Computer Science and Application 计算机科学与应用, 2022, 12(6), 1665-1673
Published Online June 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.126167

Hans汉斯

基于邻域聚合的协同过滤推荐模型

廖春节，邓鉴格，徐　涛*

西北民族大学中国民族语言文字信息技术教育部重点实验室，甘肃　兰州

收稿日期：2022年5月24日；录用日期：2022年6月22日；发布日期：2022年6月29日

摘　要

用户和待预测物的嵌入表示是推荐系统的核心，这种嵌入一般通过映射的方式获得，然而上述方法不能有效地利用到用户交互的协作信号，因此生成的嵌入不能很好地发挥协同过滤的效果。为了解决这个问题，本文研究了基于邻域聚合的协同过滤(NACF)模型的方法，该方法将用户交互集成到嵌入中，再利用嵌入传播将协作信号以高阶连通性的形式编码。最后该模型与贝叶斯个性化排序的矩阵分解(BPRMF)和图神经网络协同过滤(NGCF)在MovieLens数据集上的实验结果表明，本文的方法取得的效果更加优越。

关键词

邻域聚合，个性化推荐，协同过滤，协作信号

Collaborative Filtering Recommendation Model Based on Neighborhood Aggregation

Chunjie Liao, Jiange Deng, Tao Xu*

Key Laboratory of Information Technology and Education Ministry of Chinese Ethnic Languages and Writings, Northwest Minzu University, Lanzhou Gansu

Received: May 24th, 2022; accepted: Jun. 22nd, 2022; published: Jun. 29th, 2022

Abstract

The embedded representation of users and objects to be predicted is the core of the recommendation system. Such embedding is generally obtained through mapping. However, the above method cannot effectively utilize the cooperative signals of user interaction, so the generated embedding cannot give good play to the effect of collaborative filtering. To solve this paper stu-

*通讯作者。

文章引用: 廖春节, 邓鉴格, 徐涛. 基于邻域聚合的协同过滤推荐模型[J]. 计算机科学与应用, 2022, 12(6): 1665-1673.
DOI: 10.12677/csa.2022.126167

dies a collaborative filtering (NACF) model based on neighborhood aggregation, which integrates user interaction into embedding and encodes cooperative signals in the form of high-order connectivity by embedding propagation. Finally, the experimental results of this model, Bayesian personalized ordering matrix decomposition (BPRMF) and graph neural network cooperative filtering (NGCF) on MovieLens data set show that the proposed method achieves better results.

Keywords

Neighborhood Aggregation, Personalized Recommendation, Collaborative Filtering, Collaborative Signal

Copyright © 2022 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

面对海量的信息，用户从中获取自己所需要的信息所消耗的时间将大幅增加，而推荐系统作为解决信息过载的有效手段，主要功能是在海量的数据中为用户找出当前最合适的信息，以缓解信息过载的问题。传统的推荐系统主要分为 3 种：基于内容的推荐、基于协同过滤推荐和基于混合推荐。当前推荐系统中应用最广的算法是协同过滤，协同过滤是通过用户过往的项目评分差异来计算用户之间的相似度，然后根据用户之间的历史评分和相似度计算效用值，最后得出用户的潜在偏好。协同过滤最常见的范例是嵌入来表示用户和项目，并且基于已经嵌入的向量进行预测。后来一些研究发现通过改进用户子图结构，也就是使用它的单跳邻居来改进嵌入学习[1]，可以提高嵌入的质量。

为了深入使用具有高跳邻居的子图结构，Wang 等人提出了 NGCF，并为 CF 实现了优秀的性能。它从图卷积网络(GCN)中获得了灵感，遵循同样的原则细化嵌入的传播规则：特征变换、邻域聚合和非线性激活。虽然 NGCF 已显示出优秀的结果，但 He 等人认为其设计相当繁重和许多操作都是在没有正当理由的情况下直接从 GCN 继承的。因此，他们对 NGCF 进行了简化称其为 NACF，证明了从 GCN 继承的两种操作特征转换和非线性激活对 NGCF 的有效性没有贡献。更令人惊讶的是，删除它们会显著提高准确性。

本文以电影推荐任务为载体研究 NACF 模型。利用 MovieLens 标准数据集设计实验验证了基于邻域聚合的协同过滤电影推荐系统的有效性，主要工作如下。

第一，设计实现基于邻域聚合的协同过滤电影推荐系统。相较于传统的推荐算法，图卷积神经网络具有更为强大的数据表征学习能力，同时相较于传统的深度学习模型，图卷积神经网络通过将原始输入数据转化为以图的结构数据，不仅能够学习数据对象本身的特征信息，并且还能学习数据对象在图中的空间特征信息，从而能够进一步提升特征提取的效果[2]。因此，本文采用在 NGCF 上进行改进的 NACF 来构建电影推荐系统。

第二，基于 MovieLens 标准数据集设计对比实验。从推荐算法模型预测评分的准确率以及生成推荐列表的准确性两个维度对传统推荐算法模型与 NACF 模型的推荐结果进行对比、评价、验证电影推荐系统的有效性。

2. 基于邻域聚合的协同过滤推荐模型

NGCF 是一个用于协同过滤的沉重而繁琐的 GCN 模型[3]。正是因为这些原因，本文选择一个既拥有

GCN 基本成分又轻而有效的模型——NACF。与 NGCF 相比 NACF 更容易训练和维护，从技术上讲更容易分析模型行为，在本节中，我们首先介绍 NACF 模型，如图 1 所示。然后，我们对 NACF 进行了深入分析，以展示其设计背后的合理性。最后，我们描述了如何进行推荐模型训练。

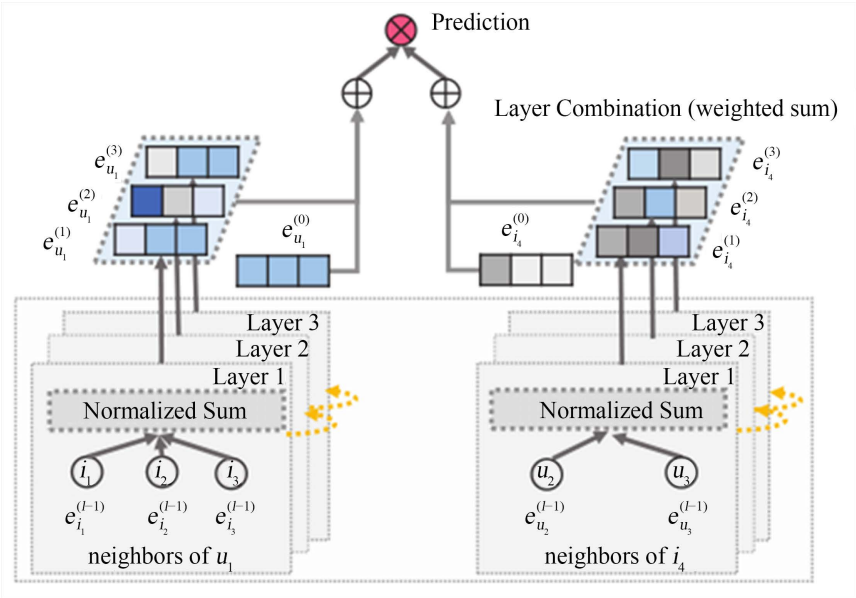

Figure 1. Collaborative filtering model architecture based on neighborhood aggregation
图 1. NACF 模型架构

2.1. 模型介绍

GCN 为了实现通过在平滑图上的特征来学习节点的表示，它迭代地执行图卷积，也就是将邻居的特征聚合为目标节点的新表示。这种邻域聚合可以抽象公式如下所示。

$$e_u^{(k+1)} = \mathrm{AGG}\left(e_u^{(k)}, \left\{e_i^{(k)} : i \in N_u\right\}\right) \tag{1}$$

AGG 是一个聚合函数(图卷积的核心)，它注意目标节点以及和目标节点相邻节点的第 k 层表示。很多方法中都用到了 AGG，比如 GIN、GraphSAGE、BGNN 等，这些方法的聚合器都用到了 AGG [4]。虽然他们在节点或者图形分类任务中效果良好，但是它们设计繁琐计算复杂成为协同过滤的负担[5]。

2.1.1. 简化的图卷积

在 NACF 中使用的是简化后的加权和聚合器，去掉了使模型变得更加复杂的特征变换和线性激活。NACF 中的传播规则被定义为。

$$e_u^{(k+1)} = \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} e_i^{(k)} \tag{2}$$

$$e_i^{(k+1)} = \sum_{i \in N_i} \frac{1}{\sqrt{|N_i|}\sqrt{|N_u|}} e_u^{(k)} \tag{3}$$

为了避免嵌入的规模随着图卷积的运算增加而增加，这里的对称归一化项 $\frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}}$ 依旧遵循 GCN 的标准设计。

图卷积运算通常需要聚合扩展的邻域，并且需要专门处理的自连接[6]。与大多数图卷积不同的是，简化后的图卷积只聚合已经链接的邻居，而不会集成目标节点本身，这是根据层组合操作基本上与自连接效果相同[7]。

2.1.2. 层组合

在 NACF 中，只有第 0 层的嵌入可以训练模型的参数，当给出表示所有用户的 $e_u^{(0)}$ 和表示所有项目的 $e_i^{(0)}$ 时，就可以通过等式(2)中定义的 LGC 计算更高层的嵌入。在经过 k 层的 LGC 之后，我们再组合每一层的嵌入就得到用户和项目的最终表示。

$$e_u = \sum_{k=0}^{k} \alpha_k e_u^{(k)}; e_i = \sum_{k=0}^{k} \alpha_k e_i^{(k)} \tag{4}$$

式中的 α_k 是大于 0 的，它表示的是第 k 层嵌入在最终嵌入中的比重。它是一个可以手动调节的参数，可以看作一个自动优化模型的参数。在实验中，将 α_k 统一设置成 $\frac{1}{K+1}$ 实验效果一般会比较好。所以，在本实验中不对 α_k 进行优化来避免使整个模型复杂化，保持模型的简单性。本文实验采用层组合来获取最终表示的原因有三个。1) 在层数增加的同时嵌入会出现过平滑的现象，所以如果只使用最后一层是不对的。2) 不同层的嵌入捕获到的信息是不同的，因此使用层组合将使表示更加准确。3) 不同层的嵌入与加权和结合可以得到图卷积与自连接的效果。

用户和项目的最终表示的内积是模型预测。

$$\hat{y}_{ui} = e_u^{\mathrm{T}} e_i \tag{5}$$

作为生成推荐排名的分数。

2.1.3. 矩阵形式

模型的矩阵形式是让用户项的交互矩阵为 $R \in R^{M \times N}$，其中 M 表示用户数量 N 表示项目数量，当 R_{ui} 为 1 时表示用户与项目交互，当 R_{ui} 为 0 时表示用户与项目不会交互。然后就得到了用户－项目图的邻接矩阵。如下所示。

$$A = \begin{pmatrix} 0 & R \\ R^{\mathrm{T}} & 0 \end{pmatrix} \tag{6}$$

$E^{(0)}$ 表示第 0 层的嵌入矩阵 $E^{(0)} \in R^{(M+N) \times T}$，$T$ 代表嵌入的大小。然后我们可以得到 LGC 的矩阵等价形式为：

$$E^{(K+1)} = \left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right) E^{(k)} \tag{7}$$

D 代表的是 $(M+N) \times (M+N)$ 对角矩阵，D_{ii} 代表邻接矩阵 A 的第 i 行向量中非零项的数量。最后，我们得到了用于模型预测的最终嵌入矩阵：

$$\begin{aligned} E &= \alpha_0 E^{(0)} + \alpha_1 E^{(1)} + \alpha_2 E^{(2)} + \cdots + \alpha_K E^{(K)} \\ &= \alpha_0 E^{(0)} + \alpha_1 \tilde{A} E^{(0)} + \alpha_2 \tilde{A}^2 E^{(0)} + \cdots + \alpha_K \tilde{A}^K E^{(0)} \end{aligned} \tag{8}$$

其中，$\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ 代表对称归一化矩阵。

2.2. 模型分析

我们将对模型进行分析来说明基于邻域聚合的协同过滤电影推荐系统设计的合理性。首先来讨论与去掉特征变换和非线性激活的 GCN (SGCN)的联系，SGCN 是一个线性 GCN 模型，它的自连接是集成到

图卷积中的，这就表明通过层组合 NACF 中拥有自连接的影响，所以 NACF 不用在邻接矩阵中添加自连接[8]。其次，我们来分析模型与个性化神经预测传播(APPNP)的关联，APPNP 是一个 GCN 的变体，它通过 PageRank 解决了过度平滑的问题；这显示出 NACF 和 APPNP 之间存在潜在等效性，NACF 在远程传播方面的过平滑是可以控制的。

二阶嵌入光滑性

由于 NACF 是线性的且具有简单性，这使得模型更容易分析，下面我们就深入地了解它是怎么平滑嵌入的。以一个两层的 NACF 来证明图模型的合理性，以用户端为例，第二层平滑了交互的项目上有重叠的用户。

$$e_u^{(2)} = \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} e_i^{(1)} = \sum_{i \in N_u} \frac{1}{|N_i|} \sum_{v \in N_i} \frac{1}{\sqrt{|N_u|}\sqrt{|N_u|}} e_i^{(0)} \tag{9}$$

可以看出，如果另一个用户 v 与目标用户 u 有共同交互关系，则 v 对 u 的平滑强度用系数表示(否则为 0)。

$$c_v \to u = \frac{1}{\sqrt{|N_u|}\sqrt{|N_v|}} \sum_{i \in N_u \cap N_v} \frac{1}{|N_i|} \tag{10}$$

这个系数可以解释为：二阶邻居 v 对 u 的影响，这个影响是由以下几点决定的，1) 共同交互的项目数量越多影响就越大；2) 交互的项目越不受欢迎越能表现用户的个性，所以交互的项目的受欢迎程度越低影响越大；3) 邻居 v 的活跃的越低影响越大。

2.3. 模型训练

LightGCN 的可训练参数仅为第 0 层的嵌入，即 $\theta = \left\{e^{(0)}\right\}$；也就是说，NACF 的复杂性与 MF 是相同的。模型采用贝叶斯个性化排名(BPR)损失，它的损失是成对的，它推崇预测一个观察到的项目要高于它没有观测到的项目：

$$L_{BPR} = -\sum_{u=1}^{M} \sum_{i \in N_u} \sum_{j \notin N_u} In\sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + \lambda \left\| E^{(0)} \right\|^2 \tag{11}$$

式中 λ 控制的是 L_2 的正则化权重们使用的是 am 优化器，并以 mini-batch 的方式使用它。

因为我们的模型中没有使用特征变换矩阵，所以模型没有使用通常在 GCNs 和 NGCF 中使用的 dropout 机制。为了解决这一点在模型的嵌入层上强制 L_2 正则化，这样就可以防止过拟合。这就凸显出 NACF 比 NGCF 更容易训练的特点。

3. 实验

3.1. 电影数据集

本文采用 MovieLens 1M 数据集，MovieLens 是由 GroupLens 研究组根据 MovieLens 网站提供的数据制作数据集，这个数据集中包含了 6000 个用户，有关于 4000 部电影的一亿条评论。MovieLens 数据集可以说是推荐系统领域中最为经典的数据集之一。数据集中有三个文件，分别是用户数据 users.dat 和电影数据 movies.dat 以及评分数据 ratings.dat。

数据集处理

在电影数据这个文件中包含了三个字段，分别是电影 ID 字段、流派字段、标题字段。在处理电影 ID 字段时不用做出改变，流派字段时一种分类的字段，因为有些电影时属于多个分类的，为了方便处理所

以将所有电影的流派转化成字符串映射到数字的字典，最后将电影的流派转成一个数字列表。最后是标题字段，处理的方法与电影流派处理的方法是类似的，值得注意的是，标题中的年份需要去掉，保持流派字段和标题字段长度相同，这样主要是为了方便神经网络的处理，空白部分用“<PAD>”对应的数字填充。

3.2. 实验环境

3.2.1. 比较方法

我们将基于邻域聚合的协同过滤(NACF)与基于贝叶斯个性化排序的矩阵分解(BPRMF)和图神经网络协同过滤(NGCF)进行比较，与 NACF 竞争的方法主要是 NGCF，因为它已经被证明优于几种方法了，包括基于 GCN 的模型 GC-MC 和 PinSage [9]，因为比较是在同一条件下同一数据集进行的所以就不与这些方法进行比较了，BPRM 是通过最大化后验概率进行学习的，它这样做的目的是为了使访问过的项目优于没有访问过的项目，它使用的是用户的点击矩阵而非评分矩阵，所以它最后得到的矩阵与原始的矩阵关联性不大。

3.2.2. 超参数设置

在本项目中，模型嵌入的大小与 NGCF 都固定为 64，嵌入参数初始化的方法是 Xavier。NACF 用 Adam 优化，默认学习率和 batch-size 分别为 0.001 和 1024。L_2 正则化系数 λ 在 $\left\{1e^{-6},1e^{-5},\cdots,1e^{-2}\right\}$ 中搜索，在一般情况下，最优值为 $1e^{-4}$。层组合系数 α_K 统一设置成 $\frac{1}{1+k}$，k 代表层数，K 的范围在 1 到 4 之间，当 k 等于 3 时结果比较理想。

3.3. 性能比较

首先我们将 BPRMF 与 NGCF 进行比较，如图 2 所示是 BPRMF 与 NGCF 的训练损失曲线，它们最后的结果相差不大但还是 NGCF 相对优秀一点，如图 3 所示是 BPRMF 与 NGCF 的召回率曲线，由图可知 NGCF 的召回率在绝大多数时候都是要高于 BPRMF 的，由图 2 和图 3 可知 NGCF 的性能要优于 BPRMF。

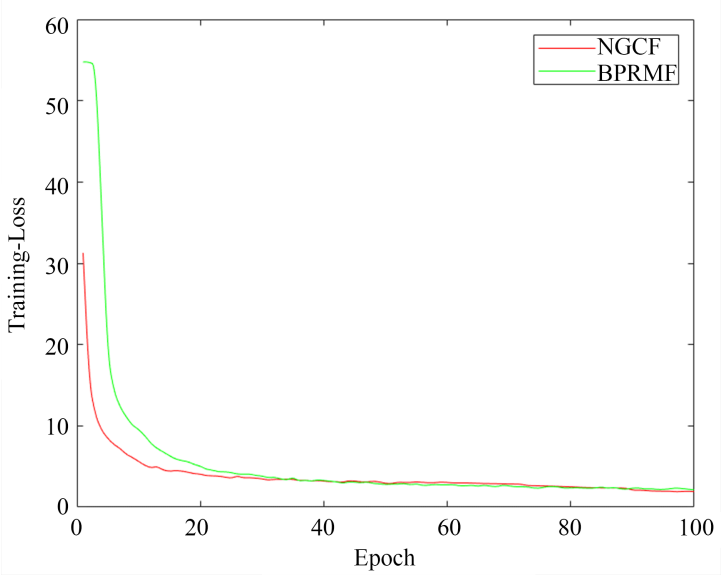

Figure 2. Loss comparison
图 2. 损失比较

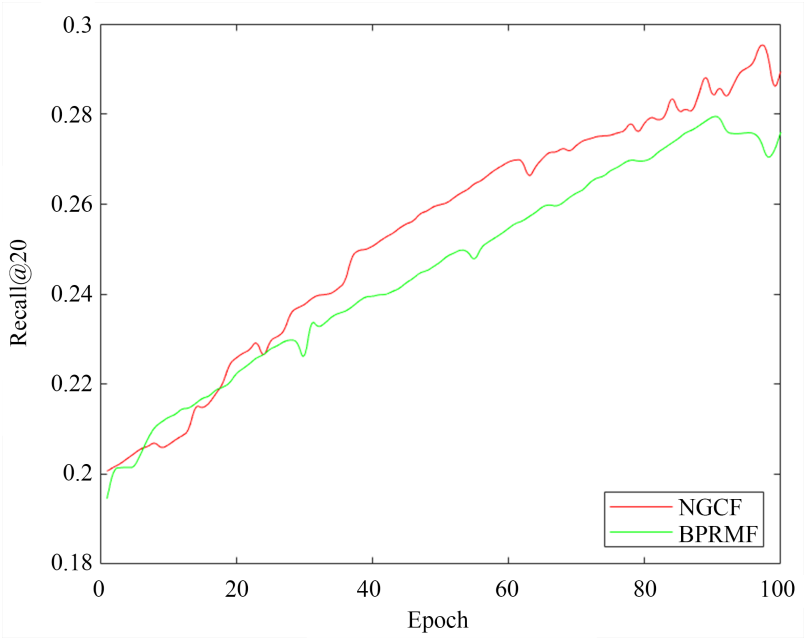

Figure 3. Comparison of recall rates
图 3. 召回率比较

其次我们将 NACF 与 NGCF 进行比较，我们在图 4 和图 5 中绘制了 NACF 的损失率和召回率的训练曲线，用来揭示我们的模型的优势以及明确训练的过程。图中显示的结果如下：

在所有情况下基于邻域的协同过滤表现都远远优于 NGCF，在训练过程中，NACF 相比于 NGCF 损失始终都很低，这说明 NACF 比 NGCF 更适合训练数据。NGCF 最高的召回率是 0.2958，而 NACF 在 4 层设置下可以达到 0.3468，相比之下基于邻域聚合的协同过滤要比 NGCF 优越很多。

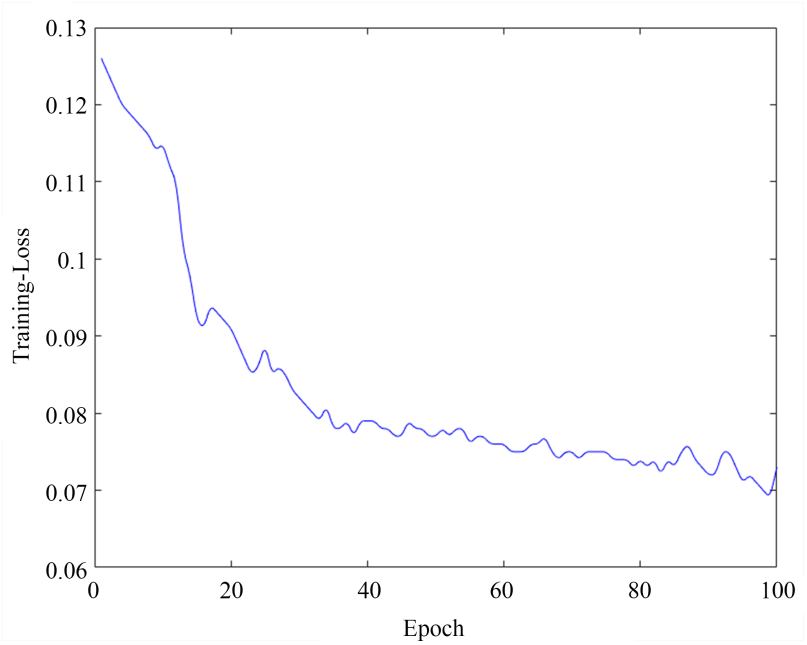

Figure 4. NACF training loss
图 4. NACF 训练损失

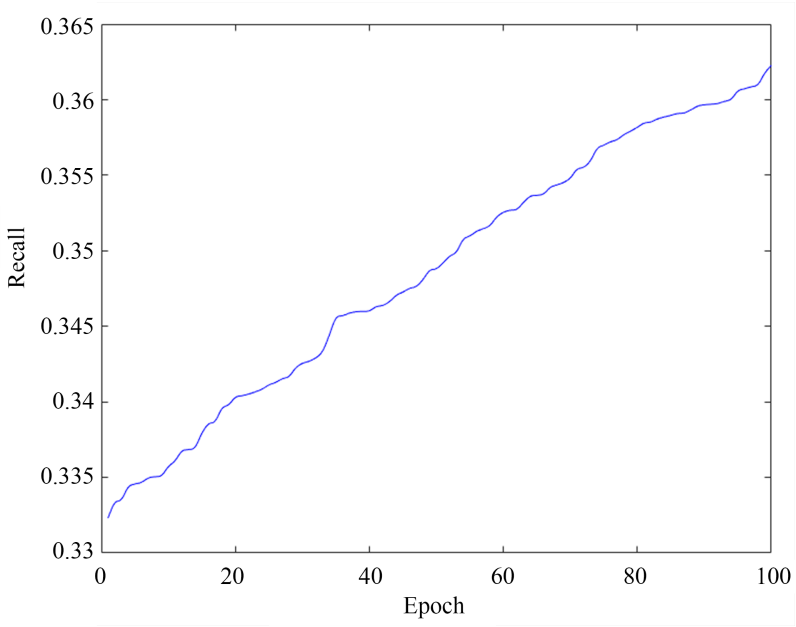

Figure 5. NACF recall
图 5. NACF 召回率

最后我们绘制了表 1，表 1 显示了与对比方法的性能比较。我们展示的是每种方法可以获得的最优效果，从表中可以看出基于邻域聚合的方法在 MovieLens 数据集上始终是优于其他两种方法的。

Table 1. Model performance comparison
表 1. 模型性能比较

模型	Recall	NDCG
BPRMF	0.27543	0.35831
NGCF	0.29427	0.36723
NACF	0.36227	0.42951

4. 结论

本文首次将 NACF 应用到电影领域，NACF 是一个在图神经网络协同过滤上改进的模型，它相比于图神经网络协同过滤有计算更加简单、计算速度更快、准确率更高的优点。将它与另外两个比较经典的算法进行比较，比较结果得出基于邻域聚合的协同过滤电影推荐系统效果要明显优于另外两种模型。

本系统仍有改进的空间，可以在层组合之间添加权重，以便为不同的用户实现个性化推荐。

基金项目

甘肃省青年科技基金(21JR1RA21)、中央高校基本科研业务费专项资金项目(31920210134)、国家档案局科技项目(2021-X-56)与甘肃省档案科技项目(GS-2020-X-07G)。

参考文献

[1] He, X.N., Deng, K., Wang, X., Li, Y., Zhang, Y.D. and Wang, M. (2020) LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *Proceedings of the* 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 25-30 July 2020, Virtual Event, China, 10 p.

[2] Zhu, H.M., Feng, F.L., He, X.N., Wang, X., Li, Y., Zheng, K. and Zhang, Y.D. (2020) Bilinear Graph Neural Network with Neighbor Interactions. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main Track*, 1452-1458.

[3] Zhao, C., Li, C.L. and Fu, C. (2019) Cross-Domain Recommendation via Preference Propagation GraphNet. In: *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, 2165-2168. https://doi.org/10.1145/3357384.3358166

[4] Zhu, H.M., Feng, F.L., He, X.N., Wang, X., Li, Y., Zheng, K. and Zhang, Y.D. (2020) Bilinear Graph Neural Network with Neighbor Interactions. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 1452-1458.

[5] Wu, F., Souza, A., Zhang, T.Y., Fifty, C., Yu, T. and Weinberger, K. (2019) Simplifying Graph Convolutional Networks. *Proceedings of the* 36*th International Conference on Machine Learning*, Long Beach, California, 20 June 2019, 6861-6871.

[6] He, X.N., Tang, J.H., Du, X.Y., Hong, R.C., Ren, T.W. and Chua, T.-S. (2019) Fast Matrix Factorization with Non-uniform Weights on Missing Data. *IEEE Transactions on Neural Networks and Learning Systems*, 7 January 2019, 1-13.

[7] Wu, L., Sun, P.J., Fu, Y.J., Hong, R.C., Wang, X.T. and Wang, M. (2019) A Neural Influence Diffusion Model for Social Recommendation. In: *Proceedings of the* 42*nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, 235-244. https://doi.org/10.1145/3331184.3331214

[8] 付裕. 基于文本的卷积网络在电影推荐系统中的应用[J]. 电脑知识与技术, 2021, 17(32): 113-116. https://doi.org/10.14004/j.cnki.ckt.2021.3235

[9] 王成龙. 可解释性电影推荐系统的研究与设计[D]: [硕士学位论文]. 北京: 北方工业大学, 2021. https://doi.org/10.26926/d.cnki.gbfgu.2021.000170