异质信息网络中元路径感知的评分协同过滤

何云飞[1] 张以文[1] 吕智慧[2),5)] 颜登程[3] 何 强[4]

[1)](安徽大学计算机科学与技术学院 合肥 230601)

[2)](复旦大学计算机科学技术学院 上海 200433)

[3)](安徽大学物质科学与信息技术学院 合肥 230601)

[4)](澳大利亚斯威本科技大学电子信息及软件工程学院 墨尔本 3122 澳大利亚)

[5)](网络信息安全审计与监控教育部工程研究中心 上海 200433)

摘 要 基于邻域的协同过滤(Neighborhood-Based Collaboration Filtering,NBCF)具有简单、可解释等优点一直备受关注且被广泛使用.然而,仅利用用户-项目的历史交互信息使得 NBCF 并不能获得很好的推荐性能.随着网络的快速发展,信息网络中包含了大量不同类型的对象和关系,越来越多丰富的语义信息可以被进一步挖掘和利用,自然构成了异质信息网络(Heterogeneous Information Network,HIN).基于 HIN 的推荐模型受到了研究者们的高度关注.相比于传统的推荐模型,基于 HIN 的推荐模型不仅能有效提高推荐性能,还能缓解冷启动和数据稀疏等问题.然而,现有基于 HIN 的推荐模型在保证模型有效性的同时往往需要学习较多的参数,参数的设定对模型性能有重要的影响.因此,本文提出一种无参数的 HIN 中元路径感知的评分协同过滤推荐模型 HRCF,其主要思想是通过 HIN 中的元路径寻找评分的邻居,然后对邻居进行加权来估计该评分.首先,在不同的元路径上产生用户(项目)之间的交换矩阵;其次,将不同元路径上的交换矩阵进行整合计算用户(项目)之间的相似矩阵;最后将用户之间的相似矩阵、用户-项目的历史评分矩阵、项目之间的相似矩阵依次相乘并归一化从而一次性估计所有的评分.为验证 HRCF 模型的有效性,本文在公开的 Douban Book 和 Yelp 数据集上进行了实验.实验结果表明,HRCF 模型的推荐精度优于目前存在的方法,且能很好地克服冷启动问题.

关键词 异质信息网络;协同过滤;推荐系统;元路径;评分预测

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.02385

Meta Path-Aware Rating Collaborative Filtering in Heterogeneous Information Network

HE Yun-Fei[1)] ZHANG Yi-Wen[1)] LV Zhi-Hui[2),5)] YAN Deng-Cheng[3)] HE Qiang[4)]

[1)](*Department of computer science and technology, Anhui University, Hefei* 230601)

[2)](*School of Computer Science, Fudan University, Shanghai* 200433)

[3)](*Institutes of physical science and information technology, Anhui University, Hefei* 230601)

[4)](*School of information technology, Swinburne University of Technology, Melbourne* 3122, *Australia*)

[5)](*Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai* 200433)

Abstract As an indispensable part of online shopping and e-commerce, recommendation system (RS) plays a vital role in helping users find products they are interested in. In many recommendation tasks, rating prediction is one of the most common and important ones. Moreover, among many recommendation models for rating prediction, Neighborhood-Based Collaborative Filtering (NBCF) has received a lot of attention and is widely used because of its advantages such as

收稿日期:2020-04-23;在线发布日期:2020-08-26.本课题得到国家重点研发计划(2019YFB1704101、2019YFB1405000)、国家自然科学基金(61872002,U1936220,61873309)资助.**何云飞**,博士研究生,主要研究方向为异质信息网络、推荐系统.E-mail:heyunfei@stu.ahu.edu.cn.**张以文**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为服务计算、推荐系统、大数据分析.E-mail:zhangyiwen@ahu.edu.cn.**吕智慧**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为边缘计算、云计算、大数据等.**颜登程**,博士,讲师,主要研究方向为复杂网络、大数据分析.**何 强**,博士,高级讲师,主要研究方向为服务计算、大数据分析.

simplicity and interpretability. The main idea of the NBCF model is that similar users have similar preferences, which is relatively intuitive. However, NBCF is not able to obtain good recommendation performance only with historical user-item interaction information. The main reason is that the historical interaction information is often very sparse and the additional information is not enough, which makes the similarity measurement between users (items) less accurate. With the rapid development of networks, the information network contains many different types of objects and relationships, and more rich semantic information can be further tapped and utilized, which naturally constitutes Heterogeneous Information Network (HIN). So, HIN-based recommendation model has received extensive attention from researchers. Compared with traditional recommendation model, HIN-based recommendation models can not only effectively improve the recommendation performance, but also alleviate the problems of cold-start and data sparsity. However, most existing HIN-based recommendation models have many parameters which need to be learned while ensuring the validity, and the parameter setting has an important impact on the model's performance. Particularity, some of the parameters in these models need to be adjusted manually, which results in a large amount of training time. Therefore, this paper proposes a parameterless meta path-aware rating collaborative filtering recommendation model in HIN (HRCF). Moreover, HRCF can be seen as an extension of NBCF in HIN, which makes HRCF inherit many advantages of NBCF. But the main difference between them is that NBCF utilizes similarity between users (items) to make recommendations, while HRCF directly measures similarity between ratings in HIN to make rating predictions. The main idea of HRCF is that ratings of similar users on similar items are close, and our statistical results in the Yelp dataset confirm this idea. In general, HRCF is to find the rating's neighborhood through the meta-path in HIN and then weight the neighbors to estimate the rating. In addition, HRCF does not need to adjust or learn any parameters, and it can be converted to more convenient and fast matrix operations to perform. Specifically, we first generate commuting matrices between users (items) on different meta-paths. Then they are integrated to calculate the similarity matrix between users (items). Finally, the similarity matrix between users, the historical rating matrix of users-items, and the similarity matrix between items are sequentially multiplied and normalized to estimate all ratings at once. In order to verify the validity of the HRCF model, this paper conducted experiments on published Douban Book and Yelp datasets. The experimental results show that the recommendation accuracy of the HRCF model is better than the existing methods, and can well overcome the cold-start problem. Furthermore, we have also shown that a given set of meta-paths on those two datasets has a positive effect on the performance of the HRCF model.

Keywords heterogeneous information network; collaborative filtering; recommendation system; meta-path; rating prediction

1 引 言

网络快速发展的同时也加剧了产品的信息过载,造成了人们越来越难以及时地找到他们感兴趣的产品[1].为缓解信息过载的问题,个性化推荐系统应运而生,且其能利用信息检索、统计、机器学习等技术快速地找到用户的偏好[2].而在众多的推荐模型中,由于基于邻域的协同过滤(Neighborhood-Based Collaborative Filtering,NBCF)具有简单、稳定、可解释等特点而使其成为最流行的推荐技术之一[3].然而,仅利用用户-项目历史交互信息来进行评分预测使得 NBCF 经常面临数据稀疏和冷启动等问题.此外,研究表明,通过扩充有用的信息到模型中是提高推荐性能的一个有效方式[4-5].为此,很多推荐模型通过利用社交网络信息,项目的内容信

息等来对用户在项目上的偏好进行建模[6].

作为一个新兴的研究方向，异质信息网络(Heterogeneous Information Network，HIN)分析给研究者们带来了很多机遇来处理各种数据挖掘任务.一个 HIN 是由多个类型的节点以及节点之间的连接组成，它不仅包含丰富的结构和语义信息，而且还能对其中复杂的对象及其关系进行灵活建模[7-8].HIN 的这些特点使其被广泛应用于分类、聚类、链接预测等任务[9].同样地，HIN 中包含的丰富信息可以扩充到推荐模型中来提高推荐性能[10].因此，越来越多基于 HIN 的个性化推荐模型被提出来.例如，Guo 等人[11]利用从社交网络和项目相关网络中学习嵌入表示来执行推荐；Shi 等人[12]提出了加权 HIN 来更精确地反映对象之间的关系并用来做个性化的语义推荐.

相比较于 NBCF 模型，基于 HIN 的个性化推荐模型不仅能在推荐性能上有显著提高，而且可以有效缓解冷启动和数据稀疏等问题.但是，存在的一些基于 HIN 或深度学习的推荐模型往往存在较多的参数需要学习或存在多个超参数需要调节，前者可通过优化器来进行优化，而后者更多的是需要人为地进行设定和调节.因此，在获得最优的模型之前我们不可避免需要花费大量时间对超参数进行调节且超参数的选择对模型的性能有重要的影响[13].例如，Dai 等人[14]通过利用用户-项目具体的网络提出了 sSVD 模型来缓解数据稀疏问题，但其存在潜在特征的维度，两个惩罚因子，带宽等超参数需要调节；Shi 等人[15]基于元路径来学习用户和项目的嵌入并将其整合到传统的矩阵分解模型中来进行评分预测，但其需要对随机游走的步长、窗宽、嵌入特征的维度、惩罚因子、嵌入特征的系数等超参数进行调节.然而，需要调节的超参数越多，则在训练模型上的时间花费就越久，反之亦然[16].

因此，结合 NBCF 和 HIN 各自的优势，本文将 NBCF 的思想应用到 HIN 中，提出了一种在 HIN 中无参数的基于评分的协同过滤模型，简称 HRCF.HRCF 的主要思想是相邻的用户在相邻的项目上的评分较为相近.如图 1 所示，给定用户 u 和项目 i，通过 HIN 中的元路径(Meta-Path)可以分别找到它们的邻居 $N(u)$ 和 $N(i)$.如果 $N(u)$ 和 $N(i)$ 之间存在评分，则称这些为评分为 r_{ui} 的邻居，记为 $N(r_{ui})$.事实上，我们还对 Yelp 数据集中的评分做了一个简单的统计，分别计算了数据集中的评分与其邻居之间的 RMSE，评分与随机挑选相同数目的邻居之间的 RMSE.如图 2 所展示的两个箱线图.从图中可以看出，评分 r_{ui} 与其邻居 $N(r_{ui})$ 之间更加相近.因此，我们将 NBCF 模型基于相似用户有相似偏好的思想在 HIN 中做了一个延伸，利用元路径来度量评分之间的相似性并通过评分的邻居来进行建模.实验结果表明，HRCF 不仅具有 NBCF 模型简单、稳定以及可解释等特点，还比一些基于 HIN 的推荐模型更加有效.特别地，在实际计算过程中，我们仅需一次性矩阵运算即可估计所有用户在每个项目上的评分.本文的贡献总结如下：

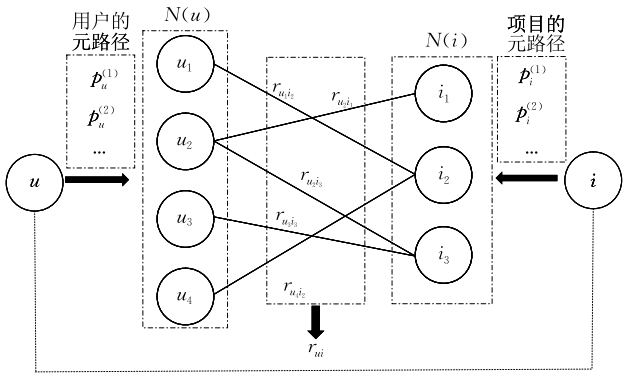


图 1　评分的邻居

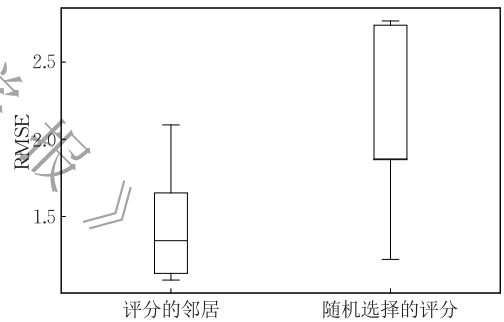


图 2　评分与邻居评分，随机选择评分之间的 RMSE 箱线图

(1)提出一种在 HIN 中元路径感知的评分协同过滤模型，简称 HRCF.HRCF 不仅继承了 NBCF 简单、可解释等优点，还将 HIN 中元路径包含的丰富语义信息扩充到模型中来提高推荐性能.相比于目前存在的一些基于 HIN 的推荐模型，HRCF 更为简单有效.

(2)提出一种基于 HIN 来度量评分之间相似性的方法.该方法首先利用元路径来度量用户(项目)之间的相似性.然后在用户和项目相互独立的条件下，通过将用户之间以及项目之间的相似性直接相乘来获得评分之间的相似性.本文提出的方法在计算过程中更为快速且易于执行.

(3)通过在 Douban Movie 和 Yelp 两个数据集上设计多个实验来验证 HRCF 模型的有效性.实验

结果表明，HRCF 模型不仅能获得很好的推荐性能，还能有效缓解冷启动问题.

本文组织结构如下：第 2 节为相关工作；第 3 节介绍与本文相关的基础知识；第 4 节主要对 HRCF 模型进行详细地介绍；第 5 节为实验结果及分析；第 6 节对本文进行总结.

2 相关工作

本节主要从 3 个方面介绍与本文相关的工作，分别是推荐系统，异质化信息网络和基于 HIN 的推荐模型.

如今，推荐系统(Recommendation Systems，RS)已经成为电子商务必不可少的一部分，它可以帮助用户从海量的产品中找到他们感兴趣的产品并被广泛应用于各行各业[17]. RS 大致上可以分为基于内容的、协同过滤以及混合推荐模型[2]. 在 RS 的众多模型中，协同过滤(Collaboration Filtering，CF)是最普通的但也是应用最为广泛的技术之一[3,18]. 此外，CF 可分为基于模型的和基于邻域的[2]. 其中，由于基于邻域的协同过滤(Neighborhood-Based CF，NBCF)简单、有效以及可解释等特点而仍然获得大量的关注[19-21]. 根据依靠的对象不同，NBCF 又可分为基于用户的[22]和基于项目的[23]，其主要思想是通过计算用户(项目)之间的相似性并利用其评分进行加权来进行预测. 因此，相似性度量在 NBCF 模型中起到了至关重要的作用. 如 McLaughlin 和 Herlocker[19]对 Pearson 相关系数进行改进，从而更好地刻画用户之间的关系，Kendall 和 Gibbons[24]基于评分的不同等级提出了 Spearman Rank Correlation(SRC)方法来度量用户之间的相似性. 但是，仅仅通过用户-项目的历史交互信息是不能充分反映出用户(项目)之间的关系，这也导致了传统的 NBCF 模型的推荐精度较低. 因此，很多研究者尝试通过扩充有用的信息来提高推荐性能[4-5,25].

随着网络的快速发展，越来越多丰富的信息可以帮助推荐模型来提高其性能. 网络信息中包含了多个对象以及对象之间的关系，它们组成了异质信息网络(Heterogeneous Information Network，HIN)，且其中蕴含丰富的语义和结构信息可以被用来灵活地建模[9,26]. 此外，HIN 分析作为一个新兴的研究方向，被广泛应用于相似性度量、链接预测、信息融合、推荐等任务中. 例如，在计算节点之间的相似性时，通过 HIN 我们可以同时考虑两个节点之间的结构信息以及不同元路径上两个节点之间的语义信息来反映节点之间的更为真实的关系. 如 Sun 等人[27]提出了基于路径的方法来度量相同类型对象之间的相似性. Wang 等人[28]将文本集视为一个 HIN 并用来度量两个文本之间的相似性. Shi 等人[29]提出了一种在 HIN 中计算节点之间相似性的框架. 现有研究实验结果均表明，基于 HIN 来度量节点之间的相似性更为准确，在相关的数据挖掘任务中表现更好.

近年来，基于 HIN 的推荐系统受到高度关注，大致可将其分为两个方面[15]：基于相似性和基于嵌入. 其中基于相似性的推荐模型主要通过 HIN 来挖掘用户之间或项目之间的相似性并作为限制与传统推荐模型相结合. 如 Yu 等人[30]利用元路径来度量节点之间的相似性并作为正则化项与矩阵分解模型进行整合. Shi 等人[12]在每条元路径上以评分的值进行区分，提出了加权的元路径，并利用该元路径来学习节点之间的相似性并作为限制整合到推荐模型中来提高性能. Luo 等人[31]则利用社交网络来计算用户之间的相似性，然后将其对用户的潜在特征进行限制来做推荐. Zheng 等人[32]则将 Luo 等人的工作进行延伸，通过元路径分别计算用户和项目的相似性，然后利用其对相似的用户和项目分别加以限制并整合到传统的矩阵分解模型中来提高推荐性能. 此外，Dai 等人[14]则从用户和项目的具体网络中学习评分之间的相似性，并用其对评分进行光滑处理后执行推荐. 对于基于嵌入的模型来说，HIN 中包含的丰富结构和语义信息可用来学习更好的嵌入表示并进行推荐. 如 Shi 等人[15]通过基于元路径的随机游走来学习用户和项目的嵌入表示并将它们整合到传统的矩阵分解模型中来执行评分预测. 此外，随着深度学习(Deep Learning，DL)的快速发展，基于 DL 的推荐模型被提出来了[33]. 同时也涌现出将 HIN 和 DL 相结合进行推荐的模型. 如 Fan 等人[34]提出了基于元路径指导的图神经网络来进行意图推荐. Shi 等人[35]利用多层感知机和注意力机制从 HIN 中学习用户和项目的嵌入表示并用于推荐. 实验结果表明，DL 和 HIN 相结合的推荐模型能获得较好的推荐性能，并且能缓解数据稀疏和冷启动等问题. 但是，一些基于深度学习的推荐模型往往缺乏可解释性且存在较多需要调节的超参数，这导致其难以执行且开销较大[16,36].

综上，本文将 NBCF 在 HIN 中进行延伸，提出

一种在 HIN 中元路径感知的评分协同过滤模型，HRCF. 该方法摒弃了 NBCF 中将用户(项目)之间的相似性作为评分的权重来进行预测，而是直接通过 HIN 中的元路径来寻找评分的邻居并计算它们之间的相似性，然后直接将邻居内的评分进行加权来进行预测. HRCF 模型在继承了 NBCF 优点的同时能够取得很好的推荐性能. 此外，HRCF 没有参数需要学习或超参数要调节，这也使得 HRCF 易于执行且无需在调参中花费大量时间.

3　相关定义

本节主要介绍一些与本文相关的定义.

定义 1.　异质信息网络[7,15,27]. 给定一个网络$\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$，其中$\mathcal{V}$和$\mathcal{E}$分别表示节点的集合和边的集合. 如果存在一个节点类型的映射$\phi(\cdot)$和一个边类型的映射$\varphi(\cdot)$分别将节点和边映射成一个具体的类型，即$\phi(v)\rightarrow\mathcal{T}(v\in\mathcal{V})$，$\varphi(e)\rightarrow\mathcal{R}(e\in\mathcal{E})$，则称其为一个信息网络. 如果$|\mathcal{T}|+|\mathcal{R}|>2$，则$\mathcal{G}$是一个异质信息网络，并且称$\mathcal{S}=\{\mathcal{T},\mathcal{R}\}$为网络模式.

定义 2.　元路径[15,27]. 路径 p 在网络模式$\mathcal{S}$中具有形式$\mathcal{T}_1\rightarrow\mathcal{T}_2\rightarrow\cdots\rightarrow\mathcal{T}_l$(缩写为$\mathcal{T}_1\mathcal{T}_2\cdots\mathcal{T}_l$)，其中$\mathcal{T}_i(i=1,2,\cdots,l)$表示特定类型的节点，$\rightarrow$表示两个节点之间的连接关系.

定义 3.　交换矩阵[27]. 给定网络模式$\mathcal{S}=\{\mathcal{T},\mathcal{R}\}$和元路径 p，此时在路径 p 下的交换矩阵表达式为

$$\boldsymbol{C}_p=\boldsymbol{A}_{\mathcal{T}_1\mathcal{T}_2}\cdots\boldsymbol{A}_{\mathcal{T}_{l-1}\mathcal{T}_l}\tag{1}$$

其中，$\boldsymbol{A}_{\mathcal{T}_i\mathcal{T}_j}(1\leq i<j\leq l)$表示具有类型$\mathcal{T}_i$的节点和类型$\mathcal{T}_j$的节点之间的邻接矩阵.

定义 4.　基于异质信息网络的推荐[15]. 给定一个异质网络$\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$，记用户集合和项目集合分别为$\mathcal{U}$和$\mathcal{I}$，用户$u(\in\mathcal{U})$在项目$i(\in\mathcal{I})$上的评分为$r_{ui}$，$\mathcal{G}$中存在的所有评分记为三元组的集合$\{\langle u,i,r_{ui}\rangle\}$. 此时，我们的主要任务是预测用户$u(\in\mathcal{U})$在其未进行评分的项目$j(\in\mathcal{I})$上的评分$r_{uj}$.

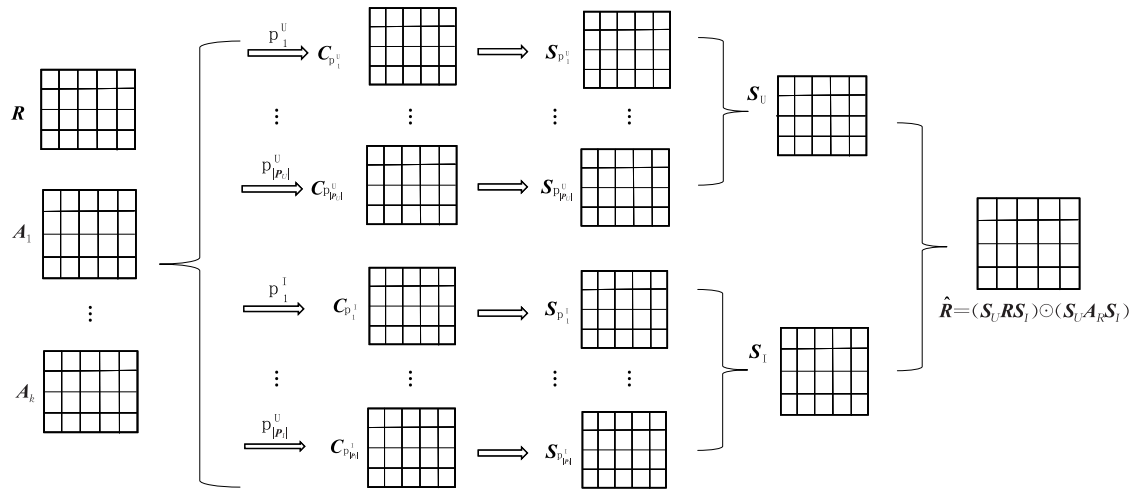
4　HRCF 模型

本节将详细介绍提出的 HRCF 模型. HRCF 主要分为 3 个部分，如图 3 所示. 首先，根据元路径集分别计算每一个路径下用户(项目)之间的交换矩阵和相似矩阵，如图 3(b)和(c)所示. 然后再根据所有的交换矩阵来计算用户(项目)之间最终的相似性矩阵$\boldsymbol{S}_U(\boldsymbol{S}_I)$，如图 3(d)所示. 最后，将用户相似矩阵$\boldsymbol{S}_U$，评分矩阵$\boldsymbol{R}$和项目相似矩阵$\boldsymbol{S}_I$依次相乘并归一化即可得到预测的评分矩阵.

为便于阐述和理解 HRCF 模型，本文使用了一些符号来进行简化. 相关符号及其说明如表 1 所示.

表 1　HRCF 中使用的符号

符号	描述	符号	描述
$\mathcal{G}$	异质信息网络	p	单个元路径
$\mathcal{V}$	节点的集合	$\boldsymbol{R}$	评分矩阵
$\mathcal{E}$	边的集合	$\hat{\boldsymbol{R}}$	预测的评分矩阵
$\mathcal{S}$	网络模式	$\boldsymbol{A}$	邻接矩阵
$\mathcal{T}$	节点类型的集合	$\boldsymbol{A}_{\boldsymbol{R}}$	用户和项目之间的邻接矩阵
$\mathcal{R}$	边类型的集合	$\boldsymbol{C}$	交换矩阵
$\mathcal{U},\mathcal{I}$	用户和项目的集合	$\boldsymbol{S}$	相似矩阵
m,n	用户的个数和项目的个数	$\langle u,i,r_{ui}\rangle$	三元组合，表示样本空间中用户在项目上的评分
$\boldsymbol{P}_U,\boldsymbol{P}_I$	用户和项目的元路径集	Ω	HIN 中所有的三元组合



(a) 输入评分矩阵及各个邻接矩阵　(b) 根据用户和项目的不同元路径计算交换矩阵　(c) 计算用户和项目在不同元路径下的相似矩阵　(d) 计算用户和项目最终的相似矩阵　(e) 评分预测

图 3　HRCF 模型的框架

4.1 交换矩阵

给定一个 HIN $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$以及元路径集合 $\boldsymbol{P}=\{\boldsymbol{P}_U,\boldsymbol{P}_I\}$,其中 $\boldsymbol{P}_U=\{p_1^U,\cdots,p_{|PU|}^U\}$表示用户的元路径集合,$\boldsymbol{P}_I=\{p_1^I,\cdots,p_{|PI|}^I\}$表示项目的元路径集合,$|PU|$和$|PI|$分别表示用户和项目的元路径的个数.然后我们利用 HIN 中的邻接矩阵$\{\boldsymbol{A}_1,\cdots,\boldsymbol{A}_k\}$来计算元路径集 $\boldsymbol{P}$ 中所有元路径的交换矩阵 $\boldsymbol{C}=\{\boldsymbol{C}_{p_1^U},\cdots,\boldsymbol{C}_{p_{|PU|}^U},\boldsymbol{C}_{p_1^I},\cdots,\boldsymbol{C}_{p_{|PI|}^I}\}$,其中 k 表示 HIN 中邻接矩阵的个数,用户的交换矩阵$\in R^{m\times m}$,项目的交换矩阵$\in R^{n\times n}$,计算流程如图 3(b)所示.特别地,每个交换矩阵中的元素表示在该元路径下对应位置上的两个节点(项目)之间的连接次数,次数越大则表示两者相关性越强,反之亦然.

4.2 相似性度量

当获得元路径集 $\boldsymbol{P}$ 对应的交换矩阵集 $\boldsymbol{C}$ 后,接下来需要计算每个元路径下节点之间的相似矩阵.这里需要指出的是,对于给定元路径 p 及其交换矩阵 $\boldsymbol{C}_p$,在计算相似矩阵 $\boldsymbol{S}_p$ 之前,我们将 $\boldsymbol{C}_p$ 中对角元素设为 0.这样做的目的主要是降低节点自身在计算它与其它节点相似性时的影响.然后,我们将交换矩阵中的每行的元素均除以该行所有元素之和即可得到给定元路径下的相似矩阵.如式(2)所示:

$$\boldsymbol{S}_p[i,j]=\begin{cases}\dfrac{\boldsymbol{C}_p[i,j]}{\sum\limits_j \boldsymbol{C}_p[i,j]}, & \sum\limits_j \boldsymbol{C}_p[i,j]\neq 0\\ 0, & \sum\limits_j \boldsymbol{C}_p[i,j]=0\end{cases} \tag{2}$$

式(2)的主要目的是对交换矩阵的每一行进行行归一化.从概率的角度来看,$\boldsymbol{S}_p$ 中的元素 $\boldsymbol{S}_p[i,j]$表示在给定元路径 p 和起始节点 i 时,$i\rightarrow j$ 的转移概率,如式(3)所示:

$$\boldsymbol{S}_p[i,j]=TP(j\,|\,i,p) \tag{3}$$

这里需要注意的是 $\boldsymbol{S}_p[i,j]\neq\boldsymbol{S}_p[j,i]$,即 $\boldsymbol{S}_p$ 是一个非对称的矩阵.

同理,根据式(2),我们可以分别计算在元路径集 $\boldsymbol{P}_U$ 中每条元路径下用户之间的相似矩阵.最后,通过取平均即可获得最终的用户相似矩阵.这里,我们默认节点与自身之间的相似性为 1 且在计算 $\boldsymbol{S}_p$ 之前已经将 $\boldsymbol{C}_p$ 的对角元素设为 0,因此最后需要加上一个单位对角矩阵,如式(4)所示:

$$\boldsymbol{S}_U=\frac{1}{|PU|}\sum_{p\in\boldsymbol{P}_U}\boldsymbol{S}_P+\boldsymbol{1} \tag{4}$$

其中,$\boldsymbol{S}_U\in R^{m\times m}$,$|PU|$表示用户的元路径个数,$\boldsymbol{1}$ 是单位矩阵($\in R^{m\times m}$),表示用户与自身的相关性为 1.类似地,可以计算项目之间最终的相似性矩阵 $\boldsymbol{S}_I$($\in R^{n\times n}$).当然,在计算最终的相似矩阵时有很多方式.这里采用取平均的方式主要有以下 3 点原因:(1)无参数需要学习或调节,可节省大量时间;(2)易于执行,运算速度快;(3)实验结果显示其能达到很好的推荐性能且结果稳定.因此,我们最终采用了直接取平均的方式来度量相似性.

接下来,根据用户和项目的相似矩阵 $\boldsymbol{S}_U$ 和 $\boldsymbol{S}_I$ 我们可以度量任意两个评分之间的相似性.给定三元组$\langle u,i,r_{ui}\rangle$,对于 HIN 中任意的一个三元组$\langle u',i',r_{u'i'}\rangle$,此时 r_{ui} 与 $r_{u'i'}$ 的相关性可看作是一个关于用户 u、u'、i 和 i'的函数,如式(5)所示:

$$S(r_{u'i'}\,|\,r_{ui})=f(u,u',i,i') \tag{5}$$

这里,我们假设用户和项目之间是相互独立的.因此,我们可将用户之间的相似性与项目之间的相似性直接相乘来直接获得这两个评分之间相似性,如式(6)所示:

$$S(r_{u'i'}\,|\,r_{ui})=\frac{\boldsymbol{S}_U(u,u')\boldsymbol{S}_I(i,i')I(\langle u',i',r_{u'i'}\rangle\in\Omega)}{\sum\limits_{\langle u_k,i_k,r_{u_ki_k}\rangle\in\Omega}\boldsymbol{S}_U(u,u_k)\boldsymbol{S}_I(i,i_k)} \tag{6}$$

其中,$\boldsymbol{S}_U(u,u')$表示 $\boldsymbol{S}_U$ 中对应位置上 u 和 u'的相似性,$\boldsymbol{S}_I(i,i')$表示 $\boldsymbol{S}_I$ 中对应位置上 i 和 i'的相似性,$I(\cdot)$是示性函数,Ω 是 HIN 中三元组合的集合.从等式(6)中可知,如果 $\boldsymbol{S}_U(u,u')$和 $\boldsymbol{S}_I(i,i')$都存在且 $I(\langle u',i',r_{u'i'}\rangle\in\Omega)$不等于 0,则表示 $r_{u'i'}$ 是 r_{ui} 的邻居且 $S(r_{u'i'}\,|\,r_{ui})$存在,反之亦然.这里需要注意的是评分之间的相似性也是不对称的,即 $S(r_{u'i'}\,|\,r_{ui})\neq S(r_{ui}\,|\,r_{u'i'})$,而且,$r_{ui}$ 可以不在样本空间内.

4.3 评分预测

HRCF 模型的评分预测函数是基于 NBCF 模型的.因此,这里我们首先给出 NBCF 模型中基于用户的评分预测函数,如式(7)[2,22]所示:

$$\hat{r}_{ui}=\frac{\sum\limits_{v\in N_i(u)}w_{uv}r_{vi}}{\sum\limits_{v\in N_i(u)}w_{uv}} \tag{7}$$

其中,w_{uv}表示用户 u 和 v 之间的相似性度量,$N_i(u)$表示和用户 u 在项目 i 上都有评分的用户集合.注意到这里 w_{uv}不会随着项目 i 的改变而变化,这也是导致 NBCF 模型性能较差的原因之一.基于上一节给出的度量评分之间相似性的方法,类似于 NBCF,我们可以估计用户 u 在项目 i 上评分,计算方法如式(8)所示:

$$\hat{r}_{ui}=\frac{\sum\limits_{\langle u',i',r_{u'i'}\rangle\in\Omega}S(r_{u'i'}\,|\,r_{ui})r_{u'i'}}{\sum\limits_{\langle u',i',r_{u'i'}\rangle\in\Omega}S(r_{u'i'}\,|\,r_{ui})} \tag{8}$$

其中，$S(r_{u'i'}|r_{ui})$表示在给定 r_{ui} 时其与 Ω 中其它评分 $r_{u'i'}$ 之间的相似性. 注意到这里只要求 $r_{u'i'}$ 属于样本空间即可，并不要求其一定是 r_{ui} 的邻居. 这是因当 $r_{u'i'}$ 不是 r_{ui} 的邻居时，$S(r_{u'i'}|r_{ui})=0$，如等式(6)所示. 为方便起见，式(8)可通过如下矩阵形式进行计算：

$$\hat{r}_{ui}=\frac{\boldsymbol{S}_U(u)\boldsymbol{R}\boldsymbol{S}_I^{\mathrm{T}}(i)}{\boldsymbol{S}_U(u)\boldsymbol{A}_{\boldsymbol{R}}\boldsymbol{S}_I^{\mathrm{T}}(i)} \tag{9}$$

其中，$\boldsymbol{S}_U(u)(\in R^{1\times m})$表示用户相似矩阵 $\boldsymbol{S}_U$ 的第 u 行，$\boldsymbol{S}_I(i)(\in R^{1\times n})$表示项目相似矩阵 $\boldsymbol{S}_I$ 的第 i 行，$\boldsymbol{A}_{\boldsymbol{R}}(\in R^{m\times n})$表示用户和项目之间的邻接矩阵. 很容易证明等式(9)的分子分母和等式(8)中的分子分母是一一对应的.

实际计算过程中，等式(9)不利于批量预测用户在未评分项目上的评分情况. 为了更便捷地一次性预测所有的评分，我们将等式(9)进行改写，如式(10)所示：

$$\hat{\boldsymbol{R}}=(\boldsymbol{S}_U\boldsymbol{R}\boldsymbol{S}_I^{\mathrm{T}})\odot(\boldsymbol{S}_U\boldsymbol{A}_{\boldsymbol{R}}\boldsymbol{S}_I^{\mathrm{T}}) \tag{10}$$

其中，$\boldsymbol{R}\in R^{m\times n}$，$\odot$表示 Hadamard 乘积.

4.4　算法复杂度分析

HRCF 算法描述如算法 1 所示. 假设用户的个数为 m，其元路径个数为$|\boldsymbol{P}_U|$，项目的个数为 n，其中元路径的个数为$|\boldsymbol{P}_I|$. 算法 1 中，计算用户最终相似矩阵的时间复杂度为 $O((|\boldsymbol{P}_U|+1)m^2)$，计算项目最终相似矩阵的时间复杂度为 $O((|\boldsymbol{P}_I|+1)n^2)$，计算评分矩阵的时间复杂度为 $O(mn)$. 因此，总时间复杂度为 $O((|\boldsymbol{P}_U|+1)m^2+(|\boldsymbol{P}_I|+1)n^2+mn)$. 此外，在计算用户最终相似矩阵时需要的空间复杂度为 $O(|\boldsymbol{P}_U|m^2)$，计算项目最终相似矩阵的空间复杂度为 $O(|\boldsymbol{P}_I|n^2)$，计算评分矩阵的空间复杂度为 $O(mn)$. 因此，HRCF 模型总的空间复杂度为 $\max\{O(|\boldsymbol{P}_U|m^2),O(|\boldsymbol{P}_I|n^2)\}$. 其中$|\boldsymbol{P}_U|$和$|\boldsymbol{P}_I|$在本文中都是很小的一个常数.

算法 1.　HRCF.

输入：评分矩阵 $\boldsymbol{R}$；邻接矩阵$\{\boldsymbol{A}_1,\cdots,\boldsymbol{A}_k\}$；用户和项目的元路径集合 $\boldsymbol{P}_U$ 和 $\boldsymbol{P}_I$

输出：预测的评分矩阵 $\hat{\boldsymbol{R}}$

1. For p IN $\boldsymbol{P}_U$ DO
2. 　根据等式(1)计算用户的交换矩阵 $\boldsymbol{C}_p$
3. 　根据等式(2)计算用户的相似矩阵 $\boldsymbol{S}_p$
4. END FOR
5. 根据式(4)计算用户的最终相似矩阵 $\boldsymbol{S}_U$
6. FOR p IN $\boldsymbol{P}_I$ DO
7. 　根据等式(1)计算项目的交换矩阵 $\boldsymbol{C}_p$
8. 　根据等式(2)计算项目的相似矩阵 $\boldsymbol{S}_p$
9. END FOR
10. 根据式(4)计算项目的最终相似矩阵 $\boldsymbol{S}_I$
11. 根据式(10)计算预测的评分矩阵 $\hat{\boldsymbol{R}}$

5　实验结果及分析

5.1　数据集

为了验证 HRCF 模型的推荐性能，我们在两个公开的数据集，Douban Book 和 Yelp(数据来源为 http://www.shichuan.org/)上设置了多个实验来进行验证. 其中 Douban Book 中有 13 024 位用户在 22 347 本书上一共给出了 792 026 个评分，该数据集同时还包含了用户的社交关系和书本的属性信息. Yelp 中有 16 239 位用户对 14 284 个商店一共打出了 198 397 个评分且该数据集同时还包含了用户的社交关系及商店的属性信息. 其次，在两个数据集上的评分均为 1～5 之间的整数，详细的统计信息和结构信息如表 2 和图 4 所示. 此外，我们在表 3 和表 4 中分别对两个数据集上给定的元路径及其含义进行了解释说明. 这里需要指出的是两个数据集的稀疏程度是不同的，其中 DoubanBook 的密度为 0.27%，Yelp 为 0.08%. 很明显，数据越稀疏越难预测. 实验环境为 python 语言，Intel Core i7-4790 3.60 GHz，16 GB RAM.

表 2　Douban Book 和 Yelp 两个数据集的统计信息

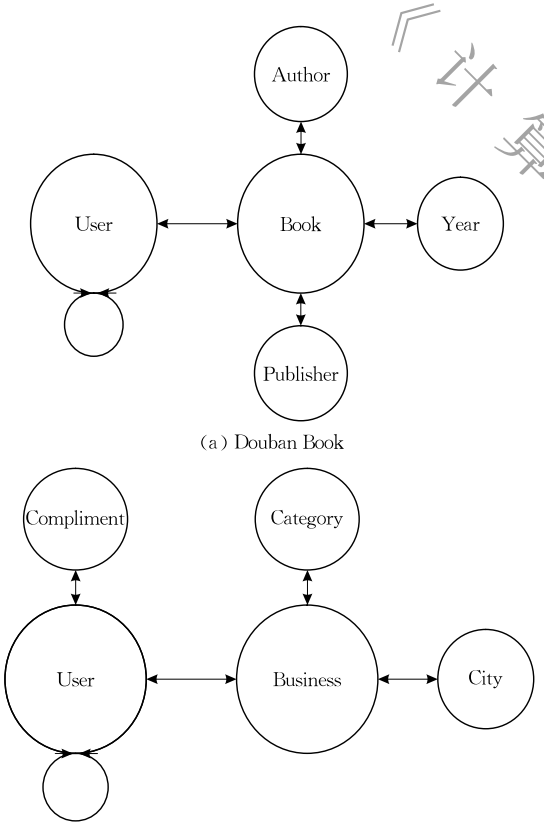
数据集	关系(X-Y)	X 的个数	Y 的个数	X-Y 的个数	X 平均的度	Y 平均的度	元路径
Douban Book	User-Book	13 024	22 347	792 026	60.8	35.4	UBU，BUB
	User-User	12 748	12 748	169 150	13.3	13.3	UBPBU，BPB
	Book-Author	21 907	10 805	21 905	1.0	2.0	UBYBU，BYB，UU，
	Book-Publisher	21 733	1815	21 733	1.0	11.9	BAB
	Book-Year	21 192	64	21 192	1.0	331.1	UBABU
Yelp	User-Business	16 239	14 284	198 397	12.2	13.9	UBU，BUB
	User-User	10 580	10 580	158 590	15.0	15.0	UBCiBU，BCiB，
	User-Compliment	14 411	11	76 875	5.3	6988.6	UU
	Business-City	14 267	47	14 267	1.0	303.6	UBCaBU，
	Business-Category	14 180	511	40 009	2.8	78.3	BCaB，UCoU

表 3 Douban Book 数据集中使用的元路径及其含义

元路径	含义
$U\leftrightarrow U$	朋友关系
$U\rightarrow B\leftarrow U$	两个用户看过同一本书
$U\rightarrow B\rightarrow P\leftarrow B\leftarrow U$	两个用户看过相同出版社的书
$U\rightarrow B\rightarrow Y\leftarrow B\leftarrow U$	两个用户看过同一年份的书
$U\rightarrow B\rightarrow A\leftarrow B\leftarrow U$	两个用户看过同一个作者的书
$B\leftarrow U\rightarrow B$	两本书被同一个用户看过
$B\rightarrow P\leftarrow B$	两本书属于同一个出版社
$B\rightarrow Y\leftarrow B$	两本书出版于同一年
$B\rightarrow A\leftarrow B$	两本书的作者为同一个人

表 4 Yelp 数据集中使用的元路径及其含义

元路径	含义
$U\leftrightarrow U$	朋友关系
$U\rightarrow B\leftarrow U$	两个用户去过同一个商店
$U\rightarrow Co\leftarrow U$	两个用户属于同一个类型
$U\rightarrow B\rightarrow Ci\leftarrow B\leftarrow U$	两个用户去过同一个城市的商店
$U\rightarrow B\rightarrow Ca\leftarrow B\leftarrow U$	两个用户去过同一类型的商店
$B\leftarrow U\rightarrow B$	两个商店被同一个用户消费过
$B\rightarrow Ca\leftarrow B$	两个商店属于同一类型
$B\rightarrow Ci\leftarrow B$	两个商店位于同一个城市



(a) Douban Book

(b) Yelp

图 4 两个数据集的数据结构

5.2 评价指标

本文采用评价指标 *MAE*(Mean Absolute Error)和 *RMSE*(Root Mean Square Error)[10,15]来比较不同模型之间的推荐性能,定义如下:

$$MAE=\frac{\sum_{\langle u,i,r_{u,i}\rangle\in\boldsymbol{\Omega}}|r_{u,i}-\hat{r}_{u,i}|}{|\boldsymbol{R}_{\text{test}}|} \tag{11}$$

$$RMSE=\sqrt{\frac{\sum_{\langle u,i,r_{u,i}\rangle\in\boldsymbol{\Omega}}(r_{u,i}-\hat{r}_{u,i})^2}{|\boldsymbol{R}_{\text{test}}|}} \tag{12}$$

其中,$r_{u,i}$和$\hat{r}_{u,i}$分别表示是真实的评分值和预测值,$\langle u,i,r_{u,i}\rangle$是样本空间$\Omega$中的三元组,$\boldsymbol{R}_{\text{test}}$是测试集,$|\boldsymbol{R}_{\text{test}}|$表示测试集中三元组的个数.

5.3 性能比较

为了更好地反映 HRCF 模型的性能,在实验中我们选择了下几种模型来进行比较,列举如下:

(1) PMF[37]:它将用户-项目的历史评分矩阵分解为两个分别关于用户和项目的低秩特征矩阵.

(2) NMF[38]:它将用户-项目的历史评分矩阵分解为两个低秩的非负矩阵.

(3) UserKNN[22]:用户在项目上的评分是根据与用户前k个相似用户在该项目上的评分学习到的.

(4) ItemKNN[23]:用户在项目上的评分是根据用户在与该项目前 K 个相似的项目上的评分学习到的.

(5) GMF[39]:广义矩阵分解,主要将用户潜在特征和项目潜在特征对应元素相乘,然后和一个权重向量做内积来执行评分预测.

(6) sSVD[14]:通过元路径寻找评分的领域并计算相似性,然后基于相似性将评分进行光滑处理,最后和传统的矩阵分解相结合来执行推荐.

(7) HERec[15]:该模型通过将从元路径学习到用户和项目的嵌入整合到传统的矩阵分解模型中来执行评分预测.

考虑到 HRCF 模型是将 CF 思想在 HIN 中进行了延伸,因此,我们选择以上模型来进行比较. 其中前 4 个模型是经典的 CF 模型,后三个是则是近期将 CF 技术在神经网络或 HIN 中做的一个延伸. 具体地说,PMF 和 NMF 是经典的基于模型的 CF;UserKNN 和 ItemKNN 是经典的基于邻域的 CF;GMF 是基于神经网络的广义矩阵分解;sSVD 是基于 HIN 来度量评分之间的相似性并利用其对评分先进行光滑处理,然后用于矩阵分解的模型;而 HERec 是基于 HIN 来学习节点的嵌入表示并整合到矩阵分解中的推荐模型. 此外,还有很多基于 HIN 的推荐模型,如 HeteMF[27]、SemRec[12]、DSR[19]. 但是,Shi 等人[14]的实验结果显示这些模型的推荐性能均差于 HERec. 因此,这里就不做过多比较.

为了公平起见,参照文献[10,30],我们将上述模型中关于矩阵分解的维度设为 10. 其次,将 UserKNN 和 ItemKNN 中邻居的个数设为 10,相

似性度量方法为 MSD(Mean Squared Difference). 此外，实验中我们将数据集分为训练集和测试集，并根据两个数据的稀疏程度将 Douban Book 和 Yelp 数据集的训练率分别设置为{0.8,0.6,0.4,0.2}和{0.9,0.8,0.7,0.6}[12,15]. 在每一个训练率下，我们随机地将数据分为训练集和测试集，并计算各个模型在测试集上的 *MAE* 和 *RMSE*.

实验结果如表 5 所示. 首先，从表中可以看出，相比于 6 个模型，HERec 和 HRCF 这两个基于 HIN 的模型有更小的 *MAE* 和 *RMSE*. 这说明 HIN 中丰富的结构和语义信息可以帮助推荐模型来提高推荐性能. 其次，在大多数情况下，HRCF 比 HERec 有更好的推荐性能. 这也说明了 HRCF 在模型的复杂度和有效性之间达到了一个新的平衡. 此外，在 Douban Book 训练率为 0.8 和 0.6 时，HERec 的性能要优于 HRCF. 造成这个现象的最主要原因是数据集的所有关系(*X*-*Y*)中，对象 *X* 或 *Y* 平均的度较低. 具体地说，HRCF 通过这些关系来寻找邻居并用于推荐，而当训练率较高时这些关系不足以支撑计算多评分之间的关系. 然而，HERec 模型通过基于元路径的随机游走可以捕捉到更远的两个节点之间的关系.

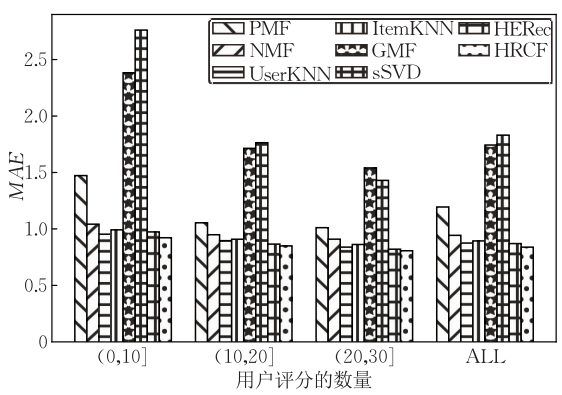
表 5　所有模型在两个数据集上进行有效性实验的结果

数据集	训练率	评价指标	PMF	NMF	UserKNN	ItemKNN	GMF	sSVD	HERec	HRCF
Douban Book	0.8	*MAE*	0.5808	0.6462	0.5842	0.5912	0.6206	0.7228	**0.5548**	0.5609
		Improve		−14.36%	+0.58%	−1.79%	−6.85%	−24.44%	+4.48%	+3.43%
		RMSE	0.7505	0.7958	0.7328	0.7462	0.8047	0.9879	**0.7023**	0.7141
		Improve		−6.04%	+2.36%	+0.57%	−7.22%	−31.63%	+6.42%	+4.85%
	0.6	*MAE*	0.6191	0.6565	0.5952	0.5994	0.7166	0.7913	**0.5617**	0.5663
		Improve		−6.04%	+3.86%	+3.18%	−15.75%	−27.81%	+9.27%	+8.53%
		RMSE	0.8112	0.8117	0.7488	0.7589	0.9581	1.1879	**0.7127**	0.7214
		Improve		−0.06%	+7.69%	+6.45%	−18.10%	−46.44%	+12.14%	+11.07%
	0.4	*MAE*	0.7340	0.6755	0.6133	0.6129	0.8318	1.0203	0.5763	**0.5724**
		Improve		+7.97%	+16.44%	+16.50	−13.32%	−39.00%	+21.49%	+21.77
		RMSE	0.9866	0.8375	0.7774	0.7824	1.1376	1.6749	0.7350	**0.7323**
		Improve		+15.11%	+21.20%	+20.70%	−15.31%	−69.76%	+25.50%	+25.78%
	0.2	*MAE*	1.1613	0.7251	0.6448	0.6470	1.4628	2.3323	0.6282	**0.5964**
		Improve		+37.56%	+44.48%	+44.29%	−25.96%	−100.84	+45.91%	+48.64%
		RMSE	1.5311	0.9055	0.8342	0.8460	1.9451	3.5409	0.8357	**0.7683**
		Improve		+40.86%	+45.52	+44.75%	−27.04%	−131.24%	+45.42	+49.82%
Yelp	0.9	*MAE*	1.0426	0.9148	0.8517	0.8732	1.3838	1.4640	0.8393	**0.8359**
		Improve		+12.26%	+18.31%	+16.25%	−32.73%	−40.42%	+19.50%	+19.83%
		RMSE	1.3823	1.1548	1.1062	1.1380	1.8464	2.2460	1.0982	**1.0882**
		Improve		+16.46%	+19.97%	+17.67%	−33.57%	−62.48%	+20.55%	+21.28%
	0.8	*MAE*	1.0874	0.9208	0.8549	0.8787	1.3907	1.5540	0.8434	**0.8321**
		Improve		+15.32%	+21.38%	+19.19%	−28.79%	−42.90%	+22.44%	+23.48%
		RMSE	1.4327	1.1622	1.1082	1.1461	1.8542	2.3439	1.1043	**1.0800**
		Improve		+18.88%	+22.65%	+20.00%	−29.42%	−63.60%	+22.92%	+24.62%
	0.7	*MAE*	1.1339	0.9397	0.8705	0.8904	1.6127	1.6963	0.8600	**0.8413**
		Improve		+17.13%	+23.23%	+21.47%	−42.23%	−49.60%	+24.16%	+25.80%
		RMSE	1.4873	1.1857	1.1305	1.1630	2.1122	2.5461	1.1236	**1.0972**
		Improve		+20.28%	+23.99%	+21.80%	−42.02%	−71.19%	+24.45%	+26.23%
	0.6	*MAE*	1.2082	0.9430	0.8751	0.8941	1.7431	1.8306	0.8706	**0.8389**
		Improve		+21.95%	+27.57%	+26.00%	−44.27%	−51.51%	+27.94%	+30.57%
		RMSE	1.5700	1.1887	1.1360	1.1693	2.2499	2.7064	1.1428	**1.0912**
		Improve		+24.29%	+27.64%	+25.52%	−43.30%	−72.38%	+27.21%	+30.50%

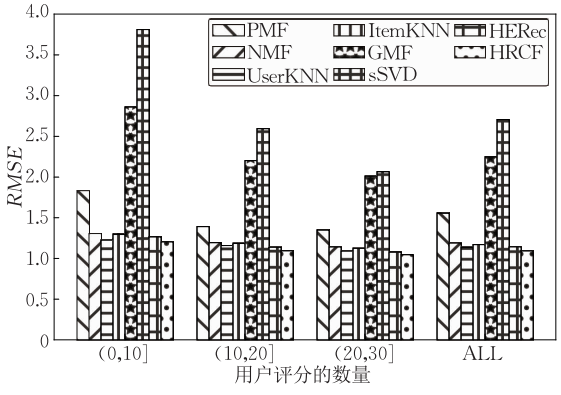
因此，HRCF 可以作为传统推荐模型和一些复杂的基于 HIN 模型的替代模型. 此外，在表中我们还展示了其它模型相对 PMF 在 *MAE* 和 *RMSE* 上的提升率.

5.4　冷启动

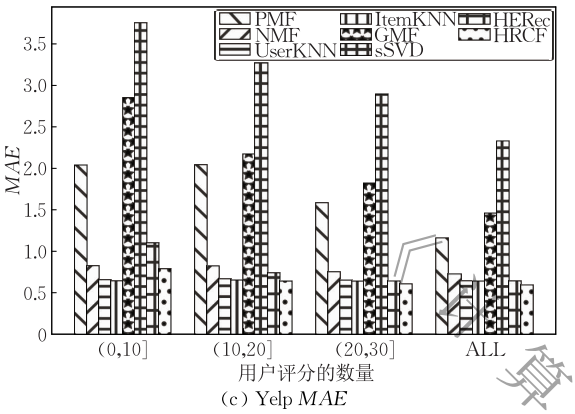
冷启动问题一直以来都是推荐模型面临的一个巨大挑战. 因此，能不能有效缓解冷启动问题是一个好的推荐模型必须具备的重要特征之一. 为了反映 HRCF 模型在处理冷启动问题上的有效性，本文将 HRCF 模型与其它模型在冷启动用户上的测试结果进行比较. 跟 Shi 等人[14]文献类似，我们首先根据用户评分的数目选出三组冷启动用户，分别为(0,10]、(10,20]和(20,30]. 然后分别计算所有模型在这三组上的 *MAE* 和 *RMSE*. 显然，值越小的 *MAE* 和 *RMSE* 代表模型在处理冷启动问题上有更好的性能，实验结果如图 5 所示.



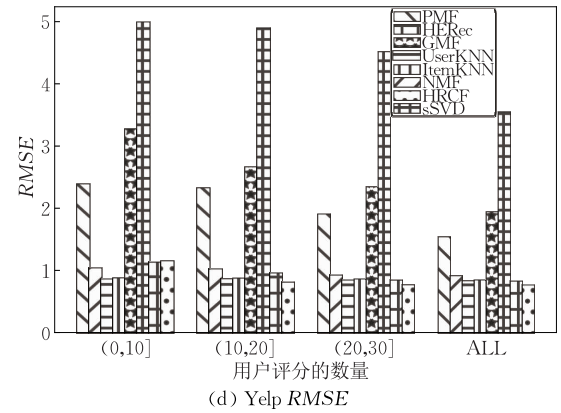
(a) Douban Book *MAE*



(b) Douban Book *RMSE*



(c) Yelp *MAE*



(d) Yelp *RMSE*

图 5 不同模型在冷启动用户上的推荐性能比较

从图 5 可以看出,总体上,所有模型在用户评分数量较多分组上有更小的 *MAE* 和 *RMSE*,反之亦然. 其次,在大多数情况下,HRCF 要优于 HERec,而 HERec 要优于其它模型,其中 GMF 和 sSVD 的性能较差. 一方面,这说明了 HIN 中丰富的结构和语义信息可以帮助推荐模型克服冷启动问题;另一方面也反映了 HRCF 模型在降低模型复杂度的同时也能有效缓解冷启动问题.

5.5 算法复杂度比较

这一节,我们给出了本文算法以及对比算法的时间复杂度,如表 6 所示. 从表中可以看出 PMF、NMF 和 GMF 的时间复杂度均为 $O(k|\Omega|d)$,其中 k 表示迭代次数,d 为潜在特征的维度,$|\Omega|$ 表示评分的数量. 此外,HERec 的时间复杂度为 $O(|P|d(m+n)+(|P|d+1)D+|P_U|m^2+|P_I|n^2)$,其中 $|P|$ 为元路径的总个数,$|P_U|$ 和 $|P_I|$ 分别表示用户和项目的元路径个数,D 表示嵌入的维度. 在本文中,表 4 中的 d、$|P|$、$|P_U|$ 和 $|P_I|$ 在本文中是很小的一个常数,D 为 128. 总的来说,在一定条件下,UserKNN 和 ItemKNN 的时间复杂度是要高于 PMF、NMF 和 GMF 的,其次是 HRCF 和 sSVD,HERec 的时间复杂度相对来讲要更大一些. 从这点来看,HRCF 模型可以在不增加太大复杂度的同时,还能使模型具有一定的有效性. 同时,这也说明 HRCF 模型在两者之间达到了一个新的平衡.

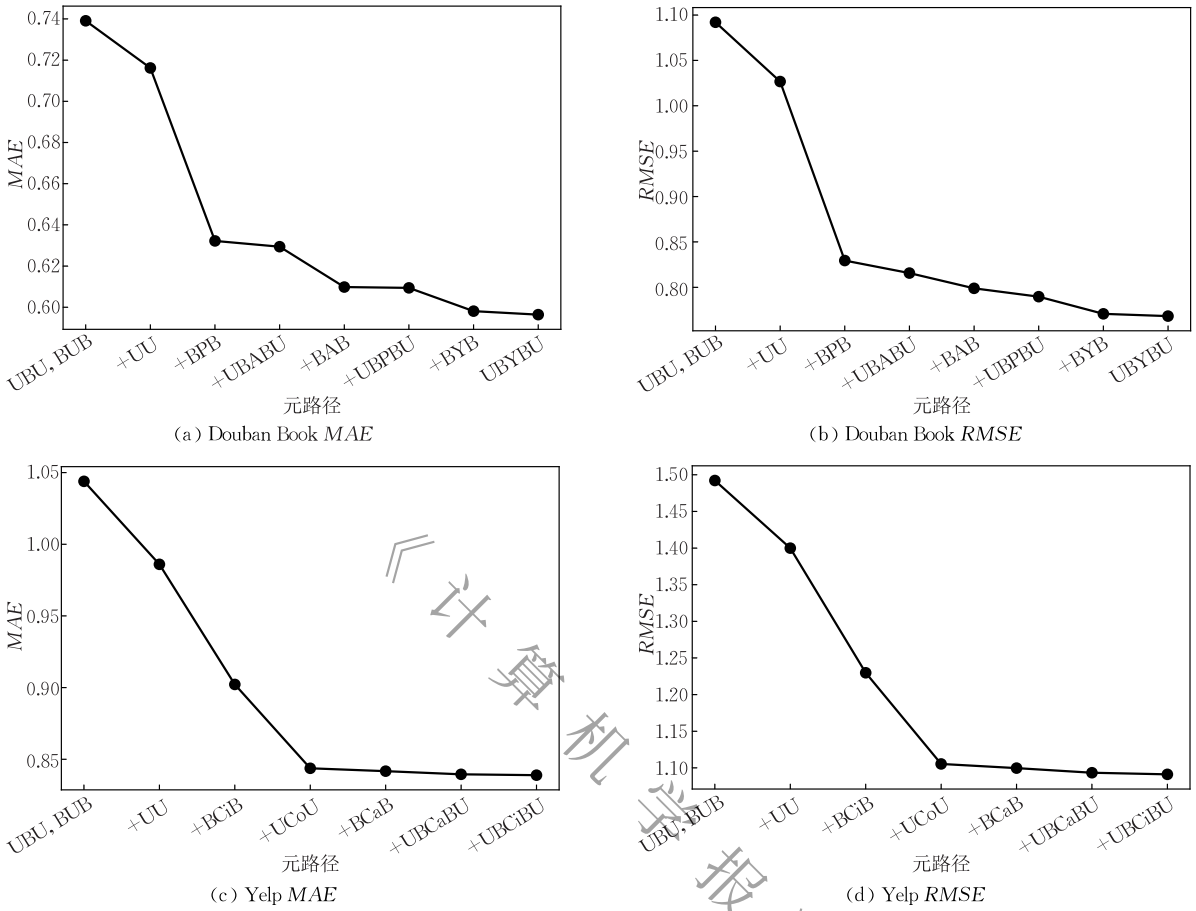
表 6 各个模型的时间复杂度

模型	复杂度
PMF	$O(k\|\Omega\|d)$
NMF	$O(k\|\Omega\|d)$
UserKNN	$O(m^2+\|\Omega\|)$
ItemKNN	$O(n^2+\|\Omega\|)$
GMF	$O(k\|\Omega\|d)$
sSVD	$O((\|\boldsymbol{P}_U\|+1)m^2+(\|\boldsymbol{P}_I\|+1)n^2+mn+k\|\Omega\|d)$
HERec	$O(\|P\|D(m+n)+k(\|P\|d+2)D)+O(\|P_U\|m^2+\|P_I\|n^2)$
HRCF	$O((\|\boldsymbol{P}_U\|+1)m^2+(\|\boldsymbol{P}_I\|+1)n^2+mn)$

5.6 元路径

在本文中,需要事先给定用户和项目的元路径集,并且文中所有的实验都是基于给定的元路径集来计算的. 事实上,如何根据具体的任务从 HIN 选择有意义且有效的元路径是一个巨大的挑战. 本节我们不讨论应该从 HIN 中选择哪些元路径,只展示在给定的元路径集中元路径对 HRCF 模型的影响. 具体地,因为 HRCF 模型需要通过用户和项目的邻居才能找到评分的邻居,所以我们在实验前会先给定一个元路径集合(包含一个用户的元路径和一个项目的元路径). 然后每次向该元路径集中不重复地添加一个元路径并计算 HRCF 模型的性能,直到该

集合包含所有的元路径. 实验结果如图 6 所示，可以看出，每增加一个元路径，HRCF 模型的性能都有所提高. 其中“+”表示添加元路径的操作. 这说明了在两个数据集上给定的元路径对 HRCF 模型的性能有积极的作用，同时也表明了 HIN 中丰富的语义和结构信息可以帮助推荐模型来提高性能.



(a) Douban Book *MAE*　(b) Douban Book *RMSE*

(c) Yelp *MAE*　(d) Yelp *RMSE*

图 6　元路径对 HRCF 模型性能的影响

此外，可以注意到在 Yelp 数据集中，随着元路径的增加，其 *MAE* 和 *RMSE* 最后都趋于稳定，而在 Douban Book 数据集中，*MAE* 和 *RMSE* 仍然还有下降的空间. 造成这个现象最主要的原因是在 Douban Book 的关系 *X*-*Y* 中，它们平均的度相对评分的数量来说较低，而 Yelp 数据集中较大.

6　总　结

本文提出了一种在 HIN 中无参数的元路径感知评分协同过滤推荐模型 HRCF. 该方法基于元路径来度量评分之间的相似性，并利用其对评分的邻居进行加权来进行预测. 事实上，HRCF 的设计与构造主要是为了使模型在易于执行的同时能保证其推荐性能，实验结果也验证了我们提出模型的有效性，即 HRCF 继承了 NBCF 优点的同时也保证了推荐精度. 在后续的工作中，我们将考虑通过图神经网络等相关技术来进一步提高模型的性能.

参　考　文　献

[1] Cao Long-Bing. Non-IID recommender systems: A review and framework of recommendation paradigm shifting. Engineering, 2016, 2(2): 212-224

[2] Francesco R, Lior R, Bracha S. Recommender Systems Handbook. New York, USA: Springer, 2015

[3] Chen Liang, Xu Yang-Jun, Xie Fen-Fang, et al. Data poisoning attacks on neighborhood-based recommender systems (CORR). http://arxiv.org/abs/1912.04109, 2020,3,10

[4] Feng Wei, Wang Jian-Yong. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1276-1284

[5] Yin Hong-Zhi, Sun Yi-Zhou, Cui Bin, et al. LCARS: A location-content-aware recommender system//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 221-229

[6] Xu Ke, Xu Yang-Jun, Min Hua-Qing, Cai Yi. Improving item ranking by leveraging dual roles influence. IEEE Access, 2018, 6: 57434-57446

[7] Sun Yi-Zhou, Han Jia-Wei. Mining heterogeneous information networks: A structural analysis approach. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20-28

[8] Ou Ming-Dong, Cui Pei, Wang Fei, et al. Comparing apples to oranges: A scalable solution with heterogeneous hashing//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, USA, 2013: 230-238

[9] Shi Chuan, Yu P S. Heterogeneous Information Network Analysis and Applications. Switzerland: Springer, 2017

[10] Shi Chuan, Liu Jian, Zhuang Fu-Zhen, Yu P S. Integrating heterogeneous information via flexible regularization framework for recommendation. Knowledge and Information Systems, 2016, 49(3): 835-859

[11] Guo Lei, Wen Yu-Fei, Wang Xin-Tong. Exploiting pre-trained network embedding for recommendations in social networks. Journal of Computer Science and Technology, 2018, 33(4): 682-696

[12] Shi Chuan, Zhang Zhi-Qiang, Ji Yu-Gang, et al. SemRec: A personalized semantic recommendation method based on weighted heterogeneous information networks. World Wide Web, 2019, 22(1): 153-184

[13] Rendle S, Krichene W, Zhang L, Andrson J R. Neural collaborative vs. matrix factorization revisited(CoRR). http://arxiv.org/abs/2005.09683, 2020

[14] Dai Ben, Wang Jun-Hui, Shen Xiao-Tong, Qu Annie. Smooth neighborhood recommender systems. Journal of Machine Learning Research, 2019, 20(16): 1-24

[15] Shi Chuan, Hu Bin-Bin, Zhao Wayne-Xin, Yu P S. Heterogeneous information network embedding for recommendation. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(2): 357-370

[16] Choi J W, Fowler R J, Vuduc R W. A roofline model of energy//Proceedings of the IEEE 27th International Parallel and Distributed Processing Symposium. Cambridge, USA, 2013: 661-672

[17] Zhang Yi-Wen, Wang Kai-Bin, He Qiang, et al. Covering-based web service quality prediction via neighborhood-aware matrix factorization. IEEE Transactions on Services Computing. DOI: 10.1109/TSC.2019.2891517, 2019(in Press)

[18] Luo Xin, Ouyang Yuan-Xin, Xiong Zhang, Yuan Man. The effect of similarity support in k-nearest-neighborhood based collaborative filtering. Chinese Journal of Computers, 2010, 33(8): 1437-1445(in Chinese)
(罗辛, 欧阳元新, 熊彰, 袁满. 通过相似度支持度优化基于k近邻的协同过滤算法, 计算机学报, 2010, 33(8): 1437-1445)

[19] McLaughlin M R, Herlocker J L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience//Proceedings of the 27th Annual International ACM SIGKDD Conference on Research and Development in Information Retrieval. Sheffield, UK, 2004: 329-336

[20] Liu Run-Ran, Jia Chun-Xiao, Zhou Tao, et al. Personal recommendation via modified collaborative filtering. Physica A: Statistical Mechanics and its Applications, 2009, 388(4): 462-468

[21] Huang Chuang-Guang, Yin Jian, Wang Jing, et al. Uncertain neighborhood's collaborative filtering recommendation algorithm. Chinese Journal of Computers, 2010, 33(8): 1369-1377(in Chinese)
(黄创光, 印鉴, 汪静等. 不确定近邻的协同过滤推荐算法. 计算机学报, 2010, 33(8): 1369-1377)

[22] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Madison, USA, 1998: 43-52

[23] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms//Proceedings of the 10th International Conference on World Wide Web Conference. Hong Kong, China, 2001: 285-295

[24] Kendall M G, Gibbons J D. Rank Correlation Methods. 4th Edition. Griffin, London, 1990

[25] Hong L, Doumith A S, Davision B D. Co-factorization machines: Modeling user interests and predicting individual decisions in twitter//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Roma, Italy, 2013: 557-566

[26] Shi Chuan, Li Yi-Tong, Zhang Jia-Wei, et al. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17-37

[27] Sun Yi-Zhou, Han Jia-Wei, Yan Xie-Feng, et al. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowent, 2011, 4(11): 992-1003

[28] Wang Cheng-Guang, Song Yan-Qiu, Li Hao-Ran, et al. KnowSim: A document similarity measure on structured heterogeneous information networks//Proceedings of the IEEE International Conference on Data Mining. Atlantic, USA, 2015: 1015-1020

[29] Shi Chuan, Kong Xian-Nan, Yu P S, Wu Bin. HeteSim, a general framework for relevance measure in heterogeneous networks. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2479-2492

[30] Yu Xiao, Ren Xiang, Gu Quan-Quan, et al. Collaborative filtering with entity similarity regularization in heterogeneous information networks//Proceedings of the 1st International Joint Conference on Artificial Intelligence Workshop on Heterogeneous Information Network Analysis. San Francisco, USA, 2013

[31] Luo Chen, Wei Pang, Wang Zhe, Lin Cheng-Hua. Hete-CF: Social-based collaborative filtering recommendation using heterogeneous relations//Proceedings of the 14th IEEE International Conference on Data Mining Series. Shenzhen, China, 2014: 917-922

[32] Zheng Jing, Liu Jian, Shi Chuan, et al. Recommendation in heterogeneous information network via dual similarity. International Journal of Data Science and Analysis, 2017, 3(1): 35-48

[33] Huang Li-Wei, Jiang Bi-Tao, Lv Shou-Ye, et al. Survey on deep learning based recommender systems. Chinese Journal of Computers, 2018, 41(7): 1619-1647(in Chinese)
(黄立威，江碧涛，吕守业等. 基于深度学习的推荐系统研究综述. 计算机学报，2018，41(7)：1619-1647)

[34] Fan Shao-Hua, Zhu Jun-Xiong, Han Xiao-Tian, et al. Metapath-guided heterogeneous graph neural network for intent recommendation//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA, 2019: 2478-2486

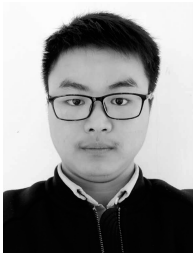
[35] Shi Chuan, Han Xiao-Tian, Song Li, et al. Deep collaborative filtering with multi-aspect information in heterogeneous networks (CORR). http://arxiv.org/abs/1909.06627, 2019,12,3

[36] Zhang Yong-Feng, Chen Xu. Explaninable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 2020, 1(14): 1-101

[37] Ruslan Salakhutdinov, Andriy Mnih. Probablistic matrix factorization//Proceedings of the 21st Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2007: 1257-1264

[38] Lee D D, Seung H S. Algorithms for non-negative matrix factorization//Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, USA, 2000: 556-562

[39] He Xiang-Nan, Liao Li-Zi, Zhang Han-Wang, et al. Neural collaborative filtering//Proceedings of the 26th International Conference on World Wide Web. Perth, Australia, 2017: 173-182



HE Yun-Fei, Ph. D. candidate. His research interests include heterogeneous information network and recommender systems.

ZHANG Yi-Wen, Ph. D., professor. His research interests include service computing, recommender systems and big data analysis.

LV Zhi-Hui, Ph. D., professor. His research interests include edge computing, cloud computing and big data.

YAN Deng-Cheng, Ph. D., lecturer. His research interests include complex network and big data analysis.

HE Qiang, Ph. D., senior lecturer. His research interests include service computing and big data analysis.

Background

As an integral part of online shopping malls and e-commerce, the recommendation system(RS) can help users find products they are interested in. And rating prediction is one of the most common and important tasks. Traditional rating prediction can usually be divided into model-based or neighborhood-based. Due to the simply, stability, and interpretable characteristics of neighborhood-based collaborative filtering (NBCF), it still attracts a lot of attention. However, these models only use historical user-item interaction information, resulting in poor recommendation performance and facing data sparseness and cold-start problems. In addition, many researchers integrate user social networks, product attribute information, etc. into the model to improve recommendation accuracy and alleviate these problems. Experimental results show that after considering additional information, the recommendation performance of this model has been effectively improved.

Therefore, it is necessary to expand useful information into the recommendation model to improve the recommendation performance. In recent years, because the heterogeneous information network contains rich structural and semantic information, a large number of HIN-based recommendation models have been proposed. HIN-based recommendation models can generally be divided into similarity-based and embedding-based. Using the meta-path in HIN, we can more realistically reflect the relationship between nodes and learn to embedding better, and then use them to make recommendations. In addition, with the rapid development of deep learning (DL), many researchers combine DL and HIN to make recommendations. However, the above models have shortcomings such as having many parameters and being difficult to implement while obtaining good recommendation performance.

To solve the above problems, this paper extends the neighborhood-based collaborative filtering in HIN, and proposes a rating-based collaborative filtering based on HIN. The model directly measures the similarity between ratings based on the meta-path and uses it as a weight to weight the rating neighbors to perform rating prediction. The experimental results on two public datasets show that our proposed model can obtain better recommendation performance and can effectively alleviate the cold-start problem.

This work is supported by the National Key R&D Program of China (2019YFB1704101、2019YFB1405000), the National Natural Science Foundation of China (61872002, U1936220, 61873309).