面向群组推荐的个性化隐私保护方法

王海艳[1,2]，陆金祥[1]

（1. 南京邮电大学计算机学院，江苏 南京 210023；2. 南京邮电大学江苏省大数据安全与智能处理重点实验室，江苏 南京 210023）

摘 要：为解决现有的隐私保护方法不能很好地满足群组推荐中用户的个性化隐私需求的问题，提出了一种面向群组推荐的基于可信客户端的个性化隐私保护框架及基于此框架的群组敏感偏好保护方法。所提方法在可信客户端收集群组内用户的历史数据以及隐私偏好需求，利用用户敏感主题相似性发现组内相似用户，通过对前 k 个用户进行随机的协同扰动，实现群组内用户的个性化隐私保护。仿真对比实验表明，所提的个性化隐私保护方法能够满足不同用户的隐私需求，具有更好的性能。

关键词：群组推荐；个性化隐私保护；随机化扰动；k-匿名

中图分类号：TP393.0

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2019183

Personalized privacy protection method for group recommendation

WANG Haiyan[1,2], LU Jinxiang[1]

1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

2. Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract: To address the problem that most of the existing privacy protection methods can not satisfy the user's personalized requirements very well in group recommendation, a user personalized privacy protection framework based on trusted client for group recommendation (UPPPF-TC-GR) followed with a group sensitive preference protection method (GSPPM) was proposed. In GSPPM, user's historical data and privacy preference demands were collected in the trusted client, and similar users were selected in the group based on sensitive topic similarity between users. Privacy protection for users who had privacy preferences in the group was realized by randomization of cooperative disturbance to top k similar users. Simulation experiments show that the proposed GSPPM can not only satisfy privacy protection requirements for each user but also achieve better performance.

Key words: group recommendation, personalized privacy protection, randomized perturbation, k-anonymous

1 引言

近年来，随着信息化、智能化的深度融合，用户在利用信息技术获得更多便利的同时，个人隐私面临着严重的威胁。2018 年，国内外发生的几起严重泄露用户隐私的事件表明，信息社会中用户的隐私保护迫在眉睫。作为个性化推荐系统的扩展，群组推荐日益受到关注。群组推荐系统是指群组中有多个用户，不同用户有各自的偏好需求，根据群组中每个用户的偏好进行推荐的系统。群组推荐主要任务是缓解群组成员之间的偏好冲突，使推荐结果尽可能满足所有群组成员的需求。群组推荐系统需要收集大量的用户历史数据来实现对群组的推荐，这些数据中可能包含一些用户的敏感信息。多数情况下推荐系统是非可信的，存在泄露用户隐私的风险，而用户为了获得个性化的推荐，只能选择信任推荐系统。因此，对群组内用户实现隐私保护，已成为群组推荐的一个研究热点。

收稿日期：2018−12−03；**修回日期**：2019−07−04

基金项目：国家自然科学基金资助项目（No.61772285）

Foundation Item: The National Natural Science Foundation of China (No.61772285)

与传统的个性化推荐系统不同，群组推荐系统首先需要考虑群组内所有用户的偏好，并通过群组内用户间的偏好共享和交互机制缩小群组内用户之间的偏好差异。在此基础上，群组推荐系统将每个群组内用户的偏好融合得到群组偏好，并根据群组偏好完成群组推荐。现阶段，面向群组推荐的用户隐私保护主要通过群组内用户偏好协同扰动的方法实现。已有的面向群组推荐的用户隐私保护方法，大都假设群组内所有用户具有相同的隐私保护需求，对群组内所有用户进行相同的隐私保护处理，而不能针对群组内用户个性化的隐私保护需求分别进行处理，具有很大的局限性。

针对以上问题，本文提出了面向群组推荐的个性化隐私保护方法，主要工作及创新点如下：1) 提出了面向群组推荐的基于可信客户端的个性化隐私保护框架（UPPPF-TC-GR, user personalized privacy protection framework based on trusted client for group recommendation），在可信客户端对群组内用户的历史偏好进行收集和分析，利用敏感主题相似度分析方法发现群组用户数据中存在的敏感数据，对群组用户的敏感偏好进行隐私保护处理；2) 基于此框架，提出了群组敏感偏好保护方法（GSPPM, group sensitive preference protection method），在可信客户端利用随机的扰动技术对群组用户偏好进行隐私保护处理，实现对群组用户敏感偏好的保护。本文提出的个性化隐私保护方法利用群组内相似用户的评分进行随机的协同扰动，实现对用户个性化隐私的保护，相对于直接添加扰动噪声的随机化扰动方法有了很大的改进与创新。

2　相关工作

2.1　个性化推荐中的隐私保护方法

现有的个性化推荐中的隐私保护方法主要使用 *k*-匿名、随机化扰动、加密等隐私保护技术，实现对用户推荐的同时有效保护用户的隐私。文献[1]在数据发布领域，提出利用 *k*-匿名方法实现对用户敏感信息的保护。文献[2]提出关系保持变换的方法来防御关于广义形式的距离保持变换的攻击。文献[3]主要针对目前比较流行的背景知识攻击和审查攻击提出了一种保护用户隐私的方法，核心思想是利用 *k*-匿名的隐私保护方法实现对背景知识攻击的拦截。文献[4]针对 *k* 近邻（*k*NN，*k*-nearest neighbor）攻击，提出分区概率邻域选择方法，以确保所选区域的安全性，同时实现针对 *k*NN 攻击的最佳预测精度。文献[5]分析了在线社交网络存在的一些隐私保护相关的问题，提出通过 *k*-匿名化实现在匿名社交网络中用户间相互通信而不会暴露其身份的隐私保护方法。文献[6]提出在网络跟踪数据中源 IP 和目的 IP 地址之间的固有图形结构，并使用 *k*-匿名防止目标主机被跟踪。文献[7]提出通过摆脱半诚实的服务提供商来设计分散的单一协议，该协议使用非常小的数据集子集，其准确性仍然等于或优于某些基线算法。文献[8]提出使用匿名身份验证来保护使用盲签名的位置隐私。文献[9]提出了一种基于信任的隐私保护朋友推荐方案。文献[10]提出一个函数发生器实体，通过该实体周期性地分布空间变换参数实现对用户位置的隐私保护。文献[11]提出了一种应用数据混淆技术的隐私保护框架，并在此框架下进一步开发了 2 种代表性的隐私保护服务质量（QoS，quality of service）预测方法。文献[12]提出了一种新型数据加密方法，该方法在时间约束下使用隐私分类方法选择性地加密数据，并使用选择性加密策略最大化隐私保护范围。文献[13]提出利用差分隐私方法作为隐私保护框架，利用目标扰动方法对矩阵中添加满足差分隐私约束的噪声得到噪声矩阵分解模型，实现对隐私数据的保护。文献[14]对差分隐私保护领域已有的研究成果进行了总结，对该技术的基本原理和特征进行了阐述，重点介绍了差分隐私的研究热点。文献[15]分析了差分隐私保护模型相对于传统安全模型的优势，对差分隐私基础理论及其在数据发布与数据挖掘中的应用研究进行综述。文献[16]提出了一种轻量级的位置感知推荐系统隐私保护框架，利用该框架服务提供者将随机处理后的历史评价信息外包给云平台，并通过安全协议在云平台的辅助下进行相似度信息的安全计算。文献[17]提出一种贪心聚类匿名方法，通过分类概化准标识属性，分别度量其信息损失，从而减小并合理评价信息损失。文献[18]提出了一种具有隐私保护的协同过滤方法，该方法主要思想是通过使用传输矩阵直接干扰原始数据集来保证用户隐私。文献[19]将 *k*-匿名隐私保护技术和区块链技术的分布式特性结合起来，提出了一种基于区块链的分布式 *k*-匿名位置隐私保护方案。文献[20]提出一种具备保护用户隐私功能的推荐系统，利用用户的历史评价和项目属性信息，结合不采集用户个人信息的协同过滤推荐算法，在实现对用户推荐的同时，保护用

户隐私。文献[21]提出将位置轮廓相似度和位置点相似度度量的过滤算法应用到协同过滤算法中，实现对用户位置隐私的保护。

以上研究发现，现有的主流隐私保护方法或者采用了 *k*-匿名的思想对用户的敏感属性进行泛化处理来保护用户的隐私，或者使用随机化的扰动方法来实现对敏感数据的保护，或者采用加密技术来实现对用户的隐私保护。但 *k*-匿名算法本身存在一系列的问题，其中比较主要的问题有同质攻击和背景知识攻击。而随机化扰动方法会带来一定程度的数据失真。因此，单纯使用 *k*-匿名的隐私保护算法或者随机化扰动方法来保护用户隐私存在一定的局限性，不能够在隐私保护和推荐准确性之间达到很好的平衡，加密技术也不适用于群组推荐这样计算量较大的应用场景。

2.2 群组推荐中的隐私保护方法

现阶段的研究成果主要是通过向原始群组偏好中添加扰动的方式实现的。文献[22]提出在群组内部对群组内用户的偏好配置文件进行扰动，将扰动后的数据进行序列化的处理后再进行传输，最后推荐服务器根据相关的数据迭代算法，从扰动后的数据中得到可用的群组偏好数据实现对群组的推荐。文献[23]提出了基于群组的隐私保护方法，将群组作为一个中间件对群组用户进行隐私保护，群组用户将个人偏好数据通过聚合策略进行聚合后，以群组的方式来进行推荐，有效地保护了群组用户的隐私数据。文献[24]提出基于可信客户端生成一组虚拟的偏好配置文件传输给推荐服务器进行推荐，以掩盖用户敏感主题，实现对用户个人隐私的保护。文献[25]从群组推荐系统的形式化定义和研究框架入手，对群组推荐系统的用户偏好获取、群组发现、偏好融合算法以及效用评价等关键技术进行前沿概述，并对群组推荐系统中的用户隐私保护问题进行了展望。文献[26]针对移动社交网络的应用场景，提出一种基于影响因子的群组推荐隐私保护方法，该方法使用模糊矩阵算法来实现对用户隐私的保护。文献[27]提出一种基于差分隐私保护算法和时间因子相结合的高效隐私保护协同过滤算法，该算法能够很好地保护用户隐私。文献[28]提出了一种新的综合考虑代理和用户属性及其偏好的私有数据信息匹配算法，该算法支持具有偏好信息的多元属性数据匹配，能够有效保障用户和子代理的安全性。文献[29]提出了一种基于位置混淆的轨迹隐私保护方法，该方法混淆用户的真实查询位置，使攻击者不能推断出用户的真实轨迹，实现对用户的隐私保护。文献[30]提出了一种基于 *k*-匿名的位置及数据隐私保护方法，并采用基于多方安全合作的方法来构造一个包含 *n* 个互相没有任何链接的用户的等价类，以保证参与用户的位置隐私。

以上研究发现，现有的群组推荐系统中的隐私保护方法大都是在客户端对群组内所有用户进行协同扰动实现隐私保护，这样的隐私保护方法实现的前提是假设群组内所有用户都具有相同的隐私保护需求。但事实上，群组内不同用户的隐私保护需求往往具有个性化差异。

本文在文献[24]模型及框架基础上，提出了面向群组推荐的组内用户个性化的隐私保护框架和方法。

3 群组敏感偏好隐私保护方法

介绍本文提出的隐私保护方法之前，首先分析群组推荐中可能存在的攻击和随机化扰动隐私保护方法。

3.1 攻击分析

群组推荐系统面临的不同于个性化推荐系统的攻击主要有审查攻击和评级攻击[21]。1) 审查攻击是指攻击者在了解到某个项目已被某个用户查看过的背景知识的前提下，可以从匿名数据集中重新识别该用户。例如，根据一些公共的数据源，如果攻击者知道用户查看了其他用户没有查看过的一些项目，则该用户的身份可以很容易地被找出。2) 评级攻击是指攻击者通过比较单个用户评分与组中其他用户的评分来识别评分异常的用户。例如，攻击者可以基于某个用户对特定项目给予高评分，而该组中的其他用户给这个项目低评分，从而识别出该用户。本文假设攻击者是具备一定背景知识的主动攻击者，攻击者会结合自身已有的背景知识和群组内用户的评分信息来识别群组内具体的用户，从而挖掘出该用户的敏感信息。

3.2 随机化扰动技术

随机化扰动技术本质上是一种数据伪装技术，其主要作用是缓解用户偏好中敏感信息泄露的风险。随机扰动技术主要是在用户原始数据上随机选取添加一个服从均匀分布或者高斯分布的扰动噪声，从而保护用户的原始数据不被窃取。例如，为了隐藏数据 a，就在 a 上添加一个随机数 r，则外

界获得的是数据 $a+r$。在本文提出的隐私保护方法中，为实现简单，给用户的真实数据加上一个随机噪声后才发给服务器，从而保护用户隐私数据。

3.3 *k*-匿名技术

匿名技术目标是防止恶意攻击者通过用户发布的信息来定位用户，对用户发布的信息进行匿名处理消除身份信息识别符，形成对用户的隐藏。常见的方法是 *k*-匿名和添加虚拟用户。设 $\mathrm{RT}(A_1, A_2, \cdots, A_n)$ 是数据表，RT 的相关关联标示符记为 QI。如果每一个在 $\mathrm{RT}[\mathrm{QI}_{\mathrm{RT}}]$ 中的序列值在 $\mathrm{RT}[\mathrm{QI}_{\mathrm{RT}}]$ 中最少出现过 k 次，则称其满足 *k*-匿名。添加虚拟用户主要是用一些合理的虚拟用户来代替部分真实用户，从而实现隐私保护。本文采用了 *k*-匿名的隐私保护思想对群组用户敏感信息进行保护，主要实现策略是通过本文提出的群组敏感偏好保护方法在群组内发现与目标用户相似的用户，在此基础上，利用相似用户的偏好对目标用户进行协同扰动，实现对目标用户敏感信息的保护。

3.4 实现隐私保护的群组推荐框架

本文提出的基于可信客户端的面向群组推荐的个性化隐私保护框架（UPPPF-TC-GR）结构如图 1 所示。其中，可信客户端是指能够实现安全目标的客户端，用户在客户端的所有操作都具有原子性、非否认性、可追究性、公平性以及隐私性等特点。非可信是指推荐服务器在进行用户数据的存储和数据建模等操作方面存在隐私泄露的可能。推荐服务器是面向所有用户或组织的，其中不法组织或攻击者可能会挖掘出用户的敏感信息，对用户的隐私产生威胁。可信客户端由 5 个模块组成，分别是行为记录模块、偏好分析模块、主题槽散列模块、敏感偏好保护模块和推荐结果选择模块，各模块的功能将在 3.5 节介绍。非可信推荐服务端由推荐算法和数据库两部分组成。

本文的隐私保护目标是在可信客户端实现对用户隐私的保护。其中，数据库存储用户历史数据和项目−主题分类文档信息，项目−主题分类文档是通过项目−主题分类树来构建的，项目−主题分类树是系统内存中存在的一个层次化的平衡多路查找树，主要用于管理推荐系统中所涉及的项目，便于将项目进一步分类。项目−主题树的实现原理类似于机器学习中的决策树模型。首先，对项目进行主题划分，其中，主题可以根据外部项目−主题分类知识库如维基百科等进行获取，或者根据实际数据集中所对应的主题集进行遍历得到；其次，考虑到一个项目具有多个属性，本文根据项目属性将项目归属到具体的主题分支下；最后，对项目−主题分类树进行剪枝操作，得到最终的项目−主题分类树。

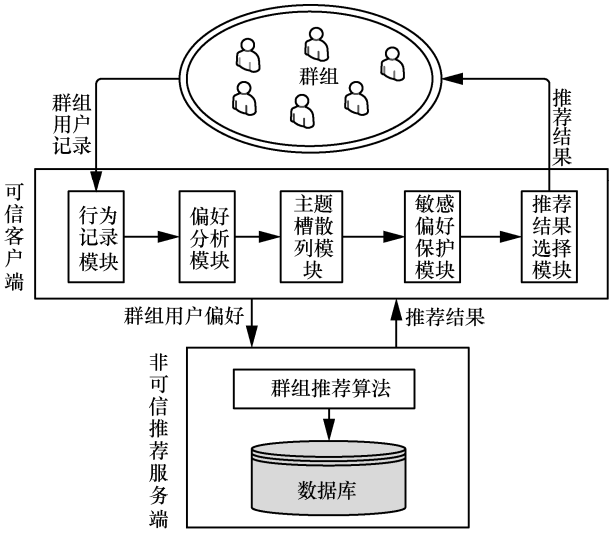

图 1　UPPPF-TC-GR 结构

下面简要分析可信客户端中各个模块的作用。行为记录模块主要负责收集群组内用户的历史记录，本文将用户行为记录的收集放在可信客户端来实现，主要是为了增强用户对于数据的可控性。偏好分析模块主要负责将收集到的群组内用户历史记录进行偏好建模，为每个用户构建属于自己的偏好配置文件。主题槽散列模块主要负责将用户偏好中的项目信息分类到相应的主题槽中。敏感偏好保护模块主要负责对主题槽中的敏感主题进行保护，实现对群组用户隐私的保护。推荐结果选择模块主要负责从推荐系统中选择出推荐结果并将其反馈给群组完成整个推荐流程。

3.5 主要模块介绍

本文提出的 UPPPF-TC-GR 将用户行为记录模块和偏好分析模块从推荐系统中分离出来，由可信客户端实现。这样做的好处是保证系统对于群组用户数据的可控性和隐私保护模块产生数据的可靠性。这 2 个模块主要是对用户历史数据挖掘、建模和分析。推荐结果选择模块主要是将推荐结果返回给群组，完成整个推荐流程。下面介绍本文提出的主题槽散列模块、敏感偏好保护模块的设计和实现以及非可信推荐服务端的群组推荐算法。

3.5.1 主题槽散列模块

主题槽散列模块主要基于项目−主题分类树和本文提出的散列槽（hashslot）来实现。主题槽散列模块的主要功能如下。首先，根据项目−主题分类树将群组用户偏好项目分类到相对应的用户 hashslot 中。其次，根据分类存储后的群组用户主题偏好进行主题重要性计算。最后，根据计算结果发现群组用户存在的敏感主题。其中，项目−主题分类树如图 2 所示。主题树具有以下特征：1) 每个叶子节点表示一个项目；2) 每个非叶子节点代表一个主题；3) 每个项目都包含在一个主题中；4) 每个主题（根节点除外）都包含在另一个主题中。得到群组用户的偏好项目相对应的偏好主题后，将其存储到 hashslot 中完成主题槽散列。其中，本文的 hashslot 结构如图 3 所示，hashslot 结构具有以下特征：1) hashslot 由多个 slot 槽组成；2) hashslot 槽的个数根据系统内存中项目所对应的主题个数来确定；3) 每个 slot 对应一个主题，存储到同一个 slot 中的数据都对应着相同的主题；4) 每个 slot 中的数据都包含着项目和评分信息。本文系统中为每个群组用户都配置了一个 hashslot 槽。经过主题槽散列模块的处理后，将群组用户的偏好配置信息都存储到相应的用户主题槽中。至此，完成了将群组内用户偏好配置信息散列到相应的用户主题槽的操作。

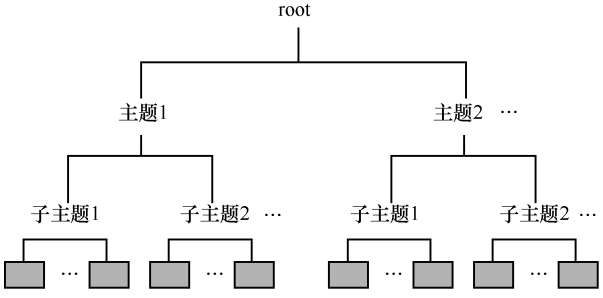

图 2 项目−主题分类树

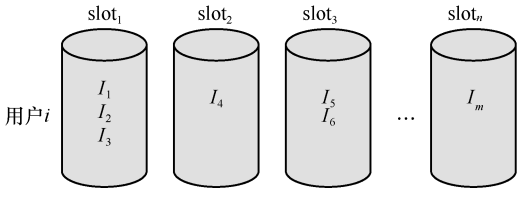

图 3 用户偏好主题槽 hashslot

3.5.2 敏感偏好保护模块

1) 相关符号描述

敏感偏好保护模块基于主题槽散模块，主要功能如下：首先，将群组用户偏好数据都散列到相对应的主题槽中；其次，根据群组用户主题槽进行主题重要性计算来发现群组用户的敏感主题；最后，根据群组用户存在的敏感主题进行隐私保护处理。为了表述方便，表 1 给出了本文用到的主要符号和定义。

表 1 相关符号说明

符号	描述	符号	描述
I	所有项目集合	$G*$	偏好项目集特征分布
U	所有用户集合	score(i)	用户对单个项目的评分
$I(m)$	第 m 个用户的项目集合	$S*$	u 的偏好主题集合
$r_{(u,i)}$	u 对项目 i 的评分	$\boldsymbol{r}_i$	群组内用户的评分向量
score(g)	u 对主题 g 的评分	$\boldsymbol{r}_i^*$	隐私保护处理后的用户评分向量
sp	敏感项目	km	主题的最高层次
sus	相似用户的数组集合	P[]	扰动之后的用户偏好项目集
∩	集合中的元素评分间逻辑与运算	$g*$	群组内用户偏好主题集
SK*	偏好主题集特征分布	$P*$	群组内用户偏好项目集

2) 相关定义描述

为了在可信客户端实现对群组内用户敏感偏好的保护，下面对本文采用的隐私保护方法中的计算式做如下定义。

定义 1 群组内用户偏好项目集。群组内用户偏好项目集是群组内用户感兴趣的所有项目的集合，可以表示为

$$I_{u*}=\{i \mid i \in I \bigcap \text{score}(i) \neq 0\} \tag{1}$$

其中，I 表示所有项目的集合，score(i) 表示用户 u 对项目 i 的评分。可以看出，群组内用户偏好项目集是由用户评分不为零的项目组成的。

定义 2 群组内用户偏好主题评分。群组内用户偏好主题评分代表用户对某个主题的喜好程度，是根据群组内用户偏好项目集评分得到的。本文只考虑群组偏好项目的直接主题，而不考虑项目的其他层次的主题，主题的层次越高，其属性就越抽象；主题的层次越低，属性就越具体。本文的保护目标是实现对具体主题的保护，因此本文可以形式化定义群组内用户偏好主题评分为

$$\text{score}(g)=\sum_{i \in w(g)} \text{score}(i) \tag{2}$$

其中，$w(g)$ 表示所有直接属于主题 g 的项目。

定义 3　群组内用户偏好主题集。是一组由用户感兴趣的所有主题所组成的集合，其可以形式化地定义为

$$g^* = \{g \mid g \in S \bigcap \text{score}(g) \neq 0\} \tag{3}$$

其中，S 代表所有主题集合，score(g) 表示用户 u 对主题 g 的评分。可以看出，群组内用户偏好主题集是由用户评分不为零的主题组成的。

定义 4　群组内用户主题槽 hashslot$_u$。代表用户 u 的偏好主题槽，用于存储用户 u 相对应主题下的项目，每个用户都由可信客户端分配相应的主题槽，单个用户的偏好主题槽的数目是由系统中项目所具备的主题类型数目来确定的，它由多个 slot 组成。

定义 5　群组内用户偏好项目集特征分布。给定一个偏好项目集SU*，其特征分布可以用下面的向量来描述。

$$\begin{aligned} &\boldsymbol{G}^* = (\text{score}(i_1), \text{score}(i_2), \cdots, \text{score}(i_n)), \\ &i_j \in \text{SU}^*,\ j = 1,2,\cdots n, \\ &\text{score}(i_j) \leqslant \text{score}(i_{j+1}),\ j = 1,2,\cdots,n-1 \end{aligned} \tag{4}$$

定义 6　群组内用户偏好主题集特征分布。给定偏好主题集合SS*，其特征分布可以使用以下向量来描述。

$$\begin{aligned} &\mathbf{SK}^* = (\text{score}(g_1), \text{score}(g_2), \cdots, \text{score}(g_n)), \\ &g_i \in \text{SS}^*,\ i = 1,2,\cdots,n \\ &\text{score}(g_i) \leqslant \text{score}(g_{i+1}),\ i = 1,2,\cdots,n-1 \end{aligned} \tag{5}$$

定义 7　用户主题重要性。描述一个敏感主题在群组内用户偏好配置文件中的重要程度，其可以形式化地定义为

$$\text{power}(s^*, u) = \frac{\text{score}(s^*)}{\text{score}(\text{hashslot}_u)} \tag{6}$$

其中，s*代表一个敏感主题，score(hashslot$_u$) 代表单个用户的所有主题评分。本文根据敏感主题重要性，对用户偏好配置文件中的敏感属性进行保护。

定义 8　项目特征相似性。表示隐私保护处理后的群组偏好配置文件和真实的群组偏好配置文件的相似度。2 个项目集合之间的特征相似性可以通过 2 个项目集合的项目特征向量的相似度和主题特征向量的相似度来进行度量。P_1* 和 P_2* 之间的特征相似性度量为

$$\begin{aligned} &\text{sim}(P_1^*, P_2^*) = a_0 \text{sim}(G_1^*, G_2^*) + \\ &\sum_{k=1}^{\text{km}} a_k \text{sim}(\text{SK}_1^*, \text{SK}_2^*) \end{aligned} \tag{7}$$

其中，a_0 和 a_k 为平衡参数，sim(G_1*,G_2*) 和 sim(SK$_1$*,SK$_2$*) 利用欧几里得距离公式进行计算。

定义 9　相似用户。根据定义 7 算出群组内所有用户的主题重要性以及目标用户需要保护的敏感属性所在的主题，将该主题重要性与组内其他用户相应主题下的主题重要性进行比较，从群组内发现 k 个相似的用户对目标用户进行敏感属性的扰动。本文定义参数 q 对用户间主题相似度进行度量，q 值越小，则 2 个用户之间的相似性越高。本文形式化将其定义为

$$q = \text{power}(g^*, u) - \text{power}(g^*, u_i) \tag{8}$$

其中，u_i指代其他用户，$i = 1,2,3,\cdots,n$，$0 < q < 1$。

3) 隐私保护方法的设计

为了保护群组内用户的隐私，根据群组用户是否有隐私保护的需求，在可信客户端对群组用户的敏感属性进行隐私保护处理。首先，用户的隐私保护需求通过显式偏好获取的方式来收集；其次，群组用户隐私保护需求在可信客户端使用隐私保护参数 Ω 进行定量分析，Ω 的取值范围为 {−1,1}，当 Ω=1 时表示用户需要进行隐私保护处理，当 Ω=−1 时表示用户不进行隐私保护处理。因此，对于有隐私保护需求的用户，考虑到其需求可能存在差异，本文采用群组敏感偏好保护方法，通过调整 k 个相似用户的大小来实现不同的隐私保护需求。

本文的隐私保护方法应尽可能地使单个用户的评分在群组内不敏感，这样在恶意攻击者具备一定的背景知识的情况下，相应的审查攻击和评级攻击就不会起作用，本文的隐私保护方法思想如图 4 所示，该思想主要利用群组用户中与目标用户相似的 k 个用户的评分信息，对目标用户的敏感属性进行协同扰动（k 是一个需要调整的参数，相似用户根据定义 9 计算），实现对用户隐私的保护。

本文提出的隐私保护算法首先根据 Ω 隐私保护参数的值进行群组内用户的分类，对需要进行隐私保护的用户分别实现隐私保护处理，为了提高整个算法效率，假设在隐私保护模块中的算法实现中预先存在项目−主题分类树。然而，项目−主题分类树的叶节点深度可能彼此不同。因此，为了便于算法的实现，可以采用以下 2 种方式对主题树进行预处理。1) 分割一些较小深度的叶节点，并构造它们

的父节点。2) 合并一些叶子更深的节点，并删除其父节点。通过这样的操作后，可以使主题树具有相同的深度。另外，本文将项目−主题分类树预先加载到内存中，以提高算法的运行效率。由于相似用户对于目标用户的敏感项目的评分不确定，因此本文的算法分为 2 种情况。1) 当找到的相似用户对敏感项目存在评分时，利用相似用户和目标用户的评分进行协同扰动，将扰动之后的评分作为这 $k+1$ 个用户对于敏感项目的评分。2) 当相似用户对敏感项目没有评分时，利用目标用户对该敏感项目的评分平均化操作后，再加上服从标准高斯分布的 w 噪声值作为这 $k+1$ 个用户对于敏感项目的评分，其中，$\sum_{i=0}^{k} w_i = 0$，$-1 \leqslant w_i \leqslant 1$。

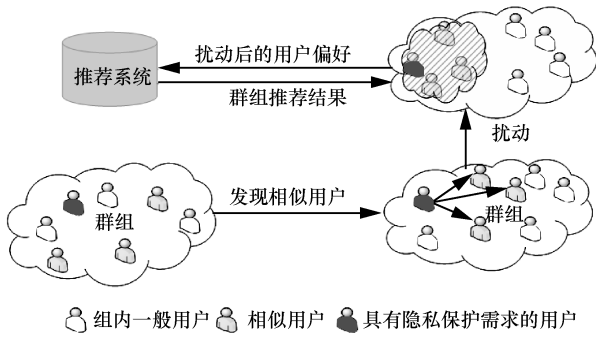

图 4 群组敏感偏好保护方法

下面介绍本文提出的群组内用户隐私保护算法，算法 1 详细介绍了群组敏感偏好保护方法的实现。在 GSPPM 中发现群组中具有隐私保护需求的用户的敏感偏好，对用户的敏感偏好进行隐私保护处理，将隐私保护处理后的群组内用户偏好传递给推荐系统进行推荐。具体方法如下：首先，对项目−评分信息进行预处理操作，主要的预处理操作方法是根据项目−主题分类树对用户项目进行主题分类；然后，根据主题重要性计算方法发现用户的敏感偏好主题，在发现敏感主题的基础上，根据主题相似性寻找与目标用户相似的前 k 个相似用户；最后，根据相似用户和目标用户对敏感项目的评分信息进行协同扰动，将扰动后的评分信息作为群组内用户的最终评分信息，从而实现对用户敏感信息的保护。

算法 1 GSPPM

输入 群组内用户的偏好项目集合 $r_1,r_2,\cdots,r_n$，相似用户的数目 k

输出 隐私保护处理后的群组内用户偏好项目 r_1*, r_2*, …, r_n*

begin *P*[] GSPPM (*G* group)

for *u* in *G*

step1 *P*[] products=*u*.getProducts()；//得到单个用户的偏好项目

step2 putProduct2hashslot(*u*,products,hashslot)；//将单个用户的偏好映射到相应的 slot 槽中

step3 SentiveSubject sp=findSensitiveSubject(*u*, hashslot)；//寻找敏感主题

step4 findSimilaryUsers(sp,hashslot,u,k)； //根据目标用户的敏感主题寻找群组内相似用户

step5 changeScore(sp,*u*,SimilaryUsers[] sus)；//根据相似用户的评分扰动目标用户的评分

return products

end for

从算法 1 中可以发现，本文采用的隐私保护方法是使用群组内相似用户的评分对目标用户评分进行协同扰动的方式来保护用户隐私的。当系统为用户评分加入过高的噪声数据时，将导致数据的大幅度失真，直接引起推荐质量的下降，推荐质量的下降又将导致用户减少对推荐系统的使用，这样的恶性循环不利于推荐系统的良性发展。当系统为用户评分加入少量的噪声数据时，数据失真不明显，不能很好地保护用户的敏感数据，攻击者可能根据背景知识推断出实际的用户信息。因此，本文在群组用户的评分信息中加入适量的噪声并结合 *k*-匿名的思想来保护群组用户的隐私。对比实验 2 显示了加入适量的噪声的隐私保护方法确实会损失部分推荐精度，但为了在推荐精度和隐私保护之间进行平衡，本文认为牺牲一定的推荐精度来实现用户隐私保护的行为是值得的。

3.5.3 群组推荐算法

在面向群组推荐系统中，本文需要考虑群组内所有用户的偏好。目前，主流的面向群组推荐的方法分为推荐融合方法和模型融合方法。模型融合方法是先根据群组内用户的用户偏好模型融合生成群组偏好模型，然后基于群组偏好模型生成群组推荐。推荐融合方法是先利用传统推荐算法对每个群组用户生成推荐，然后将所有群组用户的推荐结果融合得到群组推荐结果。本文采用模型融合的方法对群组进行推荐。对群组内需要进行隐私保护的用户，通过群组敏感偏好保护方

法实现对群组用户敏感偏好的保护后，再将扰动后的群组偏好传输给推荐服务器进行推荐，对没有隐私保护需求的用户，则按照传统的协同过滤算法进行推荐。

3.6　有效性分析和效率分析

算法有效性分析。算法的主要性能消耗在计算群组内用户的主题重要性和寻找群组内与目标用户相似的用户上，但是这部分消耗也是极少的，因为实验的群组用户数量不大，系统中存在的主题数目和群组内用户的偏好主题数目都是极小的，从项目−主题树中搜索用户偏好的效率也是极高的，因为本文对系统的项目−主题分类树进行了一系列的简化，使项目−主题分类树的深度只有三层，这样可以极高地提高算法的运行效率。

隐私保护性能分析。首先，隐私保护算法处理后的群组用户偏好项目已经不是原始的群组用户偏好项目，该群组用户偏好项目过滤了原始群组用户偏好项目中的敏感偏好，推荐系统获取到的就不是原始的群组偏好，有效地解决了 3.1 节给出的审查攻击和评级攻击。其次，根据定义 8 的项目特征相似性计算发现，扰动后的项目特征和原始的项目特征之间具有很高的相似性，说明扰动后的数据能够在保证良好的数据可用性的同时，实现对群组用户的隐私保护。最后，本文利用群组内相似用户的项目偏好进行协同扰动，实现单个用户的评分在群组中不敏感。实验结果表明，本文的隐私保护方法可以在有效地保护群组用户隐私的同时，保证推荐的质量。

4　仿真实验与效用评估

4.1　实验数据集和数据集预处理

本文实验环境是基于 Windows8 平台 64 位系统，其处理器是双核 2.5 GHz。本文算法采用 Java 语言编程实现。本文所使用的数据为学术界研究广泛使用的极具代表性的 MovieLens 数据集，其中包含 943 个用户对于 1 682 个电影的 10 万条评分，且每个用户参与评价的数目不少于 20 条。 数据集都是由 1~5 的整数值组成，数值越大表示用户越喜爱相关的项目。

为了在面向群组推荐的过程中验证本文提出方法的有效性，根据用户的基本信息对数据集中的用户进行群组划分，主要包含如下 3 种划分方式，即性别划分、按年龄划分、按职业划分。具体的划分方法如下。1) 性别：男、女。2) 年龄：小于 20 岁、21 岁～30 岁、31 岁～40 岁。3) 职业：教师、销售、程序员等。为了方便本文所提隐私方法的实现，对数据集进行了预处理。将选取数据集中的电影的标题（title）和电影的类别（genre）这 2 个属性进行项目−主题分类树的构建。具体构建操作是根据 title 和 genre 之间的对应关系，将相对应的项目叶子节点插入主题根节点下，从而实现多层次的项目−主题分类树的构建。

4.2　实验方法

本文将数据集按照 4:1 的方式划分为训练数据集和测试数据集两部分，采用了 RMSE（均方根误差）对群组推荐结果的准确性进行分析。RMSE 在群组推荐场景下的表达形式为

$$\mathrm{RMSE}=\frac{1}{n}\sqrt{\sum_{i=1}^{n}(p_i-r_i)^2} \tag{9}$$

其中，p_i 表示测试数据集中用户对项目 i 的实际评分，r_i 表示推荐系统对测试数据集中项目 i 的预测评分。由定义可知，RMSE 的值越小，则所预测的推荐结果越准确。

4.3　实验结果与分析

4.3.1　对比实验 1

对比实验 1 将群组敏感偏好保护方法（GSPPM）和随机扰动方法进行对比。表 2 给出了在不同群组大小（用户数）和群组分类方法下 2 种扰动方法 RMSE 值的对比。

根据图 5 的实验结果发现，本文提出的群组敏感偏好保护方法的 RMSE 值更小，说明本文方法的推荐准确性更高，并且当使用较细粒度的按职业进行群组划分的方法时，推荐准确性明显高于按性别进行群组划分的方法。

4.3.2　对比实验 2

对比实验 2 是用 RMSE 参数来进行度量，将加入了隐私保护方法后的群组推荐方法和文献[22]中提出的 CF-based 的群组推荐方法进行对比。其中，CF-based 的群组推荐方法采用模型融合的方式进行推荐，先对群组内用户的偏好进行偏好融合后形成群组偏好，然后再利用传统的基于项目的协同过滤推荐算法对群组进行推荐。

图 6 的实验结果表明加入隐私保护方法后，推荐系统的推荐准确性并没有出现较大幅度的失真，说明本文的方法能够在保护群组内用户隐私的同时，实现对群组用户的准确推荐。

表 2　　**不同群组大小和群组分类方法下的 RMSE 值**

群组大小	GSPPM			随机扰动方法		
	性别划分	年龄划分	职业划分	性别划分	年龄划分	职业划分
1	0.997 54	0.984 51	0.983 56	0.998 51	0.985 56	0.985 47
20	0.950 17	0.940 25	0.938 87	0.968 19	0.950 12	0.937 95
30	0.938 63	0.938 71	0.926 84	0.941 02	0.940 19	0.929 76
40	0.934 09	0.930 19	0.921 16	0.940 01	0.939 17	0.924 87
50	0.935 08	0.924 54	0.920 61	0.937 86	0.932 97	0.922 16

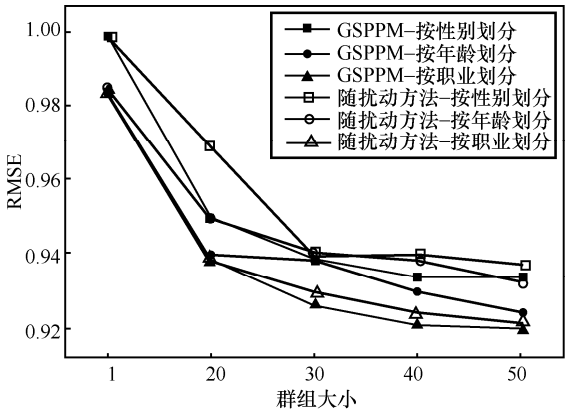

图 5　随机扰动方法和 GSPPM 的对比

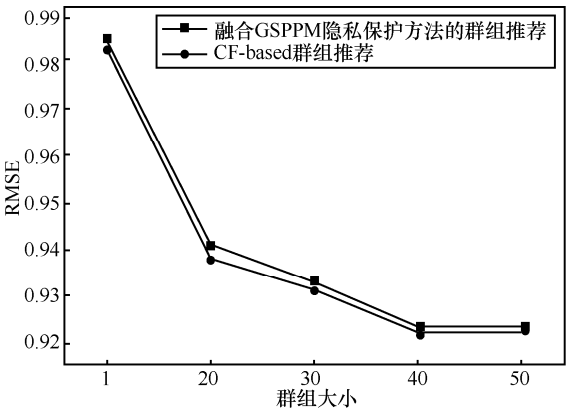

图 6　融合 GSPPM 隐私保护方法的群组推荐和 CF-based 群组推荐的对比

4.3.3　对比实验 3

对比实验 3 验证参数 k 对推荐结果的影响。对比实验 1 的结果表明，当选用职业作为划分群组的标准时，群组推荐结果的准确性明显高于以年龄和性别作为划分群组标准时的准确性，因此本文的实验是以职业进行群组划分，群组的用户数目选择为 20，在此基础上通过调整参数 k 的大小来观察其对推荐准确性的影响。

图 7 的实验结果说明在使用职业作为群组划分的前提下，当相似用户的数量达到群组数量的 $\frac{1}{4}$ 时，推荐准确性相对较高的，当使用的相似用户数量逐渐增大时，对推荐的准确性会有一定的影响，因此本文将相似用户的数量设置为当前群组大小的 $\frac{1}{4}$。

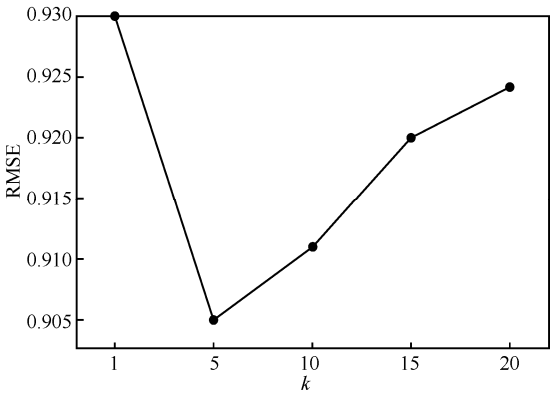

图 7　不同 k 下的 RMSE

根据以上实验表明，本文在引入隐私保护方法之后，推荐准确性没有出现较大的损失，保证推荐系统能够在一定的精度损失范围内，实现有效的推荐和保护群组内用户的敏感信息。在实验性能方面，主要在可信客户端存在一定的时间消耗，但是这样的时间消耗可以保证用户的隐私，本文认为这样的时间消耗是值得的。

5　结束语

为了解决群组推荐中的个性化隐私保护问题，本文提出了基于可信客户端的面向群组推荐的隐私保护框架，在实现面向群组准确推荐的同时，保证群组内用户个性化的隐私保护需求。然而，数据在传输过程可能存在隐私泄露的风险，后续的工作将围绕如何提升数据传输的可靠性与安全性展开研究。

参考文献：

[1] SWEENEY L. *k*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based

Systems, 2002, 10(5): 557-570.
[2] KAPLAN E , GURSOY M E , NERGIZ M E , et al. Known sample attacks on relation preserving data transformations[J]. IEEE Transactions on Dependable and Secure Computing, 2017, 1(32): 101-108.
[3] AHMED K W, MOURI I J, ZAMAN R, et al. A privacy preserving personalized group recommendation framework[C]// International Conference on Advanced Computing. IEEE, 2016: 594-598.
[4] LU Z, SHEN H. A security-assured accuracy-maximised privacy preserving collaborative filtering recommendation algorithm[C]// International Database Engineering and Applications Symposium. ACM, 2015: 72-80.
[5] LI C, PALANISAMY B, JOSHI J. SocialMix: supporting privacy-aware trusted social networking services[C]//IEEE International Conference on Web Services. IEEE, 2016: 115-122.
[6] LIU P, LI Y, SANG Y, et al. Anonymity-based privacy preserving network data publication[C]//Trustcom/BigDataSE/ISPA. IEEE, 2017: 823-828.
[7] TANG Q, WANG J. Privacy-preserving friendship-based recommender systems[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(5): 784-796.
[8] AHMADI M, GHAHFAROKHI B S. Preserving privacy in location based mobile coupon systems using anonymous authentication scheme[C]//International Iranian Society of Cryptology Conference on Information Security and Cryptology. IEEE, 2016: 60-65.
[9] GUO L, ZHANG C, FANG Y. A trust-based privacy-preserving friend recommendation scheme for online social networks[J]. IEEE Transactions on Dependable & Secure Computing, 2015, 12(4): 413-427.
[10] PENG T, LIU Q, WANG G. Enhanced location privacy preserving scheme in location-based services[J]. IEEE Systems Journal, 2014, 11(1): 219-230.
[11] ZHU J , HE P , ZHENG Z , et al. A privacy-preserving QoS prediction framework for Web service recommendation[C]//IEEE International Conference on Web Services. IEEE, 2015: 241-248.
[12] GAI K, QIU M, ZHAO H, et al. Privacy-aware adaptive data encryption strategy of big data in cloud computing[C]// IEEE International Conference on Cyber Security and Cloud Computing. IEEE, 2016: 273-278.
[13] 何明, 常盟盟, 吴小飞. 一种基于差分隐私保护的协同过滤推荐方法[J]. 计算机研究与发展, 2017, 54(7): 1439-1451.
HE M, CHANG M M, WU X F. A collaborative filtering recommendation method based on differential privacy protection [J]. Computer Research and Development, 2017, 54(7): 1439-1451.
[14] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014,37(4): 927-949.
ZHANG X J, MENG X F. Differential privacy protection for data Release and analysis [J]. Journal of Computer Science, 2014, 37(4): 927-949.
[15] 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. Differential privacy protection and its application [J]. Journal of Computer Science, 2014, 37(1): 101-122.
[16] 马鑫迪, 李辉, 马建峰, 等. 轻量级位置感知推荐系统隐私保护框架[J]. 计算机学报, 2017, 40(5): 1017-1030.
MA X D, LI H, MA J F, et al. Privacy protection framework of lightweight location-aware recommendation system [J]. Journal of Computer Science, 2017, 40(5): 1017-1030.
[17] 姜火文, 曾国荪, 马海英. 面向表数据发布隐私保护的贪心聚类匿名方法[J]. 软件学报, 2017, 28(2): 341-351.
JIANG H W, ZENG G S, MA H Y. Greedy clustering anonymous method for privacy protection of tabular data release [J]. Journal of Software, 2017, 28(2): 341-351.
[18] YANG M, ZHU T, MA L, et al. Privacy preserving collaborative filtering via the Johnson-Lindenstrauss transform[C]//Trustcom/BigDataSE/ICESS. IEEE, 2017: 417-424.
[19] 刘海, 李兴华, 雒彬, 等. 基于区块链的分布式 *K* 匿名位置隐私保护方案[J].计算机学报,2019,42(5):942-960.
LIU H, LI X H, LUO B, et al. Distributed *K* anonymous location privacy protection scheme based on block chain [J]. Journal of Computer Science, 2019, 42(5): 942-960.
[20] 林荣智, 苗耀锋. 基于用户项目特征分组的隐私保护算法[J]. 沈阳工业大学学报, 2018, 40(6): 670-675.
LIN R Z, MIAO Y F. Privacy protection algorithm based on user project feature grouping [J]. Journal of Shenyang University of Technology, 2008, 40(6): 670-675.
[21] WANG P, YANG J, ZHANG J. A strategy toward collaborative filter recommended location service for privacy protection[J]. Sensors, 2018, 18(5):1522-1541.
[22] LUO Z, CHEN Z. A privacy preserving group recommender based on cooperative perturbation[C]// International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE Computer Society, 2014:106-111.
[23] SHANG S, HUI Y, HUI P, et al. Privacy preserving recommendation system based on groups[J]. Computer Science,2013, 5(1): 219-246.
[24] WU Z, LI G, LIU Q, et al. Covering the sensitive subjects to protect personal privacy in personalized recommendation[J]. IEEE Transactions on Services Computing, 2016, 11(3):493-506.
[25] 张玉洁，杜雨露，孟祥武.组推荐系统及其应用研究[J].计算机学报，2016, 39(4): 746-760.
ZHANG Y J, DU Y L, MENG X W. Group recommendation system and its application research [J]. Journal of Computer Science, 2016, 39(4): 746-760.
[26] HE Y, ZHANG K, WANG H, et al. Impact factor-based group recommendation scheme with privacy preservation in MSNs[C]//2017 IEEE International Conference on Communications. IEEE, 2017.
[27] YIN C, SHI L, SUN R, et al. Improved collaborative filtering recommendation algorithm based on differential privacy protection[J]. The Journal of Super Computing, 2019(7):1-14.
[28] 耿魁, 万盛, 李凤华, 等. 基于隐私匹配的服务代理发现方法[J]. 通信学报, 2016, 37(8):136-143.
GENG K, WAN S, LI F H, et al. Service agent discovery method based on privacy matching [J]. Journal on Communications, 2016, 37(8): 136-143.
[29] 张少波, 刘琴, 王国军. 基于位置混淆的轨迹隐私保护方法[J]. 通信学报, 2018, 39(7):85-95.
ZHANG S B, LIU Q, WANG G J. Trajectory privacy protection method based on location confusion [J]. Journal on Communications, 2018, 39(7): 85-95.
[30] 王涛春, 刘盈, 金鑫, 等. 群智感知中基于 *k*-匿名的位置及数据隐私保护方法研究[J]. 通信学报, 2018, 39(S1):176-184.
WANG T C, LIU Y, JIN X, et al. Research on *k*-anonymise-based location and data privacy protection methods in swarm intelligence perception [J]. Journal on Communications, 2018, 39(S1): 176-184.

[作者简介]

王海艳（1974− ），女，江苏东台人，博士，南京邮电大学教授，主要研究方向为服务计算、可信计算、大数据应用与云计算技术、隐私保护技术等。

陆金祥（1993− ），男，江苏姜堰人，南京邮电大学硕士生，主要研究方向为推荐系统和隐私保护技术。