基于位置服务的分布式差分隐私推荐方法研究

郑孝遥[1,2],罗永龙[1,2],汪祥舜[1,2],孙丽萍[1,2],陈付龙[1,2],胡桂银[1,2],汪小寒[1,2]
(1. 安徽师范大学计算机与信息学院,安徽芜湖 241002; 2. 网络与信息安全安徽省重点实验室,安徽芜湖 241002)

摘 要: 随着移动互联网技术的迅速发展,传统的推荐系统已不能很好地适应基于位置的推荐服务,同时也面临隐私泄露的问题. 本文针对上述问题,首先提出一种分布式隐私保护推荐框架,并利用差分隐私保护理论,设计基于分布式框架的奇异值分解推荐算法,同时利用保序加密函数实现用户请求位置的保护. 理论分析和在两个真实的数据集上的实验表明,本文提出的方法不仅具有较强隐私保护能力,同时相比传统的几种推荐算法,也具有较好的推荐性能.

关键词: 推荐系统; 分布式框架; 位置服务; 隐私保护; 保序加密
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112 (2021)01-0099-12
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20180835

Research on Location-Based Distributed Differential Privacy Recommendation Method

ZHENG Xiao-yao[1,2], LUO Yong-long[1,2], WANG Xiang-shun[1,2], SUN Li-ping[1,2],
CHEN Fu-long[1,2], HU Gui-ying[1,2], WANG Xiao-han[1,2]
(1. *School of Computer and Information, Anhui Normal University, Wuhu, Anhui* 241002, *China*;
2. *Anhui Provincial Key Laboratory of Network and Information Security, Wuhu, Anhui* 241002, *China*)

Abstract: With the rapid development of mobile Internet technology, the traditional recommender system has not been well adapted to the location-based recommendation service, and it also faces the risk of privacy leaks. In this paper, a distributed privacy preserving recommendation framework is proposed, and a singular value decomposition recommendation algorithm based on distributed framework is designed by using the differential privacy theory. Furthermore, we use order preserving encryption function to protect user request location. Theoretical analysis and experiments on two real datasets show that the proposed method not only has stronger privacy protection ability, but also has better recommendation performance than traditional recommendation algorithms.

Key words: recommender system; distributed framework; location-based service; privacy preserving; order preserving encryption

1 引言

传统的推荐系统主要基于用户与项目的二维推荐,目前已经得到了广泛的应用,如在电子商务、新闻传播、在线零售等领域. 随着移动互联网和智能终端技术的迅速发展,基于位置服务(Loacation-Based Service, LBS)得到了普及,基于 LBS 的推荐技术开始引起学者和业界的关注. 目前移动用户可以使用智能终端中的 GPS 技术,感知自己的地理位置,同时通过向 LBS 服务提供商发送自己的位置信息,向服务提供商请求个性化的服务,最常见的有兴趣点推荐、地图导航等.

用户在请求基于 LBS 的推荐服务时,需要向服务提供商提供自己的位置信息,同时服务提供商会根据用户的历史消费记录,计算用户的偏好,从众多的项目中推荐出用户潜在感兴趣的,符合用户位置约束需求的项目. 相对于传统推荐服务,用户不仅要提供项目评分,还需要向推荐服务器提供自身的地理位置,在此过程中,用户面临着两个隐私泄露的威胁:一是用户地理

收稿日期:2018-09-27;修回日期:2020-01-07;责任编辑:王天慧
基金项目:国家自然科学基金(No. 61672039, No. 61972439, No. 61602009);安徽省自然科学基金(No. 1808085MF172, No. 1908085MF190);高校优秀青年人才支持计划重点项目(No. gxyqZD2019010)

位置隐私泄露问题;二是用户偏好信息泄露的风险. 假设某游客现在正在北京某景点游玩,其首先通过团购网站在景点周围预订晚餐,网站平台获得其所在位置,然后根据用户饮食偏好向其推荐餐厅,若游客选择清真食品消费并评分,商家和平台可以根据游客的消费推测出该游客的宗教信仰以及所处的位置. 若数据泄露给第三方,极易导致用户的位置和用户偏好隐私泄露.

目前,面向位置服务推荐系统的隐私保护方法主要分为泛化、数据扰动和加密三种类型. 文献[1,2]针对用户的位置信息进行泛化处理,采用 k 匿名方法将用户的位置信息隐藏于 k 个相近的同类用户中. 但 k 匿名方法没有对攻击者掌握的背景知识进行严格的定义,从而存在面对新型攻击时安全性较低的问题;文献[3,4]提出了数据扰动方法,将用户的历史数据进行一定程度的干扰后再将其发送给推荐服务器,从而保证对用户数据进行保护. 虽然数据扰动方法简单,但也存在保护能力不足的问题. 文献[5,6]使用同态加密方法对协同过滤推荐过程中的近邻相似度进行计算,从而保护了用户的兴趣隐私. 但同态加密算法也存在计算复杂度高,在大规模数据集中应用推荐效率低的问题. 上述隐私保护方法都是单一的针对用户位置隐私或者用户偏好隐私,同时兼顾位置和用户偏好的隐私保护的推荐方法目前涉及的研究很少. 因此本文主要针对上述问题,改进推荐系统体系架构,提出一种分布式用户隐私保护推荐框架,并利用差分隐私保护技术实现用户偏好的保护,同时利用保序加密函数实现对用户位置的保护. 最后通过两个真实数据集上实验表明,本文所提的方法在推荐准确性和安全性上具有较好的性能.

2 相关工作

2.1 推荐系统体系结构

推荐系统的体系结构主要研究用户的信息收集、兴趣偏好描述、推荐算法的执行等在系统所处的位置[7]. 目前,个性化推荐系统绝大多数都是基于服务器的,用户信息的收集、建模以及用户描述文件都集中存放在服务器端[8]. 基于服务器端的推荐系统除了增加服务器的开销并受到其功能限制,对用户的隐私存在极大的威胁. 鉴于此,一些研究者对推荐系统的体系结构提出了很多改进方法,如基于 P2P 的推荐系统[9]、基于代理的推荐系统[10,11]以及基于客户端的推荐系统[12,13]等,主要目的是将集中式的结构变为分布式的体系结构,但同时也增加了时间和通信开销. 上述的推荐系统结构主要针对传统的推荐服务,随着移动互联网技术的发展,如何构建具有位置服务的推荐系统架构有待进一步的研究. 因此有必要对传统的推荐系统结构进行改进,使其适应移动社交网络的特点,并能够保护用户隐私信息.

2.2 差分隐私

个性化推荐面临的一个重要挑战就是隐私泄露问题,目前具有隐私保护功能的推荐研究相对较少. 具有隐私保护功能的推荐技术主要分为两大类:基于安全多方计算的隐私保护推荐技术[14,15]和基于数据扰动的隐私保护推荐技术[16~18]. 目前基于安全多方计算的隐私保护推荐算法的效率不是很好,很难解决社交网络中大规模用户和项目的隐私保护问题. Dwork 在 2006 年提出的差分隐私(differencial privacy)是目前最有效的数据扰动方法之一[19],该算法通过增加噪声可以保证每一个数据集中任意一条数据的增加或删除都不会给推荐结果造成重大影响,从而能够有效的协调推荐精确度和隐私保护之间的平衡. 该隐私保护模型解决了传统模型的两个弱点:(1)给出了严格的数学定义和定量化的表示与证明;(2)无需考虑攻击者所拥有的任何可能的背景知识. 上述基于差分隐私的推荐算法都是对用户-项目评分数据添加拉普拉斯噪声,实现对用户偏好的保护. 本文提出的方法主要基于提出的分布式推荐框架,采用评分分片添加拉普拉斯噪声,实现差分隐私保护方法,保护用户偏好,提高了整个推荐环节的安全性.

2.3 推荐方法

推荐系统一般可分为三大类:基于内容的推荐系统、协同过滤推荐和混合推荐[8]. 基于内容的推荐系统主要根据用户喜欢的项目,选择与其相似度高的项目作为推荐. 协同过滤技术是推荐系统中效率较高的一种方法,协同过滤技术是根据目标用户的项目评价行为确定一组与之行为相似的推荐用户,并以推荐用户对项目的评价作为目标用户的推荐值. 混合推荐主要是为了解决单一推荐技术的不足,可以按照不同的混合策略将不同的推荐技术进行组合并完成推荐. 近期的研究表明,矩阵因子分解方法应用于协同过滤推荐,可以有效的提高了推荐系统的准确率[20~22]. 随着深度学习理论的发展,部分学者开始将神经网络技术应用于远医疗推荐、智能交通和电子商务领域,取得了较好地推荐精度和应用效果[23~26]. 但是上述研究工作都没有关注用户隐私问题,本文中主要基于矩阵因子分解算法,提出一种基于位置服务的分布式差分隐私保护推荐方法,应用差分隐私和保序加密函数实现用户偏好和位置隐私保护. Koren 等[27]将矩阵因子分解法应用于推荐系统,提出了基于奇异值分解(Singular Value Decomposition,SVD)的推荐算法,该方法能够高效率的处理大规模数据集,在与传统的协同过滤的方法比较中,其性能有较大的优势,其模型如式(1):

$$\Psi(r,p,q) = \frac{1}{2}\sum_{(u,i)\in \text{Test}}\left(r_{ui} - \boldsymbol{p}_u\boldsymbol{q}_i^{\mathrm{T}}\right)^2 + \frac{\lambda}{2}\left(\sum_u \|\boldsymbol{p}_u\|_F^2 + \sum_u \|\boldsymbol{q}_i\|_F^2\right) \quad (1)$$

式(1)中,Test 表示用户 u 对项目 i 的评价集合的训练集,$\boldsymbol{p}_u$ 和 $\boldsymbol{q}_i$ 表示用户和项目的潜在因子特征值向量,$\|\cdot\|_F$ 表示 Frobenius 范式. Ψ 是目标函数,可以通过梯度下降优化算法求得最优解. 本文主要基于 SVD 方法设计一种面向位置服务的分布式差分隐私推荐模型.

3　分布式隐私保护框架

3.1　相关符号及问题定义

(1)相关符号说明

本节将推荐中涉及的用户位置、推荐项目、及用户历史评分数据定义如下:

假设某地理区域内,共有 m 个用户和 n 个推荐兴趣点. 对任意用户 u_i,其都存在地理位置坐标是(x_i,y_i);同理,每个兴趣点 poi_j 其也存在相应的地理坐标(lon_j,lat_j). 用户对兴趣点的历史评分用用户-兴趣点评分矩阵 $\boldsymbol{R}$ 表示,其中 r_{ij} 表示用户 u_i 对兴趣点 poi_j 的历史评分,且用户评分 r_{ij} 的取值范围为$[r_{\min}, r_{\max}]$,其默认取值范围是[1,5]的等级评分. 由于用户需要根据自身的地理位置,来选择一定范围内的兴趣点,因此用户向推荐服务器请求推荐服务时,需要提供其地理位置坐标及其请求范围,本文假设用户 u_i 的请求范围是一个矩形区间$(x_i - \Delta x_{i1}, x_i + \Delta x_{i2})$,$(y_i - \Delta y_{i1}, y_i + \Delta y_{i2})$范围内.

(2)问题定义

本文提出的分布式推荐框架主要针对用户偏好隐私和位置隐私的共同保护而提出,首先采用评分分片添加拉普拉斯噪声,实现差分隐私保护方法,保护用户偏好;其次是基于保序加密,设计用户请求位置的请求泛化算法,实现推荐用户的保护需求. 下面给出差分隐私和保序加密的基本介绍.

定义 1　ε-差分隐私. 给定两个数据集 D 和 D',且 D 和 D'最多相差一条记录. 给定一个隐私算法 A,且其取值范围为 Range(A),则算法 A 在 D 和 D'上的输出结果 S($S\in$Range(A))满足式(1),则称算法 A 满足 ε-差分隐私.

$$\Pr[A(D)\in S] \leqslant e^{\varepsilon}\Pr[A(D')\in S] \quad (2)$$

其中 Pr[·]表示隐私被暴露的概率,由算法 A 的随机性决定. ε 表示隐私保护参数,其值越小表示隐私保护程度越高.

实现差分隐私保护的关键技术是添加噪声,目前最常用的是拉普拉斯机制(Laplace mechanism)和指数机制(exponential mechanism),其中 Laplace 机制适用与对数值型结果的保护,指数机制适合非数值型数据的保护. 通过添加噪声实现隐私保护算法是依赖于函数的全局敏感度和隐私保护参数 ε 的.

定义 2　函数的全局敏感度. 对于任意一个函数 f:$D\to R^d$,d 表示函数 f 的查询维度,则函数 f 的全局敏感度记为 Δf.

$$\Delta f = \max_{D,D'}\|f(D) - f(D')\|_1 \quad (3)$$

其中 $\|\cdot\|_1$ 表示 1-阶范数距离,D 和 D'表示最多相差一条记录的数据集.

由于本文是对用户的评分数据进行保护,因此采用 Laplace 机制来实现差分隐私保护算法. Laplace 分布的概率密度函数为:

$$f(x|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \quad (4)$$

μ 和 b 是变量 x 的期望与尺度参数. 为了方便产生噪声数据,一般设期望参数 $\mu=0$,则 Laplace 分布变成标准差为$\sqrt{2}b$ 的对称指数分布. 因此为实现差分隐私算法添加的噪声方法可以表示为:

$$\text{Laplace}(\Delta f/\varepsilon) \quad (5)$$

从式(5)可以看出,ε 越小,引入的噪声越大.

定义 3　位置请求秘密比较. 目标用户加密发送自己位置(x_i,y_i),位置服务器根据兴趣点的地理坐标和用户请求位置信息判定位置关系并且不泄露双方位置信息的比较.

本文使用文献[28]中提出了可比较加密的方案通过一轮交互即可得到查询结果. 该方案通过 Gen,Der,Enc 和 Cmp 四个函数实现,具体作用如下.

Gen 函数:给定一个安全参数 k 和范围参数 n,$k\in N$ 且 $n\in N$,通过输入 k 和 n,Gen 输出一个加密参数 *param* 和主密钥 *mkey*. 即

$$(param, mkey) = \text{Gen}(k,n) \quad (6)$$

Enc 函数:给定参数 *param* 和主密钥 *mkey*,输入明文 *num*,该函数可以输出密文 *ciph*.

$$ciph = \text{Enc}(param, mkey, num) \quad (7)$$

Der 函数:给定参数 *param* 和主密钥 *mkey*,输入明文 *num*,该函数可以生成令牌 *token*.

$$token = \text{Der}(param, mkey, num) \quad (8)$$

Cmp 函数:给定参数 *param*,两个密文 *ciph* 和 *ciph'* 以及令牌 *token*,该函数可以输出{-1,0,1}.

$$\text{Cmp}(param, ciph, ciph', token) \in \{-1,0,1\} \quad (9)$$

给定密文 $ciph = \text{Enc}(param, mkey, num)$ 和 $ciph' = \text{Enc}(param, mkey, num')$,则可以通过 Cmp 函数实现秘密比较.

$$\text{Cmp}(param, ciph, ciph', token) = \begin{cases} -1, & \text{if } num < num' \\ 0, & \text{if } num = num' \\ 1, & \text{if } num > num' \end{cases} \quad (10)$$

3.2 系统架构

为了防止用户的历史评分数据和位置隐私信息的泄露,本文使用分布式推荐系统架构实现对上述两种信息的隐私保护. 这种分布式的结构可以使用目前流行的云计算服务模式,把用户的评分信息采用分布式保护处理后存储在各个云端的推荐服务器中,具体如图 1 所示.

(1)分布式推荐服务器(Distributed Recommender Server,DRS)

针对单一的推荐服务器存储用户历史评分数据存在服务器被恶意攻破而导致用户隐私信息泄露或者推荐服务器本身将数据转给第三方获利的问题,采用分布式推荐服务器构架,该服务器主要负责收集用户经过隐私处理后的评分分片信息,同时响应位置服务器的推荐请求.

(2)位置服务器(Location-Based Service Server,LBSS)

位置服务器主要负责记录推荐项目的地理坐标位置,采集用户的推荐请求并收集用户请求地理位置范围,并通过用户位置范围与推荐项目位置的比对,确定推荐项目,并将用户的请求向分布式推荐服务器进行转发,请求分布式推荐服务器把各自的计算结果向用户发送.

(3)推荐用户(Recommendation Users,RU)

用户主要通过智能终端发送请求,并执行相应的隐私保护算法对评分数据进行处理.

基于图 1 中的系统架构,各对象实体的运行流程如下:

①首先用户 u_i 对消费后的推荐项目 poi_j 进行评分 r_{ij},然后执行随机切片算法,将评分根据分布式推荐服务器的个数分成 K 份 $r_{ij}=\{r_{ij}^1,r_{ij}^2,\cdots,r_{ij}^K\}$,并在每份数据上添加基于差分隐私的干扰噪声发送给每个推荐服务器.

②分布式推荐服务器 k 收到评分分片数据后,根据式(1)中的目标函数定期执行梯度下降算法,更新用户和项目的潜在因子特征值向量 $\boldsymbol{p}_i^k$ 和 $\boldsymbol{q}_j^k$.

$$\mathcal{L}^k=\frac{1}{2}\sum_{u_i}\sum_{poi_j}(\boldsymbol{r}_{ij}^k-\boldsymbol{p}_i^k(\boldsymbol{q}_j^k)^{\mathrm{T}})^2+\frac{\lambda}{2}\left(\sum_u\|\boldsymbol{p}_i^k\|_F^2+\sum_u\|\boldsymbol{q}_j^k\|_F^2\right) \tag{11}$$

③当用户 u_i 请求兴趣点推荐服务时,通过智能终端定位获取自己的地理坐标 (x_i,y_i),然后根据用户的请求范围需求,设置自己的地址请求区间 $(x_i-\Delta x_{i1},x_i+\Delta x_{i2})$,$(y_i-\Delta y_{i1},y_i+\Delta y_{i2})$ 发送给位置服务器,位置服务器通过与推荐项目的地理位置匹配,筛选出符合用户请求需求的推荐项目,并向分布式推荐服务器发送评分预测请求.

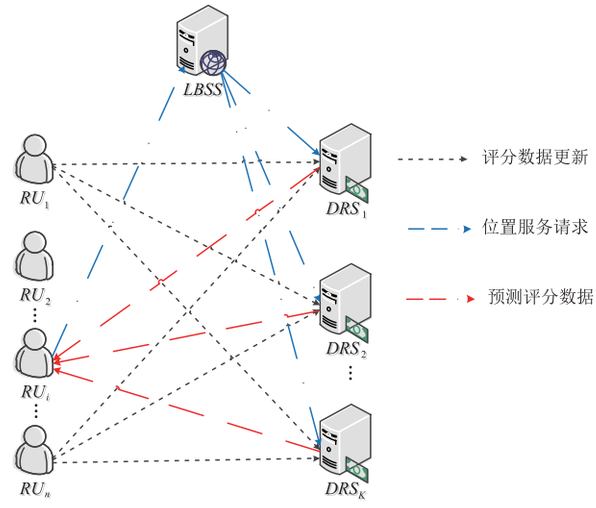


图1 分布式系统架构图

④分布式推荐服务器收到位置服务器的请求后,通过用户和项目潜在因子特征值向量计算预测评分:

$$\hat{r}_{ij}^k=\boldsymbol{p}_i\cdot\boldsymbol{q}_j^{\mathrm{T}} \tag{12}$$

每个分布式推荐服务器将自己的分片预测评分发送给用户,用户计算 $\hat{r}_{ij}=\sum_{k=1}^{K}\hat{r}_{ij}^k$.

3.3 攻击模型及设计目标

在本文设计的分布式用户隐私保护框架中,假设 DRS 和 LBSS 都是半可信的. 各个 DRS 在计算各个 RU 的评分分片时,希望与其它 DRS 交互获得完整的用户的评分信息,从而获取 RU 的偏好隐私信息. 因此,RU 的分片信息需要在传输和计算求和过程中需要进行保护. 同时 RU 在向 LBSS 请求位置服务时,需要将自己的位置范围发送给 LBSS,因此 LBSS 也可能作为攻击者获取用户的位置及其请求信息,因此对 RU 的位置进行保护也是本文的研究重点.

针对上述的攻击模型,本文致力构建一个分布式用户隐私保护框架,并设计一种基于差分隐私技术的用户位置和用户偏好保护的推荐方法,具体设计目标如下:在保证用户的推荐质量前提下,该框架能够满足保护用户的位置和兴趣偏好信息隐私.

4 分布式隐私保护推荐方法

本节主要分为两个阶段,第一阶段执行用户端的分片算法,并在各个分布式推荐服务器端执行矩阵因子分解算法更新用户和项目的潜在特征因子;第二阶段执行用户的推荐请求.

4.1 用户端分片算法设计

假设用户 u_i 对消费后的推荐项目 poi_j 的评分为 r_{ij},则在用户端执行分片算法,然后将分片评分发送给各个分布式推荐服务器. 本文提出两种随机分片算法,并

在后续实例验证中给出各个分片算法的性能分析.

(1)无约束评分随机分片算法根据分布式推荐服务器的数量 K,采用无约束的原则,将评分 r_{ij} 随机分成 K 份,并相应的发给 DRS,具体算法如算法 1.

算法 1　无约束评分随机分片算法(NCRS)

Input:r_{ij}
Output:$\{r_{ij}^1, r_{ij}^2, \cdots, r_{ij}^K\}$
Step:
1: For $t=1$ to K
2: $r \leftarrow \text{rand}(0, r_{ij})$
3: if $(r < r_{ij} - r)$ then
4: $r_{ij}^t \leftarrow r$
5: else
6: $r_{ij}^t \leftarrow r_{ij} - r$
7: End if
8: $r_{ij} \leftarrow r_{ij} - r_{ij}^t$
9: End for

(2)约束评分随机分片算法根据分布式推荐服务器的数量 K,采用等比约束的原则,将评分 r_{ij} 根据用户自身设定的比例将评分分成 K 份,并相应的发给 DRS. 具体步骤是用户首先随机初始化 K 个比例参数 $\{w_1, w_2, \cdots, w_K\}$,并使其满足 $\sum_{k=1}^{K} w_k = 1$,用户将该参数作为私密信息保存,在后续的分片算法中采用该比例参数;然后根据比例参数分割评分 r_{ij},具体步骤见算法 2.

算法 2　约束评分随机分片算法(CRS)

Input: r_{ij}
Output:$\{r_{ij}^1, r_{ij}^2, \cdots, r_{ij}^K\}$
Step:
1: $N \leftarrow 1$
2: For $t=1$ to K
3: $w \leftarrow \text{rand}(0, N)$
4: if $(w < 1 - w)$ then
5: $w_t \leftarrow w$
6: else
7: $w_t \leftarrow 1 - w$
8: End if
9: $N \leftarrow N - w_t$
10: $r_{ij}^t \leftarrow w_t \times r_{ij}$
11: End for

(3)差分隐私保护模型

为了进一步提高分布式隐私保护框架的安全性,本文在随机分片算法的基础上融入差分隐私保护方法,从而保证在分布式推荐服务器共谋的情况下,也能达到较好的隐私保护能力.

为使其满足隐私保护参数 ε-差分隐私,本文根据 Laplace 机制首先为评分数据添加噪声,其中评分的全局敏感度 $\Delta r = r_{\max} - r_{\min}$,则添加的噪声为 Laplace$(\Delta r/\varepsilon)$,并使添加噪声后的评分 r_{ij}^k 限制在$[0, r_{\max}]$,若大于 $r_{\max}$,则用 $r_{\max}$ 代替,反之,若小于 0,则用 0 代替. 然后在用户端执行随机分片算法,将评分分片数据发送给每个 DRS 后,每个 DRS 都会得到一个用户-项目分片评分矩阵 $R_k = \{r_{ij}^k\}^{m*n}$.

4.2　分布式推荐服务器端隐私保护模型

DRS 实际上获取的是添加了干扰噪声的分片矩阵. 设第 k 个 DRS 得到的评分矩阵实际上是 $\boldsymbol{R'}_k = \{r_{ij}^k\}^{m*n}$,通过算法 3 可以得到添加了隐私保护的用户和项目潜在特征向量矩阵 $\boldsymbol{P}_k^{m\times f}$ 和 $\boldsymbol{Q}_k^{n\times f}$.

在实际使用中,每个 DRS 在收到用户的分片后,定期执行 IPSGD 算法,见算法 3,更新 $\boldsymbol{P}_k^{m\times f}$ 和 $\boldsymbol{Q}_k^{n\times f}$ 矩阵,因此可以通过用户和项目潜在因子特征值向量值矩阵预测其它分片评分,即:$\widehat{\boldsymbol{R}}^k = \boldsymbol{P}_k^{m\times f}(\boldsymbol{Q}_k^{n\times f})^{\mathrm{T}}$.

算法 3　添加扰动的随机梯度下降算法(IPSDG)

Input: $\boldsymbol{R'}_k$//添加了 Laplace 噪声的分片评分矩阵
f//潜在因子矩阵的维度
λ //正则化参数;　$r_{\max}$//评分取值的最大值
Output:$\boldsymbol{P}_k^{m\times f}$ and $\boldsymbol{Q}_k^{n\times f}$
Step:
1:将评分矩阵 $\boldsymbol{R}_k'$ 的分片评分控制在$[0, r_{\max}]$
2:根据目标函数 $\mathcal{L}^k(\boldsymbol{P}^k, \boldsymbol{Q}^k) = \frac{1}{2}\sum_{u_i}\sum_{p_j}(r_{ij}^k - \boldsymbol{p}_i^k(\boldsymbol{q}_j^k)^{\mathrm{T}})^2 + \frac{\lambda}{2}\left(\sum_u \|\boldsymbol{p}_i^k\|_F^2 + \sum_u \|\boldsymbol{q}_j^k\|_F^2\right)$,利用随机梯度下降算法进行矩阵因子分解
3:　返回 $\boldsymbol{P}_k^{m\times f}$ 和 $\boldsymbol{Q}_k^{n\times f}$

4.3　位置服务器端隐私保护模型

位置服务器主要存储各个兴趣点的地理位置坐标,以及接受用户的位置服务请求. 为避免用户的位置隐私泄露,本节在用户和位置服务器之间采用可比较加密的方案,实现位置请求服务的隐私保护协议具体如下.

步骤 1　(@ RU):用户 u_i 首先生成安全参数 k 和 n,并利用 Gen 函数生成加密参数 *param* 和比较密钥 *mkey*;然后对其请求范围 $(x_i - \Delta x_{i1}, x_i + \Delta x_{i2})$,$(y_i - \Delta y_{i1}, y_i + \Delta y_{i2})$ 进行加密,进而计算得到 $\text{Enc}(x_i - \Delta x_{i1}, x_i + \Delta x_{i2})$,$\text{Enc}(y_i - \Delta y_{i1}, y_i + \Delta y_{i2})$,$\text{Der}(x_i - \Delta x_{i1}, x_i + \Delta x_{i2})$ 和 $\text{Der}(y_i - \Delta y_{i1}, y_i + \Delta y_{i2})$,用户 u_i 并将这些加密后的数据连同 *param* 和 *mkey* 一起发送给 LBSS.

步骤 2　(@ LBSS):位置服务器收到用户的位置请求后,执行可比较加密协议筛选兴趣点操作. 首先位置服务器遍历所有兴趣点,每个兴趣点 poi_j 的地理坐标

(lon_j, lat_j),并将满足筛选条件的兴趣点加入待推荐集合 $\boldsymbol{R}_P$ 中. 执行的具体比较条件如下:

$$\begin{cases}\mathrm{Cmp}(param,\mathrm{Enc}(x_i-\Delta x_{i1}),\mathrm{Enc}(lon_j),\mathrm{Der}(x_i-\Delta x_{i1}))\\ =1 \text{ or } 0\\ \mathrm{Cmp}(param,\mathrm{Enc}(x_i+\Delta x_{i2}),\mathrm{Enc}(lon_j),\mathrm{Der}(x_i+\Delta x_{i2}))\\ =-1 \text{ or } 0\\ \mathrm{Cmp}(param,\mathrm{Enc}(y_i-\Delta y_{i1}),\mathrm{Enc}(lat_j),\mathrm{Der}(y_i-\Delta y_{i1}))\\ =1 \text{ or } 0\\ \mathrm{Cmp}(param,\mathrm{Enc}(y_i+\Delta y_{i2}),\mathrm{Enc}(lat_j),\mathrm{Der}(y_i+\Delta y_{i2}))\\ =-1 \text{ or } 0\end{cases}$$

位置服务器将待推荐集合 $\boldsymbol{R}_P$ 中的兴趣点编号发送给 DRS,请求 DRS 执行预测推荐.

步骤 3 (@ DRS):每个 DRS 收到位置服务器的推荐预测请求后,执行 $\widehat{\boldsymbol{R}}^k=\boldsymbol{P}_k^{m\times f}(\boldsymbol{Q}_k^{n\times f})^{\mathrm{T}}$,并将每个预测评分分片发送给用户 RU.

步骤 4 (@ RU):用户收到推荐服务器的评分后,执行 $\hat{r}_{ij}=\sum_{k=1}^{K}\hat{r}_{ij}^k$,并从中选择 Top-N 个评分最高的推荐结果.

4.4 安全性分析

本文采用安全仿真模型来证明分布式隐私保护框架的安全性[3]. 该模型假设各参与方是半可信环境下运行的,因此只要证明用户的历史评分信息和位置请求信息是安全的,即可证明本文提出的分布式隐私保护框架是安全的.

定理 1 在半可信环境下,本文提出的分布式隐私保护框架是安全的.

首先,分析用户的评分信息,首先是对用户评分添加 Laplace 噪声,再在用户端执行分片算法并发送给各个 DRS. 在这一步存在两种风险:一是部分分片信息被窃听,二是各个 DRS 串谋,从而相互分享用户分片信息. 第一种情况下,由于客户端已经进行了添加噪声处理,并对添加噪声额评分进行了分片处理,因此窃听者无法获得有效的用户评分信息;第二种情况下,各 DRS 通过串谋已经获得用户的所有评分分片信息 $\boldsymbol{R}'_k=\{r_{ij}^k\}^{m*n}$,现在只要证明 DRS 获得所有的分片信息,也是安全的即可. 在执行算法 1 和 2 前,每个评分都添加了 Laplace$(\Delta r/\varepsilon)$的噪声,根据差分隐私并行组合性,每个 $\boldsymbol{R}'_k$分片矩阵都是满足 ε-差分隐私的,因此即使 DRS 串谋获取所有的 $\sum_k \boldsymbol{R}'_k$也是满足 ε-差分隐私的. 即用户分片信息在半可信环境下是安全的.

其次,用户在请求地理位置服务时,向 LBSS 发送应用保序加密算法加密后的请求范围 $\mathrm{Enc}(x_i-\Delta x_{i1}, x_i+\Delta x_{i2})$,$\mathrm{Enc}(y_i-\Delta y_{i1}, y_i+\Delta y_{i2})$以及相应的用于执行比较的参数 $param$, $mkey$, $\mathrm{Der}(x_i-\Delta x_{i1}, x_i+\Delta x_{i2})$, $\mathrm{Der}(y_i-\Delta y_{i1}, y_i+\Delta y_{i2})$. 由于 LBSS 无法获取用户端的安全参数 k,因此无法通过这些比较参数在多项式时间内解析出用户的明文,即用户的请求范围. 另外即使用户的请求范围被攻破,由于本文使用的是用户设置的不对称的地理范围请求方式,因此攻击者也无法精确的推断出用户的地理位置,因而用户的地理位置也是安全的.

综上所述,本文提出的分布式隐私保护框架是安全的.

5 实验评估

5.1 实验配置

5.1.1 对比算法

根据本文设置的推荐应用场景,文中选择以下四种算法与本文提出模型进行比较.

(1) UBCF Model:该模型采用基于用户的协同过滤方法实现用户项目的评分预测,不具有隐私保护功能.

(2) IBCF Model:该模型采用基于项目的协同过滤方法实现用户项目的评分预测,不具有隐私保护功能.

(3) SVD Model:该模型通过矩阵因子分解技术来获取用户和项目的潜在因子特征值向量,实现用户项目的评分预测. 该模型不具有隐私保护功能.

(4) DP-SVD Model:该模型在 SVD 推荐模型的基础上,应用差分隐私技术向用户-项目评分矩阵中添加 Laplace 噪声,实现在推荐的同时,达到保护用户评分隐私的目的,但不具有保护用户地理位置的功能.

(5) DDP-SVD Model:本文提出的分布式隐私保护模型,在实现保护用户评分隐私的同时,也能具有保护用户的地理位置.

5.1.2 实验数据集

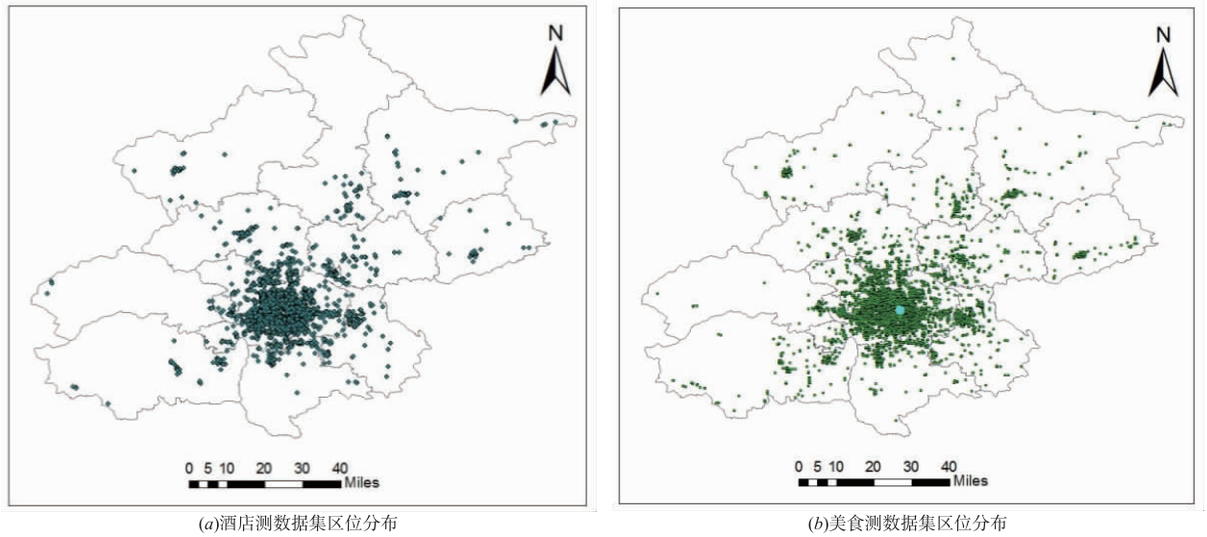
本文采用从携程网(www. ctrip. com)和大众点评网(www. dianping. com)抓取的北京市酒店和美食数据作为实验对比数据集,包括用户对项目的评价(评价等级分成 1 至 5),项目的地理坐标,具体如表 1 所示,图 2(*a*)和图 2(*b*)显示的是在北京市地图上的酒店和美食数据集分布.

表 1 测试数据集

名称	来源	用户数	项目数	评价数量	评分均值
Hotel	携程	11563	2655	2839669	3. 75
Restaurant	大众点评	124077	8874	5731040	3. 52

5.1.3 评价指标

针对上述两个数据集,本文使用数据集中的 80% 作为训练数据集,剩余的 20% 数据作为测试集. 同时在试验中使用预测精度度量的评价指标:均方根误差(RMSE)和平均绝对误差(MAE). 在本文中定义如下:



(*a*)酒店测数据集区位分布　　(*b*)美食测数据集区位分布

图2　北京市住宿和美食测试数据区位分布

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{T_{\text{test}}} (r_{ij} - \hat{r}_{ij})}{|T_{\text{test}}|}} \tag{13}$$

$$\mathrm{MAE} = \frac{\sum_{T_{\text{test}}} |r_{ij} - \hat{r}_{ij}|}{|T_{\text{test}}|} \tag{14}$$

其中 RMSE 和 MAE 的值越小,说明推荐精度越好.

5.1.4　参数设置

为了对本文提出的分布式隐私保护框架进行验证与测试,需要对文中的部分算法参数进行设置,具体见表 2.

5.2　实验分析

5.2.1　几种对比算法的推荐精度比较

本文在两个数据集上对五种算法进行对比实验,表 3 中给出了 RMSE 和 MAE 两个指标值上的实验结果. 表3的实验结果表明:(1)本文提出的DDP-SVD算法相较于传统的 UBCF 和 IBCF 算法,推荐精度上有较大提高;(2)在采用差分隐私保护机制后,相对于 SVD 算法推荐精度在 RMSE 指标上下降 12.8 % 以及 MAE 指标上下降 7.5%,与集中式的差分隐私保护推荐算法 DP-SVD 相比分别下降 1.2% 和 1.1%. 虽然本文提出的 DDP-SVD 算法相对于 DP-SVD 性能略有下降,但由于采用分布式架构,其整体的隐私保护能力更强.

表 2　参数设置表

参数名称	参数说明	参数值
ε	差分隐私保护预算参数	0.1(默认值)
f	用户与项目的潜在特征向量维度	20(默认值)
k	分布推荐服务器的数量	5(默认值)
λ	矩阵因子分解正则化参数	0.001(默认值)

表 3　推荐精度对比实验

Dataset	UBCF		IBCF		SVD		DP-SVD		DDP-SVD	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Hotel	1.579	1.293	1.421	1.218	1.012	0.895	1.146	0.957	1.160	0.968
Restaurant	1.425	1.251	1.378	1.189	0.985	0.811	1.122	0.946	1.134	0.956

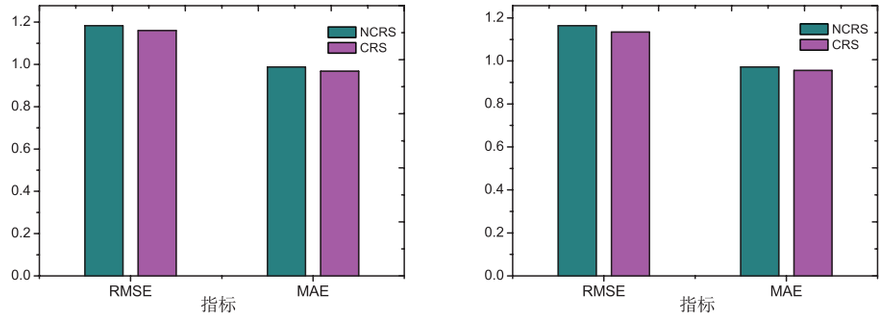
5.2.2　客户端分片算法性能分析

本文提出两种评分分片算法,一种是随机分片算法 NCRS,第二种是用户根据分布式推荐服务器的数量,随机生成比例参数对评分进行切割算法 CRS. 这两种算法的对比精度如图 3 所示. 通过在北京市 Hotel 和 Restaurant 两个数据集上的测试结果显示,采用等比约束的 CRS 算法性能优于无约束的 NCRS 算法,实验结果表明对用户评分采用规律性的切片有助于提高推荐算法的准确性. 由于用户需要保存随机比例参数,因此对用户的智能终端提出了更高的参数存储安全要求,但本文对每个分片都采用了差分隐私保护策略,即使用户泄露了随机比例参数,也能满足 ε-差分隐私安全要求.

5.2.3　隐私保护预算参数与推荐精度分析

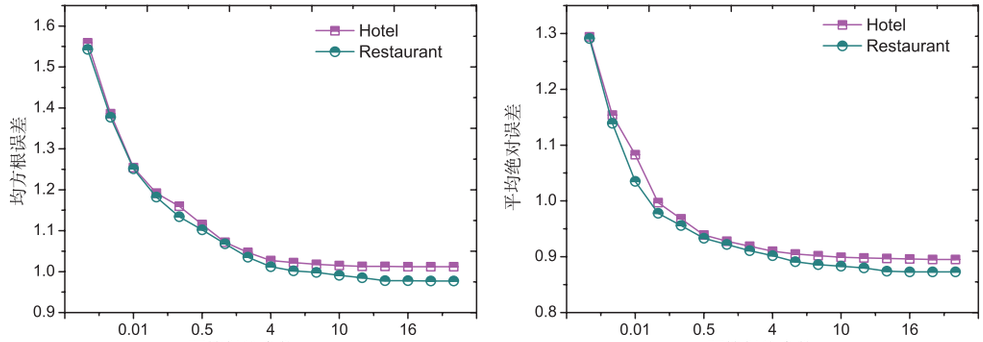
本节主要测试隐私保护预算参数 ε 对推荐精度的影响,图 4(*a*)和图 4(*b*)显示 RMSE 和 MAE 两个指标在 Hotel 和 Restaurant 两个数据集测试结果. 由图 4(*a*)

和图4(*b*)可以得出:当隐私保护预算参数 ε 取值大于6时,本文提出的 DDP-SVD 算法的推荐准确度就趋于稳定;当 $\varepsilon \leqslant 1$ 时,隐私保护强度变高,但推荐准确率也急剧下降,这是因为 ε 越小,添加的噪声也越大,导致训练处的用户和项目的潜在因子特征值向量也越偏离真实值,从而计算出的预测值与真实值偏离变大.



(*a*)Hotel数据集测试结果 (*a*)Restaurant数据集测试结果

图3 NCRS和CRS算法推荐精度对比实验



(*a*)预算保护参数对RMSE的影响 (*b*)预算保护参数对MAE的影响

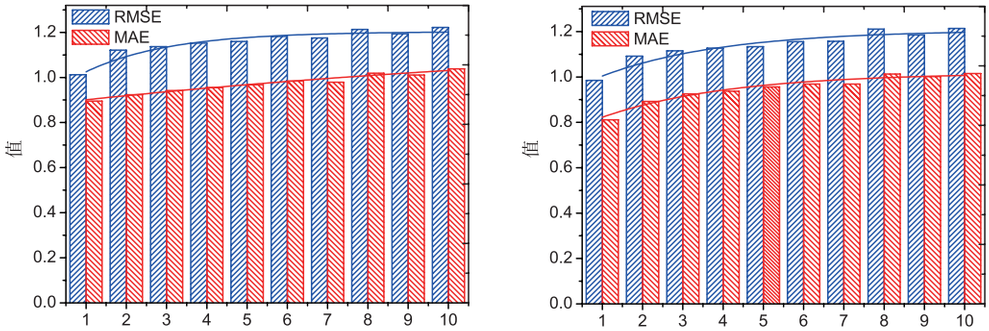
图4 预算保护参数对推荐精度的影响

5.2.4 分布式推荐服务器数量实验

本文提出的 DDP-SVD 算法是基于分布式用户隐私保护框架而构建的,分析分布式服务器的数量对推荐性能的影响也是本节实验的主要目的. 为了便于开展实验,本文选取的分布式服务器的数量分布是从1~10,并测试服务器数量与推荐性能之间的关系. 图5(*a*)和图5(*b*)显示了当服务器数量从1变化到10时,整体的推荐精确度在下降,但当数量大于5时,推荐误差呈波动上升趋势,即推荐性能出现振荡下降. 实验表明,服务器数量与推荐性能成反比,其分片数量越多,造成数据干扰越大,从而导致推荐准确率的下降.

5.2.5 隐私保护预算参数与系统安全性分析

差分隐私利用拉普拉斯机制产生随机值向评分数据中添加噪声,因此攻击者根据查询结果能以一定的概率猜出评分所在的分值区间. 本文实验的数据集采用的评分是5分制的,只要噪音落在(-0.5,0.5)之间

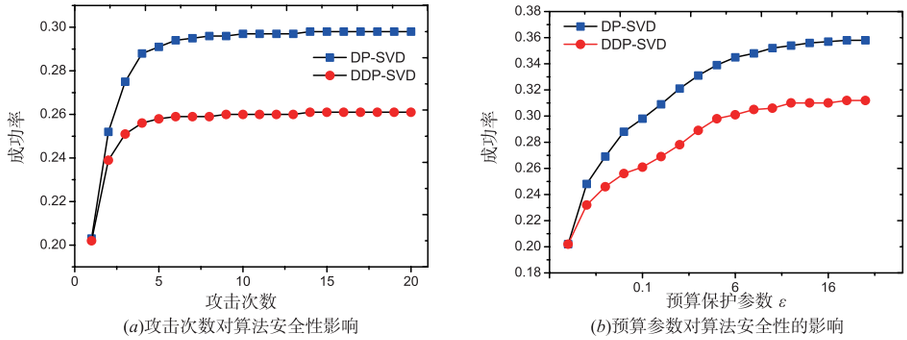


(*a*)Hotel数据集测试结果 (*b*)Restaurant数据集测试结果

图5 分布式推荐服务器数量对推荐精度的影响

就能判断出用户的评分. 因此,本文模拟重复攻击,随机查询测试数据集中的用户的评分,若查询的结果评分与真实评分的差落在(-0.5,0.5)区间,则认为攻击成功 1 次. 本次实验采取 N 次攻击查询,选择落在 5 个评分区间中查询次数最多的那个评分作为最终评分,若其与实际评分相同,认为最终攻击成功.

图 6(a)表示重复攻击次数从 1 变化到 20 时,推荐算法 DP-SVD 与本文提出的 DDP-SVD 被攻击成功的概率情况. 实验表明本文所提的分布式差分隐私推荐方法相对 DP-SVD 具有较好的安全性,主要由于本文不仅采用差分隐私技术来保护评分数据,同时在添加噪声前对评分数据进行了随机分片,进一步降低了被猜中评分的可能性. 另外,当攻击次数大于 5 后,被攻击成功的概率趋于稳定. 图 6(b)表示隐私保护预算参数 ε 算法安全性的影响,本实验采用20 次重复攻击,统计其最终攻击成功的概率. 实验结果表明随着预算参数的增长,攻击成功的概率也随之增加,算法的安全性也随之降低. 该实验表明,推荐算法可以根据相应的安全强度需求,选择相应的隐私保护预算参数值.

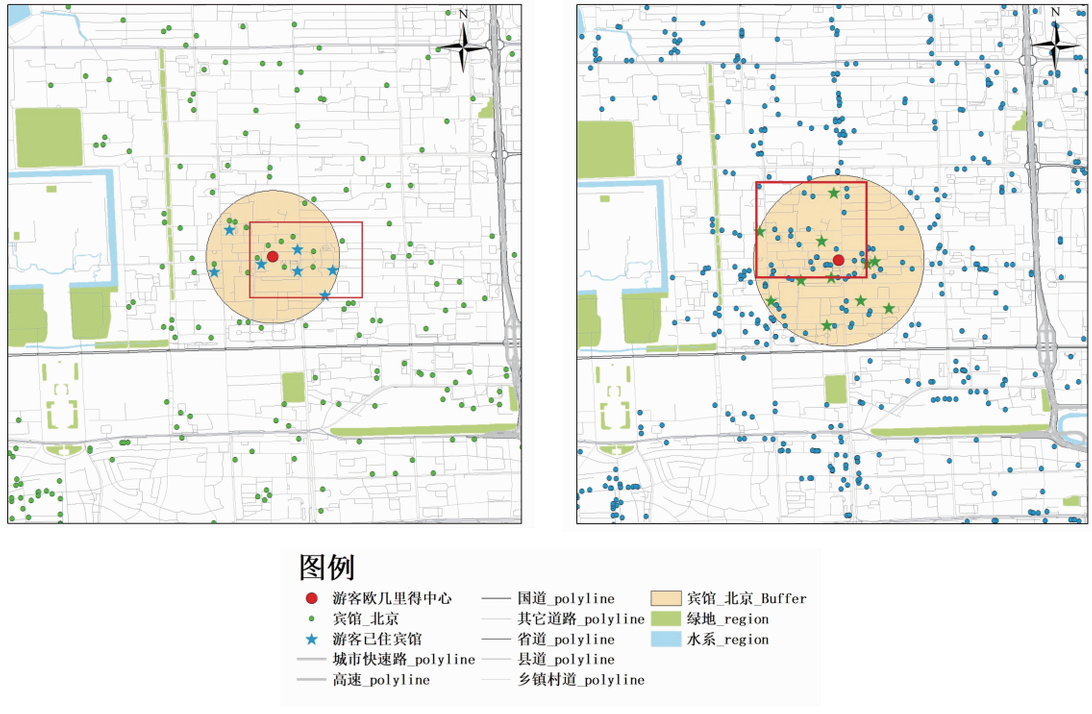


(a)攻击次数对算法安全性影响　　(b)预算参数对算法安全性的影响

图6　攻击次数与预算参数对算法安全性的分析

5.2.6　用户请求地理位置范围性能分析

本文以训练数据集中用户评价的住宿或美食地址构成的区域的欧式距离中心作为用户的请求地址来展开测试,如图 7 所示. 假设用户测试数据集中用户 u_i 设置自己的地址请求区间为$(x_i - \Delta x_{i1}, x_i + \Delta x_{i2})$,$(y_i - \Delta y_{i1}, y_i + \Delta y_{i2})$,其中$(x_i, y_i)$为其欧式距离中心,并随机



(a)Hotel数据集推荐示意　　(b)Restaurant数据集推荐示意

图7　用户请求地理位置的欧式距离中心示意图

生成 $\{\Delta x_{i1}, \Delta x_{i2}, \Delta y_{i1}, \Delta y_{i2}\}$ 的值,取值区间为[0,10000],单位为 m;设置推荐服务器的数量为 5.

通过对测试数据集中的用户进行仿真实验,推荐评分最高的 Top-K 个宾馆和餐厅给测试用户,计算得到推荐准确和召回率,如图 8 所示. 基于用户评价数据集的欧式距离中心,并采用随机生成用户请求地理范围的兴趣点推荐实验结果表明,本文提出的分布式框架差分隐私推荐方法具有较好的推荐性能和适应性.

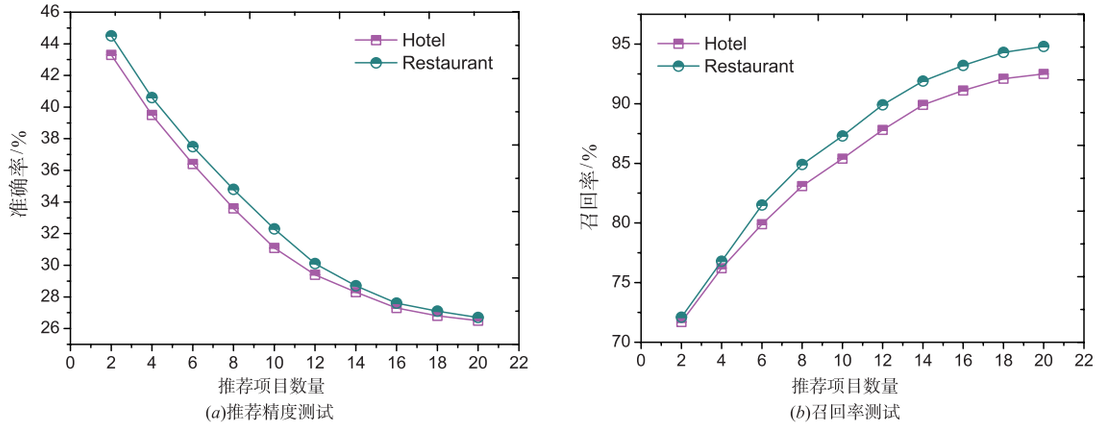


图8 基于用户随机生成的地理请求范围的推荐性能分析

6 结论

本文面向移动互联网的应用背景,提出一种分布式的隐私保护推荐框架,并在 SVD 推荐算法的基础上利用差分隐私技术实现用户评分数据的保护,同时基于保序加密函数实现了推荐用户位置隐私的保护,最后通过理论分析和两个真实的数据集上的实验表明,本文提出的基于位置服务的分布式差分隐私推荐方法相比传统的推荐方法,能够在保护用户隐私的情况下,更好地适应移动互联网背景下的兴趣点推荐;同时实验结果表明,本文提出的推荐方法也具有较好的推荐性能. 在今后的工作中,我们将进一步提高分布式框架下的计算和通信效率,优化推荐算法性能,提高推荐的准确性.

参考文献

[1] Ramakrishnan N, Keller B J, Mirza B J, et al. Privacy risks in recommender systems[J]. IEEE Internet Computing, 2001, 5(6): 54 – 62.

[2] Yu Z, Wong R K, Chi C H. Efficient role mining for context-aware service recommendation using a high-performance cluster[J]. IEEE Transactions on Services Computing, 2017, 10(6): 914 – 926.

[3] Bost R, Popa R A, Tu S, et al. Machine learning classification over encrypted data[A] Network and Distributed System Security Symposium[C]. San Diego, USA: ISOC, 2015. 4324 – 4325.

[4] Polatidis N, Georgiadis C K, Pimenidis E, et al. Privacy-preserving collaborative recommendations based on random perturbations[J]. Expert Systems with Applications, 2016, 71(C): 18 – 25.

[5] Erkin Z, Veugen T, Toft T, et al. Generating private recommendations efficiently using homomorphic encryption and data packing[J]. IEEE Transactions on Information Forensics & Security, 2012, 7(3): 1053 – 1066.

[6] Liu A, Wang W, Li Z, et al. A Privacy-preserving framework for trust-oriented point-of-interest recommendation [J]. IEEE Access, 2018, 6: 393 – 404.

[7] Wang S, Tang J, Wang Y, et al. Exploring hierarchical structures for recommender systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1022 – 35.

[8] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6): 734 – 749.

[9] Liu Z, Qu W, Li H, et al. A hybrid collaborative filtering recommendation mechanism for P2P networks[J]. Future Generation Computer Systems, 2010, 26(8): 1409 – 1417.

[10] Gilburd B, Schuster A, Ran W. k-TTP: a new privacy model for large-scale distributed environments[A]. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York, USA: ACM, 2004. 563 – 568.

[11] Neves A R D M, Álvaro Marcos G. Carvalho, Ralha C G. Agent-based architecture for context-aware and personalized event recommendation[J]. Expert Systems with Applications, 2014, 41(2): 563 – 573.

[12] Bilenko M, Richardson M. Predictive client-side profiles for personalized advertising[A] Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. San Diego, USA: ACM, 2011. 413 – 421.

[13] Veugen T, De Haan R, Cramer R, et al. A framework for secure computations with two non-colluding servers and multiple clients, applied to recommendations [J]. IEEE Transactions on Information Forensics & Security, 2015, 10(3):445 – 457.

[14] Rahman M, Ballesteros J, Carbunar B, et al. Toward preserving privacy and functionality in geosocial networks [A]. Proceedings of the 19th ACM Annual International Conference on Mobile Computing & Networking[C]. Miami, USA:ACM, 2013. 207 – 210.

[15] L Bisheng, U Hengartner. Privacy-preserving social recommendations in geosocial networks[A]. Proceedings of the 2013 Eleventh Annual International Conference on Privacy, Security and Trust (PST)[C]. Tarragona, Spain:IEEE, 2013. 69 – 76.

[16] 鲜征征,李启良,黄晓宇,陆寄远,李磊. 融合显/隐式信任协同过滤算法的差分隐私保护[J]. 电子学报,2018,46(12):3050 – 3059. .
XIAN Zheng-zheng, LI Qi-liang, HUANG Xiao-yu, LU Ji-yuan, LI Lei. Differential privacy protection for collaborative filtering algorithms with explicit and implicit trust [J]. Acta Electronica Sinica, 2018, 46(12):3050 – 3059. (in Chinese)

[17] 范利云,左万利,王英,王鑫. 一种基于差分隐私和时序的推荐系统模型研究[J]. 电子学报,2017,45(9):2057 – 2064.
FAN Li-yun, ZUO Wan-li, WANG Ying, WANG Xin. Research on recommender system model based on differential privacy and time series[J]. Acta Electronica Sinica, 2017, 45(9):2057 – 2064. (in Chinese)

[18] D Riboni, C Bettini. A platform for privacy-preserving geo-social recommendation of points of interest[A]. Proceedings of the 14th International Conference on Mobile Data Management (MDM)[C]. Washington, USA: IEEE, 2013. 347 – 349.

[19] C. Dwork. Differential privacy[J]. Lecture Notes in Computer Science, 2006, 26(2):1 – 12.

[20] Zheng X, Luo Y, Sun L, et al. A novel social network hybrid recommender system based on hypergraph topologic structure[J]. World Wide Web Journal, 2018, 21(4):985 – 1013.

[21] Xu Z, Chen L, Dai Y, et al. A dynamic topic model and matrix factorization-based travel recommendation method exploiting ubiquitous data[J]. IEEE Transactions on Multimedia, 2017, 19(8):1933 – 1945.

[22] Lian D, Ge Y, Zhang F, et al. Scalable content-aware collaborative filtering for location recommendation[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, 30(6):1122 – 1135.

[23] Wang L, Zhang W, He X, et al. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation[A]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining[C]. London, UK:ACM, 2018. 2447 – 2456.

[24] Huang Z, Shan G, Cheng J, et al. TRec:an efficient recommendation system for hunting passengers with deep neural networks[J]. Neural Computing and Applications, 2019, 31(s1):209 – 222.

[25] Seo S, Huang J, Yang H, et al. Interpretable convolutional neural networks with dual local and global attention for review rating prediction[A]. Proceedings of the Eleventh ACM Conference on Recommender Systems[C]. Como, Italy:ACM, 2017. 297 – 305.

[26] Wang Q, Yin H, Hu Z, et al. Neural memory streaming recommender networks with adversarial training[A]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining[C]. London, UK:ACM, 2018. 2467 – 2475.

[27] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8):30 – 37.

[28] Furukawa J. Request-based comparable encryption[A]. European Symposium on Research in Computer Security [C]. Egham, UK:Springer, 2013. 129 – 146

作者简介

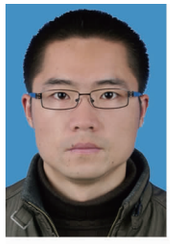
郑孝遥　男,1981 年 2 月出生,安徽芜湖人. 安徽师范大学计算机与信息学院副教授,CCF 和 IEEE 会员,研究方向为信息安全、个性化推荐等.
E-mail:zxiaoyao_2000@163.com

罗永龙(通信作者)　男,1972 年 4 月出生,安徽太湖人. 博士生导师,安徽师范大学计算机与信息学院教授,网络与信息安全安徽省重点实验室主任,研究方向为信息安全、空间数据处理等.
E-mail:ylluo@ustc.edu.cn

汪祥舜　男,1992 年 10 月出生,安徽安庆人. 现为安徽师范大学计算机与信息学院在读硕士研究生,研究方向为推荐系统和信息安全.

孙丽萍 女,1980 年 6 月生,安徽宣城人. 2000 年毕业于安徽师范大学计算机教育专业,安徽师范大学计算机与信息学院教授. 主要从事空间数据处理和智能计算等领域的研究工作.

陈付龙 男,1978 年 5 月出生,安徽霍邱人. 安徽师范大学计算机与信息学院教授,研究方向为嵌入式与普适计算,物联网安全等.



胡桂银 男,1980 年 12 月出生,安徽舒城人. 现为安徽师范大学计算机与信息学院教师,研究方向为信息安全.

汪小寒 女,1978 年 12 月出生,安徽枞阳人,安徽师范大学计算机与信息学院副教授,主要研究方向为智能计算和信息安全.