计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900152

融合用户属性与项目流行度的用户冷启动推荐模型

韩立锋　陈　莉
西北大学信息科学与技术学院　西安 710127
(lifeng_han@126.com)

摘　要　冷启动一直是推荐系统领域中被密切关注的问题,针对新注册用户冷启动的问题,文中提出了一种融合用户人口统计学信息与项目流行的推荐模型。首先对训练集用户进行聚类,将训练集用户划分为若干类。然后计算新用户与所属类别中其他用户之间的距离,选择其近邻用户集,在评分计算时综合考虑项目流行度对推荐效果的影响,进而为目标用户推送感兴趣的节目。最后在经典推荐系统数据集中对所提模型进行验证。实验结果表明,该模型明显优于传统协同过滤算法,并在一定程度上解决了冷启动问题。

关键词:推荐系统;用户冷启动;社会统计学信息;协同过滤;项目流行度
中图法分类号　TP391

User Cold Start Recommendation Model Integrating User Attributes and Item Popularity

HAN Li-feng and CHEN Li
School of Information Science & Technology, Northwest University, Xi'an 710127, China

Abstract　Cold start has always been a closely watched issue in the field of recommendation systems. Aiming at the problem of cold start for newly registered users, this paper proposes a recommendation model that integrates user demographic information and item popularity. The training set users are divided into several categories by clustering the training set users, and then the distance between the new user and other users in the category is calculated, and the neighboring user set is selected. When calculating the score, we consider comprehensively the impact of popularity, and then push the programs of interest to target users. Finally, the proposed model is verified on the classic recommendation system data set. The results show that the model is significantly better than the traditional collaborative filtering algorithm and has a certain mitigation effect on the cold start problem.

Keywords　Recommended system, User cold star, Social statistics information, Collaborative filtering, Item popularity

互联网亦或移动互联网行业得到了前所未有的发展,由此产生了包括购物数据在内的海量信息,信息过载成为困扰信息制造者和信息消费者的一大难题[1]。个性化推荐,旨在按照个人消费兴趣和产品或服务的固有特性预测用户的消费偏好,进而为用户推荐其可能喜欢的内容,这有利于在一定程度上解决信息过载等一系列问题[2-3]。

协同过滤算法为个性化推荐中最广泛应用的算法,其假设使用者会对与他历史行为相近的其他使用者喜爱度高的项目感兴趣。该算法根据用户的各种日志数据,包括用户点击、收藏、购买等用户行为,计算系统中用户之间的关联,获取用户兴趣模型,从而给用户推荐邻居可能感兴趣的项目集合,这在一定程度上表现出了较高的准确性。

然而,推荐系统在实际应用中经常遭受冷启动问题的困扰。冷启动问题一直被研究者密切关注,一般根据场景可以分为新项目冷启动问题及新用户冷启动问题[4]。基于内容的个性化推荐,通过比对项目的属性以及一些相关信息来计算项目和项目之间的相似性,从而使得项目冷启动问题得到缓解。相比而言,新用户问题更难处理,新用户指在系统中拥有极少评分或者没有评分、标签等相关数据的用户[5]。对于新用户,系统获取到的信息有限,只有年龄、性别、职业等一些普遍的社会统计学相关信息,这在无形之中给个性化推荐任务带来了很大挑战。

针对新用户冷启动问题,国内外学者进行了大量研究,早期的冷启动研究中主要的解决策略有随机推荐策略、平均值策略、流行度策略、信息熵策略等,它们存在覆盖率低等缺点,降低了用户的个性化体验[6]。后来,学者们对传统推荐方法进行了相应的改进,提出了混合推荐策略、融合多源数据推荐策略、动态情景推荐策略等。Chen 等[7]针对微博类 App 中存在的用户冷启动问题,提出了一种基于混合聚类的个性化推荐算法,该算法不仅提高了推荐准确度,一定程度地解决了冷启动问题,同时提升了个性化推荐的多样性结果。Yang 等[8]通过上下文信息。一定程度上解决了新闻推荐系统中新

到稿日期:2020-09-21　返修日期:2020-10-18　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:陕西省重点研发计划项目(2019ZDLGY10-01)
This work was supported by the Key R&D Program of Shaanxi Province(2019ZDLGY10-01).
通信作者:陈莉(chenli@nwu.edu.cn)

注册用户因缺失新闻浏览历史行为所导致的冷启动问题，用户满意度有着较为明显的提高。Sedhain 等[9]在协同过滤算法中加入用户的人口统计数据（性别、年龄、位置）以及社交网络信息（Facebook 好友或好友链接），证明了使用社交网络信息进行冷启动推荐有着巨大的预测能力。Zhao 等[10]在矩阵分解的基础上，利用社交网络中的信任关系作为辅助数据，提出了一种社交网络中的系统矩阵分解（a recommendation system in social network with Feature Transfer and Probabilistic Matrix Factorization，RTMF）技术，该技术可以有效使用辅助数据并对现有推荐算法进行明显优化。Tang 等[11]提出了构建“元情景”的应对策略，对不同的情景信息进行了相关分类，当新用户进入系统后，通过让新用户与系统进行交互来为用户反馈适当的情景作为其上下文环境，由此产生推荐内容。

上述算法可在一定程度上解决冷启动问题，但仍存在一定的局限性，如在加入用户人口统计学数据集中的社交网络数据的混合类算法中，需要大规模地训练数据，其稳定性较差，同时推荐的可解释性也有待加强；在融合多源数据的推荐中，需要学习更多的参数，且推荐效果依赖于历史数据；在动态情景推荐中，需要多种实时数据，且用户覆盖率较低。

近年来，深度学习技术已经在人工智能的各领域取得了突破性进展，为推荐系统的发展带来了新的机遇。Lei 等[12]提出了一种深度协作神经网络模型（Deep Cooperative Neural Network，DeepCoNN），提高了推荐质量。Chen 等[13]提出了一种 LP-DSA（Location-aware Personalized News Recommendation with Deep Semantic Analysis）模型，提高了新闻推荐性能。Gong 等[14]提出了一种基于注意力的卷积神经网络（CNN）来进行微博中的 Hashtag 推荐。Lei 等[15]基于深度学习方法研究了图像推荐的问题。Aaron 等[16]研究了如何利用深度学习模型来解决音乐推荐系统中的冷启动问题。Wang 等[17]提出了一种基于注意力的 LSTM 来进行 Hashtag 推荐。

尽管深度学习技术相比传统个性化推荐方法，能够在学习用户和项目隐表示过程中具有更好的性能，但此类方法仍然无法改变数据稀疏、冷启动等一系列问题。特别是对于一个新型场景下的个性化系统，因为缺乏相对丰富的大规模数据，只能依赖于系统沉淀的用户注册信息对新用户进行个性化推荐，但是由于单一的用户注册信息的粒度较粗，特别是针对流行度较高的商品，系统仍不能在真正意义上满足用户的个性化需求。为此，本文提出了基于用户属性信息与项目流行度相结合的推荐算法。首先使用用户的社会统计学信息进行聚类，然后在各个类别中找出新用户与其他用户，特别是活跃用户的相似度，记录与新用户最相似的用户的历史评分，再融合项目流行度进行评分修正，将打分最高的 Top N 个项目推荐给新注册的用户。

1 相关工作

1.1 聚类

聚类指根据某种规则（如距离准则）将某个数据集切分成不同的簇，使得同一个簇内的数据尽可能相似，即聚类后同一类数据尽可能聚集到同一簇，而不同类的数据尽可能分离。

假设 $X=\{x_1,x_2,\cdots,x_n\}$ 是待分析的样本集合。X 中的每个样本通常使用有限个属性值来度量。聚类的目的就是按照各样本间的距离计算每个样本所对应的不同类别，最终将 $x_1,x_2,\cdots,x_n$ 划分成 k 个不相交的模式子集 $x_1,x_2,\cdots,x_k$，并满足：1）每个样本只属于某一类；2）每个子类都非空。

为了实现对数据对象有效聚类，国内外专家学者提出了很多改进算法，Han 等对聚类算法进行了分类，总结归纳了聚类算法的 5 种类型，分别为基于划分的聚类、基于层次的聚类、基于密度的聚类、基于网络的聚类、基于模型的聚类[11]。其中，最常用的聚类算法有 K-means[18]，STING[19]（Statistical Information Grid），CLIQUE[20]（Clustering in quest）和 CURE[21]（Clustering Using Representative）。

K-means 算法是目前应用较为广泛的聚类算法。在个性化推荐场景中，K-means 算法能够通过对用户社会统计学数据进行计算，来划分出若干不同类别。

1.2 基于用户的协同过滤

协同过滤算法基于“群体智慧”的思想对信息进行建模，传统的协同过滤算法主要包括基于用户的协同过滤（user-based collaborative filtering，UBCF）、基于项目的协同过滤（item-based collaborative filtering，IBCF）两大类。

UBCF 的核心是用户，通过收集用户感兴趣的信息（如用户对于项目的评分等）进行建模，进而找到目标用户的近邻用户，最终给目标用户产生推荐结果集。一般包括以下 3 个步骤。

（1）数据表示。在 UBCF 中，评分信息是一个 $m*n$ 阶矩阵，第 i 行第 j 列的值代表用户 i 对项目 j 的评分，评分值一般是 1～5 之间的整数。用户项目评分数据矩阵如下：

	$Item_1$	$\cdots$	$Item_k$	$\cdots$	$Item_n$
$User_1$	$R_{1,1}$	$\cdots$	$R_{1,k}$	$\cdots$	$R_{1,n}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$User_j$	$R_{j,1}$	$\cdots$	$R_{j,k}$	$\cdots$	$R_{j,n}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$User_m$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$R_{m,n}$

（2）寻找最近邻用户。本阶段主要通过计算目标用户 u 与其他用户之间的相似度（记为 $sim(u,v)$），来找出与目标用户最相似的“最近邻”用户集。在进行相似度计算时，一般有余弦相似度、相关相似度等多种计算方式。

1）余弦相似性。用户 u,v 之间的相似度就是两个向量 $\vec{u},\vec{v}$ 夹角的余弦值，余弦值越大，则表示两个用户越相似。余弦相似性的计算公式如下：

$$sim(u,v)=\cos(\vec{u},\vec{v})=\frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\|\vec{v}\|} \tag{1}$$

2）相关相似性。设用户 u,v 同时评过分的集合为 I，则用户 u,v 之间的 $sim(u,v)$ 可通过 Pearson 系数进行计算。Pearson 相关系数主要用于衡量变量与变量之间的线性关系，目前也是推荐系统使用较为广泛的相似度度量方法。其计算公式如下：

$$sim(u,v)=\frac{\sum_{i\in I}(R_{u,i}-\bar{R}_u)(R_{v,i}-\bar{R}_v)}{\sqrt{\sum_{i\in I}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{i\in I}(R_{v,i}-\bar{R}_v)^2}} \tag{2}$$

通过计算目标用户与其他用户之间的相似性，将与目标用户最为相似的其他用户作为目标用户的最近邻用户集。

(3)推荐生成。设用户 u 的最近邻用户集用 U 表示，用户 u 对项目 i 的预测评分为 $P_{u,i}$，可通过用户 u 对最近邻集合 U 中用户的评分加权相似度的值得到。其计算公式如下：

$$P_{u,i}=\overline{R}_u+\frac{\sum_{v\in U} sim(u,v)\cdot(R_{v,i}-\overline{R}_v)}{\sum_{v\in U}|sim(u,v)|} \tag{3}$$

其中，$sim(u,v)$表示用户之间的相似性，$R_{v,i}$表示用户 v 对项目 i 的评分，$\overline{R}_u$ 和 $\overline{R}_v$ 分别表示用户 u、用户 v 对所有项目的平均评分。通过如上计算，选择评分最高的 Top N 个项目推荐给目标用户。

2 融合用户属性与项目流行度的纯冷启动个性化推荐模型

传统的协同过滤技术根据用户历史的评分信息，通过寻找近邻用户集，为目标用户产生推荐。对于一个新用户，在没有任何历史评分信息的情况下，由于无法通过传统协同过滤技术进行相似性计算来确定近邻用户，因此无法为新用户提供项目推荐。

针对此类问题，本文提出了一个新的基于人口统计学特征信息的个性化推荐模型。人口统计学特征包括人的性别、年龄、职业、健康状况、婚否、学历、收入等个人信息。但受到个人隐私问题的影响，我们无法获取个人的全部人口统计学特征信息。而人口统计学特征信息中相对重要的性别、年龄、职业等信息，为新系统注册的必要信息，获取相对容易，只需要从系统日志文件中进行采集即可。同时，实践证明，在电影个性化推荐中，我们经常会有这样的体验：不同年龄、性别、职业的用户所喜好的电影类型存在较为明显的差异。因此，在电影个性化推荐系统中，可以使用用户的人口统计学特征挖掘用户与用户之间的潜在关联，即具有相似年龄、性别、职业等特征的用户极有可能具有相同的购买兴趣[22-25]。对于新注册用户，即冷用户，尽管缺少评分信息，但是可以根据新注册用户的年龄、性别、职业等人口统计学信息找到新用户的相似用户，然后通过这些相似用户的喜好，给目标用户推荐其喜欢的电影。本文算法的框图如图 1 所示。

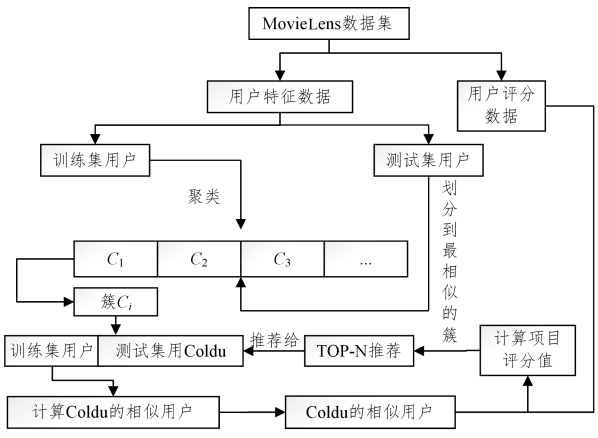

图 1 本文算法的框图

Fig. 1 Block diagram of our algorithm

图 1 中，首先根据 MovieLens 数据集，将用户划分为训练集用户及测试集用户。对于训练集用户，该数据集中包括其评价的电影信息及用户个人信息；而对于测试集用户，该数据集中只包括其个人人口统计学信息。然后，根据人口统计学特征，即年龄、性别、职位等信息，对训练集用户进行聚类，将其划分为若干簇。接着，对于测试集用户，计算出目标用户所处的类别(簇)。在该类中找出该目标用户的近邻用户，最后将评分最高的 N 项电影推荐给目标用户。

2.1 用户聚类

对现有用户的分类是基于“物以类聚，人以群分”的基本假设，根据用户的属性将相似用户划分为同一簇，而不同用户则被划分在不同的簇中。本文首先对用户人口统计学信息(见表 1)进行预处理，然后使用 K-means 算法，针对用户人口统计学信息对用户进行聚类分析，最后得到不同的类型(簇)。

表 1 用户人口统计学信息

Table 1 User demographic information

用户 ID	年龄	性别	职业
1	24	M	Technician
2	53	F	Other
3	23	M	Writer
...	...	...	...
943	22	M	Student

在处理用户人口统计学信息时，首先可以根据年龄段对年龄信息进行处理，根据生物学年龄划分标准并结合本文数据集的特点，可以将用户年龄划分为 7 组数据(用数值 1－7 代替)：年龄小于 18，年龄为 18～24，年龄为 25～34，年龄为 35～44，年龄为 45～49，年龄为 50～55，年龄大于或等于 56。

对于职业信息，原 Users 表中有 21 个分类，根据我国职业分类标准，按照企事业负责人、专业技术人员、服务业商业、文娱从事着、教育行业、家政以及其他分为 7 类，使用数字 1－7 代替。

对于性别，男(M)、女(F)分别使用数字 1，0 代替。处理后的人口统计学信息如表 2 所列。

表 2 处理后的用户人口统计学信息

Table 2 Processed user demographic information

用户 ID	年龄	性别	职业
1	2	1	2
2	6	0	7
3	2	1	4
...	...	...	...
943	2	1	5

对处理后的数据进行归一化处理，计算式如下：

$$x_{scale}=\frac{x-x_{min}}{x_{max}-x_{min}} \tag{4}$$

对于进行归一化处理后的用户人口统计学信息，可以对用户进行 K-means 聚类分析，算法描述如算法 1 所示。

算法 1 K-means 算法

输入：归一化后的用户人口统计学信息表，聚类个数 K

输出：k 个聚类

1. 将 k 个聚类 $c_1, c_2, \cdots, c_k$ 初始化为空，记为集合 $C=\{c_1, c_2, \cdots, c_k\}$；

2. 从处理后的用户人口统计学信息表中得到 n 个用户，记为集合 $U=\{u_1, u_2, \cdots, u_n\}$；

3. 从 n 个用户中随机选取 K 个用户充当各个簇的中心点{μ_1，μ_2，…，μ_k}；
4. 计算其他用户与簇中心用户之间的距离 dist($u^{(i)}$，μ_j)，依次计算每个用户 u 与 K 个聚类中心 μ_k 的相似度，将每个用户划入最近的簇 c_k 中，$u^{(i)} \in \mu_{rearest}$；
5. 根据簇中已有用户，重新计算簇中心；
6. 重复步骤 2、步骤 3，直到聚类不再发生改变为止。

2.2 推荐生成

根据算法 1 对用户进行聚类，原有用户被划分为不同的簇。对于新用户，首先计算其与各簇中心点的距离，距离较近的就为该用户的类别。在确定类别后，通过计算新用户与该簇中所有用户之间的距离，找到最近邻用户。根据最近邻用户对项目的评分，计算出新用户的项目评分，从而给目标用户推荐 Top N 个项目集合。结合用户人口统计学的 UBCF 算法如算法 2 所示。

算法 2 结合用户人口统计学的 UBCF 算法

输入：用户 u(用户 ID、年龄、性别、职业)、用户-项目评分矩阵 **R**，参数 α 和 β，近邻用户数目 K′，推荐产品数量 N

输出：用户 u 的 Top N 推荐列表

1. 计算用户 u 与聚类后各簇中心点 μ 的距离。距离最近的即为该用户的所属类别，并记录所属类别 C。
2. 根据式(5)计算目标用户 u 与 c 中每个用户 u′之间的相似性。其中，S(u，u′)为性别相似性，A(u，u′)为年龄相似性，O(u，u′)为职位相似性，参数 α，β，(1−α−β)分别代表性别、年龄、职位相似性的权重。
3. 选取相似性值最大的 k′个用户作为目标用户 u 的邻居用户，记为 $N_i=\{i_1, i_2, \cdots, i_{k'}\}$。同时，合并所有邻居 N_i 的项目集合 C。
4. 对于所有项目 j∈C，根据式(6)计算用户 u 对项目 j 的评分。其中，P_{ix} 为目标用户 i 对项目 x 的预测评分，$\bar{R}_X$ 为项目 x 的平均评分，M_k 为用户 i 的近邻集，$\bar{R}_j$ 为用户 j 的近邻用户的平均评分，R_{jx} 为用户 j 对项目 x 的评分，sim(i，j)为用户之间的相似性。
5. 将 C 中所有项目评分按照从大到小的顺序进行排序，将分值最高的前 N 个项目作为用户 u 的推荐列表。

$$sim(u,u')=\alpha S(u,u')+\beta A(u,u')+(1-\alpha-\beta)O(u,u') \tag{5}$$

$$P_{ix}=\bar{R}_x+\frac{\sum\limits_{j\in M_{k'}} sim(i,j)\cdot(R_{jx}-\bar{R}_j)}{\sum\limits_{j\in M_{k'}}|sim(i,j)|} \tag{6}$$

2.3 融合项目流行度的个性化推荐修正

算法 2 弥补了新用户没有项目评分的不足，并能根据新用户人口统计学信息找到其近邻用户，根据近邻用户所评分的项目，推测新用户对未知项目的评分。但是，在此过程中，对于任意项目，其权重始终保持不变，这并不合理。因为从社会学角度来看，所有事物都遵循着“马太效应”，即强者更强，弱者更弱。在很多电商网站，商品也服从这种分布特征，即越是热门的商品越容易被更多人购买，相比之下，越是冷门的商品则越不容易被人关注，这一现象也被称为“长尾分布”。显然这并不符合个性化推荐的初衷，因为对于流行度高的商品的购买行为，很难体现用户的“个性化”需求，相比之下，对于一些冷门商品的购买，则更能代表用户真正的喜好。因此，在个性化推荐中，应该充分体现用户个性化兴趣的数据，同时削弱用户非个性化兴趣的数据。为此，在算法 2 中，本文引入与项目流行度相关的权重因子：

$$normPopItem_i=\frac{popItem_i-\min Pop}{\max Pop-\min Pop} \tag{7}$$

$$weightItem_i=1-normPopItem_i \tag{8}$$

其中，$normPopItem_i$ 代表项目流行度的最终值，为了使其取值范围保持在[0，1]之间，可以对其进行归一化处理；$popItem_i$ 代表项目 i 的流行度，其数值等于项目 i 的所有被评分数量。max Pop 代表最流行项目的被评分数量，min Pop 代表最不流行项目的被评分数量。而 $weightItem_i$ 则与 $normPopItem_i$ 负相关，即项目流行度越高，权重因子就越小，反之亦然，权重因子取值范围也保持在[0，1]之间。

通过计算得到流行度权重因子，可将式(6)进行如下修正：

$$P_{ix}=\bar{R}_x+\frac{\sum\limits_{j\in M_{k'}} sim(i,j)\cdot(R_{jx}\cdot weightItem_x-\bar{R}_j)}{\sum\limits_{j\in M_{k'}}|sim(i,j)|} \tag{9}$$

修正后的评分预测值考虑了项目流行度的影响。然后，将评分按照从大到小的顺序进行排序，将分值最高的前 N 个项目作为新用户 u 的推荐列表，完成推荐。

3 实验结果及分析

3.1 实验数据集与衡量标准

由于推荐系统领域中纯冷启动用户问题较为新颖，目前相关公开数据集并不多，同时由于我们解决此类问题时需要使用人口统计学特征，因此本文采用美国明尼苏达计算机学院的 GroupLens 小组提供的 MovieLens 数据集，该数据集包括用户的性别、年龄、职业等人口统计学特征数据。

此外，为了真实模拟出新用户，本文在仿真实验中将 80%的用户数据作为训练集，将 20%的用户数据作为测试集。对于训练集用户，数据包括用户个人人口统计学信息及所评分电影信息，对于测试集，数据只包括用户个人统计学信息。

本文实验的软件环境如下：Windows10-64bits，Anaconda3、Python3.7。硬件环境如下：CPU 为英特尔 Core i7-8750H @ 2.20GHz 六核、内存为 16GB。

在推荐系统中，能够衡量推荐质量的标准很多，如精度、召回率、F1 度量等。本文采用推荐精度、平均绝对误差 MAE 对推荐的准确性进行评价。推荐精度指给目标用户推荐的产品在用户所喜欢的产品列表中所占的比例，如式(10)所示：

$$Precision=\frac{\sum\limits_{u\in U} hit_u}{N*|U|} \tag{10}$$

其中，hit_u 表示给目标用户推荐的产品中用户喜欢的产品数量，N 表示推荐的产品总数量。平均绝对误差衡量的是预测评分与实际评分之间的差距，MAE 越小，误差越小，推荐准确性就越高。其公式如下：

$$MAE=\frac{\sum\limits_{i=1}^{N}|P_{ui}-R_{ui}|}{N} \tag{11}$$

其中，P_{ui} 表示目标用户 u 对项目 i 的预测评分，而 R_{ui} 表示目标用户 u 对项目 i 的实际评分。

3.2 参数设置及结果分析

本文算法模型通过融合用户人口统计学特征与项目流行度对新用户进行推荐。在此过程中，首先对用户进行聚类，然后通过新用户个人人口统计学特征与已有用户进行相似度计

算，再根据此结果，结合项目流行度，选取相似度最大的若干近邻用户，通过相关计算对目标用户进行评分预测。在此过程中，影响最终算法准确性的参数有：聚类数目 k、性别属性对相似度影响的权重 α、年龄属性对相似度影响的权重 β、近邻用户数目 k'。

本文主要验证以上各参数对推荐精度的影响，同时将本文算法与其他协同过滤算法进行比较并加以分析。

(1)聚类数目 k 对推荐精度的影响

聚类数目 k 对算法有很好的促进作用。如果 k 值太小，如当 $k=1$ 时所有用户被当作一类，缺乏个性；如果 k 值过大，如类型数量等于用户数量，则等同于每个用户都是一个规定类型，缺乏共性。因此，k 在一定程度上体现了用户群体的兴趣偏好。由图 2 可知，当 $k=12$ 时推荐精度最高。

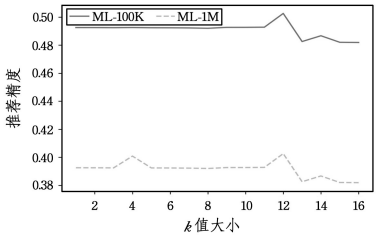

图 2 不同 k 值对精度的影响

Fig. 2 Impact of different k on accuracy

(2)性别和年龄参数 α 和 β 对推荐精度的影响

由图 3—图 5 可以看出，不同性别、不同年龄、不同职业的用户对电影评分有着较大的差异。可见，用户与用户之间，由于年龄、性别、职业等属性的不同，喜好也有较大差异。若要进行冷用户个性化推荐，可以首先计算目标用户与其他用户由于属性不同所造成的相似度差异。

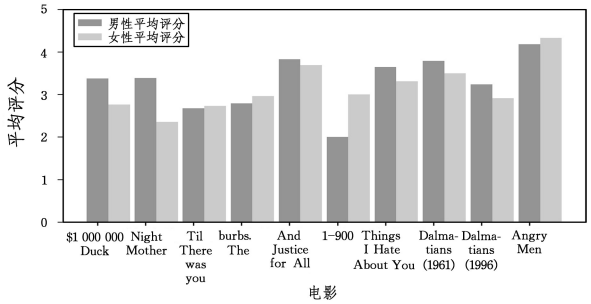

图 3 不同性别用户对 10 部电影平均评分的对比

Fig. 3 Comparison of average ratings of ten movies by users of different genders

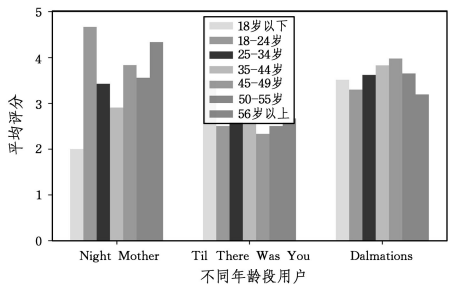

图 4 不同年龄用户电影平均评分的对比

Fig. 4 Comparison of average movie ratings of users of different ages

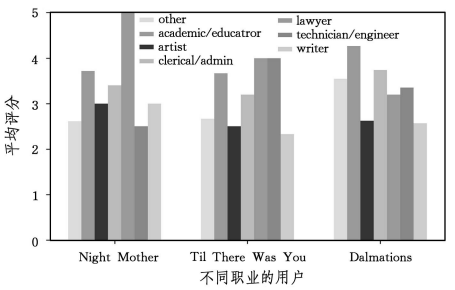

图 5 不同职业用户电影平均评分的对比

Fig. 5 Comparison of average ratings of movies by users of different professions

本文在计算目标用户 u 与聚类后各簇中其他用户的相似度时，引入了权重因子 α，β，$(1-\alpha-\beta)$，分别表示年龄、性别、职位对相似度的影响程度。现分别取 7 组权重值，分别为 $\{0.6,0.1,0.3\}$，$\{0.1,0.6,0.3\}$，$\{0.3,0.1,0.6\}$，$\{0.5,0.1,0.4\}$，$\{0.1,0.5,0.5\}$，$\{0.4,0.1,0.5\}$，$\{0.34,0.33,0.33\}$。表 3 列出了各权重值对精度的影响。

由表 3 可以看到，无论是 ML-100K 数据集，还是 ML-1M 数据集，当年龄、性别、职业权重为 $\{0.6,0.1,0.3\}$ 时，推荐精度最高。其次，在各权重组合中也能明显看出，当年龄权重较大时，推荐精度也较高。这说明年龄相近的用户具有相似产品喜好的可能性更大。而性别和职位等用户统计学特征并没有明显影响推荐精度，这在一定程度上也表明，用户对产品的喜好程度与性别、职位等特征并没有特别大的关联。

表 3 社会统计学信息对精度的影响

Table 3 Impact of social statistics information on accuracy

	ML-100k	ML-1M
$\{0.6,0.1,0.3\}$	**0.5026**	**0.4135**
$\{0.1,0.6,0.3\}$	0.4893	0.4002
$\{0.3,0.1,0.6\}$	0.4956	0.4117
$\{0.5,0.1,0.4\}$	0.4771	0.3882
$\{0.1,0.5,0.4\}$	0.4558	0.3633
$\{0.4,0.1,0.5\}$	0.4726	0.3857
$\{0.34,0.33,0.33\}$	0.4957	0.4076

(3)近邻用户数目 k' 对推荐精度的影响

近邻数目的选取对推荐精度也有很重要的影响。假设聚类数目为 12，年龄、性别、职位各影响参数分别为 $\{0.6,0.1,0.3\}$，推荐产品数目为 10，通过实验观察近邻数目为 5，10，15，20 时推荐精度的变化情况，实验结果如图 6 所示。

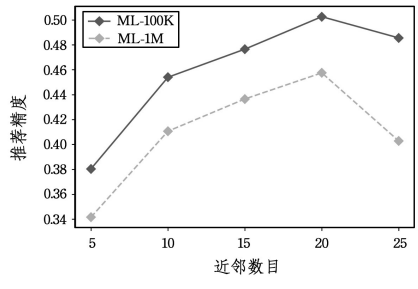

图 6 近邻数目对精度的影响

Fig. 6 Impact of number of neighbors on accuracy

从图 6 中可以看到，当近邻数目增大时，推荐精度刚开始

呈现上升趋势，当近邻数目达到 20 时，推荐精度达到最高值。此后，推荐精度有下降趋势。可见，在实际个性化推荐中，并不是近邻数目越大越好。此外，随着近邻数目的不断增加，系统所花费的时间也不断增长，而当近邻数目达到临界值后，推荐质量也在逐渐下降。因此，个性化推荐应该从推荐质量和时间消耗两个维度进行均衡，以找到最佳近邻数目。

(4)本文算法与传统协同过滤算法的性能比较

为了检验算法的有效性，本实验将本文算法与传统协同过滤算法 UBCF 和 IBCF、基于用户兴趣和关联项目的算法 CF-UICI[26]、一种应用 CF 缓解推荐系统新用户冷启动的方法[27]、MIPFGWC-CS 算法[28]、NHSM 算法[29]进行比较。

本文算法将近邻用户数量从 5 变化为 50，步长为 5。本文使用的数据集为 ML-100K，本文算法中聚类数目取值 12，年龄、性别、职位对应的影响因子为{0.6，0.1，0.3}。图 7 给出了随着近邻用户数量变化，本文算法和其他对比算法的 *MAE* 值的变化。

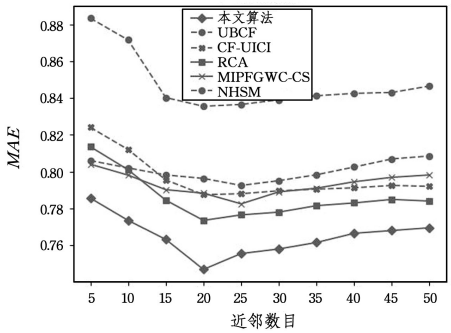

图 7 不同算法的性能比较

Fig. 7 Performance comparison of different algorithms

从图 7 可以看出，相比传统的 UBCF 及 IBCF，以及基于用户兴趣和关联项目的协同过滤，本文提出的个性化推荐模型取得了较低的 *MAE* 值，这表明本文模型有较高的推荐准确度和质量，同时也说明了本文模型对于缓解冷启动问题具有一定的作用。

但不难看出，本文算法只使用了人口统计学特征信息中的年龄、性别、职业 3 个维度作为目标用户相似用户计算的依据，未能使用更加全面有效的用户信息，如用户学历、收入等更能代表用户偏好信息的特征，因此在相似用户计算中也会存在一些偏差。本文算法具有一定的扩展性，后期可以在不影响用户隐私的前提下，通过收集更多用户的详细信息来为用户进行更精准的个性化推荐。因此，本文算法依然具有可优化空间。

结束语 针对新用户冷启动问题，本文考虑融入人口统计数据，并通过聚类、相似度计算等方式为目标用户找到合适的近邻用户集。在对目标用户进行项目推荐时，本文还考虑了项目流行度的影响。实验结果表明，本文提出的融合用户属性信息与项目流行度的个性化推荐模型在一定程度上提高了推荐准确性，一定程度地解决了冷启动问题。考虑到现有推荐数据公开数据集中用户属性的单一性，在未来的工作中，将充分考虑社交网络等跨领域信息，从而找到新注册用户的真正近邻集，为新注册用户推荐更加符合其喜好的项目，提高推荐的准确性。

参 考 文 献

[1] EDMUNDS A，MORRIS A. The problem of information overload in business organisations：a review of the literature[J]. International Journal of Information Management，2000，20(1)：17-28.

[2] RESNICK P，VARIAN H R. Recommender systems[J]. Communications of the ACM，1997，40(3)：56-58.

[3] VERBERT K，MANOUSELIS N，OCHOA X，et al. Context-aware recommender systems for learning：a survey and future challenges[J]. IEEE Transactions on Learning Technologies，2012，5(4)：318-335.

[4] 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 and Data Engineering，2005，17(6)：734-749.

[5] WANG J，ZHAO H Y，CHEN Q K，et al. Representative Item Selection for Cold Start Users[J]. Journal of Chinese Computer Systems，2019，40(8)：1589-1594.

[6] DUAN D K，FU X F. Research on user cold start problem in hybrid collaborative filtering algorithm[J]. Computer Engineering and Applications，2017，53(21)：151-156.

[7] CHEN K H，HAN P P，WU J. User Clustering Based Social Network Recommendation[J]. Hinese Journal of Computers，2013，36(2)：349-359.

[8] YANG X M，SUN Y，WANG M J，et al. Research on the Problem of User Cold-start in News Recommendation Systems[J]. Journal of Chinese Computer Systems，2016，37(3)：479-482.

[9] SEDHAIN S，SANNER S，BRAZIUNAS D D，et al. Social Collaborative Filtering for Cold-start Recommendations[C]// Proceedings of the 8th ACM Conference on Recommender systems. New York：ACM Press，2014：345-348.

[10] ZHAO Z L，WANG C D，WAN Y Y，et al. FTMF：Recommendation in social network with Feature Transfer and Probabilistic Matrix Factorization[C]// International Joint Conference on Neural Networks (IJCNN). Piscataway：IEEE Press，2016：847-854.

[11] TANG L，JIANG Y X，LI L，et al. Ensemble contextual bandits for personalized recommendation[C]// Proceedings of the 8th ACM Conference on Recommender Systems. New York：ACM Press，2014：73-80.

[12] LEI Z，NOROOZI V，YU P S. Joint Deep Modeling of Users and Items Using Reviews for Recommendation[C]// Tenth Acm International Conference on Web Search & Data Mining. New York：ACM Press，2017：425-434.

[13] CHEN C，MENG X，XU Z，et al. Location-aware Personalized News Recommendation with Deep Semantic Analysis[J]. IEEE Access，2017，5：1624-1638.

[14] GONG Y，ZHANG Q. Hashtag Recommendation Using Atten-

tion-Based Convolutional Neural Network[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 2782-2788.

[15] LEI C, DONG L, LI W, et al. Comparative Deep Learning of Hybrid Representations for Image Recommendations [C]// Computer Vision & Pattern Recognition. Piscataway: IEEE Press, 2016: 2545-2553.

[16] AARON V, DIELEMAN S, SCHRAUWEN B. Deep content-based music recommendation[J]. Advances in Neural Information Processing Systems, 2013, 26: 2643-2651.

[17] WANG Y, QU J, LIU J, et al. What to Tag Your Microblog: Hashtag Recommendation Based on Topic Analysis and Collaborative Filtering[C]// Asia-Pacific Web Conference. Switzerland: Springer Press, 2014: 610-618.

[18] HAN J W, KAMBER M, PEI J. Data Mining: Concepts and Techniques: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2011, 5(4): 1-18.

[19] WAGSTAFF K L, CARDIE C, ROGERS S, et al. Constrained K-means Clustering with Background Knowledge[C]// international conference on machine learning. San Francisco: Margan Kaufmann Press, 2001: 577-584.

[20] WANG W, YANG J, MUNTZ R. STING: A statistical information grid approach to spatial data mining[C]// Proceedings of 23rd International Conference on Very Large Data Bases. 1997: 186-195.

[21] CHENG Z P, ZHOU D, WANG C. CLINCH: Clustering Incomplete High-Dimensional Data for Data Mining Application[C]// Web Technologies Research and Development — APWeb 2005. Berlin: Springer Press, 2005: 88-99.

[22] HE L, HU P. Cold start recommendation model based on user multi-dimension trust[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2018, 30(6): 827-834.

[23] QUINLAN J R. Induction of Decision Trees[J]. Machine Learning, 1986, 1(1): 81-106.

[24] BOBADILLA J, ORTEGA F, HERNANDO A, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46: 109-132.

[25] WANG Y, WAN X Y, TAO Y Z, et al. Collaborative filtering recommendation algorithm based on K-medoids item clustering [J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2017, 29(4): 521-526.

[26] YE F, ZHANG H. A collaborative filtering recommendation based on users' interest and correlation of items[C]// 2016 International Conference on Audio, Language and Image. Piscataway: IEEE Press, 2016: 515-520.

[27] LIKA B, KOLOMVATSOS K, HADJIEFTHYMIADES S. Facing the cold start problem in recommender systems[J]. Expert Systems with Applications, 2014, 41(4): 2065-2073.

[28] SON L H, MINH N T H, CUONG K M, et al. An application of fuzzy geographically clustering for solving the cold-start problem in recommender systems[C]// Proceeding of 5th IEEE International Conference of Soft Computing and Pattern Recognition (SoCPaR2013). Piscataway: IEEE Press, 2013: 44-49.

[29] LIU H, HU Z, MIAN A, et al. A new user similarity model to improve the accuracy of collaborative filtering[J]. Knowledge-Based Systems, 2014, 56: 156-166.

HAN Li-feng, born in 1980, Ph.D, is a member of China Computer Federation. His main research interests include recommendation system, knowledge graph, business intelligence, etc.

CHEN Li, born in 1963, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include intelligent information processing, data mining, network security, etc.