

FastGR: 一种基于神经协同过滤的群组推荐算法

尚文倩^{1,2}, 曹 原^{1,2}

(1. 中国传媒大学媒体融合与传播国家重点实验室, 北京 100024)

(2. 中国传媒大学计算机与网络空间安全学院, 北京 100024)

[摘要] 群组推荐问题的关键在于如何对组内各成员不同的偏好进行融合来适应所有成员的需求。基于神经协同过滤框架和注意力机制的群组推荐算法从数据中动态地学习融合策略, 相较于传统基于预定义策略的方法明显提升了推荐效果, 但模型训练及推理时间较长。本文在此基础上重构了群组偏好融合模块, 引入卷积神经网络来提取群组成员的特征, 从而实现偏好融合。在公开数据集上的实验表明, 本文算法比现有的算法具有更优的精度, 训练速度提高了 14 倍。

[关键词] 群组推荐算法, 卷积神经网络, 深度学习, 偏好融合, 神经协同过滤

[中图分类号] TP181 **[文献标志码]** A **[文章编号]** 1672-1292(2022)02-0029-06

FastGR: A Group Recommendation Algorithm Based on Neural Collaborative Filtering

Shang Wenqian^{1,2}, Cao Yuan^{1,2}

(1. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)

(2. School of Computer Science and Cyber Sciences, Communication University of China, Beijing 100024, China)

Abstract: The key problem of group recommendation is how to integrate the different preferences of group members to meet the needs of all members. The group recommendation algorithm based on neural collaborative filtering framework and attention mechanism dynamically learns fusion strategy from the data, significantly improving the recommendation effect compared with the traditional predefined strategy based method, but the model training and infer time is longer. In order to achieve the convergence of preferences, we reconstruct the group preference fusion module by adopting the convolution neural network to extract the feature of the group members. Experiments on open data sets show that the algorithm in this paper has better accuracy and improved the training speed by 14 times than that of the current algorithm.

Key words: group recommendation, convolutional neural network, deep learning, preference fusion, neural collaborative filtering

大数据时代, 推荐系统作为一种解决信息过载问题的有效手段, 被广泛应用于各大电商平台、新闻客户端、流媒体应用等。在为用户提供个性化推送的同时, 也提高了平台的收益。

目前, 推荐系统的研究大多面向单一用户^[1]。国内外学者提出了很多经典模型, 如 FM^[2]、Wide&Deep^[3]、DeepFM^[4]、DIN^[5]等, 在工业界取得了杰出的成果。然而这些模型无法直接应用于群组推荐问题。

在日常生活中, 社交网络、电商团购等平台将用户由相同的兴趣、地理位置、社会关系等特性聚集在一起形成群组, 群组内成员的偏好千差万别。在深度神经网络未流行之前, 一般采用预定义的策略(如均值策略)^[6]解决群组偏好融合问题, 但该方法不够灵活。文献[6]提出的 AGREE 模型基于表示学习方法计算出用户和物品的嵌入向量后利用注意力机制在推荐不同的物品时动态地为组内用户分配不同的权重, 相较于预定义策略显著提升了推荐效果, 达到了业界顶尖水平。但在执行群组推荐任务时对不同物品需

收稿日期: 2021-08-31。

基金项目: 国家重点研发计划项目(2018YFB0803701-1)、中国传媒大学中央高校基本科研业务费专项资金项目。

通讯作者: 尚文倩, 博士, 教授, 研究方向: 机器学习。E-mail: shangwenqian@cuc.edu.cn

要为组内每一位成员计算其注意力分数,群组偏好融合算法时间复杂度高,是制约整个模型效率提升的瓶颈。

本文设计了一种基于神经协同过滤的群组推荐算法(fast group recommendation, FastGR),在不损失模型推荐精度的前提下,优化群组偏好融合算法,加快训练与推理速度,从而提升模型效率。

1 相关工作

文献[7]提出的神经协同过滤(neural collaborative filtering, NCF)使用多层感知器代替基于矩阵分解协同过滤算法中的内积操作,其主要思想是通过神经网络训练用户和物品的嵌入向量,学习匹配函数,引入非线性特征,增强了模型的表达能力。FastGR 基于 NCF 框架训练群组内用户嵌入向量和物品嵌入向量,并提出一种一维全域卷积(global convolution)聚合群组成员偏好特征得到群组嵌入向量,无需预训练即可达到现有主流模型的性能且训练速度显著加快。文献[8]认为群组推荐可分为 3 个步骤:群组形成、群组建模和群组预测推荐,根据偏好融合发生阶段及融合内容的不同,将偏好融合方法分为 3 类:偏好模型融合、推荐结果融合和评分融合。

推荐系统中用户和物品的特征主要为高维 ID 类特征,其所导致的特征数据稀疏性、模型训练难度挑战就成为驱动表示学习的最直接源动力。具体地,在输入数据(例如用户 ID、物品 ID)预处理后,由嵌入层把稀疏的 ID 转化为固定长度的嵌入向量(embedding),即用向量 $X=(X_1, X_2, \dots, X_N)$ 表示某一实体,方便模型处理。目前最先进的 AGREE 采用了基于表示学习方法的 NCF 框架学习群组与物品交互行为,使用注意力机制对群组偏好进行融合。本文在 AGREE 开源代码上实验时发现模型执行群组推荐任务过程中耗时较长,存在优化空间。

TextCNN^[9-10]是一种文本分类模型,通过引入卷积神经网络^[11],靠卷积核窗口抽取特征,将语句中的词向量聚合为句子向量,具有对文本浅层特征的抽取能力强、速度快及对语序不敏感的特性。

受 TextCNN 启发,本文在 AGREE 的基础上利用 NCF 学习群组内成员的嵌入向量,提出了一种全域卷积的方式融合群组偏好特征。

2 基于神经协同过滤的群组推荐模型框架

FastGR 与 AGREE 的不同之处在于群组偏好融合模块,用卷积层替代了注意力层。FastGR 模型整体结构如图 1 所示。

图中, u_i 和 v_j 分别为用户 i 和物品 j 的嵌入向量。Group(l)代表第 l 个组,由 n 个索引为 u_k 的群组成员组成,将组内成员嵌入向量 u_k 横向拼接成二维矩

阵 $\begin{bmatrix} u_{k_{l,1}} \\ u_{k_{l,2}} \\ \vdots \\ u_{k_{l,n}} \end{bmatrix}$ 经卷积层提取特征后,由单层全连接网络融

合得到群组 l 的嵌入向量 g_l 。池化层通过哈达玛积^[3]来计算群组 l 对物品 j 的交互嵌入向量后,将其与用户 i 及物品 j 的嵌入向量拼接起来,如式(1)所示:

$$e_0 = \varphi_{\text{pooling}}(g_l(j), v_j) = \begin{bmatrix} g_l(j) \odot v_j \\ g_l(j) \\ v_j \end{bmatrix}. \quad (1)$$

通过共享隐含层,用户-物品与群组-物品的学习任务相互促进,协同优化。最后分别在预测层输出群组 l 对物品 j 的分数 y_{lj} 、用户 u 对物品 j 的分数 r_{uj} 。

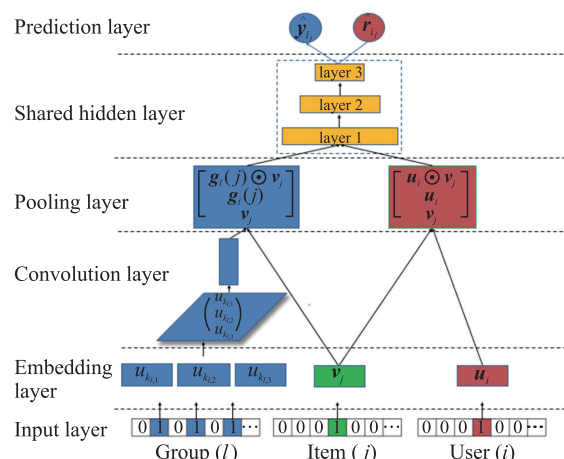


图 1 基于神经协同过滤的 FastGR 模型结构

Fig. 1 Structure of FastGR based on NCF

目标函数采用推荐系统中常用的基于回归的成对损失方法(regression-based pairwise loss):

$$\mathcal{L}_{\text{group}} = \sum_{(l,j,s) \in O'} (y_{ljs} - \hat{y}_{ljs})^2 = \sum_{(l,j,s) \in O'} (\hat{y}_{lj} - \hat{y}_{ls} - 1)^2, \quad (2)$$

式中, O 代表训练集;三元组 (l,j,s) 表示群组 l 与物品 j 有过交互行为,而与物品 s 无交互行为(负样本).

3 基于一维卷积的偏好融合

3.1 用户嵌入向量聚合

组内成员用户偏好融合算法决定了群组推荐结果. AGREE 模型基于表示学习框架得到组内成员与物品的嵌入向量后,利用注意力机制可以动态学习群组中用户所占的比重,最后由成员嵌入向量乘上该比重完成用户嵌入向量聚合. 这种偏好融合方法相较于传统基于预定义策略更加灵活,但在时间复杂度上还有很大的优化空间.

受 TextCNN 思想启发,一条语句由多个词语组成,而一个群组也是由多个用户组成,因此群组成员偏好的融合可以类似于句子中的多个词语嵌入向量聚合为句子嵌入向量. FastGR 卷积层结构如图 2 所示.

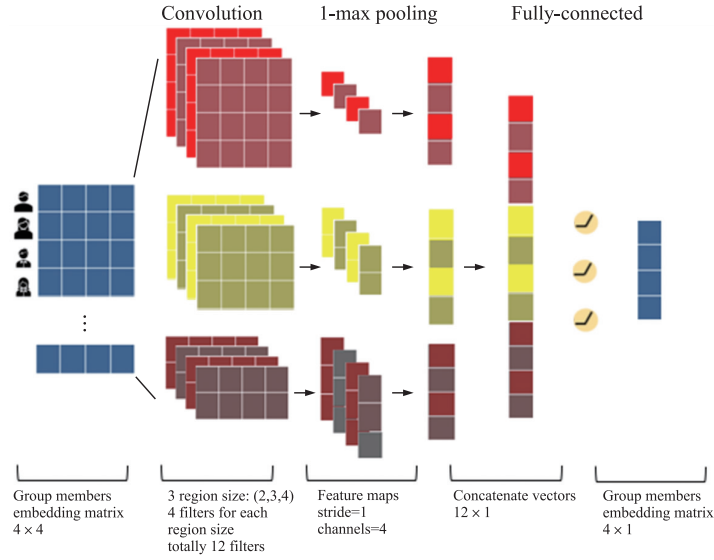


图 2 FastGR 群组成员偏好融合结构图

Fig. 2 Structure of FastGR group members prefer fusion

在 NCF 框架中可以由用户与物品的交互行为学习到用户的嵌入向量 u_i (其他框架如 YoutubeDNN^[11-13] 等也可生成用户嵌入向量),之后将组内成员看作构成句子的词语,通过纵向拼接组内成员的嵌入向量 u_i 得到一个二维矩阵 $u_{1:n}$,如式(3)所示:

$$u_{1:n} = u_1 \oplus u_2 \oplus \dots \oplus u_n. \quad (3)$$

假设某个组有 $n(n \geq 2)$ 位成员,与 TextCNN 相似,采用多个窗口 ($n = \{2, 3, 4 \dots n\}$) 的卷积核在成员嵌入向量堆积成的矩阵上做卷积运算得到用户 i 的 Feature Map c_i ,如式(4)所示,其中 $w \in \mathbf{R}^{nk}$,窗口大小为 $n \times k$:

$$c_i = f(w \cdot u_{i:i+h-1} + b). \quad (4)$$

使用 1-max pooling^[9] 对提取到的 Feature Map 进行降采样 $c = \max\{c\}$,以解决卷积核大小不同带来的 Feature Map 尺寸不一致问题.

考虑到卷积神经网络对局部特征敏感,无法有效捕获全局数据之间的长距离特征,FastGR 卷积层采用全域卷积,最大卷积核尺寸 ($n \times k$, k 为嵌入向量维度) 与组成员矩阵保持一致(如图 2 中的红色卷积核),并借鉴文献[12]的思想,FGCNN^[12] 使用重组层进行特征生成缓解这一问题. 具体地,将池化后的 Feature Map 展平成一个向量,然后使用单层的全连接层进行特征组合,全连接层可以解决成员间不同排列顺序导致的模型性能不稳定问题,增加了模型的鲁棒性:

$$g_i = f(w^T \cdot c + b). \quad (5)$$

由于卷积神经网络权重共享机制^[14-15],卷积层提取到的特征具有平移不变性,对位置信息不敏感,且

采用全域卷积提取整体特征,最后由全连接层进行特征重组,因此拼接顺序不影响特征提取^[16-20].

融合得到的群组嵌入向量 \mathbf{g}_l 可与物品嵌入向量 \mathbf{V}_j 拼接,经过多层隐含层,如式(6)进行高阶特征交互,最后由 sigmoid 激活函数映射到 0 到 1 之间,可以看作群组 l 对物品 j 的分数,如式(7)所示:

$$\begin{cases} \mathbf{e}_1 = \text{ReLU}(\mathbf{W}_1 \mathbf{e}_0 + \mathbf{b}_1), \\ \mathbf{e}_2 = \text{ReLU}(\mathbf{W}_2 \mathbf{e}_1 + \mathbf{b}_2), \\ \vdots \\ \mathbf{e}_h = \text{ReLU}(\mathbf{W}_h \mathbf{e}_{h-1} + \mathbf{b}_h). \end{cases} \quad (6)$$

$$\hat{y}_{lj} = \text{sigmoid}(\mathbf{w}^T \mathbf{e}_h). \quad (7)$$

式中, \mathbf{w}_h 、 \mathbf{b}_h 、 \mathbf{e}_h 为分别为第 h 层的权重矩阵、偏置向量和输出神经元.

3.2 群组成员偏好融合策略比较

基于注意力机制的 AGREE 和基于一维全域卷积的 FastGR 在聚合群组内用户嵌入向量的算法如下所示,本节比较两种模型在速度上的优势.

算法 1 AGREE 组内成员嵌入向量聚合

```

Input: Group inputs and item inputs
Output: Group embedding
1 for  $i=1$  to batch_size do;
2   get members for each group;
3   get interact items for each member;
4   members_embeds = embedding_layer( members );
5   item_embeds = embedding_layer( items );
6   attention_score = attention_layer( concat( members_embeds, item_embeds ) );
7   group_embeds = members_embeds  $\times$  attention_score;
8 end for;
9 group_embeds = concat( group_embeds );
10 return group_embeds.

```

算法 2 FastGR 组内成员嵌入向量聚合

```

Input: Group inputs
Output: Group embedding
1 for  $i=1$  to batch_size do;
2   get members for each group;
3   padding members to fixed length;
4   append members to members_list;
5 end for;
6 members_matrix = embedding_layer( members_list );
7 group_embeds = conv_layer( members_matrix );
8 return group_embeds.

```

二者训练时均采用 mini-batch 方法,上述算法伪代码即为单个小批次模型计算群组嵌入向量的步骤.

通过研究 AGREE 的群组偏好融合算法可以发现,制约模型训练速度的关键在于计算每一位组内成员的注意力分数,该操作在循环内部执行,循环次数为 batch size,是一个超参数,此处设定为 256. 对于每一次循环占用的时间假设为 t :

$$t = t_m + t_i + t_a + t_g, \quad (8)$$

式中, t_m 为计算每个组的成员列表占用的时间; t_i 为嵌入层计算物品嵌入向量占用的时间; t_a 为注意力层计算注意力分数占用的时间; t_g 为嵌入层计算组嵌入向量占用的时间.

由此可计算出一个批次占用的时间 T_{AGREE} :

$$T_{\text{AGREE}} = \text{batch_size} \times t. \quad (9)$$

而 FastGR 仅需在循环内得到每个组的成员列表,并补齐至相同长度(此处采用补零填充),然后在循环外部根据组成员列表计算这一批次内所有组的成员嵌入向量后直接通过卷积层提取特征,一个小批次占用的时间 T_{FastGR} 为:

$$T_{\text{FastGR}} = \text{batch_size} \times t_m + t_c, \quad (10)$$

式中, t_c 为卷积层占用的时间,只需要计算一次.

可见, T_{FastGR} 远小于 T_{AGREE} , 因而 FastGR 模型的速度要快于 AGREE 模型, 训练耗时显著降低.

4 模型评价指标

为了更公平地与 AGREE 模型比较, 本文在模型的各个维度上(如优化方法、学习率、负采样率等)都与 AGREE 保持一致, 采用命中率(hit ratio, HR)和归一化折扣累积增益(normalized discounted cumulative gain, NDCG)作为模型的评价指标. HR 衡量测试集中的项目是否出现在模型预测的 top-K 列表里, HR 值越大, 说明模型推荐命中的越多, 反映了模型的准确度. NDCG 衡量排序质量, 即测试项目出现在 top-K 列表中位置越靠前得分越高.

$$\text{HR@K} = \frac{\text{Number of Hits@K}}{GT}, \quad (11)$$

$$\text{NDCG@K} = Z_K \sum_{i=1}^K \frac{2^{r_i} - 1}{\log 2(i+1)}, \quad (12)$$

式中, GT 表示所有测试项目的集合; Z_K 是归一化系数, 将 DCG 的值保持在 0~1 之间; r_i 表示处于位置 i 的推荐结果的相关性, 若命中则 r_i 为 1, 否则为 0.

5 实验结果及分析

在公开数据集上对比了 NCF、GREE (AGREE 移除注意力模块)、AGREE、FastGR 群组推荐任务的性能和效率, 以验证本文方法的可行性与高效性.

5.1 数据集

CAMRa2011 为 AGREE 采用的公开数据集, 包含了个人用户和群组观看电影的评分记录, 通过将有过评分的项目标记为 1 即正样本、随机采样若干个未观看过的电影标记为 0 即负样本进行训练.

5.2 实验和结果

本文基于 PyTorch 实现 FastGR 模型, 其他基线模型 NCF、GREE 和 AGREE 为文献[6]的开源版本.

模型优化方法为 RMSProp, 嵌入层采用 Xavier 均匀初始化策略^[6], top-K = 5, 采用学习率固定步长衰减策略, 验证方法同 AGREE 一致采用 leave-one-out 进行评估. 如表 1 所示, 在相同实验条件下, FastGR

表 1 在 CAMRa2011 上的实验结果

Table 1 The performance results on CAMRa2011

模型	group_loss	Hit Ratio	NDCG
NCF	0.290 5	0.580 3	0.389 6
GREE	0.289 0	0.588 3	0.387 1
AGREE	0.288 3	0.588 3	0.395 5
FastGR	0.277 9	0.593 1	0.398 3

的 HR 提高了 0.81%, NDCG 提高了 0.70%. 如图 3 所示, 本文分别从群组推荐任务的 loss、HR、NDCG 3 个指标比较两个模型 20 轮次迭代训练过程, 可以看出 FastGR 在群组推荐任务中模型收敛更快, 且性能上要优于 AGREE. 从表 2 可以看出, FastGR 在 CPU 和 GPU 上效率均优于 AGREE, 不同操作系统上训练速度

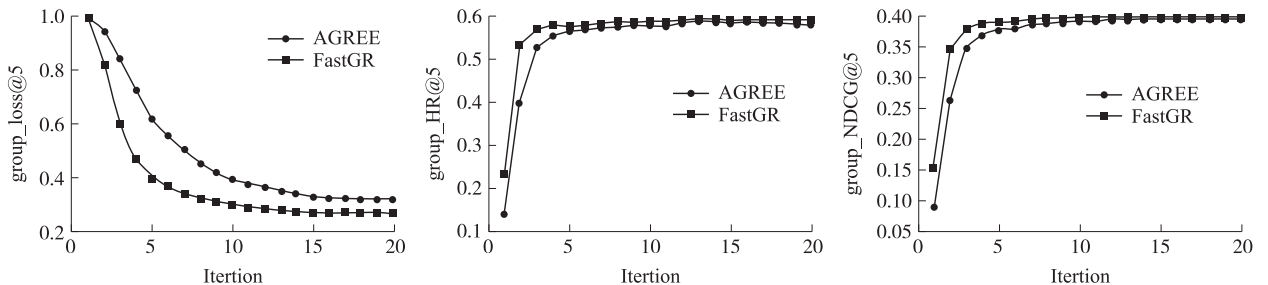


图 3 模型性能比较

Fig. 3 Performance comparison on CAMRa2011

均提升 14 倍左右。

如图 4 所示,由于用户嵌入向量聚合模块重构后时间复杂度降低,模型训练速度得到了提升。这是由于精简了模型结构并大大减少了运算量,这一改进使得模型部署上线工程难度降低,显著提升了模型效率。

表 2 在 CAMRa2011 上的训练与推理时间

Table 2 Efficiency results on CAMRa2011				
模型	CPU Training Time/h	GPU Training Time/h	CPU Prediction Time/s	GPU Prediction Time/s
AGREE	22.27	7.615	25.0	47.1
FastGR	1.623	0.3555	15.1	14.8

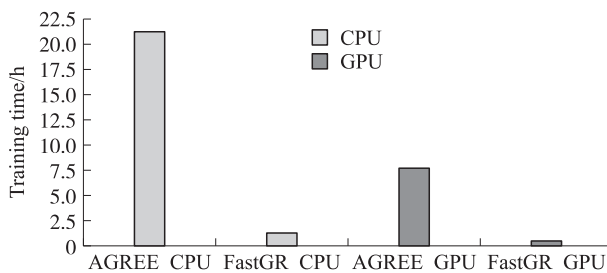


图 4 模型效率比较

Fig. 4 Efficiency comparison on CAMRa2011

6 结论

如何为群组推荐系统设计更快速、更合理的偏好融合算法,本文在现有方法的基础上提出了一种基于神经协同过滤和一维全域卷积的群组推荐算法 FastGR,通过神经协同过滤框架学习用户的嵌入向量,利用一维卷积来提取群组成员的特征从而实现偏好融合。公开数据集上的实验结果表明,本文方法在群组推荐准确度尤其是效率方面优于现有方法。

[参考文献] (References)

- [1] ZHANG J,GAO C,JIN D P,et al. Group-buying recommendation for social e-commerce[C]//Proceedings of the 2021 IEEE 37th International Conference on Data Engineering(ICDE). Chania,Greece:IEEE,2021.
- [2] RENDLE S. Factorization machines[C]//2010 IEEE International Conference on Data Mining. Sydney,Australia:IEEE,2010.
- [3] CHENG H T,KOC L,HARMSSEN J,et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston,USA:ACM,2016.
- [4] GUO H,TANG R,YE Y,et al. DeepFM:a factorization-machine based neural network for CTR prediction[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne,Australia:AAAI Press,2017.
- [5] ZHOU G R,SONG C R,ZHU X Q,et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London,UK:ACM,2018.
- [6] CAO D,HE X N,MIAO L H,et al. Attentive group recommendation[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor,USA:ACM,2018.
- [7] HE X N,LIAO L Z,ZHANG H W,et al. Neural collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web. Perth,Australia:IWWWCS,2017.
- [8] 吴云昌,刘柏嵩,王洋洋,等. 群组推荐分析与研究综述[J]. 电信科学,2018,34(12):71-83.
- [9] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[C]//Proceedings of the 2013 International Conference on Learning Representation(ICLR2013). Scottsdale,USA:ICLR,2013.
- [10] ZHANG Y,WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820,2015.
- [11] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.
- [12] LIU B,TANG R M,CHEN Y Z,et al. Feature generation by convolutional neural network for click-through rate prediction[C]//Proceedings of the World Wide Web Conference 2019. San Francisco,USA:ACM,2019.
- [13] COVINGTON P,ADAMS J,SARGIN E. Deep neural networks for youtube recommendations[C]//Proceedings of the 10th ACM Conference on Recommender Systems. Boston,USA:ACM,2016.
- [14] CAO D,HE X N,MIAO L H,et al. Social-enhanced attentive group recommendation[J]. IEEE Transactions on Knowledge and Data Engineering,2021,33(3):1195-1209.

(下转第 47 页)