计算机科学
COMPUTER SCIENCE

**基于注意力机制和门控网络相结合的混合推荐系统**

郭亮，杨兴耀，于炯，韩晨，黄仲浩

引用本文

郭亮，杨兴耀，于炯，韩晨，黄仲浩. 基于注意力机制和门控网络相结合的混合推荐系统[J]. 计算机科学, 2022, 49(6): 158-164.

GUO Liang, YANG Xing-yao, YU Jiong, HAN Chen, HUANG Zhong-hao. Hybrid Recommender System Based on Attention Mechanisms and Gating Network[J]. Computer Science, 2022, 49(6): 158-164.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500013

# 基于注意力机制和门控网络相结合的混合推荐系统

郭 亮 杨兴耀 于 炯 韩 晨 黄仲浩
新疆大学软件学院 乌鲁木齐 830008
(gliang@stu.xju.edu.cn)

**摘 要** 将用户评论和用户评分相结合来提升推荐系统的性能是推荐系统当前主流的研究方向,但是当用户评论数据稀疏时,现有的大多数推荐系统的性能会出现一定幅度的下降。针对这一问题,文中提出了一种结合注意力机制和门控网络形成的混合推荐系统(Attention Mechanism and Gating Network-based Recommendation System,AMGNRS)。该模型利用志趣相投的用户所产生的辅助评论来缓解用户评论的稀疏性问题,首先将多种混合注意力机制相结合来提高提取用户评论及评分的特征的效率,然后通过门控网络自适应地融合提取的特征并选出与用户偏好最相关的特征,最后利用神经因子分解机的高阶线性相互作用来推导评分预测。将所提模型与当前表现优异的模型在3个真实数据集上进行了对比实验,结果表明,所提模型显著地缓解了数据的稀疏性问题,验证了其有效性。

**关键词**:推荐系统;注意力机制;门控网络;语义信息;协同过滤
**中图法分类号** TP391

## Hybrid Recommender System Based on Attention Mechanisms and Gating Network

GUO Liang,YANG Xing-yao,YU Jiong,HAN Chen and HUANG Zhong-hao
School of Software,Xinjiang University,Urumqi 830008,China

**Abstract** Combining user reviews with user ratings to improve the performance of recommender system is the current mainstream research direction of recommender system. However,when user review data is sparse,the performance of most existing recommender systems will degrade to a certain extent. To solve this problem,this paper proposes a hybrid recommendation system (AMGNRS),which combines attention mechanism and gating networking based recommendation system. It use auxiliary comments generated by like-minded users to alleviate the sparsity of user comments. Firstly,a variety of mixed attention mechanism are combined to impove the feature extracting efficiency of user comments and grading. Then features are extracted by adaptive fusion of gated network,and features most relevant to user preference are selected. Finally,the higher order linear interaction of the neural factorization machine is used to derive the score prediction. By comparing the model with the current model with excellent performance on three real data sets,the results show that the problem of data sparsity is significantly alleviated and the effectiveness of the model is verified.

**Keywords** Recommender system,Attention mechanism,Gated network,Semantic information,Collaborative filtering

## 1 引言

近年来,随着互联网特别是移动互联网的快速发展,互联网信息也呈现出爆炸式增长态势。面对海量的信息,用户从中获取自己需要的内容所耗费的时间成本大幅增加,而推荐系统作为解决信息过载问题的有效手段,引起了学术界和工业界的广泛关注[1]。

推荐系统可以在海量的数据中通过筛选、过滤,为用户提供当前最适合的信息,以缓解信息过载问题[2]。传统的推荐系统主要分为3种[3]:基于内容的推荐、协同过滤推荐和混合推荐。基于内容的推荐主要根据用户做出的选择或者已经评分过的项目,通过已有的数据挖掘与其他项目类似的项目,将其作为用户推荐。协同过滤推荐[4]是当前推荐系统中应用最广的算法,只需要使用用户过往的评分记录就可进行有效的推荐。首先通过用户过往的项目评分差异来计算用户之间的兴趣相似度,然后根据用户之间的历史评分和相似度来

到稿日期:2021-05-02 返修日期:2021-09-08
基金项目:国家自然科学基金(61862060,61966035,61562086);新疆维吾尔自治区教育厅项目(XJEDU2016S035);新疆大学博士科研启动基金项目(BS150257)
This work was supported by the National Natural Science Foundation of China(61862060,61966035,61562086),Education Department Project of Xinjiang Uygur Autonomous Region(XJEDU2016S035) and Doctoral Research Start-up Foundation of Xinjiang University(BS150257).
通信作者:杨兴耀(yangxy@xju.edu.cn)

计算效用值，最后得出用户的潜在偏好。由于单一的推荐算法都存在着各自的不足，混合推荐将不同的推荐算法进行组合，往往能够产生比单种推荐算法更好的推荐效果。

深度学习理论的快速发展，使得深度学习在许多领域（如计算机视觉和语音识别等）取得了巨大的成功，学术界以及工业界正逐渐将深度学习扩展到更多的领域。例如用户在Netflix上观看的电影有80%来自推荐系统[5]；一种新的基于深度神经网络的推荐算法也早已被用于YouTube上的视频推荐[6]；Yahoo几年前就已经提出了基于RNN的新闻推荐系统[7]。所有的这些模型早已在工业界得到广泛应用，并显示出了比传统模型更为精确的结果。

尽管通过协同过滤以及与深度学习等新兴的方法相结合提高了推荐系统的效率，但数据的稀疏性导致当前的推荐算法难以为记录很少的用户提供高质量的服务，严重阻碍了推荐算法效率的进一步提高。用户评论直接描述了用户给出当前评分的直接原因和对该项目的使用体验，直接体现了用户偏好以及项目特征，很好地缓解了数据稀疏性问题。当前的一些研究表明，考虑用户评论的推荐算法的表现通常比仅考虑用户与项目交互记录的协同过滤方法更好[8-10]。尽管基于用户评论的推荐方法已经被证明可以有效缓解数据的稀疏性问题，但是仍然存在一些固有的局限性，例如，其很难进一步化解基于用户评论数据的稀疏性，目前存在一种趋势是，大多数用户对所参与的项目不进行任何评论或者撰写的评论过于简短，无法完全反映用户的兴趣。

本文提出的结合注意力机制和门控网络形成的混合推荐系统，通过建立辅助评论来缓解数据稀疏性，使用多种注意力机制和门控网络所形成的神经网络模型更有效地提取了用户评论文本中的特征。本文的主要贡献如下：

(1)提取与用户兴趣相同的其他用户评论，建立用户的辅助评论文档，弥补了用户评论简短和数据稀疏性问题。

(2)设计具有层次结构的混合注意力机制，通过3种注意力机制彼此之间的渐近关系，使模型更有效地提取评论的特征。结合门控网络，自适应地将评论特征和用户交互特征融合在一起并动态调整特征的重要性。

(3)在真实数据集上进行实验，证明了AMGNRS模型的有效性，实验结果表明AMGNRS优于当前流行的推荐算法。

## 2　相关工作

### 2.1　基于协同过滤的特征交互

协同过滤主要通过对用户过往的行为数据的深度挖掘来发现该用户的偏好，基于不同的偏好将用户分组，将品味相似的项目推荐给组内的其他用户。特征交互是预测用户偏好的重要组成部分。Bao等[8]将用户评论的潜在主体信息与交互信息连接起来以进行评分预测。Wu等[11]引入了FM来实现特征的高级交互。Cheng等[12]利用多层感知器（Multilayer Perceptron，MLP）代替传统的MF来实现用户-项目对的多层非线性相互作用。另一方面，近期图神经网络在提取基于图的数据关系中显示出了巨大的潜力[13]。He等[14]基于图神经网络提出了LightGCn模型，该模型利用用户项目图来学习用户和项目之间的潜在特征，从而对用户进行项目推荐。

### 2.2　基于注意力机制的推荐

近年来，注意力机制在自然语言处理、图像识别等领域得到广泛应用并取得了令人瞩目的成就。注意力机制基于以下特征：人们观察事物时，通常会专注于一个区域而忽略周围其他的区域[15]，因此注意力机制被用于从评论中识别重要的单词，以提高推荐的准确性。Gong等[16]利用基于CNN的注意力机制来提取微博文本中最重要的单词，以进行标签推荐。Liu等[17]使用带有CNN的双重注意力机制来突出显示用户的喜好和商品的特征，并将这些单词可视化以解释推荐的结果。NARRE[18]采用注意力机制来调整并行神经网络模型中评论的权重。Chen等[19]在不同的特征交互阶段利用注意力机制来实现特征的动态交互。

### 2.3　基于用户反馈评论的推荐

近年来，通过提取用户评论的特征来降低数据的稀疏性得到了广泛关注。一些研究提出采用主题建模的技术从文本评论中提取更有价值的特征，例如Tan等[20]基于潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）从评级提升评论文本中提取主题特征作为潜在特征。如果用户对项目的评分为$r$，则只需要在对应的评论中充分提取特征$r$次，就可以形成评级提升评论文档，以便LDA可以轻松地提取评分较高的评论中表达的主题特征，这些提取的主题特征随后被集成到MF框架中已得到项目的潜在特征中。Liu等[21]提出了一个统一的框架NRCA，该框架同时对用户和项目的评论进行文档级和评论级的特征提取，并通过用户表示学习的交叉注意力模型为不同的目标项目选择不同的信息词和评论，以提高推荐的效率。Huang等[22]提出了一个基于情感相似度的个性化评论推荐模型A2SPR，该模型从评论中分析用户的偏好，利用用户的细粒度情感和项目相关性来提高用户的相似度，降低推荐的误差。Zheng等[23]提出了新的并行CNN模型，它们共同从用户评论中学习用户和项目的潜在特征，然后完成用户偏好的预测。以上方法比仅基于用户项目交互的传统模型具有更好的性能，并被证明可以在一定程度上缓解数据的稀疏性问题。

### 2.4　基于门控网络的推荐

门控网络首先在自然语言处理中得到了广泛应用。Dauphin等将门限机制与CNN相结合，提出新的线性门控单元，其可以用来标识长文本，不仅降低了模型的梯度弥散，还能更好地处理非线性的问题[24]。后来门控网络拓展到其他领域，Chen等提出了由特性门控模块、实例门控模块以及项目产品模块组成的门控网络，该模型不仅可以对长期用户兴趣进行建模，也可以更好地对短期用户兴趣进行建模[25]。

本文提出的AMGNRS模型通过辅助评论来进一步缓解用户评论的数据稀疏性问题，使用不同类型的注意力机制来更有效地获取评论的特征。自我注意力（Self-attention）忽略了两个单词之间的距离，直接建立它们之间的相互作用，并且为后面的高阶注意力（Higher-order Attention）提取更重要的

语义之间的信息提供了基础；共同注意力(Co-attention)利用高阶注意力提取出来的特征学习用户与项目特征对的动态交互。门控网络通过调整不同特征的权重来动态地融合评论和交互数据，以提高推荐效果的准确性。

## 3 AMGNRS 算法

AMGNRS 模型包含两个组件。1)特征提取组件。特征提取组件负责提取评论以及评分中的特征。传统模型提取的特征可能包含影响用户偏好的无关信息，本文提出的多类型注意力机制相结合的方式，通过学习评论中相关语义信息的动态交互，可以使模型专注于相关信息，尽可能避免无关信息的影响。2)动态交互组件。动态交互组件由门控网络和神经因子分解模型(Neural Factorization Machines，NFM)组成，其中门控网络负责将评论与评分特征动态地融合，并且调整相关影响因素的权重；NFM 通过特征的高阶非线性作用得出最终的预测分数。AMGNRS 的模型框架如图 1 所示。

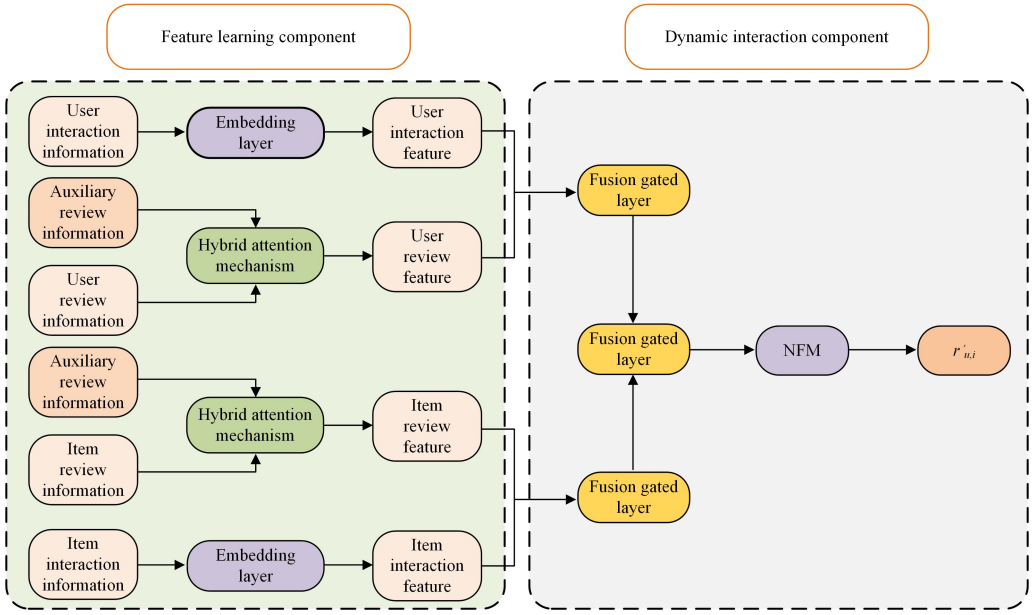



图 1 AMGNRS 模型的框架图

Fig. 1 Framework of AMGNRS model

### 3.1 辅助评论

在常见的基于用户评论的推荐系统中，当用户评论的文本数据稀疏时，推荐系统无法提取出有用的特征；在评论文本简短而又不完整时，推荐模型会遭受性能损失。本文旨在从辅助评论文本中提取依赖于用户项目对的特征，以缓解评论文本的数据稀疏性。

辅助评论是由其他用户撰写的与目标用户在同一项目中具有相同的评分的评论。对于用户评论的每个项目，随机选取该项目中一个志趣相投的用户的评论，将其添加到用户的辅助评论中。因此，辅助评论不会受到热门评论数量的影响。考虑到大多数用户倾向于给予较高的评分，本文将给出相似评分的用户视为志趣相投的用户，并且偏好较高的评分，即如果没有相同的评分，本文将考虑给出相似评分(优先+1，否则-1)的用户作为其志趣相投的用户，将其对应的评论作为辅助评论。

### 3.2 评论文本特征提取

本文基于评论的特征提取使用了 3 种不同的注意力机制。首先，带有位置的自我注意力机制通过建立评论单词之间的长期依赖关系，为后续提取文本特征提供了基础。然后，高阶注意力层通过特征之间的多重交互捕获重要的语义信息。在此基础上，共同注意力层对用户特征和项目特征之间的动态交互进行建模，以捕获用户项目对的相关性。本文针对用户评论、项目评论以及辅助评论的特征提取操作是相同的，因此下文以用户评论为例进行特征提取。

#### 3.2.1 基于位置的 Self-attention 层

在用户评论中，单词与单词之间通常存在着复杂的依赖性，因此，本文使用基于位置的自我注意力机制来提取评论单词之间的长期依赖关系。

自我注意力机制的计算式如下：

$$Attention(\boldsymbol{F}^{Q},\boldsymbol{F}^{K},\boldsymbol{F}^{V})=\mathrm{softmax}\left(\frac{\boldsymbol{F}^{Q}(\boldsymbol{F}^{K})^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{F}^{V} \tag{1}$$

其中，矩阵 $\boldsymbol{F}$ 是使用词向量模型对用户评论进行映射而得到的词向量矩阵，$\boldsymbol{F}\in\boldsymbol{R}^{d\times l}$；$d$ 为文本嵌入维数；$l$ 为单一用户评论所包含的单词个数；$\boldsymbol{F}^{Q}$，$\boldsymbol{F}^{K}$，$\boldsymbol{F}^{V}$ 分别表示注意力机制中的查询、键以及值；$d$ 表示矩阵 $\boldsymbol{F}^{K}$ 的维数，用于防止内积取最大值。

为了保持输入顺序的位置信息，本文采用正弦-余弦位置编码将序列的时间信息添加到自我注意力机制中。通过定义位置矩阵 $\boldsymbol{P}\in\boldsymbol{R}^{d\times l}$ 来编码序列位置信息，并通过在不同的位置使用正弦和余弦函数来计算 $\boldsymbol{P}$。

$$\begin{aligned}\boldsymbol{P}_{pos,2j}&=\sin(pos/10\,000^{2j/d})\\ \boldsymbol{P}_{pos,2j+1}&=\cos(pos/10\,000^{2j/d})\end{aligned} \tag{2}$$

其中，$pos$ 是位置，$j$ 是词向量每一维的大小。本文将 $\boldsymbol{P}$ 添加到所提取的特征后，然后添加一个完全连接的前馈网络，前馈

网络由具有两个线性变换的 ReLU 激活函数组成：

$$O=\max(0,(F+P)\boldsymbol{W}_o^1+b_1)\boldsymbol{W}_o^2+b_2 \tag{3}$$

其中，$O$ 表示带有位置注意力的特征，$\boldsymbol{W}_o^1$ 和 $\boldsymbol{W}_o^2$ 表示权重矩阵，$b_1$ 和 $b_2$ 表示偏差。

3.2.2　Higher-order attention 层

为了更好地区分评论文本中不同单词的重要性，并尽量减少无关词语的干扰，本文利用内核大小为 1 的二阶卷积运算来学习具有长期依赖性的局部注意力表示。

$$Z_i^{l\text{-att}}=\delta(w_i^*\ W_{(:,k)}^{l\text{-att}}+b_k^{l\text{-att}}),k\in[1,f_{l\text{-att}}] \tag{4}$$

其中，$\boldsymbol{W}_{(:,k)}^{l\text{-att}}\in R^{d\times l}$ 为卷积权重矩阵，$f_k^{l\text{-att}}$ 为偏差，$\delta$ 表示 ReLU 激活函数，$[1,f_{l\text{-att}}]$ 表示卷积核的数量。为了更好地实现重要特征的过滤，需要执行第二次卷积。经过第二次卷积可以得到第二个注意力权重 $\boldsymbol{Z}_i^{l\text{-att-2}}$，$\boldsymbol{Z}_i^{l\text{-att-2}}$ 越大，代表这个单词的重要性越高，越小则代表该单词的重要性越低。位置 $i$ 的单词可表示为：

$$\boldsymbol{c}_i=w_i\Theta\boldsymbol{Z}_i^{l\text{-att-2}} \tag{5}$$

其中，$w_i$ 表示在位置 $i$ 的单词，$\boldsymbol{Z}_i^{l\text{-att-2}}$ 表示第二次卷积后的注意力权重，$\boldsymbol{c}_i$ 为当前位置单词的词向量表示，$\Theta$ 表示逐元素相乘。评论文本单词的长期依赖性提取可以表示为：

$$\begin{aligned}\boldsymbol{U}&=[\boldsymbol{c}_1^u,\boldsymbol{c}_2^u,\boldsymbol{c}_3^u,\cdots,\boldsymbol{c}_{m_u}^u]\\ \boldsymbol{V}&=[\boldsymbol{c}_1^i,\boldsymbol{c}_2^i,\boldsymbol{c}_3^i,\cdots,\boldsymbol{c}_{n_i}^i]\end{aligned} \tag{6}$$

其中，$\boldsymbol{c}_k^u$ 和 $\boldsymbol{c}_j^i$ 分别表示用户和项目评论文本中位置 $k$ 和 $j$ 处的单词的上下文特征向量，$m_u$ 和 $n_i$ 分别为用户特征和项目特征的长度。

3.2.3　Co-attention 层

本文通过共同注意力层实现用户与项目的动态交互，以及捕获用户偏好特征和项目特征之间的相关性。受 Li 等[26]的启发，本文使用三线函数来构造用户-项目的相关性矩阵 $\boldsymbol{A}\in R^{m\times n}$。

$$\boldsymbol{A}=W_s[\boldsymbol{U}\boldsymbol{W}_u+\boldsymbol{V}\boldsymbol{W}_v+(\boldsymbol{U}\boldsymbol{W}_{u,v})\odot\boldsymbol{V}] \tag{7}$$

其中，$\boldsymbol{W}_u$，$\boldsymbol{W}_v$ 和 $\boldsymbol{W}_{u,v}$ 是实现矩阵变换并得出最终用户和项目潜在特征向量的注意力矩阵，$W_s$ 代表要训练的参数。$\boldsymbol{A}$ 中的每个元素表示用户-项目特征对的相关性，而每行 $\boldsymbol{A}$ 中的值表示 $\boldsymbol{U}$ 中的每个特征 $\boldsymbol{c}_k^u$ 与 $\boldsymbol{V}$ 中特征的相关性。类似地，每列 $\boldsymbol{A}$ 中的值表示 $\boldsymbol{V}$ 中每个特征 $\boldsymbol{c}_j^i$ 与 $\boldsymbol{U}$ 中特征的相关性。本文将 $\boldsymbol{A}$ 的行和列进行平均池化操作，生成用户和项目向量，表达式为：

$$\begin{aligned}g_k^u&=meanpooling(\boldsymbol{A}[k,:])\\ \underline{g_j^i}&=meanpooling(\boldsymbol{A}[:,j])\end{aligned} \tag{8}$$

然后将获得的向量进行归一化处理。

$$\begin{aligned}t_k^u&=g_k^u\cdot c_k^u\\ t_j^i&=g_j^i\cdot c_j^i\end{aligned} \tag{9}$$

其中，$t_k^u$ 表示具有用户项目相关性的用户评论文档第 $k$ 个位置的文本特征，$t_j^i$ 表示具有用户项目相关性的项目评论文档第 $j$ 个位置的文本特征。

3.2.4　卷积池化层

为了从评论文本中提取更多的抽象特征并减小不相关的噪声影响，本文在共同注意力层后面添加了一个卷积层和平均池化层，表达式如下：

$$\begin{aligned}h_h^j&=\delta(\boldsymbol{W}_a^j U_{h:h+\omega-1}^u+b_a^j)\\ h_h&=\text{mean}(h_1^j,\cdots,h_{l_u^j-\omega+1}^j)\\ \boldsymbol{h}^u&=[h_h,\boldsymbol{h}_{\text{Aux}}]\end{aligned} \tag{10}$$

其中，$\boldsymbol{W}_a^j$ 表示第 $j$ 个卷积滤波器的卷积权重，$\omega$ 为滑动窗口的大小，$b_a^j$ 为偏差，$h_h^j$ 表示位置为 $h$ 的第 $j$ 个卷积滤波器的特征。同理，$\boldsymbol{h}_{\text{Aux}}$ 表示从用户辅助评论文本中提取的特征，$\delta$ 表示 tanh 激活函数。由此可以获得最终的具有用户评论的文本特征 $\boldsymbol{h}^u$。同样地，可以获得项目评论的文本特征。最后，使用非线性函数将文本特征映射到一维空间以完成推荐任务。

$$\begin{aligned}\boldsymbol{h}^u&=\delta(\boldsymbol{W}^u\boldsymbol{h}^u+b^u)\\ \boldsymbol{h}^i&=\delta(\boldsymbol{W}^i\boldsymbol{h}^i+b^i)\end{aligned} \tag{11}$$

其中，$\boldsymbol{W}^u$ 和 $\boldsymbol{W}^i$ 分别表示权重矩阵，$b^u$ 和 $b^i$ 表示偏差，$\delta$ 表示 tanh 激活函数。

### 3.3　从项目评分中提取特征

通过 one-hot 将用户和项目的 id 进行编码，形成项目特征向量 $\boldsymbol{v}_i^I$ 和用户特征向量 $\boldsymbol{v}_u^U$，然后通过潜在因子矩阵 $\boldsymbol{P}_u\in R^{M\times K}$ 和 $\boldsymbol{Q}_i\in R^{N\times K}$ 将特征向量映射成低维向量。

$$\begin{aligned}p_u&=\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{v}_u^U\\ q_i&=\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{v}_i^I\end{aligned} \tag{12}$$

其中，$p_u$ 和 $q_i$ 分别表示用户 $u$ 和 $i$ 项目的交互特征。

### 3.4　动态交互

本文中的动态交互指将从评论文本和用户评分中提取的特征进行动态特征融合和评分预测。动态交互包括融合门控层(Fusion Gated Layer)和过滤门控层(Filter Gated Layer)。

3.4.1　融合门控层

本文提出的融合门控层动态地集成了评分特征和评论特征。用户特征和用户评分特征融合公式如下：

$$\begin{aligned}s_u'&=\delta(w_u^f(h^u+p_u)+b_u)\\ s_u^f&=s_u'h^u+(1-s_u')p_u\end{aligned} \tag{13}$$

其中，$s_u'$ 是门控网络的更新门，$s_u^f$ 负责将用户评论特征和用户评分特征进行融合。推荐算法使用评论信息时，通常会受到一些限制。1)评论信息的质量参差不齐。评论通常包含一些无用的评论，这些评论可能会影响到最后的推荐效果。2)很容易受到所提取的评论特征的影响。由于融合后的特征包含评论特征，并且可能会引入噪声，因此本文使用滤波器将融合后的特征和原始交互特征相结合。滤波函数的范围是 0～1，如果融合后的特征有利于提高模型性能，则滤波函数的值越大，否则滤波函数的值越小，噪声干扰也越小。滤波器 $s_u''$ 与用户交互特征 $p_u$ 的定义如下：

$$\begin{aligned}s_u''&=\delta(w_u^p p_u+w_u^f s_u^f)\\ s_u^{ff}&=s_u''(\tanh(w_u^{ff}s_u^f+b_u^{ff}))\\ \boldsymbol{u}_u&=p_u+s_u^{ff}\end{aligned} \tag{14}$$

其中，$w_u^p$，$w_u^f$，$w_u^{ff}$ 和 $b_u^{ff}$ 为参数，$s_u^{ff}$ 是经过噪声过滤而被保留下的特征，$\boldsymbol{u}_u$ 是增强后的用户特征，本文用同样的方法增强项目特征 $v_i$。

3.4.2　过滤门控层

由于用户通常会关注项目的特定部分，因此本文采用了

基于用户的过滤门控层来控制传播到用户偏好预测任务的特征。

$$v_i^F = v_i \odot \delta(u_u w_u + v_i w_i + b_F) \tag{15}$$

其中，$v_i^F$ 表示经过过滤后的特征，$\delta$ 表示激活函数，$\boldsymbol{w}_u$ 和 $\boldsymbol{w}_i$ 为权重矩阵，$b_F$ 为偏差。受 He 等[27]的启发，本文使用直观的连接操作来集成用户和项目特征。

$$\boldsymbol{z} = [u_u, v_i^F] \tag{16}$$

其中，$\boldsymbol{z}$ 表示用户-项目特征，[,]表示通过隐藏层组合这两个特征。本文使用 NFM 来获取特征之间的高阶非线性相互作用，目标函数表示为：

$$r'_{u,i}(z) = m_0 + \sum_{j=1}^{|z|} m z_j + f(z) \tag{17}$$

其中，$z_j \in z$ 是特征 $j$ 的值，$m_0$ 表示全局偏差，$m_j$ 表示潜在特征向量的系数，$f(z)$ 表示对特征的高阶交互作用进行建模，即：

$$f(z) = \frac{1}{2}[(\sum_{j=1}^{|z|} z_j v_j)^2 - (\sum_{k}^{|z|} z_k v_k)^2] \tag{18}$$

其中，$z_j, z_k \in z$ 表示第 $j$ 个和第 $k$ 个用户-项目的特征值，$\boldsymbol{v}_j$，$\boldsymbol{v}_k \in \boldsymbol{R}_s$ 表示特征 $j$ 和 $k$ 的嵌入向量，$s$ 为嵌入维数。本文通过具有多个隐藏层的 MLP 来获取特征之间的高阶相互作用，最终的预测目标函数为：

$$\begin{aligned} r'_{u,i}(z) &= m_0 + \sum_{j=1}^{|z|} m_j z_j + \Gamma(f(z)) \\ \Gamma(f(z)) &= h^{\mathrm{T}} \delta_L(W_L(\cdots \delta_1(\sim W_1 f(z) + b_1)\cdots) + b_L) \end{aligned} \tag{19}$$

其中，$r'_{u,i}$ 表示预测得分，模型参数 $\theta = \{m_0, \{m_j, v_j\}, h, \{W_L, b_L\}\}$，而 $\delta_L$ 表示 ReLU 激活函数，与 FM 相比，添加的参数 $\{W_L, b_L\}$ 主要用于学习特征的高阶交互作用，$L$ 表示 NFM 的隐藏层数量。

### 3.5 模型训练

在训练 AMGNRS 参数时，本文将平方损失作为模型的损失函数。

$$J = \sum_{(u,i)\in \boldsymbol{R}} (r'_{u,i} - r_{u,i})^2 + \lambda_\Theta \| \Theta \|^2 \tag{20}$$

其中，$\boldsymbol{R}$ 为用户-项目评分矩阵，$r_{u,i}$ 是用户 $u$ 对项目 $i$ 的真实评分，$r'_{u,i}$ 为预测评分，$\Theta$ 表示所有的参数，$\lambda_\Theta \| \Theta \|^2$ 用作正则化，以防止模型过度拟合。

## 4 实验

### 4.1 数据集

为了测试模型的有效性，本文利用 Amazon 的真实数据集进行实验。为了减少用户评论的长尾理论，按照 Chen 等[18]采用的数据预处理步骤来调整评论的长度。数据集信息如表 1 所列。

表 1 数据集

Table 1 Dataset

| Dataset | # User | # Item | # Ratings | Word per user | Word per item |
|---|---|---|---|---|---|
| Food | 14681 | 8713 | 151254 | 176 | 168 |
| Video | 24303 | 10672 | 234577 | 147 | 131 |
| Phones | 27879 | 10429 | 194493 | 162 | 169 |

### 4.2 评价指标

本文使用均方误差(MSE)和平均绝对误差(MAE)作为 AMGNRS 的评价指标，其数值越小，表示模型的准确率越高。

$$MSE = \frac{1}{N} \sum_{(u,i)\in \boldsymbol{R}} (r_{u,i} - r'_{u,i})^2 \tag{21}$$

$$MAE = \frac{1}{N} \sum_{(u,i)\in \boldsymbol{R}} | r_{u,i} - r'_{u,i} | \tag{22}$$

### 4.3 实验对比

为了验证模型的有效性，将 AMGNRS 模型与 NARRE，DeepCoNN，D-Attn 在 3 种不同的数据集上进行对比，3 种算法介绍如下。

(1)NARRE[18]：将 CNN 和注意力机制相结合，学习用户评论的文本特征，并将评论和评分通过共享网络集合为一个统一的模型，从而产生更好的推荐效果。

(2)DeepCoNN[23]：通过两个并行的 CNN 网络分别提取用户兴趣特征和项目特征，最后使用 FM 模型来完成商品推荐的高级交互。

(3)D-Attn[28]：首先将模型局部和全局的注意力机制与并行 CNN 神经网络相结合，然后使用结合的网络模型获得用户评论的文本特征，最后使用点积完成评分预测。

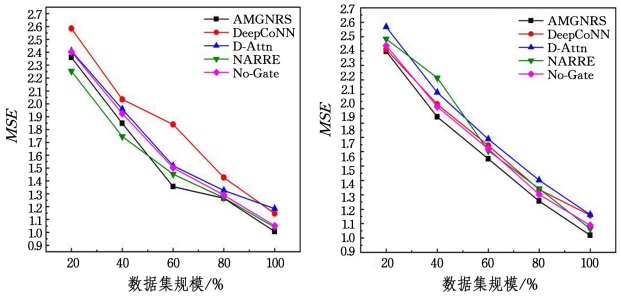
同时，为了更好地从自身来验证模型的有效性以及更好地凸显模型的推荐效果，将未采用门控网络的 No-Gate 作为 AMGNRS 的变体进行实验，并参与最后的结果对比。

### 4.4 实验结果与分析

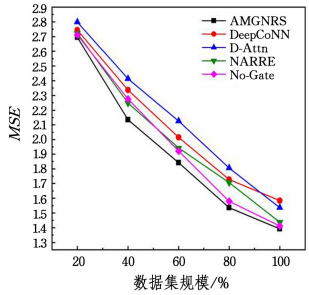
在实验中，将每个数据集随机分为 3 个部分，其中，训练集占 80%，测试集占 10%，验证集占 10%。实验将数据集规模分为 20%，40%，60%，80%和 100%来分别进行对比，数据集规模越小，用户评论数量就越少。

图 2 和图 3 给出了所提 AMGNRS 模型和对比模型在 MSE 和 MAE 上的对比结果。实验表明，AMGNRS 整体上优于对比模型。随着数据集规模的增加，AMGNRS 的优势逐渐减小，这是因为随着数据集的增加，数据的稀疏性问题得到了进一步的缓解，模型可以从用户评论文本中提取到更多的潜在特征，使得模型可以得到更为精确的结果。当数据集规模达到最大时，与对比模型 DeepCoNN 和 D-Attn 相比，AMGNRS 在 MAE 和 MSE 的评价标准下也有最少 3%的提升，这是因为 DeepCoNN 和 D-Attn 仅使用了用户评论这一单一的交互数据，而用户的评论文本信息仅包含用户的特定偏好和项目的相关属性，因此无法完全捕获用户的偏好，从而导致推荐有失偏颇。而且相对于 D-Attn 模型使用基于局部和全局的注意力机制来提取用户评论特征，AMGNRS 采用了多种混合注意力机制，可以捕获更为全面的用户和项目特征。NARRE 模型使用注意力机制来提取评论的有用性并结合用户评分来预测用户偏好，但是 NARRE 仅在评论级别提取特征，忽略了对评论进行进一步的细化，而本文不仅通过 AMGNRS 模型对评论级别进行了进一步的细化，从更多的词语中提取用户项目特征，还通过辅助评论强化了用户和项目的特征。No-Gate 作为 AMGNRS 的变体，缺少门控网络对特征的动态交互和特征选择，忽略了不同特征之间的重要性

以及在特征学习中产生的噪声，使得推荐的效果差于AMGNRS。



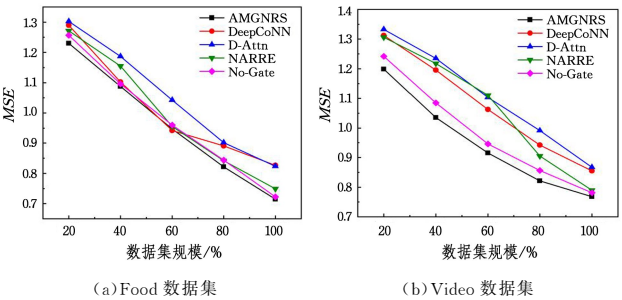

(a)Food 数据集　(b)Video 数据集



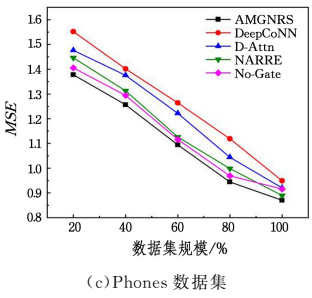

(c)Phones 数据集

图 2　MSE 评测结果

Fig. 2　Evaluation results of MSE




(a)Food 数据集　(b)Video 数据集




(c)Phones 数据集

图 3　MAE 评测结果

Fig. 3　Evaluation results of MAE

**结束语**　本文提出了一种结合注意力机制和门控网络的混合推荐系统 AMGNRS。AMGNRS 利用其他用户撰写的评论作为辅助评论文本来缓解推荐系统的稀疏性问题，通过结合多种注意力机制和门控网络，自适应地整合评论和评分信息来捕获用户的偏好，最终实现提升推荐性能的目的。在3个公开的数据集上与多种主流算法进行对比，结果表明AMGNRS模型优于现有的主流推荐模型。但本文仅在 Amazon 的相关购物数据集下对 AMGNRS 模型进行测试，并未涉及当下流行的视频推荐、文本推荐等其他推荐，因此该模型在推荐适用性上可能存在局限性。在未来的工作中，将进行进一步的研究，以更有效地缓解数据稀疏性问题，并在当前的AMGNRS模型的基础上进行改进，考虑用户偏好随时间的变化而发生的变化，通过合并时间序列来更好地预测用户未来的喜好。

## 参 考 文 献

[1] KIM D, PARK C, OH J, et al. Convolutional Matrix Factorization for Document Context-Aware Recommendation[C]// ACM Conference. ACM, 2016: 233-240.

[2] HU J H, LI P. Collaborative Topic Regression Model Integrating Users' Social Relations[J]. Computer Engineering and Application, 2018, 54(19): 151-157, 171.

[3] HUANG L W, JIANG B T, LU S Y, et al. Review of recommender systems based on deep learning [J]. Chinese Journal of Computers, 2018, 41(7): 1619-1647.

[4] HUANG J T, CHEN J B, CHEN P H. Recommendation algorithm combining preference degree and network structure[J]. Computer Engineering and Applications, 2019, 55(10): 9-15.

[5] GOMEZ-URIBE C A, HUNT N. The Netflix Recommender System: Algorithms, Business Value, and Innovation[J]. ACM Transactions on Management Information Systems, 2015, 6(4): 1-19.

[6] COVINGTON P, ADAMS J, SARGIN E. Deep Neural Networks for YouTube Recommrndation[C]// ACM Conference on Recommender Systems. ACM, 2016: 191-198.

[7] OKURA S, TAGAMI Y, ONO S, et al. Embedding-based News Recommendation for Millions of Users [C]// the 23rd ACM SIGKDD International Conference. ACM, 2017: 1933-1942.

[8] BAO Y, FANG H, ZHANG J. TopicMF: simultaneously exploiting ratings and reviews for recommendation[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence. 2014: 2-8.

[9] CHEN C, ZHANG M, LIU Y, et al. Neural Attentional Rating Regression with Review-level Explanions [C] // 2018 World Wide Web Conference. 2018: 1583-1592.

[10] TAN Y, ZHANG M, LIU Y, et al. Rating-boosted latent topics: Understanding users and items with ratings and reviews[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016: 2640-2646.

[11] WU L B, QUAN C, LI C L, et al. A Context Aware User-Item Representation Learning for Item Recommendation[J]. ACM Transactions on Information Systems, 2019, 37(2): 1-29.

[12] CHENG Z, DING Y, HE X, et al. A^3NCF: An Adaptive Aspect Attention Model for Rating Prediction[C]// Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018: 3748-3754.

[13] LIU F, XUE S, WU J, et al. Deep Learning for Community Detection: Progress, Challenges and Opportunities[C]// International Joint Conference on Artificial Intelligence. 2020: 4981-4987.

[14] HE X, DENG K, WANG X, et al. Lightgcn: Simplifying and powering graph convolution network for recommendation[C]// Proceedings of the 43st International ACM SIGIR Conference

on Research and Development in Infrmation Retrieval. ACM, 2020.

[15] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]// International Conference on Machine Learning. 2015: 2048-2057.

[16] GONG Y, ZHANG Q. Hashtag recommendation using attention-based convolutional neural network[C]// Proceedings of the 25th International Joint Conference on Artifificial Intelligence. 2016: 2782-2788.

[17] LIU D, LI J, DU B, et al. DAML: Dual Attention Mutual Learning between Ratings and Reviews for Item Recommendation[C]// Proceedings of the 25th ACM SIGKDD International Conference. ACM, 2019: 344-352.

[18] CHEN C, ZHANG M, LIU Y, et al. Neural Attentional Rating Regression with Review-level Explanations[C]// Proceedings of the 2018 World Wide Web Conference. 2018: 1583-1592.

[19] CHEN J, ZHUANG F, HONG X, et al. Attention-driven Factor Model for Explainable Personalized Recommendation[C]// Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval—SIGIR'18. 2018: 909-912.

[20] TAN Y, ZHANG M, LIU Y, et al. Rating-boosted latent topics: Understanding users and item items with ratings and reviews [C]// Proceedings of the Twenty-Fifth International Joint Conference on Artifificial Intelligence. 2016: 2640-2646.

[21] LIU H T, WANG W. Neural Unified Review Recommendation with Cross Attention[C]// Preceedings of the 43rd International ACM SIGIR Conference on Reaearch and Development in Information Retrieval. 2020: 1789-1792.

[22] HUANG C, JIANG W, WU J, et al. Personalized Review Recommendation based on User'Aspect Setiment[J]. ACM Transaction on Internet Technology, 2020(4): 1-26.

[23] ZHENG L, NOROOZI V, YU P S. Joint Deep Modeling of Users and Items Using Reviews for Recommendation[C]// Proceedings of the 10th ACM International Conference on Web Search and Data Mining. 2017: 425-434.

[24] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[C]// Proceedings of the 34th International Conference on Machine Learning (ICML'17). 2017: 933-941.

[25] CHEN H, LIN Z, DING G, et al. GRN: gated relation network to enhance convolutional neural networkfor named entity recognition[C]// Proceedings of the 33th AAAI Conference on Artificial Intelligence. 2019: 6236-6243.

[26] LI X, SONG J, GAO L, et al. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8658-8665.

[27] HE X, CHUA T S. Neural Factorization Machines for Sparse Predictive Analytics[C]// International ACM SIGIR Conference. ACM, 2017: 355-364.

[28] SEO S, HUANG J, YANG H, et al. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction[C]// Proceedings of the 11th ACM Conference on Recommender Systems. 2017: 297-305.

**GUO Liang**, born in 1997, postgraduate. His main research interests include data mining and recommender system.

**YANG Xing-yao**, born in 1984, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include recommender system and trust computing.

(责任编辑:李亚辉)