5]指出 k^m -匿名的隐私保护力度低于 k -匿名,提出一种针对于事务数据的 k -匿名模型.Xu 等人[13]同样指出由于事务数据的高维性导致匿名后数据的效用性严重不足,通过对攻击者的背景知识进行限定,假设攻击者最多拥有 p 个非敏感项,提出一种(h,k,p)-coherence 隐私模型,采用全局消除的方式以保留更多的信息,有效的提升了数据的效用性.

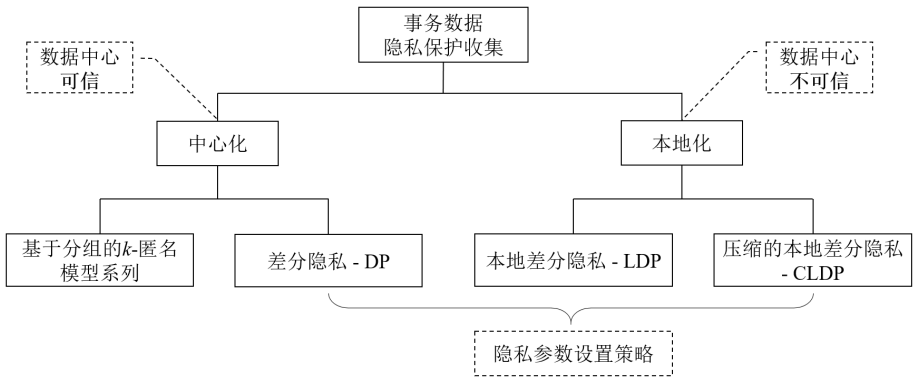


Fig.1-1 The research of privacy preserving in transaction data collection and publishing

图 1-1 事务数据收集与发布的隐私保护研究

由于 k-匿名对攻击者背景知识作了过多的假设,且消除与泛化导致大量的信息损失,导致高维的事务数据隐私保护隐私性与效用性不足.差分隐私由于其强隐私性与高效用性,逐渐正为事务数据隐私保护的研究热点.Chen 等人[14]基于加拿大蒙特利尔市的真实轨迹数据提出了一种轨迹数据的差分隐私发布方法.该方法首先构建一棵含有噪音的前缀树,然后根据父节点的计数必须大于等于子节点计数和的约束对每个节点的噪音进行优化,取得非常好的效用性.欧阳佳等人[15] 通过构建事务数据库的完整 Trie 项集树,然后基于压缩感知技术对项集树添加满足差分隐私约束的噪音得到含噪 Trie 项集树,最后在含噪树上进行频繁项集挖掘任务,能有效保持数据效用性.针对分布式结构下的隐私保护事务数据发布,欧阳佳等人[16]将结果效用性优化与差分隐私约束相结合,构建分布式非线性规划模型.然后,基于全局与局部数据设计两种解决方案安全求解该分布式模型.

针对事务数据的频繁项集挖掘任务,文献[17-20]通过对项集的计数添加噪音,提出了一系列满足差分隐私约束的频繁项集挖掘算法.

1.1.2 本地化

上述中心化的隐私保护方法假设数据中心是可信的,然而真实环境下,由于利益的驱使,数据中心上的真实数据难免会由于各种原因泄露出去.基于本地差分隐私模型[21, 22]的本地化方法假设数据中心是不可信的,用户在本地对数据进行随机响应[23]后发送给数据中心,因此数据中心收集的用户数据不是真实的,但保留了统计性质.随机响应技术也常用于事务数据的隐私保护发布,文献[24]针对个性化隐私要求,将随机响应响应技术应用于频繁项集挖掘,文献[25]基于随机响应技术,提出一种 γ -增强机制,保证攻击者对某个隐私信息的先验与后验概率在一个指定的区间内.

本地差分隐私模型已成功应用于工业界中,包括 Google 的 Chrome 浏览器[21, 22],Apple 的 iOS、微软的 Windows 10 系统中[26]都基于 LDP 收集用户的隐私数据,并进行统计与分析,包括均值计算、直方图统计等.

本地差分隐私模型在不同数据类型的收集中也有广泛的研究与应用,如类型数据[27, 28]、位置数据[29, 30]、事务数据[2, 31-33].特别地,文献[33]提出一种两阶段的频繁项集估计机制,第一阶段为分布估计阶段,而第二阶段为分布估计改进阶段,文献[2]认为文献[33]对每个项均有一个隐私预算,加入了过多的噪音,文献[2]基于指数机制提出了一种高效的候选项集抽样算法,其中每个候选项集的抽样概率与原始数据的交集有关,如果有交集,则效用性函数为 1,否则为 0.

压缩的本地差分隐私模型[29]将距离度量的概念引入差分隐私模型,基于距离矩阵随机响应一个值.文献[29]提出了 Geo-indistinguishability 概念,即位置不可识别,从真实位置的相邻位置中,基于与真实位置的距离矩阵,随机响应一个地点,能有效保护用户的真实位置,且保证位置信息的效用性,在一些对精确位置要求不高的应用中(如:天气预报)取得了很好的效果.类似的,文献[34]基于距离矩阵提出了压缩的本地差分隐私(Condensed Local Differential Privacy, CLDP)概念,对序数与非序数的小样本数据的长度与内容提供了有效的保护,在小样本的统计应用中,效用性较高,且实验效果优于 LDP 模型.

1.2 隐私参数设置策略

差分隐私系列模型的隐私参数 ε 或 α 是隐私模型的重要参数,用于决定噪声的添加量以及度量隐私保护的程度.隐私参数的设置依赖于实验或经验,大多数情况下没有明确的指导语义.

Lee 等人提出了一种 ρ 差分可识别(ρ-Differential Identifiability)[35]的概念,可以基于 ρ 差分可识别设置隐私参数 ε ,添加的噪声限定攻击者在获得分析结果后推断目标受害者敏感值的概率不高于 ρ.但是 ρ 差分可识别依赖于个体的先验分布,并需要假设预先知道所有可能的值及数目.欧阳佳等人[36]基于(ρ_1, ρ_2)隐私模型[37]提出一种启发式的隐私参数 ε 设置策略,分析隐私参数 ε 与(ρ_1, ρ_2)的内在联系,实现噪声量的添加由(ρ_1, ρ_2)决定.文献[34]中通过最大后验置信度(Maximum Posterior Confidence, MPC)将 CLDP 与 LDP 联系起来,提出一种基于 LDP 的隐私参数 ε 设置 CLDP 的隐私参数 α 的方法,但在本地差分隐私模型下并没有解决隐私参数的启发式设置问题.

1.3 相关工作总结

综上所述,差分隐私系列模型在隐私保护研究领域得到了广泛的发展与应用,其中本地差分隐私模型由于对数据中心假设不可信,已在工业界中得到推广.压缩的本地差分隐私模型由于将距离的概念引入,也开始得到相应的关注与研究,而隐私参数的设置问题研究相对较少.本文受文献[2]启发,基于文献[34]提出的 CLDP 模型,将候选项集与原始数据的相似度作为距离函数的分值,提出了一种新的基于压缩的本地差分隐私模型的事务数据收集方法,并基于 MPC 攻击模型提出一种新的隐私参数的设置策略,用于从启发式的角度设置隐私参数.

2 预备知识

2.1 事务数据（Transaction Data）

事务数据是一种无结构化数据,如表 2-1 所示.与关系型数据相比其共同点是事务数据库 D 同样由记录 $t_1,t_2,t_3,\ldots,t_n$ 组成;不同点是每条记录称为事务,为任意项的集合,其中 I 为整个项集域,定义为 $I=\{a_1,a_2,a_3,\ldots,a_d\}$,$d=|I|$为项集域的长度.事务数据的例子有很多,如:包含多个搜索关键词的 WEB 查询记录;包含已购买物品的购物记录;包含 URL 的点击流等等.对事务数据可以进行各种数据挖掘任务,如关联规则挖掘、用户行为预测、推荐系统、信息检索等其他基于 WEB 的相关应用.其中应用最广泛、最典型的是频繁项集挖掘.项的集合称为项集,包含 k 个项的项集称为 k-项集.项集的支持计数是指项集的出现频率,即包含项集的事务数.频繁项集是指支持计数大于等于设定的最小支持计数的项集.

Table 2-1 Example of Transaction Data

表 2-1 事务数据示例

交易 ID	购买物品
T100	牛奶,面包,火腿
T200	面包,辣椒酱
T300	面包,鸡蛋
T400	牛奶,面包,辣椒酱
T500	牛奶,鸡蛋

2.2 本地差分隐私（Local Differential Privacy, LDP）

差分隐私(Differential Privacy, DP)假设可信的数据收集者不会窃取或泄露用户的隐私信息,然而实际应用中这种假设并不成立,特别是当前数据即价值的时代,在利益的驱使下,用户的隐私信息很难得到保证.因此,本地差分隐私模型(LDP)应运而生,每个用户基于 LDP 对拥有的数据进行随机化,然后将随机化后的数据发送给数据收集者,即数据收集者无法直接访问用户的真实数据,这就从根源上保证了用户的信息安全.形式上,令 D 代表数据集,$\Re$ 为随机算法,其输入为数据 t,输出为 t^*,LDP 形式定义如下:

定义 1 (ε-本地差分隐私[22]). 一个随机算法 $\Re$ 满足 ε-本地差分隐私要求,当且仅当对于任意的输入 t 与 t' 以及对于任意的输出 t^*,下面的不等式成立:

$$\Pr\left[\Re(t)=t^*\right]\le e^{\varepsilon}\times\Pr\left[\Re(t')=t^*\right] \tag{2-1}$$

直观意义上,当得到的结果为 t^*,数据收集者无法以很高的置信度(通过 ε 控制)推断出输入数据是 t 还是 t',这点与中心差分隐私有根本的区别,中心差分隐私的输入为只相差一条数据的邻近数据集.随机响应(Randomized Response, RR)技术是一种主流的扰动机制,数据收集者在没有得到用户真实值的情况下,同样能准确地获得相应的统计信息.

2.3 压缩的本地差分隐私（Condensed Local Differential Privacy, CLDP）

令 U 定义为所有可能的值(或项),定义函数 $d:U\times U\to[0,\infty)$ 为距离度量函数,其输入为两个值 $v_1,v_2\in U$,函数 d 的结果用来度量这两个值的距离,要求函数 d 满足以下性质:非负性、同一性、对称性、三角不等式等.基于该函数,CLDP 有如下定义.

定义 2 (α-压缩的本地差分隐私[34]). 当 $\alpha>0$ 时,一个随机算法 Φ 满足 α-CLDP,当仅且当对于任意的输入 $v_1,v_2\in U$:

$$\forall y\in Range(\Phi):\frac{\Pr\left[\Phi(v_1)=y\right]}{\Pr\left[\Phi(v_2)=y\right]}\le e^{\alpha\cdot d(v_1,v_2)} \tag{2-2}$$

其中 $Range(\Phi)$ 代表随机算法 Φ 所有可能的输出.

从定义中可以发现,CLDP 的隐私性由参数 α 以及项之间的距离进行控制,本质上是对 LDP 引入距离的概念.特别地,当任意两个项之间的距离为 1 时,即 $d(v_1,v_2)=1$,令 $\alpha=\varepsilon$,则 CLDP 就退化成 LDP.可见 LDP 是 CLDP 的一种特殊情形.

2.4 隐私攻击模型（Threat Model）

本地差分隐私机制的主要目的是对用户的隐私数据提供保护.尽管如此,不可信的第三方数据收集者仍然可以从扰乱的数据集中推断出用户真实的隐私信息.攻击者观察到完整的扰乱数据 y 后,则最坏情况下推断出用户真实数据的最大后验置信度(Maximum Posterior Confidence, MPC)为:

$$MPC=\max_{v,y\in CandI}\Pr[v|y]=\max_{v,y\in CandI}\frac{\pi(v)\cdot\Pr[f(v)=y]}{\sum_{z\in CandI}\pi(z)\cdot\Pr[f(z)=y]} \tag{2-3}$$

其中,*CandI* 为项集域 *I* 的候选项集,$\pi(v)$ 为 *v* 的先验知识,*f* 为隐私机制,如 LDP 或 CLDP.MPC 量化了攻击者的攻击能力,对其进行限制为 $MPC\le\rho$,可以推断出 LDP 或 CLDP 的隐私参数与 ρ 的联系,从而可以从直观意义上设置隐私参数.

3 事务数据的本地差分隐私收集方法

事务数据集的本地随机响应远比一般的数据集要复杂,主要原因有如下两个: (1) 项集域的候选项集是指数级的; (2) 事务数据的长度往往是不相等的,每个用户拥有的数据长度范围为[0, *m*],其中 *m* 为事务数据可能的最大长度,导致事务数据的长度信息[34]无法保护.

Fig.3-1 The Overall Research Idea

图 3-1 整体研究思路

本节是论文的核心,将介绍本文提出的事务数据收集方法.首先提出第一种方法 TDC_CLDP_Cand,该方法先计算所有候选项集与 *t* 之间的分值,并得到抽样概率,然后从候选项集中选择一个,但该方法存在的主要问题是候选项集的数量是指数级的,其数量为 $2^{|I|}$-1,其中 *I* 为项集域.为解决指数级的抽样问题,本文提出第二种方法 TDC_CLDP,该方法的思想来自于 TDC_CLDP_Cand,但不直接从候选项集随机抽取一个项集,而是将候选项集的样本空间划分为 *k*+1 个部分,然后基于分值函数的抽样概率从 *k*+1 个部分中随机抽取一个,最后基于这个确定的子空间,继续抽取生成事务数据并发送给数据中心.接下来详细介绍两种方法.整体研究思路如图 3-1 所示.

3.1 TDC_CLDP_Cand算法

令 *t* 表示用户的事务数据,每个用户在本地从 *I* 的所有候选项集 *CandI* 中基于 CLDP 隐私模型随机抽取一个候选项集 *s*, CLDP 模型的关键是距离函数 *dist*,该距离函数必须是一个距离度量,即需要满足非负性、对称性、三角不等式等特性.本文定义的距离函数 *dist* 为候选项集 *s* 与 *t* 的相异度,如式(3-1)所示.

$$dist(t,s)=\max(|t|,|s|)-\sum_{i=1}^{\max(|t|,|s|)}[t_i=s_i]=\sum_{i=1}^{\max(|t|,|s|)}|t_i-s_i| \tag{3-1}$$

其直观意义为 *s* 与 *t* 的最大长度减去二者的相似度,本质上是 *t* 与 *s* 之间的**曼哈顿距离(Manhattan Distance), 曼哈顿距离是一个有效的距离度量.**

Table 3-1 The TDC_CLDP_Cand Algorithm

表 3-1 TDC_CLDP_Cand 算法

输入： 用户的事务数据 t,所有候选项集 $CandI$
输出： 扰乱后的事务数据 t'
1： **for** $s \in CandI$ **do**：
2： // **计算 t 与每个候选项集 s 的分数**
3： $score(s) = e^{\frac{-\alpha \cdot dist(t,s)}{2}}$
4： **end for**
5： **随机选择一个候选项集 t' ,其概率为：**
6： $\Pr[t' \text{ is sampled}] = \frac{score(t')}{\sum_{s \in CandI} score(s)}$
7： **return** t'

式(3-1)中 t_i 与 s_i 代表 t 与 s 中的第 i 项; 当 t_i 与 s_i 相等时,$[t_i = s_i] = 1$,反之为 0.如:当 t={abc},s={ae}时,将 t 与 s 分别转成字符串形式得到 t=[11000],s=[10001],则 t 与 s 的相异度为 2,相似度为 3.值得注意的是:当 t 与 s 长度不等时,可以通过加 0 进行补全.用户的事务数据 t 与每个候选项集 s 的分值函数为:

$$score(s) = e^{\frac{-\alpha \cdot dist(t,s)}{2}} \tag{3-2}$$

从式(3-2)定义的分值函数可以发现,每个 s 的分值与 t,s 的曼哈顿距离反比,该定义合理且符合直观语义.因为曼哈顿距离代表数据之间的差异性,差异性越大,分值越小;差异性越小,分值越大.基于所有 s 对 t 的分值函数,指数机制的抽样概率为:

$$\Pr[s' \text{ is sampled}] = \frac{score(s')}{\sum_{s \in CandI} score(s)} \tag{3-3}$$

可见该抽样概率与距离度量函数有关.基于上述原理,本文提出第一种算法 TDC_CLDP_Cand,如表 3-1 所示.但该算法需要遍历所有的候选项集,并计算其距离与分值,该过程是指数级的,其时间复杂度为 $O(2^d)$,即与项集域的大小呈指数关系,实验表明,当 $d \geq 16$ 时,很难得到结果.

3.2 TDC_CLDP算法

TDC_CLDP_Cand 算法存在的主要问题是候选项集的样品空间随着项集域的大小 d 呈指数级增长,导致时间复杂度非常高.本文基于 TDC_CLDP_Cand,提出 TDC_CLDP 算法,该算法是对 TDC_CLDP_Cand 的具体实现,其优点是可以避免直接从候选项集的样品空间中进行抽样, TDC_CLDP 方法如表 3-2 所示.

TDC_CLDP 基于 CLDP 隐私模型,其输入为:用户的事务数据 t,项集域 I,CDLP 的隐私参数 α ,输出的事务数据长度 k;输出为:事务数据 t',其长度为 $k = |t'|$ .TDC_CLDP 首先对 t 进行预处理得到 $\hat{t}$,使得 $|\hat{t}| = m$,然后从所有的候选项集中随机选择一个 t' ,t' 被选中的概率为：

$$\exp\left(\frac{-\alpha \cdot dist(t,t')}{2}\right) \Big/ \Omega \tag{3-4}$$

其中 Ω 为概率泛化因子,其定义如式(3-5)所示.

$$\Omega = \underbrace{e^{\frac{-\alpha}{2} \cdot 0} \cdot C_d^k}_{*} + \underbrace{\sum_{inter=1}^{k} \left(e^{\frac{-\alpha}{2} \cdot (k-inter)} \cdot \left(C_m^{inter} \cdot C_d^{k-inter} \right) \right)}_{**} \tag{3-5}$$

Table 3-2 The TDC_CLDP Algorithm

表 3-2 TDC_CLDP 算法

输入: 用户的事务数据 $t \in D, |t| \le m$,项集域 $I = \{a_1, \cdots a_d\}$,隐私参数 α ,输出的事务数据长度 k

输出: 扰乱的事务数据 $t', |t'| = k$,且满足 α -CLDP

1: $t' = \varnothing$, $d = |I|$

2: ▷ **对 t 作等长处理**

3: $\hat{t} = t$, $\hat{I} = \{a_1, \cdots a_d\} \bigcup \{a_{d+1}, \cdots a_{d+m}\}$ // 已有项集域的项为: $a_1, \cdots a_d$, 补充的项为 $a_{d+1}, \cdots a_{d+m}$,属于噪音项.

4: **for** $i = 0; i < m - |t|; i = i + 1$ **do:**

5: $\hat{t} = \hat{t} \bigcup a_{d+i+1}$

6: **end for**

7: ▷ **对 $\hat{t}$ 作随机化处理**

8: $\Omega = \underbrace{e^{\frac{-\alpha}{2}0} \cdot C_d^k}_{*} + \underbrace{\sum_{inter=1}^{k} \left(e^{\frac{-\alpha}{2}(k-inter)} \cdot \left(C_m^{inter} \cdot C_d^{k-inter} \right) \right)}_{**}$ // 其中,*部分表示相似度为 0,**部分表示相似度大于 0

9: $r = uniform\_random(0.0, 1.0)$

10: $inter = 0$

11: $p = e^{\frac{-\alpha}{2}0} \cdot \frac{C_d^k}{\Omega}$ // 相似度为 0 的候选项集被选中的概率

12: **while** $p < r$ **do:**

13: $inter = inter + 1$

14: $p = p + \frac{e^{\frac{-\alpha}{2}(k-inter)} \cdot \left(C_m^{inter} \cdot C_{d-1}^{k-1-inter} \right)}{\Omega}$ // 相似度大于 0 的候选项集被选中的概率

15: **end while**

16: $t' = t' \bigcup sample(\hat{t}, inter)$ // 以等概率方式从 $\hat{t}$ 中不放回随机抽取 *inter* 个项, 这部分为保留的真实数据

17: $t' = t' \bigcup sample(\hat{I} - \hat{t}, k - inter)$ // 以等概率方式从 $\hat{I} - \hat{t}$ 中不放回随机抽取 $k - inter$ 个项, 这部分为引入的噪音数据

18: **return** t'

Table 3-3 Subspace and Corresponding Sampling Probability

表 3-3 样本子空间及对应的抽样概率

样本子空间	每个子空间的概率
$\sum_{i=1}^{\max(\|t\|,\|s\|)} [t_i = s_i] = 0$, $inter = 0$	$p_{inter} = \frac{e^{\frac{-\alpha \cdot (k-inter)}{2}} \cdot \left(C_m^{inter} \cdot C_{d-1}^{k-1-inter} \right)}{\Omega}$
$\sum_{i=1}^{\max(\|t\|,\|s\|)} [t_i = s_i] = 1$, $inter = 1$	
…	
$\sum_{i=1}^{\max(\|t\|,\|s\|)} [t_i = s_i] = k$, $inter = k$	
说明：每个子空间中的候选项集与 t 的相似度大小相同	

式(3-5)中的*部分代表相似度为 0 的子空间,**部分代表相似度大于 0 的子空间.由于其样本空间大小为 C_{d+m}^{k},当 d 与 m 很大时,运行的时间与空间是指数级的.因此,TDC_CLDP 将候选项集总的样本空间划分为 k+1 个样本子空间,每个样本子空间中的候选项集与 t 的相似度大小 *inter* 是相同的,其范围为[0, k],如表 3-3 所示.

TDC_CLDP 方法中第 3~6 步对事务数据作等长处理,使得所有的事务数据 t 的长度均为 m,添加的项从集合 $\{a_{d+1},\cdots a_{d+m}\}$ 中选择,第 9 步基于 *uniform _ random*(0.0,1.0) 生成概率 r,第 12~15 基于 r 与 p_{inter} 确定子样本空间,最后 16~17 步利用 *sample* 抽样函数随机从 $\hat{t}$ 中抽取项的数目为 *inter*,从 $\hat{I}-\hat{t}$ 中抽取项的数目为 *k-inter*,拼接后返回给用户,由用户发送给服务器. TDC_CLDP 方法的时间复杂度为 $O(d)$,与项集域的大小呈线性关系.

定理 3-1. 基于 CLDP 模型与指数机制实现的 TDC_CLDP 方法满足 $\alpha-CLDP$ 约束.

证明：因为 TDC_CLDP 方法是对 TDC_CLDP_Cand 方法的具体实现,本质过程为指数机制的具体实现,针对原始数据 t ,以概率 p 随机选择一个候选项集 t' ,概率 p 为:

$$p=\frac{score(t')}{\sum_{s\in CandI} score(s)} \tag{3-6}$$

命名 TDC_CLDP 方法的随机机制为 Φ ,则基于 t 得到 t' 的概率为:

$$\Pr\left[\Phi(t)=t'\right]=\frac{e^{\frac{-\alpha\cdot dist(t,t')}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t,z)}{2}}} \tag{3-7}$$

为证明 TDC_CLDP 满足 $\alpha-CLDP$ 约束,需要证明式(3-8)正立.

$$\frac{\Pr\left[\Phi(t_1)=t'\right]}{\Pr\left[\Phi(t_2)=t'\right]}\le e^{\frac{\alpha\cdot dist(t_1,t_2)}{2}} \tag{3-8}$$

基于指数机制的定义以及式(3-7),可以得到二者的比值为:

$$\frac{\Pr\left[\Phi(t_1)=t'\right]}{\Pr\left[\Phi(t_2)=t'\right]}\le\frac{\dfrac{e^{\frac{-\alpha\cdot dist(t_1,t')}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}}{\dfrac{e^{\frac{-\alpha\cdot dist(t_2,t')}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_2,z)}{2}}}}=\underbrace{\frac{e^{\frac{-\alpha\cdot dist(t_1,t')}{2}}}{e^{\frac{-\alpha\cdot dist(t_2,t')}{2}}}}_{*}\cdot\underbrace{\frac{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_2,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}}_{**} \tag{3-9}$$

首先处理*部分,可以得到:

$$\frac{e^{\frac{-\alpha\cdot dist(t_1,t')}{2}}}{e^{\frac{-\alpha\cdot dist(t_2,t')}{2}}}=e^{\frac{\alpha\cdot\left[dist(t_2,t')-dist(t_1,t')\right]}{2}} \tag{3-10}$$

由于 *dist* 为距离度量,满足三角不等式的性质,即 $dist(t_2,t')-dist(t_1,t')\le dist(t_1,t_2)$ 成立,因此可以得到*部分的结论:

$$\frac{e^{\frac{-\alpha\cdot dist(t_1,t')}{2}}}{e^{\frac{-\alpha\cdot dist(t_2,t')}{2}}}\le e^{\frac{\alpha\cdot dist(t_1,t_2)}{2}} \tag{3-11}$$

接下来处理**部分,先对分子进行处理得到式(3-12):

$$\frac{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_2,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}=\frac{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_2,z)+\alpha\cdot dist(t_1,z)-\alpha\cdot dist(t_1,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}} \tag{3-12}$$

对式(3-12)应用三角不等式, $dist(t_1,z)-dist(t_2,z)\le dist(t_1,t_2)$,得到式(3-13):

$$\frac{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_2,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}\le\frac{\sum_{z\in CandI} e^{\frac{\alpha\cdot dist(t_1,t_2)-\alpha\cdot dist(t_1,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}} \tag{3-13}$$

观察式(3-13)发现 $\alpha\cdot dist(t_1,t_2)$ 与 z 没有关系,可以直接从求和运算中提出,得到式(3-14)为**部分的结论:

$$\begin{aligned}\frac{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_2,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}&\le\frac{e^{\frac{\alpha\cdot dist(t_1,t_2)}{2}}\cdot\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}{\sum_{z\in CandI} e^{\frac{-\alpha\cdot dist(t_1,z)}{2}}}\\&\le e^{\frac{\alpha\cdot dist(t_1,t_2)}{2}}\end{aligned} \tag{3-14}$$

综合*部分与**部分的结论,完成定理 3-1 的证明:

$$\frac{\Pr\left[\Phi(t_1)=t'\right]}{\Pr\left[\Phi(t_2)=t'\right]}\le e^{\frac{\alpha\cdot dist(t_1,t_2)}{2}}\cdot e^{\frac{\alpha\cdot dist(t_1,t_2)}{2}}=e^{\alpha\cdot dist(t_1,t_2)} \tag{3-15}$$

定理 3-1 证明完毕.

3.3 支持度计数的分布估计

项支持度计数的估计是差分隐私事务数据收集的重要任务,是评价隐私保护数据收集方法的重要指标,项支持度计数定义为包含该项的事务数据的数目,即 $P_a=\#\{t_i|a\in t_i,i\in[1,n]\}$.考虑项 a_i 以及事务数据 t,如果 $a_i\in t$,收集到的事务数据 t' 中同样包含 a_i 的概率定义为真正率 *TPR*(True Positive Rate):

$$TPR=\frac{e^{\frac{-\alpha}{2}\cdot(k-1)}\cdot\left(C_d^{k-1}\right)+\sum_{inter=2}^{k}\left(e^{\frac{-\alpha}{2}\cdot(k-inter)}\cdot\left(C_{m-1}^{inter-1}\cdot C_d^{k-inter}\right)\right)}{\Omega} \tag{3-16}$$

反之,如果 $a_i\notin t$,而收集到的事务数据 t' 中包含 a_i 的概率定义为错正率 *FPR*(False Positive Rate):

$$FPR=\frac{e^{\frac{-\alpha}{2}\cdot 0}\cdot C_{d-1}^{k-1}+\sum_{inter=1}^{k-1}\left(e^{\frac{-\alpha}{2}\cdot(k-inter)}\cdot\left(C_m^{inter}\cdot C_{d-1}^{k-1-inter}\right)\right)}{\Omega} \tag{3-17}$$

由于本文采用了不同的隐私模型 CLDP 与距离函数,因此 TPR 与 FPR 与文献[2]是完全不同的,推导过程的原理也不同.是两种不同的方法,是对 PrivSet 方法的改进.接下来分析如何对项的支持度计数进行估计.

考虑项 $a_i\in I$ 以及含有 n 个用户的事务数据集 $D=\{t_1,t_2,t_3,...,t_n\}$,数据收集者得到的事务数据集为 $D'=\{t_1',t_2',t_3',...,t_n'\}$,假设其真实的项分布为 $P_a=\{P_{a_1},P_{a_2},P_{a_3},...,P_{a_d}\}$,则 a_i 在隐私事务数据 t' 的期望频率为:

$$\begin{aligned}E\left[F_{a_i}\right]&=E\left[\#\{t_i'|a_i\in t_i'\}\right]\\&=n\cdot P_{a_i}\cdot TPR+n\cdot\left(1-P_{a_i}\right)\cdot FPR\end{aligned} \tag{3-18}$$

其中前半段 $n\cdot P_{a_i}\cdot TPR$ 表示保留下来真实的 a_i,而后半段 $n\cdot\left(1-P_{a_i}\right)\cdot FPR$ 代表噪音.因此, P_{a_i} 可以估计为:

$$\tilde{P}_{a_i}=\frac{1}{n}\cdot\frac{F_{a_i}-n\cdot FPR}{TPR-FPR} \tag{3-19}$$

$\tilde{P}_a$ 是对 P_a 的无偏估计,证明过程如式(3-20)所示.

$$\begin{aligned}E\left[\tilde{P}_a\right]&=E\left[\frac{1}{n}\cdot\frac{F_a-n\cdot FPR}{TPR-FPR}\right]\\&=\frac{1}{n\cdot\left(TPR-FPR\right)}\cdot E\left[F_a\right]-\frac{FPR}{TPR-FPR}\\&=\frac{1}{n\cdot\left(TPR-FPR\right)}\cdot\left\{n\cdot P_a\cdot TPR+n\cdot\left(1-P_a\right)\cdot FPR\right\}-\frac{FPR}{TPR-FPR}\\&=P_a\end{aligned} \tag{3-20}$$

F_{a_i} 为数据收集者得到 D' 统计得到,本文采用文献[2]提出的频数估计算法统计 F_{a_i} 并进一步得到 $\tilde{P}_{a_i}$,如表3-4 所示,注意其中 FPR 与 TPR 不同于文献[2].

Table 3-4 The Frequency Estimation Algorithm

表 3-4 项的频数估计算法

输入:数据收集者得到的 $D'=\left\{t'_1,t'_2,t'_3,...,t'_n\right\}$

输出: 项的频率分布估计 $P_{a_i}=\left\{P_{a_1},P_{a_2},P_{a_3},...,P_{a_d}\right\}$

```
1:  P̃_a = {0}^|I| , F_a = {0}^|I|
2:  for t' ∈ D' do:
3:      for a_i ∈ t' do:
4:          F_{a_i} = F_{a_i} + 1  // 从 D' 中统计每个项的频率
5:      end for
6:  end for
7:  for i = 1 to |I| do:
8:      P̃_{a_i} = (1/n) · (F_{a_i} − n·FPR) / (TPR − FPR)
9:  end for
10: return P̃_a
```

4 理论分析与隐私参数设置

本节首先分析项的频数估计的错误边界,该边界是 TDC_CLDP 中设置 k 的重要依据.过程与文献[2]类似,但本文对整个推导过程做了更详细的说明;其次考虑到最大后验置信度的攻击模型,本文通过约束 MPC 的上界为 ρ,找到了 CLDP 的隐私参数 α 与 ρ 的关系,由于参数 ρ 的设置具有启发式意义,即限定攻击者的攻击能力的上界,因此提出一种基于 ρ 的隐私参数启发式设置策略.

4.1 项的频数分布估计的错误边界

基于项的 *TPR* 与 *FPR*,接下来分析项分布估计的均方差(Mean Squared Error, *MSE*),考虑一个项 a_i,式(3-19)

中的随机变量 $\tilde{P}_a$ 是 F_a 的线性变换,而 F_a 又是 n 个伯努利随机变量之和,其中 $n\cdot P_a$ 是伯努利实验中以概率 TPR 成功的数目,而 $n-n\cdot P_a$ 是伯努利实验中以概率 FPR 成功的数目.又因为,重复 n 次独立的伯努利试验的方差为:$Var(X)=E(X^2)-E(X)^2=n\cdot p\cdot(1-p)$,其中 p 为实验成功的概率,则 F_a 的方差为:

$$Var(F_a)=n\cdot P_a\cdot TPR\cdot(1\text{-}TPR)+(n-n\cdot P_a)\cdot FPR\cdot(1\text{-}FPR) \tag{4-1}$$

因为随机变量 $\tilde{P}_a$ 是 F_a 的线性变换,根据离散型随机变量方差的线性运算性质,令 a,b 为常数:

$$Var(a\cdot X+b)=a^2\cdot Var(x) \tag{4-2}$$

可以将 $\tilde{P}_a$ 的线性变换整理为:

$$\tilde{P}_a=\frac{1}{n\cdot(TPR-FPR)}\cdot F_a-\frac{FPR}{TPR-FPR} \tag{4-3}$$

则 $\tilde{P}_a$ 的方差为:

$$\begin{aligned}Var(\tilde{P}_a)&=\frac{n\cdot P_a\cdot TPR\cdot(1\text{-}TPR)+(n-n\cdot P_a)\cdot FPR\cdot(1\text{-}FPR)}{n^2\cdot(TPR-FPR)^2}\\&=\frac{P_a\cdot TPR\cdot(1\text{-}TPR)+(1-P_a)\cdot FPR\cdot(1\text{-}FPR)}{n\cdot(TPR-FPR)^2}\end{aligned} \tag{4-4}$$

则项分布估计的均方误差为:

$$\begin{aligned}E\left[(\tilde{P}_a\text{-}P_a)^2\right]&=E\left[(\tilde{P}_a\text{-}E[\tilde{P}_a]+E[\tilde{P}_a]\text{-}P_a)^2\right]\\&=E\left[(\tilde{P}_a\text{-}E[\tilde{P}_a])^2\right]+E\left[E[\tilde{P}_a]-P_a\right]^2+2E\left[(\tilde{P}_a-E[\tilde{P}_a])\cdot(E[\tilde{P}_a]-P_a)\right]\\&=Var(\tilde{P}_a)+(E[\tilde{P}_a]-P_a)^2\end{aligned} \tag{4-5}$$

式(3-20)证明 $\tilde{P}_a$ 是 P_a 的无偏 p 估计,则:

$$MSE(\tilde{P}_a)=E\left[(\tilde{P}_a\text{-}P_a)^2\right]=Var(\tilde{P}_a) \tag{4-6}$$

则总的均方差为:

$$\begin{aligned}Error\ Bound&=\sum_{i\in[1,d]}E\left[(\tilde{P}_{a_i}-P_{a_i})^2\right]=\sum_{i\in[1,d]}Var(\tilde{P}_{a_i})\\&=\frac{\sum_{i\in[1,d]}P_{a_i}\cdot TPR\cdot(1\text{-}TPR)+\left(d-\sum_{i\in[1,d]}P_{a_i}\right)\cdot FPR\cdot(1\text{-}FPR)}{n\cdot(TPR-FPR)^2}\end{aligned} \tag{4-7}$$

4.2 隐私参数的启发式设置策略

差分隐私模型中,隐私参数用于控制隐私性与效用性,使二者达到平衡.但隐私参数的设置目前没有较好的指导策略,大多通过实验或者经验来完成.CLDP 隐私模型的最大后验置信度(MPC)攻击模型可定义为:

$$MPC=\Pr[v|y] \tag{4-8}$$

则 $MPC\le\rho$ 具有一定的启发式意义,即设置隐私参数为 ε,得到的响应结果为 y,通过 y 推断出原始值 v 的风险概率的上界为 ρ,找出 ρ 与 ε 的关系后,便可基于具有启发意义的 ρ 设置隐私参数 ε 的值.对式(4-8)展开得到式(4-9).

$$
\begin{aligned}
MPC = \Pr\left[v|y\right] &= \frac{\pi(v)\cdot \Pr\left[f(v)=y\right]}{\sum_{z\in CandI}\pi(z)\cdot \Pr\left[f(z)=y\right]} \\
&= \frac{\pi(v)\cdot \frac{e^{\frac{-\alpha\cdot d(v,y)}{2}}}{\Omega}}{\sum_{z\in CandI}\pi(z)\cdot \frac{e^{\frac{-\alpha\cdot d(z,y)}{2}}}{\Omega}} = \frac{\pi(v)\cdot \frac{e^{\frac{-\alpha\cdot d(v,y)}{2}}}{\Omega}}{\pi(v)\cdot \frac{e^{\frac{-\alpha\cdot d(v,y)}{2}}}{\Omega} + \sum_{z\in CandI, z\neq v}\pi(z)\cdot \frac{e^{\frac{-\alpha\cdot d(z,y)}{2}}}{\Omega}} \\
&= \frac{1}{1+\sum_{z\in CandI, z\neq v}\frac{\pi(z)}{\pi(v)}\cdot e^{\frac{\alpha\cdot(d(v,y)-d(z,y))}{2}}}
\end{aligned}
\tag{4-9}
$$

令所有项集的先验概率 π 均为：$1/C_{d+m}^{k}$,当 $d(v,y)=0$, $d(z,y)=k$ 时, $d(v,y)-d(z,y)=-k$,为最小,则不等式可以变形为:

$$
MPC=\frac{1}{1+\sum_{z\in CandI, z\neq v}\frac{\pi(z)}{\pi(v)}\cdot e^{\frac{\alpha\cdot(d(v,y)-d(z,y))}{2}}} \le \frac{1}{1+\left(C_{d+m}^{k}-1\right)\cdot e^{\frac{\alpha}{2}\cdot -k}} \tag{4-10}
$$

又因为 $k \le d$,则：

$$
MPC \le \frac{1}{1+\left(C_{d+m}^{k}-1\right)\cdot e^{\frac{\alpha}{2}\cdot -k}} \le \frac{1}{1+\left(C_{d+m}^{1}-1\right)\cdot e^{\frac{\alpha}{2}\cdot -d}} \tag{4-11}
$$

令：

$$
\frac{1}{1+\left(C_{d+m}^{1}-1\right)\cdot e^{\frac{\alpha}{2}\cdot -d}} \le \rho \tag{4-12}
$$

即最大后验置信度不超过 ρ ,通过对公式(4-11)运算后,得到式(4-13)：

$$
\alpha \le -\ln\left(\frac{1-\rho}{\rho}\cdot\frac{1}{C_{d+m}^{1}-1}\right)\cdot\frac{2}{d} \tag{4-13}
$$

通过式(4-13),建立了隐私 α 与 ρ 的关系,令：

$$
\alpha = -\ln\left(\frac{1-\rho}{\rho}\cdot\frac{1}{C_{d+m}^{1}-1}\right)\cdot\frac{2}{d} \tag{4-14}
$$

能够使得 $MPC \le \rho$.

5 实验结果与分析

本节我们在实验环境中评价本文提出的本地差分隐私事务数据收集方法 TDC_CLDP,并与 PrivSet 方法进行比较.首先考查在不同参数设置的情况下,(d,m)与(ε,α)的不同值对 TDC_CLDP 与 PrivSet 所产生的 k 值与错误边界(*Error Bound*)的影响.同时文献[33]与文献[22]都是针对事务数据基于本地差分隐私模型进行随机响应的优秀方法,这二者的基本思想类似,都是从事务数据随机抽取一个项,进行随机响应后发送给服务器,即服务器最终只会收到一个 Bit 的数据,本文将这类方法统称为 Binary Randomized Response(简称:BRR 方法),实验中也与 BRR 方法进行了对比;其次,在人造数据集与真实数据集中对比 TDC_CLPD 与 PrivSet 对项的频数分布估计;第三,在不同的参数设置下,分析 TDC_CLDP 在 TopK 频繁项集挖掘任务中的效用性;第四,分析不同的 MPC 上界 ρ 对隐私参数 α 的影响;最后,从整体上分析与对比 TDC_CLDP 与 PrivSet 的区别及改进.

5.1 实验设置与运行环境

为与 PrivSet 方法进行比较,本文用的数据集与实验环境与该方法相同.通过模拟生成的事务数据集尽可能

与现实的数据集接近,数据集中的用户数为1000,项集域长度d的范围为4~200,事务最大长度m的范围为2~150,隐私参数ε,α的范围为0.01~3.0.不同的参数组合均进行模拟实验1000次,结果取平均值.每次模拟生成的事务数据集的中每个项均是通过从I中以概率m/d随机抽取.本文提出的TDC_CLDP与PrivSet类似,都与具体的数据集无关.模拟生成的事务数据集具有广泛的代表性,为了进一步验证TDC_CLDP的应用场景,本文在真实数据集MovieLens进行了相同的实验,从MovieLens数据集中选取1000个用户,以及前200个使用频繁的项.

实验的运行环境为Intel(R) Core(TM) i7-7660U CPU @ 2.50GHz 2.50 GHz,16.00GB内存,64位Win10操作系统,实现算法的编程语言为Python3.6.3.整个实验过程如图5-1所示.

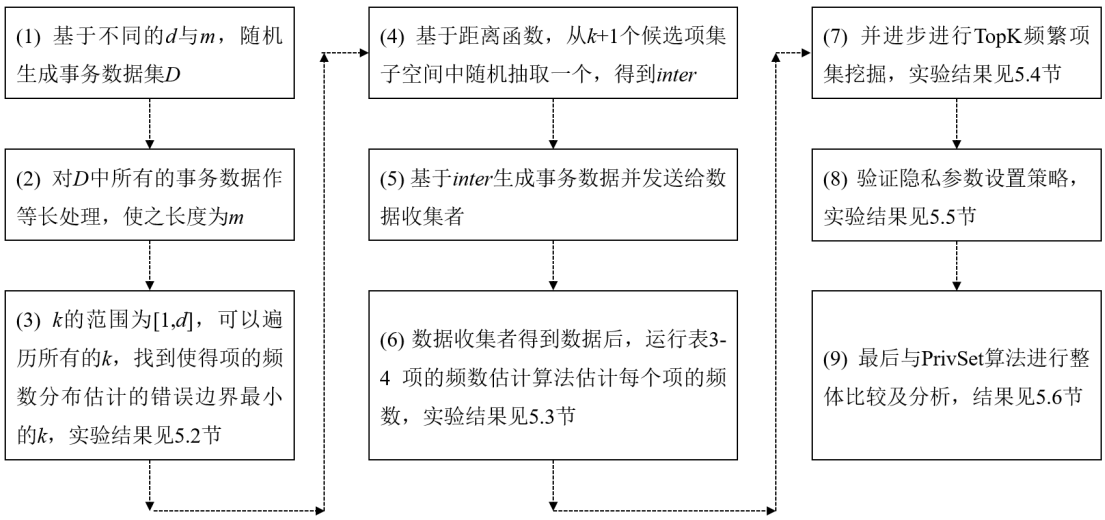


Fig.5-1 The ExperimentalPprocedure

图5-1 实验过程

5.2 k值与项的频数分布估计的错误边界

TDC_CLDP方法生成的事务数据的长度是参数k,而可能的k值范围为[1,d],由于项支持度计数的估计是整个算法的核心,它的错误边界是最重要的评价指标,其值越小越好.在d与m确定的前提下,错误边界只与k有关,可以通过遍历[1,d],寻找一个k值使得错误边界最小.因此,k值的确定以及计算项支持度计数错误边界的最小值是TDC_CLPD的关键.本文通过实验对比了TDC_CLPD与PrivSet中k值与支持计数估计错误边界随不同参数变化的情况.由于BRR算法只随机选择了一个项, k值没有意义,因此只需比较TDC_CLPD与BRR方法的支持计数估计错误边界,实验结果如表5-1所示,其中每一对(d,m)值对应三行数据.

从表5-1中可以发现: (1) (d,m)的值相同时,随着隐私参数(ε为PrivSet与BRR的参数,α为TDC_CLDP的参数)的增长,错误边界都是下降的,符合理论推断.因为随着隐私参数的增长,隐私性越来越低,而效用性会越来越高,错误边界也会越来越小;(2) 从k值的角度来看,一方面TDC_CLDP与PrivSet两种方法的k值均会随着隐私参数的增长而减小,原因是噪声项越来越少;另一方面PrivSet的k值非常小,最终随着隐私参数的增长会变成1,这是因为PrivSet方法主要用于单个项的支持度计数估计,相当于1-频繁项集的支持度计数估计,因此k值为1是符合理论的,而TDC_CLDP的k值一般会比较大,这是因为TDC_CLDP方法为了满足频繁项集挖掘任务,采用了不同的距离度量函数,保留了更多的信息;(3) 当(d,m)较小时,PrivSet方法的Error Bound要比TDC_CLDP方法小,但TDC_CLDP方法还是比BRR要优.而当(d,m)增长到一定程度时(如d=16,m=8),TDC_CLDP方法的Error Bounds比PrivSet方法要小.综上分析,PrivSet适合于d与m较小的场景,主要用于1-频繁项集的支持度计数分布估计,而TDC_CLDP方法适用于d与m较大的场景,主要用于TopK频繁项集挖掘,包括1-频繁项集的支持度计数分布估计任务.

Table 5-1 The Error Boundary of k Value and Frequency Distribution Estimation

表 5-1 k 值与频数分布估计的错误边界

(d, m)	方法	$\varepsilon,\alpha=0.01$		$\varepsilon,\alpha=0.1$		$\varepsilon,\alpha=0.4$		$\varepsilon,\alpha=1$		$\varepsilon,\alpha=2$	
		k	*Error Bound*	k	*Error Bound*	k	*Error Bound*	k	*Error Bound*	k	*Error Bound*
(4,2)	BRR	-	960000	-	9600	-	600	-	96	-	24
	PrivSet	1	299004	1	2904	1	167	1	24	1	6
	TDC_CLDP	3	**666666**	3	**6666**	3	**416**	3	**66**	2	**16**
(8,4)	BRR	-	7679999	-	76799	-	4799	-	767	-	191
	PrivSet	1	1315623	1	12783	1	737	1	108	1	29
	TDC_CLDP	6	**1613333**	6	**16133**	6	**1008**	5	**160**	5	**39**
(16,2)	BRR	-	2879999	-	28799	-	1799	-	287	-	71
	PrivSet	4	1404490	4	13827	3	830	2	116	1	20
	TDC_CLDP	9	**2568896**	9	**25696**	8	**1601**	7	**252**	5	**59**
(16,4)	BRR	-	12799998	-	127998	-	7998	-	1278	-	318
	PrivSet	2	2880450	2	28169	2	1663	1	229	1	45
	TDC_CLDP	10	**2888004**	10	**28884**	9	**1803**	8	**285**	7	**68**
(16,8)	BRR	-	61439998	-	614398	-	38398	-	6142	-	1534
	PrivSet	1	5501702	1	53460	1	3086	1	457	1	127
	TDC_CLDP	12	**3526666**	12	**35266**	12	**2204**	11	**350**	10	**85**
(32,8)	BRR	-	102399997	-	1023997	-	63997	-	10237	-	2557
	PrivSet	2	11996243	2	117231	2	6907	1	948	1	192
	TDC_CLDP	20	**6084008**	20	**60848**	19	**3796**	17	**601**	14	**145**
(32,16)	BRR	-	491519996	-	4915196	-	307196	-	49148	-	12284
	PrivSet	1	22485227	1	218502	1	12624	1	1879	1	531
	TDC_CLDP	24	**7363333**	24	**73633**	23	**4597**	22	**731**	20	**179**
(64,8)	BRR	-	184319994	-	1843194	-	115194	-	18426	-	4602
	PrivSet	4	25296086	4	248035	3	14606	2	2007	1	359
	TDC_CLDP	36	**11202249**	35	**112011**	33	**6984**	29	**1103**	23	**263**
(64,16)	BRR	-	819199993	-	8191993	-	511993	-	81913	-	20473
	PrivSet	2	48925531	2	477949	2	28134	1	3852	1	791
	TDC_CLDP	40	**12482015**	39	**124817**	38	**7788**	34	**1234**	29	**298**
(64,32)	BRR	-	3932159992	-	39321592	-	2457592	-	393208	-	98296
	PrivSet	1	90897749	1	883327	1	51057	1	7619	1	2167
	TDC_CLDP	48	**15041666**	48	**150416**	46	**9391**	44	**1493**	40	**365**
(128,16)	BRR	-	1474559988	-	14745588	-	921588	-	147444	-	36852
	PrivSet	4	103015973	4	1009489	3	59292	2	8128	1	1461
	TDC_CLDP	72	**22721165**	71	**227184**	66	**14166**	58	**2238**	46	**535**

5.3 项支持度计数的分布估计

项支持度计数的分布估计是指数据收集者收集数据后进行的主要数据分析任务,评价项支持度计数分布估计的好坏主要有两个指标,分别是总的误差(**L_1 Norm Error**)以及最大绝对误差(**L_{Max} Norm Error**),其中总的误差公式为:

$$\left|\tilde{P}_a - P_a\right| = \sum_{i\in I}\left|\tilde{P}_{a_i} - P_{a_i}\right| \tag{5-1}$$

最大绝对误差为:

$$\left|\tilde{P}_a - P_a\right|_\infty = \max_{i\in I}\left|\tilde{P}_{a_i} - P_{a_i}\right| \tag{5-2}$$

为与 PrivSet 类似,本文同时采用文献[38]中提出概率单纯形方法对项支持度计数分布估计 $\tilde{P}_a$ 进行优化.实

验结果如图 5-2I、5-2II 与图 5-3I、5-3II 所示.图 5-2I、II 与图 5-3I、II 的实验结果进一步验证了表 5-1 中蕴含的结论,当(d,m)固定时,随着隐私参数的增长,总误差与最大绝对误差是一直下降的,且本文提出的 TDC_CLDP 方法要比 PrivSet 方法表现更优,总误差与最大绝对误差均是较小的.另一方面,随着(d,m)的增长,TDC_CLDP 方法对比 PrivSet 的优势逐渐增大,越来越好.特别是图 5-3I 中所示的最大绝对误差,当 d 与 m 的值增长到 64 与 32 时,TDC_CLDP 方法已明显占优,这个结果进一步验证了 TDC_CLDP 方法适用于 d 与 m 均较大的场景.

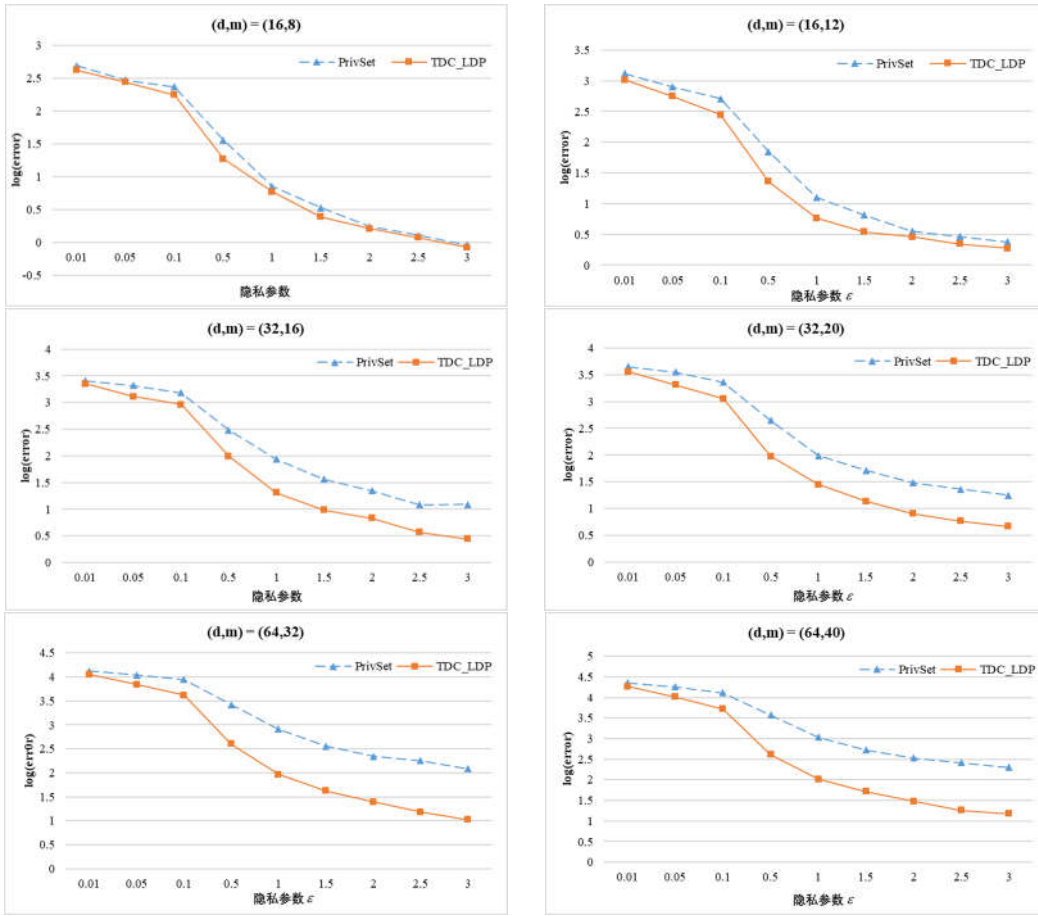
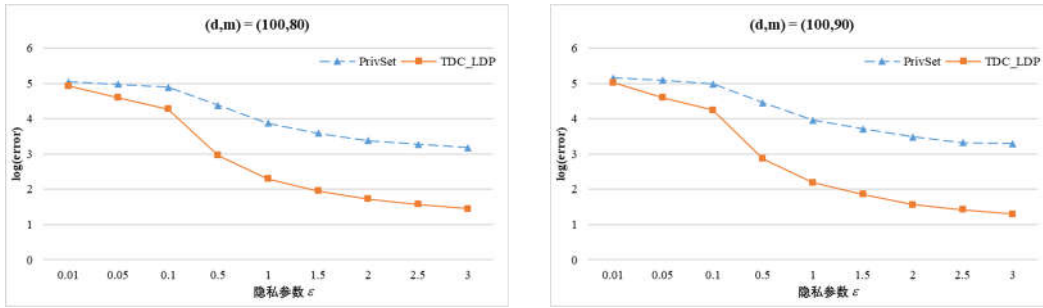
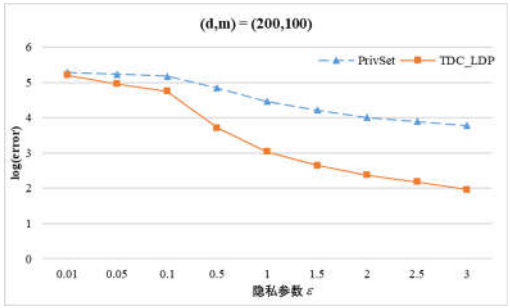


Fig.5-2 I The Distribution Estimation of Item Support Count for L_1 Norm Error with Synthetic Data

图 5-2 I 项支持度计数的分布估计, L_1 Norm Error-人造数据





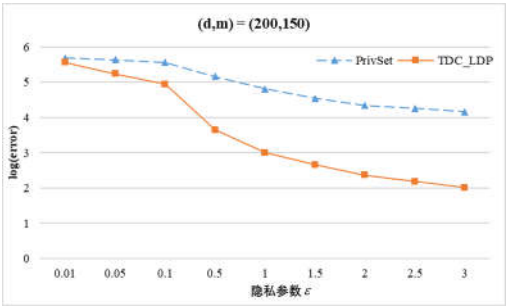
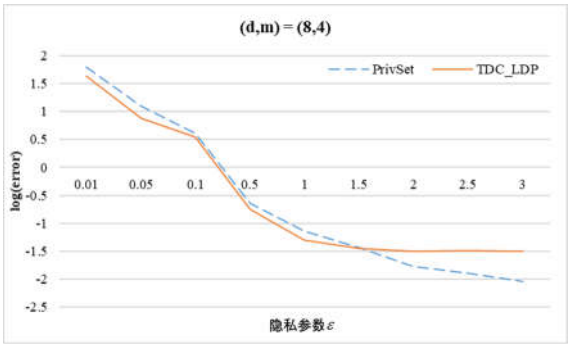
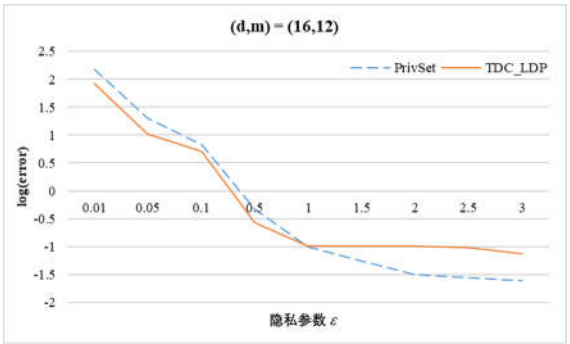
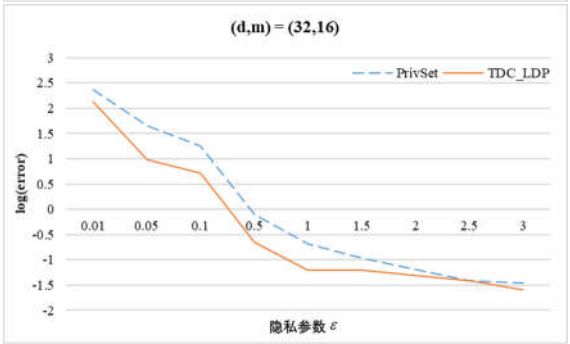


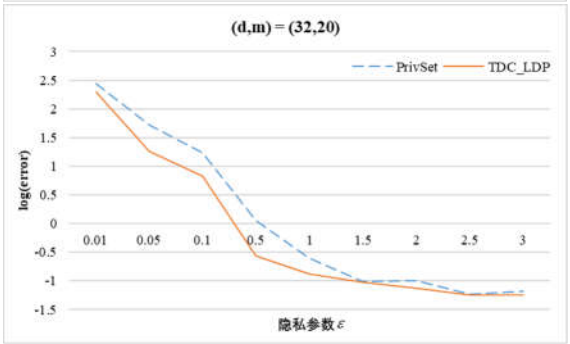
Fig.5-2 II The Distribution Estimation of Item Support Count for L_1 Norm Error with MovieLens

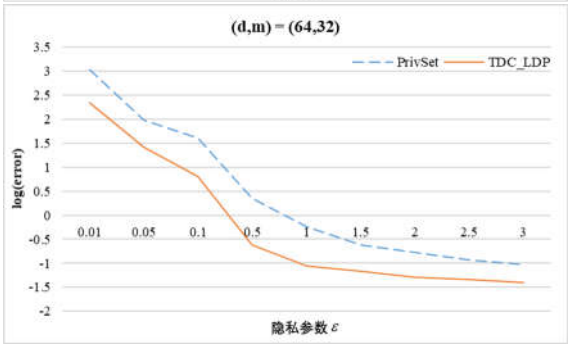
图 5-2 II 项支持度计数的分布估计, L_1 Norm Error-真实数据 MovieLens











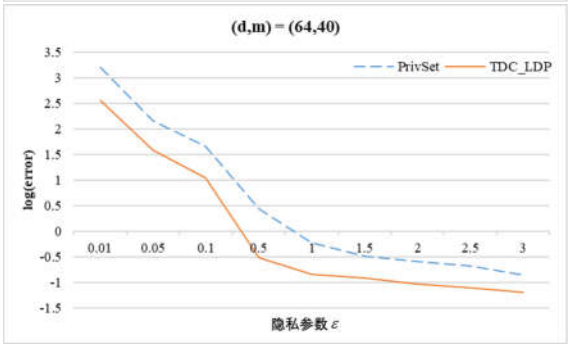


Fig.5-3 I The Distribution Estimation of Item Support Count for L_{Max} Norm Error with Synthetic Data

图 5-3 I 项支持度计数的分布估计, L_{Max} Norm Error-人造数据

5.4 TopK频繁项集挖掘

本文提出的 TDC_CLDP 方法与 PrivSet 最大的不同是,TDC_CLDP 方法生成的事务数据集除了保留足够多的统计信息外,还尽可能多的保留了项之间的关联信息,这些关联信息可用于关联规则、频繁项集挖掘、TopK 频繁项集挖掘等任务,本节通过实验重点验证了 TDC_CLDP 对 TopK 频繁项集挖掘的效用性.

令事务数据集真实的 TopK 频繁项集为 F_k ,基于 TDC_CLDP 生成的事务数据集 D' 的 TopK 频繁项集为 F_k' ,则二者之间的绝对误差为:

$$absolute\ error = \left|F_k - F_k'\right| \tag{5-3}$$

相对误差为:

$$relative\ error = \frac{\left|F_k - F_k'\right|}{F_k} \tag{5-4}$$

实验结果如图 5-4、图 5-5、图 5-6 所示.其中图 5-4 显示当隐私参数固定为 0.01,d 与 m 取不同值时,TopK 频繁项集任务随着 k 增长其绝对误差的变化趋势;图 5-5 显示当隐私参数固定为 0.1,d 与 m 取不同值时,TopK 频繁项集任务随着 k 增长其**绝对误差**的变化趋势;图 5-6 显示当 d 与 m 固定不变,隐私参数取不同值时,TopK 频繁项集任务随着 k 增长其**相对误差**的变化趋势;

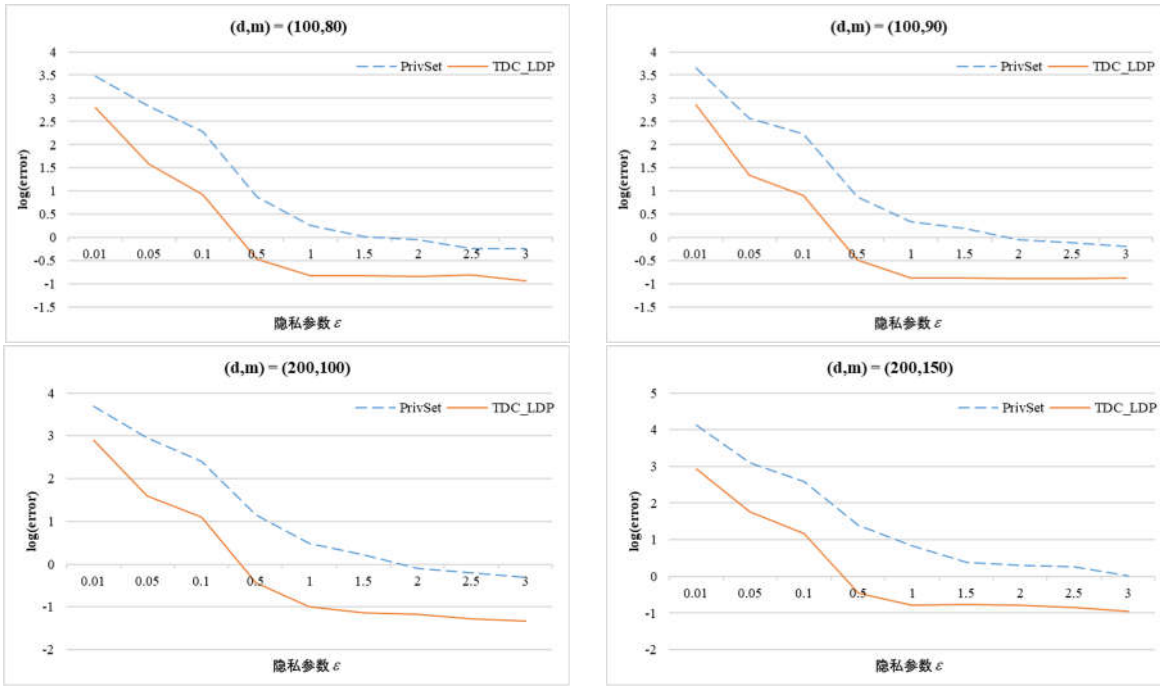
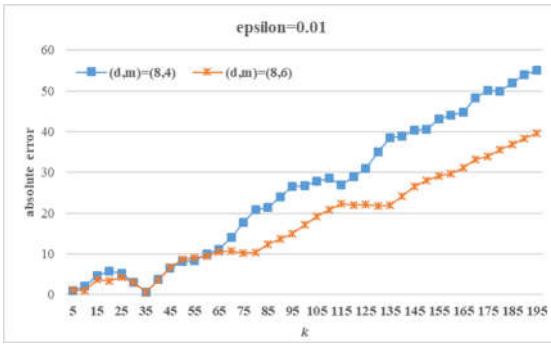


Fig.5-3 II The Distribution Estimation of Item Support Count for L_{Max} Norm Error with MovieLens Data

图 5-3II 项支持度计数的分布估计, L_{Max} Norm Error-真实数据 MovieLens

从图 5-4 与图 5-5 中可以发现:(1) 在隐私参数固定的前提下,随着 k 的增长,TopK 频繁项集挖掘任务的绝对误差呈现相同的趋势,即 k 越大,绝对误差也相对较大,其原因是随着 k 的增长,频繁项集的数量也随之增长,由于随机扰乱的原因,必然会造成一定的效用性损失;另外,图中指出当 k 较小时,结果明显较优,TDC_CLDP 能保留大多数的频繁项集. (2) 从参数 d 与 m 的角度分析,d 固定,随着 m 的增长,绝对误差会随之降低,这说明 TDC_CLDP 在用户拥有较长事务数据的场景会有比更好的表现.



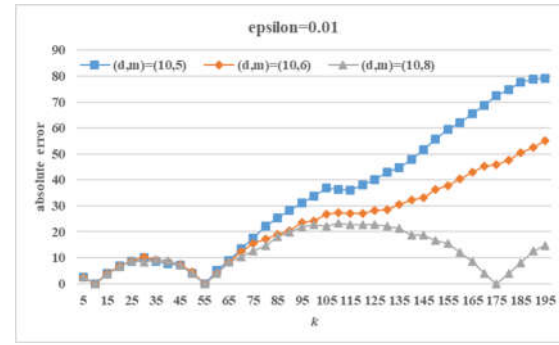
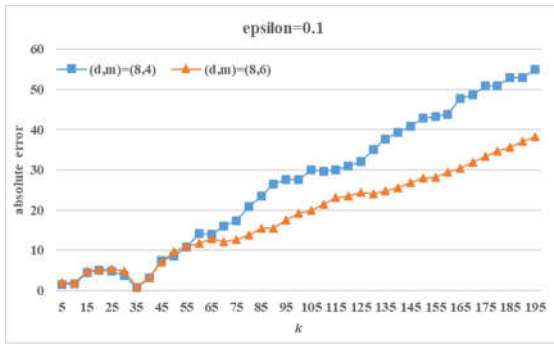


Fig.5-4 TopK frequent itemset, ε=0.01,d and m vary

图 5-4 TopK 频繁项集, ε=0.01,d 与 m 取不同值



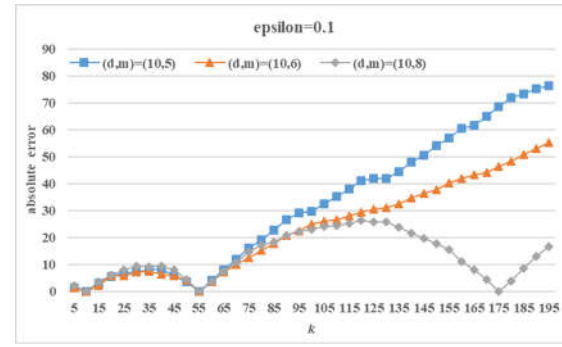
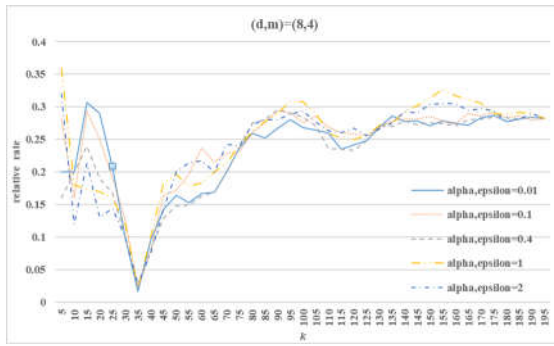


Fig.5-5 TopK frequent itemset, ε=0.1,d and m vary

图 5-5 TopK 频繁项集,隐私参数固定为 0.1,d 与 m 取不同值



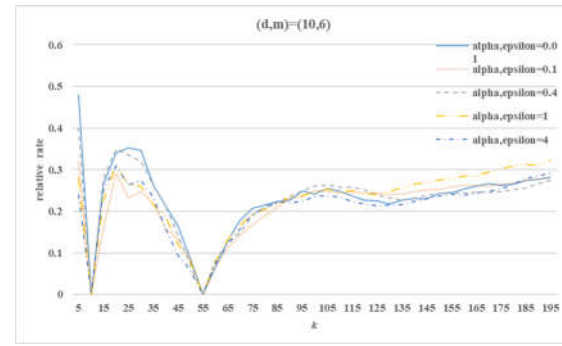


图 5-6 TopK frequent itemset, d and m fixed, ε vary

图 5-6 TopK 频繁项集,固定 d 与 m 不变,隐私参数取不同值

从图 5-6 发现,固定 d 与 m 时,随着隐私参数的变化,TopK 频繁项集挖掘任务的相对误差没有呈现出明显的趋势,但随 k 变化整体趋势是相同的,因此可以得出结论,相对误差并不受隐私参数的影响,只随着 d 与 m 的变化呈现改变趋势.

5.5 隐私参数设置策略

隐私参数是差分隐私模型中的重要参数,本文提出的启发式隐私参数设置策略基于 MPC 攻击模型推断出 ρ 与 α 之间的关系如式(4-13).

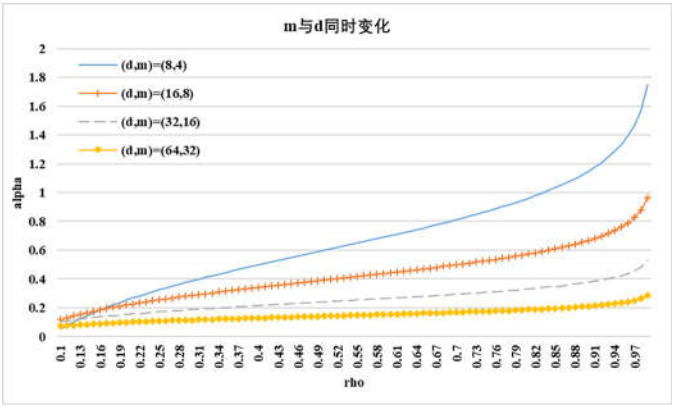


Fig. 5-7 The relationship between ρ and α,m and d vary

图 5-7 ρ 与 α 的关系,m 与 d 同时变化

理论分析表明,基于 MPC 的上界 ρ 设置的隐私参数并不影响算法的隐私性与效用性,同时为隐私参数的设置提供很好的语义解释.隐私参数的启发式设置策略的实验结果如图 5-7、图 5-8、表 5-2 所示.

图 5-7 显示了隐私参数 α 随着 ρ 增长所呈现的趋势,从图 5-7 中可以发现:(1) 随着 ρ 的增长,α 也显示出增长的趋势,这是因为 ρ 越大,即 MPC 越高,用户的信息越容易被推断出来,说明数据扰乱的程度越低,因此 α 随之增长是合理的,α 越大,添加的噪声越少;(2) 当(d,m)越小时,α 随 ρ 增长的趋势越明显,曲线越陡峭,当(d,m)越大时,α 随 ρ 增长的趋势曲线越平缓,这个实验结论是合理的,因为(d,m)较大时,数据蕴含的隐私信息越多,在 ρ 相同的情况,需要添加更多的扰乱信息,即要求 α 更小.

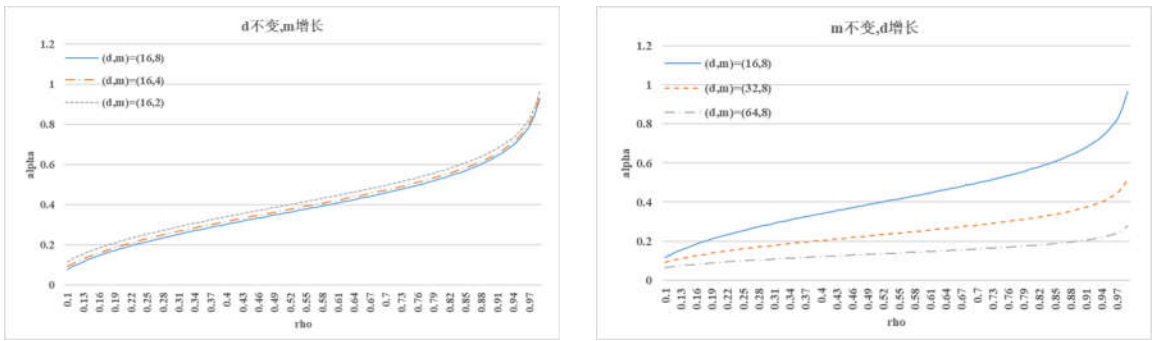


Fig. 5-8 The relationship between parameter d and m, fix one and vary the other

图 5-8 参数 d 与 m 的关系, 分别固定其中一个,另一个变化

图 5-8 的左图显示了固定 d 不变,m 增长时隐私参数 α 随着 ρ 增长所呈现的趋势,右图显示了固定 m 不变,d 增长时隐私参数 α 随着 ρ 增长所呈现的趋势.从图 5-8 中可以发现:(1) 整体的变化趋势与图 5-7 的实验结果相同; (2) 图 5-8 左图所示,当 d 不变,m(=2,4,8)的增长较为缓慢时,三根曲线变化趋势相同,且区别不大.特别地,在 ρ 相同的前提下,m 增大时,α 是减小的; (3) 图 5-8 右图所示,当 m 不变,d(=16,32,64)的增长较快时,三根曲线变化趋势相同,但区别是比较大的(这点与左图不同).特别地,在 ρ 相同的前提下,d 增大时,α 也是增大的.

综上图 5-7 与图 5-8 所示实验结果,可以得出结论,当(d,m)较大时,α 随 ρ 的变化比较明显,进一步验证了 TDC_CLDP 适用于(d,m)较大的场景.表 5-2 显示了在不同(d,m)的场景下,可以依据 ρ 指导隐私参数 α 的设置,由于 ρ 具有一定的语义,因此,隐私参数 α 的设置也具有语义性.

Table 5-2 The relationship between privacy α parameter and upper bound ρ of MPC

表 5-2 隐私参数 α 与 MPC 上界 ρ 的关系

ρ	α				
	(d,m)=(16,8)	(d,m)=(32,8)	(d,m)=(32,16)	(d,m)=(64,8)	(d,m)=(64,16)
0.1	0.12	0.09	0.1	0.06	0.07
0.2	0.22	0.14	0.15	0.09	0.09
0.3	0.29	0.18	0.19	0.11	0.11
0.4	0.34	0.2	0.22	0.12	0.12
0.5	0.39	0.23	0.24	0.13	0.14
0.6	0.44	0.25	0.27	0.15	0.15
0.7	0.5	0.28	0.29	0.16	0.16
0.8	0.57	0.32	0.33	0.18	0.18
0.9	0.67	0.37	0.38	0.2	0.21

5.6 TDC_CLDP对比PrivSet的改进

本文受文献[2]启发,对 PrivSet 方法进行改进,二者的区别主要有三个方面: (1) 采用的隐私模型不同; (2) 所用的距离函数不同; (3) TDC_CLDP 提出了新的隐私参数设置策略.具体如表 5-3 所示.从表 5-3 中可以发现,本文的优势主要体现在以下几个方面: (1) 本文采用的隐私模型 CLDP 是 LDP 的一般形式,距离函数的引入,更加适合于面向频繁项集挖掘的事务数据收集任务,且语义上更加明显; (2) 启发式隐私参数的设置方式,使得隐私参数的设置更加直观; (3) TDC_CLDP 应用范围更加广泛,主要包括事务数据分析,项的频数估计、频繁项集挖掘、TopK 频繁项集挖掘、关联规则挖掘等; (4) 总的来说,理论与实验结果表明 TDC_CLDP 整体上优于 PrivSet.

Table 5-3 Difference and improvement of TDC_CLDP VS. PrivSet

表 5-3 TDC_CLDP 对比 PrivSet 的区别及改进

名称	算法	区别与改进	本文优势
隐私模型	TDC_CLDP	CLDP	**LDP 是 CLDP 的特殊情况,CLDP 是 LDP 的泛化**
	PrivSet	LDP	
隐私参数 隐私上界	TDC_CLDP	隐私参数: α , $\Pr\left[\Phi\left(v_1\right)=y\right]\le e^{-\alpha\cdot d\left(v_1,v_2\right)}\cdot\Pr\left[\Phi\left(v_2\right)=y\right]$	**将距离的概念引用本地差分隐私,语义上更明显**
	PrivSet	隐私参数: ε , $\Pr\left[\Re\left(t\right)=t^*\right]\le e^{\varepsilon}\times\Pr\left[\Re\left(t'\right)=t^*\right]$	
隐私参数设置策略	TDC_CLDP	基于 MPC 的上界 ρ	**隐私参数的设置更加直观**
	PrivSet	-	
距离函数	TDC_CLDP	项集的曼哈顿距离,即相异度: $\sum_{i=1}^{\max(\lvert t\rvert,\lvert s\rvert)}\lvert t_i-s_i\rvert$	**保留更多信息,具有更好的数据效用性**
	PrivSet	是否有交集: $\left[\hat{t}\cap\hat{s}\ne\varnothing\right]$	
抽样概率	TDC_CLDP	$e^{\frac{-\alpha\left\{\sum_{i=1}^{\max(\lvert t\rvert,\lvert s\rvert)}\lvert t_i-s_i\rvert\right\}}{2}}\Big/\Omega$	**相异度越低,相似度越高,抽中概率越大,即效用性越好**
	PrivSet	$e^{\varepsilon\left[\hat{t}\cap\hat{s}\ne\varnothing\right]}/\Omega$	

Table 5-3 Difference and improvement of TDC_CLDP VS. PrivSet **[Continued]**

表 5-3 TDC_CLDP 对比 PrivSet 的区别及改进 **[续表]**

名称	算法	区别与改进	本文优势
引入距离函数后,规范化因子、真正率、假正率有所区别			
Ω	TDC_CLDP	$e^{\frac{-\alpha}{2}\cdot 0}\cdot C_d^k+\sum_{inter=1}^{k}\left(e^{\frac{-\alpha\cdot(k-inter)}{2}}\cdot\left(C_m^{inter}\cdot C_d^{k-inter}\right)\right)$	
	PrivSet	$C_d^k+\exp(\varepsilon)\cdot\left(C_{d+m}^k-C_d^d\right)$	
TPR	TDC_CLDP	$\frac{e^{\frac{-\alpha\cdot k}{2}}\cdot\left(C_d^{k-1}\right)+\sum_{inter=2}^{k}\left(e^{\frac{-\alpha\cdot(k-inter)}{2}}\cdot\left(C_{m-1}^{inter-1}\cdot C_d^{k-inter}\right)\right)}{\Omega}$	
	PrivSet	$\frac{e^{\varepsilon}\cdot C_{d+m-1}^{k-1}}{C_d^k+e^{\varepsilon}\cdot\left(C_{d+m}^k-C_d^k\right)}$	
FPR	TDC_CLDP	$\frac{e^{\frac{-\alpha}{2}\cdot 0}\cdot C_{d-1}^{k-1}+\sum_{inter=1}^{k-1}\left(e^{\frac{-\alpha\cdot(k-inter)}{2}}\cdot\left(C_m^{inter}\cdot C_{d-1}^{k-1-inter}\right)\right)}{\Omega}$	
	PrivSet	$\frac{C_{d-1}^{k-1}}{\Omega}+\exp(\varepsilon)\cdot\frac{k\cdot\left(C_{d+m}^k-C_d^k\right)-m\cdot C_{d+m-1}^{k-1}}{d\cdot\Omega}$	
适用范围	TDC_CLDP	主要用于数据分析,项的频数估计、频繁项集挖掘、TopK 频繁项集挖掘、关联规则挖掘等	
	PrivSet	主要用于统计,项的频数估计、1-频繁项集挖掘	
结论	理论与实验结果表明：TDC_CLDP 整体上优于 PrivSet		

6 结束语

事务数据是一种重要的数据类型,具有广泛的应用场景,如推荐系统、购物分析、用户行为分析等.由于事务数据产生于用户真实的购物或浏览行为,其中含有大量用户的隐私信息,人们对隐私信息也越来越关注.因此,研究在保护用户隐私的前提下收集用户的数据显得至关重要.本文提出一种面向频繁项集挖掘的本地差分隐私事务数据收集方法,基于压缩的本地差分隐私模型,设计了一种新的距离度量函数与抽样方法,理论依据充实,实验效果对比PrivSet方法更优,同时考虑到隐私参数的设置困难,本文还提出一种基于最大后验置信度的启发式隐私参数设置策略,使得隐私参数的设置能够在语义的指导下进行.接下来工作方向有以下 3 个: (1) 考虑将本方法应用于轨迹数据的隐私保护收集; (2) 分析与对比本地差分隐私与压缩的本地差分隐私的本质区别; (3) 实验表明 TDC_CLDP 方法适用于(d,m)较大的场景,主要原因是不同的距离函数导致了错误边界的不同,需要进一步从理论的角度分析其原因; (4) 本文设计的分值函数是一个曼哈顿距离,该距离的直观意义为相异度,即项不相同的数目,基于该分值抽样时引入了额外的噪音,即将不存在的项(项为0的部分)也看成存在(即将为0的所有项全部转为1),后续工作需要解决该问题,提升算法的效用性.

References:

[1] Ye Q, Meng X, Zhu M, Huo Z. Survey on Local Differential Privacy . Ruan Jian Xue Bao/Journal of Software. 2018, 7(29): 159-183(in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5364.htm.

[2] Shaowei W, Liusheng H, Yiwen N, et al. PrivSet: Set-Valued Data Analyses with Local Differential Privacy. In: IEEE Conference on Computer Communications.Honolulu, HI, USA: IEEE, 2018. 1088-1096.

[3] Sweeney L. k-Anonymity: A Model For Protecting Privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems. 2002, 5(10): 557-570.

[4] Machanavajjhala A, Kifer D, Gehrke J, et al. L-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2017, 1(1): 1-3.

[5] He Y, Naughton J F. Anonymization of Set-Valued Data via Top-Down, Local Generalization. Proceedings of the VLDB Endowment. 2009, 2(1): 934-945.

[6] Gergely A, Jagdish P A, Claude C. Probabilistic k^m-anonymity Efficient Anonymization of Large Set-Valued Datasets. In: 2015 IEEE International Conference on Big Data (Big Data).California: IEEE, 2015. 1164-1173.

[7] Liu J. Publishing Set-Valued Data Against Realistic Adversaries. Journal of Computer Science and Technology. 2012, 27(1): 24-36.

[8] Wang S, Tsai Y, Kao H, et al. Anonymizing Set Valued Social Data. In: Green Computing and Communications.New York: IEEE, 2010. 809-812.

[9] R. Chen B C F N. Privacy-preserving Trajectory Data Publishing by Local Suppression. Information Sciences. 2013, 231(10): 83-97.

[10] Ghinita G, Tao Y, Kalnis P. On the Anonymization of Sparse High-Dimensional Data. In: Proceedings of the 25nd International Conference on Data Engineering. Piscataway: IEEE, 2008.

[11] Cao J, Karras P, Raïssi C, et al. ρ-uncertainty: Inference-Proof Transaction Anonymization. Proceedings of the VLDB Endowment. 2010, 3(1-2): 1033-1044.

[12] Terrovitis M, Mamoulis N, Kalnis P. Privacy-Preserving Anonymization of Set-Valued Data. Proceedings of the VLDB Endowment. 2008, 1(1): 115-125.

[13] Xu Y, Wang K, Fu A, et al. Anonymizing transaction databases for publication. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 767-775.

[14] Chen R, Fung B C M, Desai B C. Differentially Private Trajectory Data Publication. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2012. 213-221.

[15] Ouyang J, Yin J, Liu SP, Liu YB. An Effective Differential Privacy Transaction Data Publication Strategy. Journal of Computer Research and Development, 2014, 51(10): 2195-2205 (in Chinese with English abstract).

[16] Ouyang Jia, Yin Jian, Liu Shaopeng. Differential Privacy Publishing Strategy for Distributed Transaction Data . Ruan Jian Xue Bao/Journal of Software. 2015, 26(06): 1457-1472. http://www.jos.org.cn/1000-9825/4576.htm

[17] Su S, Xu S, Cheng X, et al. Differentially Private Frequent Itemset Mining via Transaction Splitting. IEEE Transactions on Knowledge and Data Engineering. 2015, 27(7): 1875-1891.

[18] Li N, Qardaji W H, Su D, et al. PrivBasis: Frequent Itemset Mining with Differential Privacy. PVLDB. 2012, 5(11): 1340-1351.

[19] Lee J, Clifton C. Top-k frequent itemsets via differentially private FP-trees. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM, 2014. 931-940.

[20] Xiong X, Chen F, Huang P, et al. Frequent Itemsets Mining With Differential Privacy Over Large-Scale Data. IEEE Access. 2018, 6: 28877-28889.

[21] Fanti G, Pihur V, Erlingsson U. Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries. Proceedings on Privacy Enhancing Technologies. 2016, 2016(3): 41-61.

[22] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In: Proceeding CCS '14 Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, Arizona, USA: ACM, 2014. 1054-1067.

[23] Warner S L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association. 1965, 60(309): 63-69.

[24] Sun C, Fu Y, Zhou J, et al. Personalized Privacy-Preserving Frequent Itemset Mining Using Randomized Response. The Scientific World Journal. 2014, 2014: 1-10.

[25] Evfimievski A, Srikant R, Agrawal R, et al. Privacy Preserving Mining of Association Rules. Information Systems. 2004, 29(4): 343-364.

[26] Ding B, Kulkarni J, Yekhanin S. Collecting Telemetry Data Privately. In: Neural Information Processing Systems. California: Curran Associates, 2017. 3574-3583.

[27] Kairouz P, Bonawitz K, Ramage D. Discrete Distribution Estimation under Local Privacy. IEEE Transactions on Information Theory. 2016, 8(64): 5662-5676.

[28] Duchi J C, Jordan M I, Wainwright M J. Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Nevada: Curran Associates, 2013. 1529-1537.

[29] Andrés M, Bordenabe N, Chatzikokolakis K, et al. Geo-indistinguishability: differential privacy for location-based systems. In: Proceedings of the 2013 ACM SIGSAC Confe-rence on Computer & Communications Security. Berlin: ACM, 2013. 901-914.

[30] Bordenabe N, Chatzikokolakis K, Palamidessi C. Optimal Geo-Indistinguishable Mechanisms for Location Privacy. In: ACM, 2014. 251-262.

[31] Hsu J, Khanna S, Roth A. Distributed Private Heavy Hitters[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 7391, 461-472.

[32] Bassily R, Smith A. Local, Private, Efficient Protocols for Succinct Histograms. In: ACM, 2015. 127-135.

[33] Zhan Q, Yin Y, Ting Y, et al. Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. In: Computer and Communications Security.Hofburg Vienna: ACM, 2016. 192-203.

[34] Mehmet Emre G, Acar T, Stacey T, et al. Secure and Utility-Aware Data Collection with Condensed Local Differential Privacy. arXiv preprint arXiv:1905.06361. 2019.

[35] Lee J, Clifton C. Differential Identifiability. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM, 2012. 1041-1049.

[36] Ouyang J, Xiao ZH, Liu SP. Heuristic privacy parameter setting strategy for differential privacy model . Application Research Of Computers. 2019, 36(01): 250-253(in Chinese with English abstract).

[37] Evfimievski A, Gehrke J, Srikant R. Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the 22th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.New York: ACM, 2003. 211-222.

[38] Wang W, Carreira-Perpiñán M Á. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. Mathematics. 2013.

附中文参考文献:

[1] 叶青青,孟小峰,朱敏杰.本地化差分隐私研究综述.软件学报.2018,7(29):159-183.http://www.jos.org.cn/1000-9825/5364.htm.

[15] 欧阳佳,印鉴,刘少鹏,刘玉葆. 一种有效的差分隐私事务数据发布策略.计算机研究与发展,2014,51(10):2195-2205.

[16] 欧阳佳,印鉴,刘少鹏.一种分布式事务数据的差分隐私发布策略.软件学报,2015,26(6):1457-1472.http://www.jos.org.cn/1000-9825/4576.htm

[36] 欧阳佳,肖政宏,刘少鹏,等. 差分隐私模型的启发式隐私参数设置策略.计算机应用研究,2019,36(1):250-253.