doi:10.3969/j.issn.1000-436x.2014.10.023

面向频繁模式挖掘的差分隐私保护研究综述

丁丽萍[1]，卢国庆[1,2]
(1. 中国科学院 软件研究所 基础软件国家工程研究中心，北京 100190；2. 中国科学院大学，北京 100190)

摘 要：频繁模式挖掘是数据挖掘的一个基本问题，其模式本身和相应计数都有可能泄露隐私信息。当前，差分隐私通过添加噪音使数据失真，有效实现了隐私保护的目的。首先介绍了差分隐私保护模型的理论基础；其次，详细综述了差分隐私下 3 种典型的频繁模式挖掘方法的最新研究进展，并进行对比性分析；最后对未来的研究方向进行了展望。

关键词：差分隐私；隐私保护；频繁模式；数据挖掘
中图分类号：TP309.2; TP392 **文献标识码**：A **文章编号**：1000-436X(2014)10-0200-10

Survey of differential privacy in frequent pattern mining

DING Li-ping[1], LU Guo-qing[1,2]
(1. National Engineering Research Center of Fundamental Software, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Frequent pattern mining is an exploratory problem in the field of data mining. However, directly releasing the discovered frequent patterns and the corresponding true supports may reveal the individuals' privacy. The state-of-the-art solution for this problem is differential privacy, which offers a strong degree of privacy protection by adding noise. Firstly, the theoretical basis of differential privacy was introduced. Then, three representative frequent pattern mining methods under differential privacy were summarized and compared in detail. Finally, some future research directions were discussed.

Key words: differential privacy; privacy protection; frequent pattern; data mining

1 引言

频繁模式挖掘(FPM, frequent pattern mining)[1]是数据挖掘研究中的一个重要课题，其目的是找出频繁出现在数据集中的模式(如项集、子序列或子结构)，是关联规则、相关性分析、分类、聚类和其他数据挖掘任务的基础。随着大量数据不断的收集和存储，频繁模式挖掘可以为推荐系统、个性化网站和顾客购买习惯分析等许多应用提供帮助。然而频繁模式本身的内容以及计数信息都有可能泄露用户隐私信息或者披露用户的真实身份。

传统的隐私保护方法大多基于 k-匿名及其扩展分组模型，这些模型普遍存在 2 个主要缺陷：1)需要特殊的背景知识和攻击假设；2)无法提供一种有效且严格的方法来证明其隐私保护水平。此外，新型攻击的出现，如组合攻击、前景知识攻击等，都对上述模型形成了巨大的挑战。

差分隐私(DP, differential privacy)是 Dwork 在 2006 年提出的一种新的基于数据失真的隐私保护模型[2]。该方法能够解决传统隐私保护模型的 2 大缺陷[3]：1)定义了一个相当严格的攻击模型，不关心攻击者拥有多少背景知识，即使攻击者已掌握除某一条记录之外的所有记录信息(即最大背景知识假设)，该记录的隐私也无法被披露；2)对隐私保护水平给出了严谨的定义和量化评估方法。实施差分隐私主要考虑以下 2 个方面的问题[4]：1)设计隐私保护算法满

收稿日期：2014-03-03；**修回日期**：2014-04-20
基金项目：国家科技重大专项基金资助项目(2012ZX01039-004)；中国科学院战略性科技先导专项基金资助项目(XDA06010600)
Foundation Items: The National Science and Technology Major Program of China(2012ZX01039-004); The Strategic Technology Pilot Program of the Chinese Academy of Sciences(XDA06010600)

足差分隐私，以确保不泄露隐私；2) 如何减少数据失真带来的误差，以提高数据可用性。

本文着眼于频繁模式挖掘领域，对差分隐私保护技术最新研究进展和方向进行综述。一方面，对差分隐私的理论基础进行介绍；另一方面，对差分隐私下的频繁模式挖掘方法进行详细阐述，并从关键技术、优缺点、适用范围等方面进行综合对比分析。目前，差分隐私保护技术在频繁模式挖掘领域，主要集中在频繁项集挖掘(FIM, frequent itemset mining)、频繁序列挖掘(FSM, frequent sequence mining)和频繁子图挖掘(FGM, frequent subgraph mining)3 种模式类型，本文着重介绍差分隐私在这 3 种模式类型的应用。

2　差分隐私保护模型

2.1　差分隐私定义

差分隐私[2,5~8]是基于数据失真的隐私保护技术，通过向查询或者分析结果中添加噪音使数据失真，确保在某一数据集中插入或者删除某一条记录的操作不会影响任何查询的输出结果，从而达到隐私保护的目的。差分隐私的形式化定义如下。

定义 1　(ε-差分隐私[2]) 对于所有差别至多为一个记录的 2 个数据集 D_1 和 D_2，给定一个隐私算法 K，Range(K) 表示 K 的取值范围。若算法 K 满足 ε-差分隐私，则对于所有 $S\in \text{Range}(K)$，有

$$\Pr[K(D_1)\in S]\leqslant \exp(\varepsilon)\Pr[K(D_2)\in S] \quad (1)$$

其中，概率 $\Pr[E_s]$表示事件 E_s 的披露风险，即隐私被披露的风险，由算法 K 的随机性所控制。隐私预算 ε 表示隐私保护水平，ε 越小隐私保护程度越高，一般取值为 $\{0.01, 0.1, \ln 2, \ln 3\}$。

2.2　噪音机制

噪音机制是实现差分隐私的主要技术，而不同噪音机制下满足差分隐私的算法所需噪音大小与全局敏感性(global sensitivity)密切相关。

定义 2　(全局敏感性[9]) 对于任意一个函数 $f: D\to R^d$，f 的全局敏感性定义为

$$\Delta f=\max_{D_1,D_2}\left\|f(D_1)-f(D_2)\right\|_1 \quad (2)$$

其中，D_1 和 D_2 差别至多为一个记录，d 表示函数 f 的查询维度，R 表示所映射的实数空间。全局敏感性只是函数 f 的性质，与数据集 D 无关。

常用的噪音机制分别为拉普拉斯机制(Laplace mechanism)[9]与指数机制(exponential mechanism)[10]，其他噪音机制包括高斯机制[11]、几何机制[12]、矩阵机制[13]、函数机制[14]等。

定理 1　(拉普拉斯机制[9])对于任意一个函数 $f: D\to R^d$，若算法 K 的输出结果满足等式(3)，则 K 满足 ε-差分隐私保护。

$$K(D)=f(D)+<Lap_1(\Delta f/\varepsilon),\cdots,Lap_d(\Delta f/\varepsilon)> \quad (3)$$

其中，$Lap_i(\Delta f/\varepsilon)(1\leqslant i\leqslant d)$ 是相互独立的拉普拉斯变量，噪音大小与 Δf 成正比，与 ε 成反比，即函数 f 的全局敏感性越大，所需噪音越大。拉普拉斯机制主要处理一些输出结果为实数型的算法。

由式(3)可知，$K(D)$ 中第 $i(1\leqslant i\leqslant d)$ 个元素拉普拉斯噪音对应的标准差为

$$error_i=\mathrm{E}\left|Lap(\Delta f/\varepsilon)\right|=\frac{\sqrt{2}\Delta f}{\varepsilon} \quad (4)$$

定义 3　(指数机制[10]) 给定一个打分函数 $u: (D\times O)\to R$，若算法 K 满足等式(5)，则 K 满足 ε-差分隐私。

$$K(D,u)=\{r\,|\,\Pr[r\in O]\propto \exp(\frac{\varepsilon u(D,r)}{2\Delta u})\} \quad (5)$$

其中，Δu 为打分函数 $u(D,r)$ 的全局敏感性。指数机制的关键技术是如何设计打分函数 $u(D,r)(r\in O)$，其中，r 表示从输出域 O 中所选择的输出项。由式(5)可知，打分越高，被选择输出的概率越大。指数机制主要处理一些输出结果为非数值型的算法。

2.3　隐私性分析

差分隐私本身蕴含 2 个重要的组合性质[15]：序列组合性(sequential composition)和并行组合性(parallel composition)。

性质 1　(序列组合性[15]) 给定数据库 D，设 K_i 为任意一个随机算法 $(1\leqslant i\leqslant n)$ 满足 ε_i-差分隐私，则 K_i 算法在 D 上的顺序操作满足 $\sum\varepsilon_i$-差分隐私。

性质 2　(并行组合性[15]) 给定数据库 D，设 K_i 为任意一个随机算法 $(1\leqslant i\leqslant n)$ 满足 ε_i-差分隐私，则 K_i 算法在 D 的一系列不相交操作满足 $\max(\varepsilon_i)$-差分隐私。

上述 2 种性质在证明算法是否满足差分隐私以及隐私预算 ε 的合理分配过程中起着重要作用。

1) 隐私保护算法是否满足 ε-差分隐私。通常采用定义 1、性质 1 和性质 2 来证明所设计的隐私保护算法是否满足 ε-差分隐私。

2) 隐私预算 ε-的合理分配。ε 代表着隐私保护水平，一旦 ε 被耗尽，将破坏差分隐私，隐私算法也就失去了意义。隐私预算的分配过程需要考

虑性质 1 和性质 2，常用的分配策略包括平均分配、线性分配、指数分配、自适应分配以及混合分配等。

2.4 可用性度量

满足差分隐私的保护算法在保护隐私的同时，需要兼顾噪音对数据可用性的影响，通常数据可用性通过理论和具体应用 2 个方面来度量。

1) 理论角度。常采用 $(\alpha,\beta)-$userfullness [16]技术来度量差分隐私算法的可用性。

定义 4 ($(\alpha,\beta)-$userfullness [16]) 对于差分隐私算法 K，给定一个操作 Q 和数据集合 D，对于 $\widehat{D}=K(D)$，若满足等式(6)，则算法 K 满足 $(\alpha,\beta)-$userfullness 。

$$\Pr[|Q(\widehat{D})-Q(D)|\leqslant a]>1-\beta \tag{6}$$

2) 具体应用。常用的差度量方法包括：相对误差、绝对误差、欧拉函数以及 F-measure 等。度量方法的选择，需要着重考虑具体数据操作，如计数操作常采用相对误差，top-k 频繁模式挖掘常采用准确率、召回率、F_1-score 等。

3 差分隐私下的频繁模式挖掘方法

频繁模式本身的内容以及计数信息都有可能泄露用户隐私信息或者披露用户的真实身份。如医疗病例信息，可以获得病人所患何种疾病；搜索日志，可以获得用户搜索的行为模式等敏感信息。差分隐私下的频繁模式挖掘方法主要保护频繁模式本身的内容及其计数信息不被披露。

文献[4]指出差分隐私下的数据保护框架通常有 2 种。1) 交互式框架。数据分析者面向原始数据集，执行满足差分隐私的查询算法得到噪音结果。需要关注如何设计满足差分隐私的查询算法，如频繁模式挖掘算法。2) 非交互式框架。数据拥有者发布满足差分隐私的数据集相关信息，数据分析者根据发布的数据集提交查询任务得到噪音结果。需要关注如何设计高效的发布算法，既满足差分隐私又具有较高的数据可用性。

差分隐私的研究工作均是基于以上 2 个框架来展开的，图 1 列出了当前差分隐私下频繁模式挖掘方法的研究进展。本文着重介绍差分隐私下的频繁项集挖掘、频繁序列挖掘和频繁子图挖掘 3 种模式类型。

3.1 频繁项集挖掘

频繁项集挖掘[17]最初应用于事务数据库中发现关联规则，并没有考虑数据记录内项之间的关系，是最简单的频繁模式挖掘类型。Apriori[17]和 FP-growth[18]算法是发现频繁项集的基本算法。

TF[19]方法是差分隐私下 FIM 方法的典型代表。该方法借鉴截断频率(truncated frequency)思想提出了 2 种满足差分隐私的 top-k 频繁项集挖掘策略。TF 方法的基本思路概括为如下 2 步：1)从所有长度不大于 l 的候选项集合 C 中选择 top-k 频繁项集；2)对 k 个项集的真实计数分别添加拉普拉斯噪音。2 种挖掘策略，分别应用指数机制(记为 PT-Exp)和拉普拉斯机制(记为 PT-Lap)实现第 1 步，第 2 步实现相同，具体操作如表 1 所示。

TF 方式的第 2 步相对容易实现，只要对 k 个项集的真实计数添加大小为 $Lap(2k/\varepsilon n)$ 的拉普拉斯

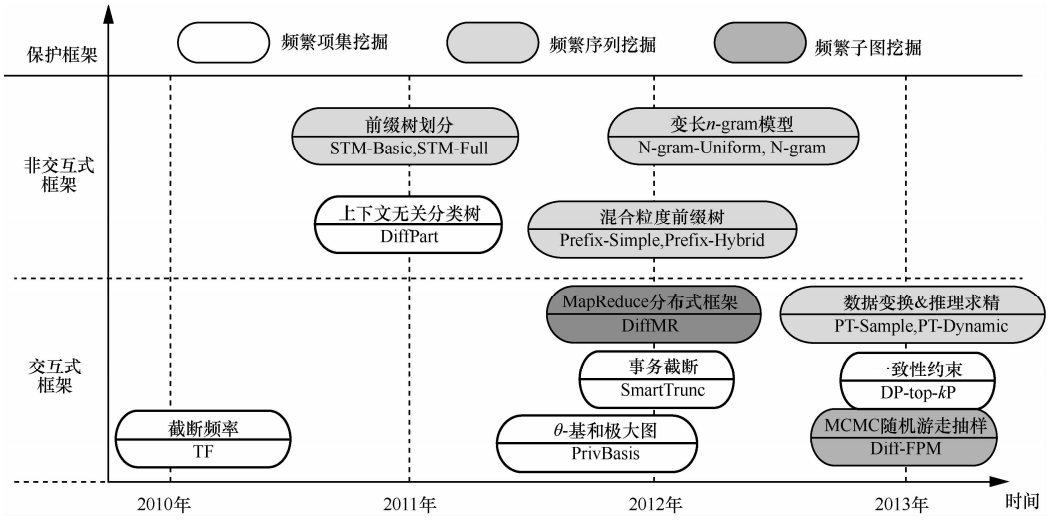

图 1 差分隐私下的频繁模式挖掘方法研究进展

噪音即可。第 1 步实现的关键在于如何从候选项集合 C 中选择 top-k 频繁项集。难点在于集合 C 呈指数规模，即 $|C|\approx|I|^l$，其中，$|I|$ 表示项集域的大小。如果直接遍历 C，计算量非常大且不能保证结果的准确性。TF 应用截断频率技术将集合 C 进行划分处理，即任意项集 p，其截断频率 $\hat{f}(p)=\max(f(p),f_k-\gamma)$，$f_k$ 表示 C 中第 k 大频繁项真实计数，γ 为调节参数。集合 C 划分为

$$C=\begin{cases}S_0, & f>f_k-\gamma\\ S_1, & f\leqslant f_k-\gamma\end{cases} \tag{7}$$

即 C 划分为集合 S_0 和 S_1，操作过程中 S_1 黑盒处理（$|S_1|=1$），S_0 正常处理（$|S_0|$ 且 $|S_0|\ll|C|$），因此集合 C 规模降为 $|S_0|+1\ll|C|$，有效缩小了候选项集范围。此外，TF 方法采取均匀分配策略分配隐私预算，即 ε 均分为二，分别为以上 2 步操作。

表 1　　TF 方法的 2 种挖掘策略比较

步骤	PT-Exp 策略	PT-Lap 策略
1	以概率 $\Pr[p]\propto\exp\left(\frac{\varepsilon n}{4k}\hat{f}(p)\right)$ 从 C 中无放回执行 k 次选择操作得到 top-k 频繁项集	对 C 中所有项集添加大小为 $Lap(4k/\varepsilon n)$ 拉普拉斯噪音后重新获取 top-k 频繁项集
2	对 k 个项集的真实计数分别添加大小为 $Lap(2k/\varepsilon n)$ 拉普拉斯噪音	对 k 个项集的真实计数分别添加大小为 $Lap(2k/\varepsilon n)$ 拉普拉斯噪音

图 2(c)给出了 TF 采用 PT-Exp 策略，即经过指数机制筛选及拉普拉斯噪音添加后生成 l=2，top-3 频繁项集的示例。尽管 TF 应用截断频率技术有效缩减了候选项集合 C 的规模，但由于调节参数满足 $\gamma>4kl\ln(|I|/\varepsilon n)$，随着 k 或者 l 值的增加，可能使得 $f_k-\gamma\leqslant 0$，进而弱化剪枝条件 $\hat{f}(p)=\max(f(p),f_k-\gamma)$ 甚至失效。

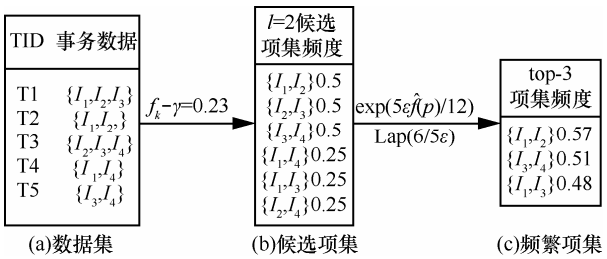

图 2　TF 方法 PT-Exp 策略发布流程

为了弥补 TF 方法的不足，PrivBasis[20]方法结合 θ-基和映射技术实现 top-k 频繁项集挖掘。该方法的基本思路为：1) 数据集 D 中找到所有满足计数不小于 θ 的频繁项 F，即把 D 映射到集合 F 中；2) 基于 F 构建所有 θ-频繁对 P，结合集合 F 和 P 构建 θ-基集合 B；3) 根据集合 B 创建候选频繁项集合 $C(B)$，并对集合 $C(B)$ 中所有项集的真实计数分别添加拉普拉斯噪音。

PrivBasis 方法的关键问题是如何构建 θ-基集合 B。其借鉴极大团(maximal clique)思想提出了一种结合集合 F 和 P 构建集合 B 的方法。该方法把 F 和 P 作为节点和边生成图 $G(F,P)$，找出图 $G(F,P)$ 的所有极大团。每个极大团可视为一个 θ-基，最后合并所有 θ-基生成 θ-基集合 B。图 3 给出了 θ-基集合 B 的一个生成实例，数据集 D 对应的 θ-基集合为 $\{I_1,I_3,I_4\}$，而基于集合 B 可以生成 top-k 频繁项集。例如，l=2，top-3 频繁项集为 $\{I_1,I_3\}$、$\{I_1,I_4\}$ 和 $\{I_3,I_4\}$。

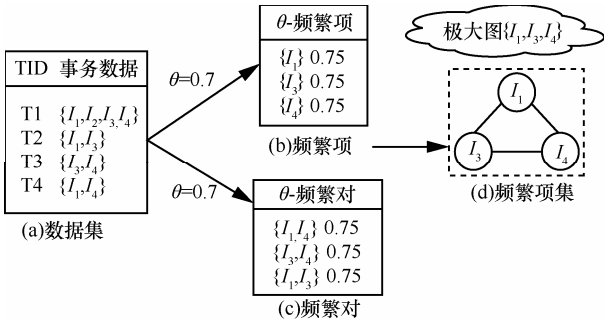

图 3　θ-基集合生成过程

此外，TF 方法输出 top-k 频繁项集对应的噪音计数蕴含有一致性约束：1) k 个项集按照计数降序排列；2) 计数应为整数。若输出结果违背该约束，会导致项集可用性很差。文献[21]提出了 DP-topkP 方法，采用后置处理技术对噪音计数添加求精处理，保证输出结果符合该一致性约束，获得较好的可用性。文献[22]依据长事务记录会导致较高查询全局敏感性的缺陷，提出了一种事务截断的贪婪方法 SmartTrunc，该方法利用阈值和动态权重频率对数据集中每条记录进行局部转换，截断长记录来降低全局敏感性，进而提高项集的可用性。然而，该方法仅适用于分布比较极端的数据集(如数据集包含大量的短事务记录)。

上述几种方法均是交互式框架下的频繁项集挖掘方法，文献[23]提出了一种非交互式框架下的频繁项集挖掘方法 DiffPart。该方法基于上下文无关分类树，结合自顶向下的树划分方法来发布集值

型数据集，支持频繁项集挖掘。该方法的基本思路可以概括为：1)构建上下文无关分类树泛化高维集值型数据集；2) 依据分类树，自顶向下从树根节点开始迭代执行子分割划分并最终生成不同的叶子节点；3)根据叶子节点及其噪音计数重构并发布扰动集值数据集 $\tilde{D}$。由于记录数据之间的依赖性，分割父节点的记录到不同孩子节点过程中，树结构本身可能泄露记录的计数信息。例如，图 4 中 4 条记录 $\{T1,T2,T7,T8\}$ 被随机划分给非叶子节点 v_1，如果直接划分，则计数值 4 就会被泄露。DiffPart 方法采用拉普拉斯机制保护非叶子节点子分割和叶子节点中的计数信息，即扰动节点 v_1 得到噪音计数为 $N_{v_1}=4+Lap(1/\varepsilon_{v_1})$，其中 ε_{v_1} 为节点 v_1 所分得的隐私预算。

DiffPart 方法的关键问题是如何设定非叶子节点的子分割划分条件和叶子节点的发布条件。该方法采用拉普拉斯噪音的标准差作为非叶子节点子分割的分割阈值和叶子节点的发布阈值。对于非叶子节点 v_i，若其噪音计数满足不等式 $c(v_i)\geqslant\sqrt{2}C_2\cdot h(v_i)/\bar{\varepsilon}$，则分割该节点，其中，$C_2$ 为常数，$h(v_i)$ 为 v_i 所在划分树的层数，$\bar{\varepsilon}$ 为节点 v_i 所分得的隐私预算。而对于叶子节点 v_i，若其噪音计数满足 $c(v_i)\geqslant\sqrt{2}C_1/(\varepsilon/2+\bar{\varepsilon})$，则发布该节点，其中，$C_1\in[1,C_2]$ 为常数，$\bar{\varepsilon}$ 为非叶子节点子分割划分剩余隐私预算。

为了合理地分配隐私预算，DiffPart 方法提出了一种自适应分配策略。即给定预算 ε 被均分为二，其中，$\varepsilon/2$ 分给非叶子节点子分割过程，$\varepsilon/2$ 用来叶子节点发布。根节点到叶子节点被称为一条分割链，分割链上的隐私预算分配满足序列组合特性，而分割链间满足并行组合特性。在每条分割链上，分割过程中剩余的预算均被追加到叶子节点。因此，每条分割链上的隐私预算 ε 分配如下

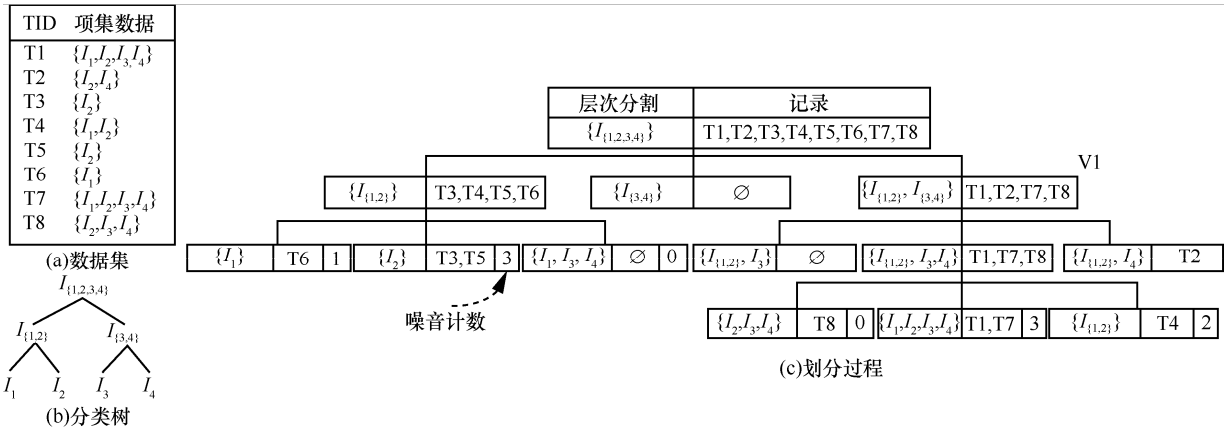

图 4　DiffPart 方法分类树结构

表 2　　**差分隐私下频繁项集挖掘方法对比分析**

方法名称	主要技术	主要优点	主要缺点	ε 分配	算法性能
TF	截断频率	实现简单，精度高	处理较大 k 值或 l 时性能与效率比较差；未考虑记录本身长度带来的影响	指数机制、拉普拉斯机制；ε 均分为 2 份，平均分配	$O(k'\|D\|)$
PrivBasis	θ-基映射	性能和效率优于 TF	难以兼顾隐私保护与模式可用性的不足；未考虑记录本身长度带来的影响	拉普拉斯机制；ε 平均分配	$O(w\|D\|)$
DP-topkP	一致性约束	一致性约束输出结果，精度较高	处理较大 k 值或 l 性能与效率比较差；未考虑记录本身长度带来的影响	指数机制、拉普拉斯机制；ε 被划分为 2 份，平均分配	$O(k'\|D\|)$
SmartTrunc	事务截断	规约数据集，精度高	扩展性比较差	拉普拉斯机制；ε 平均分配	$O(\|D\|\|L\|)$
DiffPart	分类树划分	发布精度高，扩展性高	仅支持计数查询；没有考虑不同项之间的语义关联	拉普拉斯机制；ε 均分为 2 份，自适应分配	$O(\|D\|\|L\|)$
DiffMR	一致性约束	支持 MapReduce 分布式查询，实现简单	只支持简单计数查询；处理较大 k 值或 l 时性能与效率比较差	指数机制、拉普拉斯机制；ε 线性分配	$O(k'\|D\|)$

$$\varepsilon \geqslant \underbrace{\left(\frac{\varepsilon}{2}\frac{1}{n_1}\right)}_{\text{第一次划分}} + \underbrace{\left(\frac{\varepsilon}{2}\left(1-\frac{1}{n_1}\right)\frac{1}{n_2}\right)}_{\text{第二次划分}} + \cdots + \underbrace{\left(\frac{\varepsilon}{2}\prod_{i=1}^{m-1}\left(1-\frac{1}{n_i}\right)\frac{1}{n_m}\right)}_{\text{第一次划分}} + \frac{\varepsilon}{2} \tag{8}$$

其中，m 表示一条分割链上所有节点，n_i 表示第 i 层上某子分割到叶子节点所需的最大划分步数 $(n_i \geqslant n_{i+1}, n_m = 1)$。差分隐私下频繁项集挖掘方法对比分析详见表 2[19~24]。

3.2　频繁序列挖掘

频繁序列挖掘[25]是指挖掘相对时间或其他顺序出现频率较高的模式。根据模式的特征，序列数据可以分成时间序列(如股票交易数据、交通轨迹数据)、符号序列(如顾客购买序列、Web 点击流)和生物学序列(如 DNA、蛋白质序列)。

FSM 与 FIM 的区别在于后者描述的是事务内部的关联模式，如在一次购物中所购买物品之间的关联关系。而前者则描述事务之间的关联模式，如同一顾客在多次购物中所购物品之间可能存在的某种关联关系。显然，频繁序列集合是频繁项集的超集，因此如何有效地缩小候选序列集合是实现差分隐私过程中重点关注的问题。

文献[26]借鉴 DiffPart[23]方法的树划分思想，第一次提出了结合自顶向下前缀树(prefix tree)发布轨迹数据集的方法 STM-Full，支持频繁序列挖掘。该方法的基本思路可以概括为：1)对原始轨迹数据集执行一系列计数查询，添加噪音构建扰动前缀树 PT；2) 发布重构轨迹数据集 $\tilde{D}$。

前缀树本身蕴含有如下一致性约束：1) 对任意一条根节点到叶子节点的路径 p，满足 $\forall v_i \in p, |\mathrm{tr}(v_i)| \leqslant |\mathrm{tr}(v_{i+1})|$，其中，$\mathrm{tr}(v_i)$ 表示节点 v_i 计数，v_i 是 v_{i+1} 的孩子节点；2) $\forall v, |\mathrm{tr}(v)| \geqslant \sum_{u \in \mathrm{children}(v)} |\mathrm{tr}(u)|$。若直接基于扰动前缀树重构轨迹数据集（记为 STM-basic），有可能违背一致性约束导致序列的可用性很差。STM-Full 方法提出了一种一致性约束推理策略，通过评估扰动前缀树中所有节点(不包括根节点)对应的噪音计数来满足一致性约束，增强发布精度。

此外，每层子分割是根据分类树结构随机产生的，因此会生成大量计数为 0 的空节点，而这些空节点不但消耗隐私预算，还会影响最终的发布精度。为了限制产生过多的空节点，STM-Full 方法应用独立的布尔测试和二项分布，设计了一种面向拉普拉斯机制的统计抽样过程用于生成 k 个空节点。设 m 为某一子分割生成的空节点总数，抽取 k 个空节点的操作记为二项分布 $B(m, p_\theta)$，其中，$p_\theta = \exp(-\bar{\varepsilon}\theta)/2$ 表示取到空节点且噪音计数大于阈值 θ 的概率，其分布函数为

$$p(x) = \begin{cases} 0, & \forall x < 0 \\ 1 - \exp(\bar{\varepsilon}\theta - \bar{\varepsilon}x), & \forall x \geqslant 0 \end{cases} \tag{9}$$

其中，$\bar{\varepsilon}$ 表示选择操作所分得的隐私预算，阈值 $\theta = 2\sqrt{2}/\bar{\varepsilon}$，即为拉普拉斯噪音标准差的 2 倍。

由于 STM-Full 方法只发布了轨迹的位置信息，而忽视了轨迹中每个位置所携带的时间戳，导致发布的序列可用性低。同时，随着前缀树长度的增长，划分到每一子分割的序列数量会急剧减小，严重影响发布序列的可用性。

为了弥补 STM-Full 方法的不足，文献[27]提出了一种基于变长 n-gram 模型发布序列数据集的方法 N-gram。该方法的基本思路为：1) 抽取原始序列数据集 D 中所有变长的 $n_{\max}-\mathrm{gram}$ ($n_{\max}$ 用来限制 n-gram 长度)及其对应的噪音计数，构建扰动前缀树 T；2) 依据一致性约束优化 T 中所有 n-gram；3) 采用马尔科夫假设(Markov assumption)迭代计算较短的 n-gram 获得所有可得的较长 n-gram，并更新前缀树 T 发布重构序列数据集 $\tilde{D}$。其中，3)中更新后的前缀树 T 可直接用于频繁序列挖掘，而重构序列数据集 $\tilde{D}$ 可用于各种类型的查询任务，如计数查询。

第 1)步对应前缀树 T 的构建过程直接操作原始序列数据集，需要考虑隐私保护问题即隐私预算的分配，而后两步均是基于扰动前缀树 T 执行的相应操作，故无需考虑隐私保护问题。因此在前缀树构建过程中，需要考虑如何合理分配隐私预算。若采用平均分配策略，即给定隐私预算 ε 均分为 $n_{\max}$ 份，分配给前缀树每一层(记为 N-gram-uniform)，这种会存在较大的隐私预算浪费。如图 5 中树的深度是 3 但 $n_{\max} = 5$，至少存在 $2\varepsilon/5$ 隐私预算浪费。为了合理地分配隐私预算，N-gram 提出了一种自适应分配策略，根据已知树节点的噪音计数，应用马尔科夫假设估计根节点到叶子节点路径的长度，依据估计长度分配剩余的隐私预算 $\bar{\varepsilon}$。某节点 v 可行的隐私预算分配按照式(10)计算得到，其中，$P_{\max}$ 表示父节点 v 划分得到孩子节点的最大概率，$c(v)$ 为节点 v 对应的噪音计数。

$$\varepsilon_v = \frac{\bar{\varepsilon}}{\min(\log_{P_{\max}} \frac{\theta}{c(v)}, n_{\max} - i)} \tag{10}$$

例如图 5 中树的第一层分得 $\varepsilon/5$，对于节点 v_1 计算得到 $P_{\max}=\frac{10}{4+10+9}=0.43$，$h_{v_1}=1$，第二层分得 $\varepsilon_v=\frac{\varepsilon-\varepsilon/5}{1}=4\varepsilon/5$，隐私预算 ε 耗尽故划分终止。

N-gram 方法的关键问题在于变长 $n_{\max}$ -gram 前缀树 T 的构建。其结合自顶向下的分类树进行划分，且需要同时满足如下 3 个条件：1)树的深度小于 $n_{\max}$；2) 噪音计数不小于阈值 θ；3)隐私预算 ε 未耗尽。图 5 给出了一个前缀树 T 的生成实例，前缀树满足 $n_{\max}$=5，θ=3，其中 v_1 是隐私预算 ε 耗尽划分终止，v_2、v_3 皆是噪音计数小于 θ 划分终止。

此外，文献[28~30]采用 $(\alpha,\beta)-$userfullness 技术进一步分析 STM-Full[26]方法中的扰动前缀树 PT，计算得到执行频繁前缀序列挖掘中，某节点 v_1 噪音计数对应误差 α 满足 $O(1/\varepsilon_i\log 1/\beta)$，而频繁序列挖掘中 v_1 对应误差 α 满足 $O(\sqrt{\sum_{i=0}^{n-1}1/\varepsilon_i^2}\log 1/\beta)$，其中，$\varepsilon_i$ 为节点 v_1 所分得的隐私预算。容易发现频繁序列挖掘输出误差较大，导致所发布的序列可用性较低。为了有效提高序列可用性，文献[29]提出了一种两阶段方法 PT-Sample[29]，如图 6 所示。1) 构建噪音扰动后的前缀树 PT，可直接用于频繁前缀序列挖掘；2)对原始序列数据集执行分箱离散化，输出添加拉普拉斯噪音的 top-k 频繁序列。

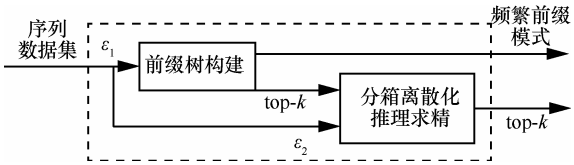

图 6　PT-Sample 方法流程

PT-Sample 方法最大贡献在于第 2 阶段对原始序列数据集执行的数据变换操作。该方法以字符串序列数据集为例，采用字符串指纹匹配方法对原始数据集执行分箱离散化，其中指纹库来自于前缀树 PT 遍历得到的候选 top-k' 频繁序列，极大地降低候选频繁序列集合的规模，有效提高了 top-k 频繁序列的准确性。

差分隐私下频繁序列挖掘方法对比分析详见表 3[26,27,29,31]。

3.3　频繁子图挖掘

频繁子图挖掘[32]在许多实际应用中具有巨大

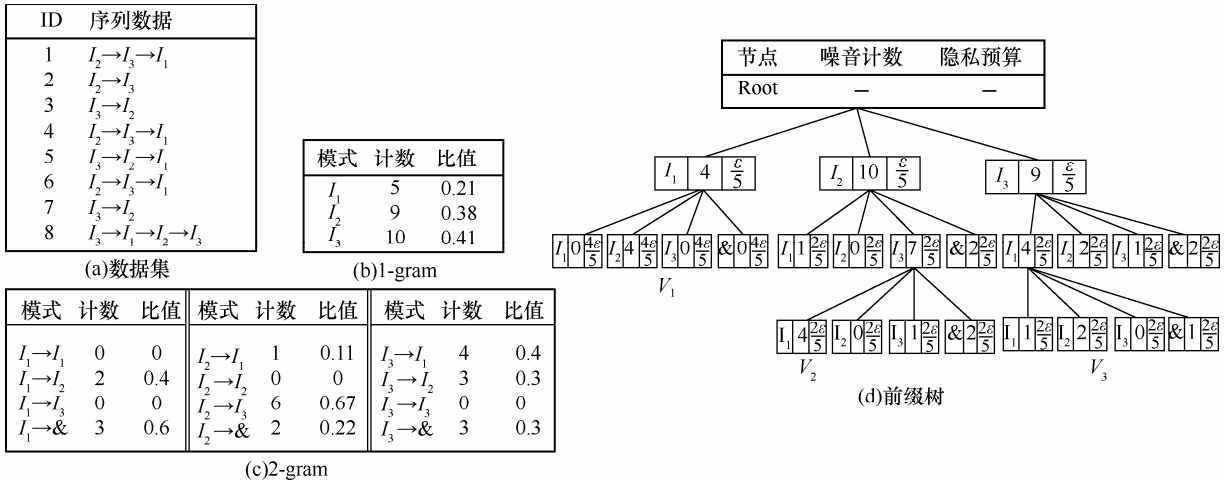

图 5　N-gram 方法前缀树结构

表 3　　**差分隐私下频繁序列挖掘方法对比分析**

方法名称	主要技术	主要优点	主要缺点	ε 分配	算法性能
STM-Full	前缀树划分	支持多维数据依赖计数查询；精度高	忽视轨迹时间戳信息，实用性差；前缀树冗余	拉普拉斯机制；ε 线性分配	$O(\lvert D\rVert L\rvert)$
Prefix-Hybrid	混合粒度前缀树划分	支持多维数据依赖计数查询；精度高	忽视轨迹时间戳信息，实用性差；前缀树冗余	拉普拉斯机制；ε 划分为两份，混合分配	$O(\lvert D\rVert L\rvert)$
N-gram	变长 *N*-gram 模型	支持多维数据依赖计数查询；精度高	容易受到序列维度影响；有一定信息丢失	拉普拉斯机制；ε 自适应分配	$O(\lvert D\rVert L\rvert)$
PT-Sample	数据变换和推理求精	支持频繁前缀树和序列模式挖掘；精度高	数据变换过程中存在信息丢失	拉普拉斯机制；ε 划分为两份，混合分配	$O(\lvert D\rVert L\rvert^{h_{\max}+1})$

的潜在价值，如语义网络、行为建模、生物网络分析、化学化合物分类和大分子分析。

给定图 $G=(V,E)$，$G_s=(V_s,E_s)$，当且仅当 $V_s\subseteq V$ 且 $E_s\subseteq E$，称图 G_s 为图 G 的子图。如果 2 个图 $G_1=(V_1,E_1)$ 和 $G_2=(V_2,E_2)$ 具有相同的拓扑结构，则称这 2 个图是同构的，即存在一个 V_1 到 V_2 的映射关系，使得 E_1 中的每一条边唯一对应 E_2 中的一条边，反之亦然。子图同构问题是指判断 G_1 是否存在一个子图和 G_2 同构，也就是说判断 G_2 是否为 G_1 所包含。由于子图同构问题是 NPC 问题，使得子图计数操作非常困难，进而导致在频繁子图挖掘过程中输出域（候选频繁子图集）无法直接得到。因此，如何快速有效地构建输出域是差分隐私下的频繁子图挖掘亟需要解决的问题。

文献[33]借鉴 TF[19]方法 PT-Exp 策略，结合含随机游走的蒙特卡洛(MCMC)抽样方法，提出了一种满足差分隐私的 top-k 频繁子图挖掘方法 Diff-FPM。该方法的基本思路为：1) MCMC 方法与指数机制相结合，执行 k 次随机游走选择 top-k 频繁子图；2) 对 k 个子图的真实计数添加拉普拉斯噪音。

Diff-FPM 方法的关键问题在于如何合理的定义和实施 MCMC 随机游走，建立一条收敛的马尔科夫链，获得满足差分隐私的 top-k 频繁子图。其中，MCMC 方法的基本思想是通过建立一个平稳分布为 $\pi(x)$ 的马尔科夫链来得到样本。该方法选择 MH(metropolis-hastings)算法执行随机游走，需要事先确定马尔科夫链对应的状态空间、转移概率和平稳分布。1) 状态空间定义为偏序完整图(POFG)，状态空间元素记为 POFG 中所有节点唯一表示的子图，边唯一表示一个子图到邻近子图的转换；2) 平稳分布，定义为

$$\pi(x)=\frac{\exp\left(\varepsilon_1 u(x)/2\Delta u\right)}{\sum_{x\in X}\exp\left(\varepsilon_1 u(x)/2\Delta u\right)} \tag{11}$$

由选择某一子图为样本输出的概率和指数机制相结合得到，其中 x 表示输出域，$u(x)$ 为打分函数；3)转移概率的定义，Diff-FPM 采用启发式计算，通过添加可调参数区别频繁子图和非频繁子图。设 q_{xy} 为状态 x 下一步跳转的目的子图为 y 的概率，后经过转移概率

$$a_{xy}=\min\left(\frac{\exp\left(\varepsilon_1 u(y)/2\Delta u\right)q_{yx}}{\exp\left(\varepsilon_1 u(x)/2\Delta u\right)q_{xy}},1\right) \tag{12}$$

计算确定是否执行该跳转。例如，当前状态为 *A-A-D*，简单假设 $q_{xy}=1/N(x)$，其中 $N(x)$ 表示节点 x 邻接子图总数。设 *A-D* 被选为下一步跳转的目的子图，邻接子图总数 $N(x)=5$，$N(y)=10$，则转移频率 $a_{xy}=\min\left(\frac{\exp(3/2)\cdot(1/10)}{\exp(2/2)\cdot(1/5)}\right)=0.82$，即状态 *A-A-D* 有 82%的概率跳转到 *A-D*。

Diff-FPM 方法实施前提是 MCMC 随机游走建立马尔科夫链极的极限分布渐近于平稳分布 $\pi(x)$，故存在偏离平稳分布的情况，此时只能满足 $(\varepsilon,\delta)-$差分隐私[34]，可用性也会受到影响。因此如何高效地兼顾隐私保护和数据可用性是未来亟需要解决的问题。

4　下一步工作

4.1　差分隐私下的数据挖掘技术研究

数据挖掘的目的是从大量的数据中抽取或者学习到有价值的模型或者规则。由于敏感信息隐含在模型或者规则中，隐私保护的数据挖掘技术值得广泛关注，诸如分类、聚类等数据挖掘关键技术与差分隐私保护模型的有效结合是需要研究的问题。

1) 分类技术在数据预测分析中起着关键作用，该技术的目的是找出描述和区分数据类或概念的模型，而分类模型的典型代表是决策树。已有方法 SuLQ-based ID3[5]、DiffP-*C*4.5[35]和 DiffGen[36]均采用了信息增益(information gain)来选择分割属性，并递归地构建满足差分隐私的决策树。但当数据集分类属性较多时，均存在分类精度降低、效率降低或者隐私预算耗尽的风险。因此，如何对具有高维度分类属性的数据集进行分类，以及如何设计有效的隐私预算分配策略是未来的研究方向。

2) 聚类同样是数据分析的主要技术，是把数据对象划分成多个簇的过程。而在聚类过程中数据隐私同样有可能泄露，例如，均值(means)、中心点(center)与中值(median)等。已有 Pk-means[37,38]和 Pk-median 方法[39]，以及 GUPT[40]数据分析系统，均结合抽样与聚集技术输出满足差分隐私的聚类结果，但实际应用型比较差。当数据集合很大时，k 值的选择是 NP 问题，选择 k 值操作有可能泄露真实的数据点，并且每次选择均要消耗隐私预算。因此，如何利用指数机制挑选上述 2 种方法的 k 值是未来的研究方向。

4.2 差分隐私下的位置隐私保护

移动通信和传感设备等位置感知技术的发展，为人们的生活、商业运作以及科学研究带来了巨大收益。然而位置数据存在泄露个人信息的风险，这是因为位置数据既直接包含用户的隐私信息，又隐含了用户的个性习惯、健康状况、社会地位等其他敏感信息。

尽管目前为止差分隐私的研究已经涉及到位置数据有关的用户位置模式挖掘 PDBSCAN[41,42]，但主要针对离线数据，不能直接应用于位置隐私保护。同时由于差分隐私对于攻击者背景知识的假设十分保守，而大数据时代攻击者获取背景知识的渠道和途径又十分广泛，将差分隐私应用到位置隐私保护中，是未来有潜力的研究方向。

4.3 差分隐私下的大数据分析研究

大数据是当前学术界和产业界的研究热点，但目前大数据在收集分析过程中面临着诸多安全风险，隐私问题是人们公认的关键问题之一，差分隐私下的大数据隐私保护值得未来深入研究。

1) 数据采集与预处理。当今现实世界的数据库极易受噪声、缺失值和不一致数据的侵扰，噪声数据将导致低质量的挖掘结果。当前差分隐私下的频繁模式挖掘技术研究，都是在理想数据的环境下进行的，文献[30]提到针对噪声数据模式挖掘的想法，但还未出现相应研究成果。因此，如何实施有效的数据清理进而提高数据可用性是未来的研究方向。同时数据集成、规约和变换技术的有效利用，会极大地降低候选频繁模式集进而保证较好的数据可用性，也是未来的研究方向。

2) 数据分析。在计算架构方面，MapReduce 等并发处理结架构得到广泛应用。Diff-FPM[24]方法支持 MapReduce 框架下满足差分隐私的频繁项集挖掘，但只是面对简单事务数据。Airavat[43]聚集分析系统虽然实现差分隐私，但仅提供聚集分析结果，不支持基于数据挖掘和机器学习的任务分析以及请求过大的任务分析，因此并发处理结构下的差分隐私应用是未来的研究方向；在查询和索引方面，NoSQL 半结构化或非结构化数据得到更多关注，差分隐私的相关应用也值得关注；在数据分析和处理方面，结合差分隐私的大数据模式挖掘、回归分析、个性化推荐等是未来的研究方向。

5 结束语

本文对差分隐私下频繁模式挖掘方法的研究进展进行了阐述和分析，虽然已经取得了一定的进展，但针对已有方法的不足，还有许多优化工作值得深入开展。同时差分隐私作为一种新的隐私保护模型，拥有相当广阔的研究空间，有很多挑战性的问题需要解决。

参考文献：

[1] HAN J, KAMBER M, PEI J. Data Mining: Concepts and Techniques[M]. Morgan kaufmann, 2006.

[2] DWORK C. Differential privacy[A]. Proc of the 33rd International Colloquium on Automata, Languages and Programming[C]. Berlin: Spinger-Verlag,2006.

[3] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014,37(1):101:122.
XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014,37(1): 101-122.

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

[5] BLUM A, DWORK C, McSHERRY F, *et al*. Practical privacy: the SuLQ framework[A].Proc of the 24th ACM SIGMOD International Conference on Management of Data/Principles of Database Systems[C]. NewYork: ACM Press, 2005.128-138.

[6] BLUM A, LIGETT K, ROTH A. A Learning theory approach to non-interactive database privacy[A]. Proc of the 40th Annual ACM Symposium on Theory of Computing(STOC)[C]. Victoria, British Columbia, Canada, 2008.351-360.

[7] DWORK C. A firm foundation for private data analysis[J]. Communications of the ACM, 2011, 54(1):86-95.

[8] NGUYEN H H, KIM J, KIM Y. Differential privacy in practice[J]. Journal of Computing Science and Engineering, 2013, 7(3): 177-186.

[9] DWORK C, MCSHERRY F, NISSIM K, *et al*. Calibrating noise to sensitivity in private data analysis[A]. TCC[C]. 2006.265-284.

[10] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[A]. FOCS[C]. 2007.94-103.

[11] DWORK C, NAOR M, VADHAN S. The privacy of the analyst and the power of the state[A]. 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)[C]. 2012. 400-409.

[12] GHOSH A, ROUGHGARDEN T, SUNDARARAJAN M. Universally utility-maximizing privacy mechanisms[A]. Proceedings of the 41st Annual ACM Symposium on Theory of Computing[C].ACM, 2009. 351-360.

[13] LI C, HAY M, RASTOGI V, *et al*. Optimizing linear counting queries under differential privacy[A]. Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems[C]. 2010. 123-134.

[14] ZHANG J, ZHANG Z, XIAO X, *et al*. Functional mechanism: regression analysis under differential privacy[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1364-1375.

[15] MCSHERRY F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis WWW[A]. SIGMOD Conference[C].

2009. 19-30.

[16] KIFER D, MACHANAVAJJHALA A. No free lunch in data privacy[A]. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data[C]. 2011.193-204.

[17] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[A]. Proc 20th Int Conf Very Large Data Bases[C]. 1994. 487-499.

[18] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation[J]. ACM SIGMOD Record, 2000, 29(2):1-12.

[19] BHASKAR R, LAXMAN S, SIMTH A, *et al*. Discovering frequent patterns in sensitive data[A]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)[C]. Washington, DC, USA, 2010. 503-512.

[20] LI N, QARDAJI W, SU D, *et al*. Privbasis: frequent itemset mining with differential privacy[A]. Proceedings of the 38th Conference of Very Large Databases (VLDB)[C]. Istanbul, Turkey, 2012. 1340-1351.

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

[22] ZENG C, NAUGHTON J, CAI J Y: On differentially private frequent itemset mining[J]. PVLDB, 2012,6(1):25-36.

[23] CHEN R, MOHAMMED N, FUNG B C M, *et al*. Publishing set-valued data via differential privacy[A]. Proceedings of the 37th Conference of Very Large Databases(VLDB)[C]. Seattle, Washington, USA, 2011.1087-1098.

[24] HAN X, WANG M, ZHANG X, *et al*. Differentially private top-*k* query over MapReduce[A]. Proceedings of the 4th International Workshop on Cloud Data Management[C]. 2012.25-32.

[25] AGRAWAL R, SRIKANT R. Mining sequential patterns[A]. Proceedings of the 11th International Conference on Data Engineering IEEE[C]. 1995.3-14.

[26] CHEN R, FUNG B, DESAI B C. Differentially private trajectory data publication[EB/OL]. http://arxiv.org/abs/1112.2020.

[27] CHEN R, ÁCS G, CASTELLUCCIA C. Differentially private sequential data publication via variable-length *n*-grams[A]. CCS[C]. 2012. 638-649.

[28] BONOMI L, XIONG L, CHEN R, *et al*. Frequent grams based embedding for privacy preserving record linkage[A]. CIKM[C]. 2012. 1597-1601.

[29] BONOMI L, XIONG L. A two-phase algorithm for mining sequential patterns with differential privacy[A]. CIKM[C]. 2013.269-278.

[30] BONOMI L. Mining frequent patterns with differential privacy[J]. PVLDB, 2013, 6(12): 1422-1427.

[31] CHEN R, FUNG B, DESAI B C, *et al*. Differentially private transit data publication: a case study on the montreal transportation system[A]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. ACM, 2012: 213-221.

[32] HAN J, CHENG H, XIN D, *et al*. Frequent pattern mining: current status and future directions[J]. Data Mining and Knowledge Discovery, 2007, 15(1): 55-86.

[33] SHEN ET, YU T. Mining frequent graph patterns with differential privacy[A]. KDD[C]. 2013. 545-553.

[34] DWORK C, KENTHAPADI K, MCSHERRY F, *et al*. Our Data, Ourselves: Privacy Via Distributed Noise Generation[M]. Springer Berlin Heidelberg, 2006.486-503.

[35] FRIEDMAN A, SCHUSTER A. Data mining with differential privacy[A]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. ACM, 2010.493-502.

[36] MOHAMMED N, CHEN R, FUNG B, *et al*. Differentially private data release for data mining[A]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. 2011.493-501.

[37] NISSIM K, RASKHODNIKOVA S, SMITH A. Smooth sensitivity and sampling in private data analysis[A]. Proceedings of the 39th Annual ACM Symposium on Theory of Computing[C]. 2007. 75-84.

[38] 李杨, 郝志峰, 温雯. 差分隐私保护 *k*-means 聚类方法研究[J]. 计算机科学, 2013, 40(3): 287-290.
LI Y, HAO ZF, WEN W. Research on differential privacy preserving *k*-means clustering[J]. Computer Science, 2013, 40(3): 287-290.

[39] GUPTA A, LIGETT K, MCSHERRY F, *et al*. Differentially private combinatorial optimization[A]. Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics[C]. 2010.1106-1125.

[40] MOHAN P, THAKURTA A, SHI E, *et al*. GUPT: privacy preserving data analysis made easy[A]. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data[C]. 2012.349-360.

[41] HO S S, RUAN S. Differential privacy for location pattern mining[A]. Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS {SPRINGL)[C]. Chicago, IL, USA, 2011.17-24.

[42] HO S S. Preserving privacy for moving objects data mining[A]. 2012 IEEE International Conference on Intelligence and Security Informatics (ISI) [C]. 2012. 135-137.

[43] ROY I, SETTY S T V, KILZER A, *et al*. Airavat: security and privacy for MapReduce[A]. Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation[C]. USENIX Association, 2010. 20-20.

作者简介：

丁丽萍（1965-），女，山东青州人，中国科学院软件研究所研究员、博士生导师，主要研究方向为数字取证、系统安全与可信计算。

卢国庆（1989-），男，山东章丘人，中国科学院软件研究所硕士生，主要研究方向为差分隐私保护、数据挖掘。