CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 43 卷第 2 期 2021 年 2 月
Vol. 43, No. 2, Feb. 2021

文章编号:1007-130X(2021)02-0370-11

基于注意力机制的深度学习推荐研究进展*

陈海涵,吴国栋,李景霞,王静雅,陶　鸿
(安徽农业大学信息与计算机学院,安徽 合肥 230036)

摘　要:近年来,注意力机制 AM 被广泛应用到基于深度学习的自然语言处理任务中,基于注意力机制的深度学习推荐也成为推荐系统研究的一个新方向。探讨了注意力机制的结构和分类标准,从基于注意力机制的 DNN 推荐、CNN 推荐、RNN 推荐、GNN 推荐 4 个方面分析了现有融合注意力机制的深度学习推荐研究的主要进展和不足,阐明了其中的主要难点,最后指出了多特征交互的注意力机制推荐、多模态注意力机制深度学习推荐、融入注意力机制的多种深度神经网络混合推荐和注意力机制的群组推荐等基于注意力机制的深度学习推荐未来的主要研究方向。

关键词:注意力机制;深度学习;推荐系统
中图分类号:TP391.3　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2021.02.023

Research advances on deep learning recommendation based on attention mechanism

CHEN Hai-han, WU Guo-dong, LI Jing-xia, WANG Jing-ya, TAO Hong
(School of Information & Computer, Anhui Agricultural University, Hefei 230036, China)

Abstract: In recent years, Attention Mechanism (AM) has been widely used in natural language processing tasks based on deep learning. Deep learning recommendation based on attention mechanism has become a new direction in the research of recommendation system. This paper discusses the structure and classification standard of attention mechanism, and analyzes the main progress and shortcomings of the existing deep learning recommendation researches based on attention mechanism from four aspects: DNN recommendation, CNN recommendation, RNN recommendation and GNN recommendation. The main difficulties in the research are illustrated. Finally, the paper points out the future direction of deep learning recommendation including multi-feature interaction attention mechanism recommendation, multi-modal attention mechanism recommendation, hybrid recommendation for multiple deep neural networks based on attention mechanism, and group recommendation based on attention mechanism.

Key words: attention mechanism; deep learning; recommendation system

1　引言

当前,深度学习在计算机视觉、自然语言处理和语音识别等领域得到了广泛的应用,许多学者也将其用于推荐系统研究。针对传统协同过滤算法中存在的数据稀疏性和冷启动问题,深度学习具有良好的对数据集本质特征进行学习的能力,一定程度上克服了推荐过程中的数据稀疏问题。但是,深度学习具有黑盒特性,很难对推荐系统的最终决策

* **收稿日期**:2020-04-03;**修回日期**:2020-05-26
基金项目:国家自然科学基金(31671589);安徽省重点研发计划(201904a06020056);智慧农业技术与装备安徽省重点实验室开放基金(APKLSATE2019X003)
通信作者:吴国栋(wugd@ahau.edu.cn)
通信地址:230036 安徽省合肥市安徽农业大学信息与计算机学院
Address: School of Information & Computer, Anhui Agricultural University, Hefei 230036, Anhui, P. R. China

做出解释,而没有解释性的推荐是缺乏说服力的,会对提升用户的信任度带来负面影响。因此,如何在提高推荐性能的前提下,提升深度学习推荐系统的可解释性和透明度受到了工业界与学术界的广泛关注。

注意力机制通过对关注事物的不同部分赋予不同的权重,从而降低其它无关部分的作用。从注意力机制可解释性的角度看,它允许直接检查深度学习体系的内部工作,通过可视化输入与对应输出的注意权重,达到增强深度模型可解释性的效果[1]。在推荐算法中融入注意力机制,对每个潜在因素或特征的重要性进行区分,在提升推荐性能的同时,也提高了推荐系统内部的可解释性。本文主要分析了基于注意力机制的深度神经网络 DNN(Deep Neural Network)、卷积神经网络 CNN(Convolutional Neural Network)、循环神经网络 RNN(Recurrent Neural Network)和图神经网络 GNN(Graph Neural Network)等几种深度学习推荐的研究进展,指出了各自的优点与不足,并指出了相关研究难点与未来主要研究方向。

2 注意力机制及其分类

注意力机制是一种模拟人脑注意力的模型,最初由 Treisman 等人[2]提出,其本质是利用注意力的概率分布,捕捉某个关键输入对输出的影响[3]。以 Bahdanau 等人[4]提出的注意力机制模型为例,求解注意力的计算过程可以抽象为 3 个阶段,如图 1 所示。

图 1 中,注意力机制的 3 个阶段包括:计算打分函数阶段,主要根据解码器(Decoder)端和编码器(Encoder)端隐状态进行相似度计算;计算对齐函数阶段,主要通过归一化处理,将输出的相关性值进行数值转换;计算生成上下文向量函数阶段,主要对输入序列进行加权求和。

按照注意力机制在图 1 中 3 个阶段的不同变换,得到注意力机制的不同类型。根据不同的打分函数,将注意力机制分为加法注意力、乘法注意力、自注意力[5]、多头注意力[6]和分层注意力[7];根据不同的对齐函数,注意力机制可分为全局注意力和局部注意力[8];根据不同的生成上下文向量函数,得到硬注意力与软注意力[9]。

其中,图 1 的核心步骤是注意力分数 $a'_{t,j}$ 的计算,$\boldsymbol{X}_T$ 是输入序列,$\boldsymbol{h}_j$ 是 Encoder 端第 j 个词的隐向量,$\boldsymbol{s}_{t-1}$ 是 Decoder 端在 $t-1$ 时刻的隐状态,$\boldsymbol{y}_{t-1}$ 表示 $t-1$ 时刻的目标词,$\boldsymbol{C}_t$ 表示上下文向量。

3 基于注意力机制的深度学习推荐相关研究

将注意力机制融入深度学习推荐过程中,主要思路是先利用各类深度学习模型学习用户或项目的隐特征,结合注意力机制学习隐特征的权重;其次构建优化函数对参数进行训练,得到用户和项目隐向量;最后利用隐向量信息得到项目排序结果,对用户进行推荐。对于不同的深度学习模型,本文将基于注意力机制的深度学习推荐研究主要分为 4 类,如表 1 所示。

3.1 基于注意力机制的 DNN 推荐方法

DNN 即深度神经网络,由多层感知机 MLP(Multi-Layer Perceptron)发展而来,但 DNN 比 MLP 的激活函数种类更多,层数更深,其网络层数可以达到一百多层乃至更高,一定程度上改善了 MLP 优化函数的梯度消失和局部最优解问题。

针对当前的音乐推荐系统只能从不同歌曲中学习到相同的上下文权重问题,张全贵等人[10]利用注意力机制给每个用户的历史交互歌曲动态分

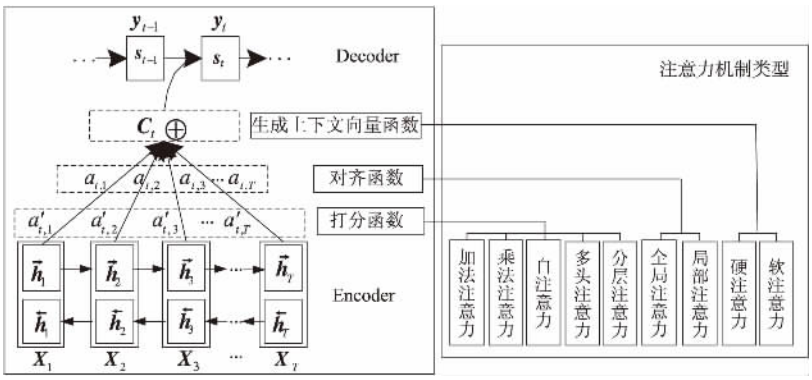

Figure 1 Structure and classification of attention mechanisms

图 1 注意力机制结构及其分类

Table 1 Research on deep learning recommendation based on attention mechanism

表 1 基于注意力机制的深度学习推荐主要研究

应用领域	作用	主要文献	数据集
注意力机制的 DNN 推荐	有重点地提取隐式反馈数据中潜在特征	文献[10-13]	MSD、MovieLens、Pinterest、Alibaba
注意力机制的 CNN 推荐	在卷积层之前或之后融入注意力层,增强重点信息的贡献	文献[15-20]	Twitter、Microblog、Amazon Articles
注意力机制的 RNN 推荐	在 RNN 网络的输出上增加注意力层,抓住文本的序列特征	文献[22-26]	Microblogs、Amazon、Yelp
注意力机制的 GNN 推荐	在图注意网络中自动确定社会效应对用户偏好的影响权重	文献[28-30]	Epinions、WeChat、Douban、Yelp、Delicious

配不同的注意力权重,得到更符合用户偏好的推荐结果,增加了对用户偏好分析的可解释性。沈冬东等人[11]加入平滑系数减轻对长历史活动用户的惩罚,并通过多层感知机参数化注意力函数改进注意力网络,解决了传统 ItemCF(Item Collaboration Filter)算法难以充分挖掘数据间隐含信息的问题。针对传统推荐算法未充分提取用户行为中的隐式反馈特征问题,郭旭等人[12]利用自注意力机制生成用户短期动态项目的向量化表示,提高了推荐质量,但该方法对用户的向量化表示比较粗糙,未考虑融入用户的画像属性。

文献[13]为了解决基于矩阵分解的协同过滤算法不能获取用户历史交互中复杂的非线性特征问题,构建了 DeepCF-A(Deep Collaborative Filtering model based on Attention)模型,提取线性与非线性特征。DeepCF-A 模型如图 2 所示。具体步骤主要有:

(1)线性特征提取。将用广义矩阵分解模型 GMF(Generalized Matrix Factorization)方法提取到的用户隐向量 $\boldsymbol{p}_u^{\mathrm{G}}$ 和项目隐式向量 $\boldsymbol{q}_i^{\mathrm{G}}$ 做元素积操作,得到线性特征 $\boldsymbol{\varphi}_{\mathrm{GMF}}$,如式(1)所示:

$$\boldsymbol{\varphi}_{\mathrm{GMF}} = \boldsymbol{p}_u^{\mathrm{G}} \odot \boldsymbol{q}_i^{\mathrm{G}} \tag{1}$$

(2)非线性特征提取。在 MLP 中融入注意力机制得到用户和项目间历史交互数据的非线性特征 $\boldsymbol{\varphi}_{\mathrm{MLP\text{-}A}}$,如式(2)所示:

$$\boldsymbol{\varphi}_{\mathrm{MLP\text{-}A}} = a_g(\boldsymbol{W}_g^{\mathrm{T}}(a_{g-1}(\cdots a_1(\boldsymbol{W}_1^{\mathrm{T}} f_{\mathrm{a}}\left(\begin{bmatrix}\boldsymbol{p}_u^{\mathrm{A}}\\ \boldsymbol{q}_i^{\mathrm{A}}\end{bmatrix}\right)+\boldsymbol{b}_1)+\cdots))+\boldsymbol{b}_g) \tag{2}$$

其中,$\boldsymbol{p}_u^{\mathrm{A}}$ 和 $\boldsymbol{q}_i^{\mathrm{A}}$ 是用 MLP-A 方法提取到的嵌入层用户和项目隐式特征向量,f_{a} 是注意力机制层的映射函数,a_g 是深度神经网络中第 g 层的激活函数,$\boldsymbol{W}_g$ 和 $\boldsymbol{b}_g$ 是深度神经网络第 g 层的权重和偏置。

(3)注意力机制层。在非线性特征提取部分,将嵌入层的 m 维特征向量 $\boldsymbol{X}_m$ 送入 Softmax 函数,

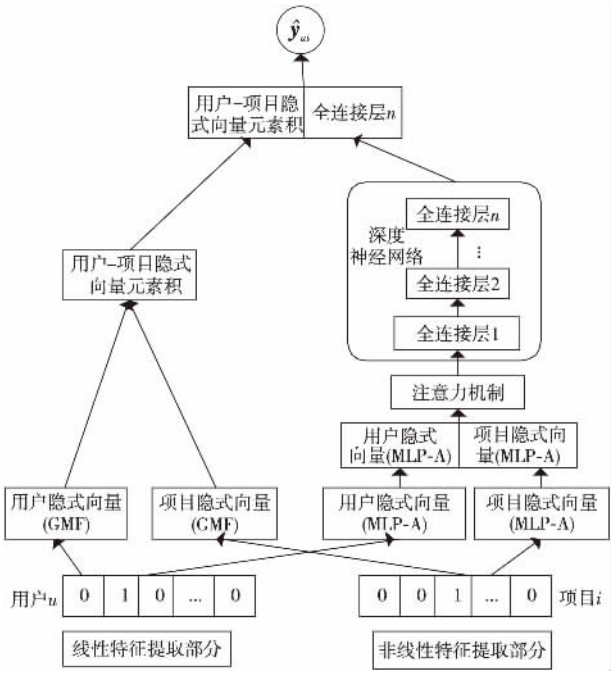

Figure 2 DeepCF-A recommendation model

图 2 DeepCF-A 推荐模型

得到每个维度特征的关注度 $\boldsymbol{A}_m$,如式(3)所示;再将 $\boldsymbol{A}_m$ 与相应维度的特征向量对应相乘,得到更新权重的特征向量 $\boldsymbol{A}_{\mathrm{out}}$,如式(4)所示:

$$\boldsymbol{A}_m = Softmax(\boldsymbol{X}_m) \tag{3}$$

$$\boldsymbol{A}_{\mathrm{out}} = \boldsymbol{A}_m \odot \boldsymbol{X}_m \tag{4}$$

(4)推荐。连接 $\boldsymbol{\varphi}_{\mathrm{GMF}}$ 与 $\boldsymbol{\varphi}_{\mathrm{MLP\text{-}A}}$,送入 Sigmoid 函数,得到用户 u 与项目 i 交互的预测评分 $\hat{\boldsymbol{y}}_{ui}$,如式(5)所示。再将得到的预测评分进行排序,选取排名靠前的项目对用户实施推荐。

$$\hat{\boldsymbol{y}}_{ui} = \sigma\left(\begin{bmatrix}\boldsymbol{\varphi}_{\mathrm{GMF}}\\ \boldsymbol{\varphi}_{\mathrm{MLP\text{-}A}}\end{bmatrix}\right) \tag{5}$$

该模型提升了协同过滤方法处理隐式反馈数据的效果,适用于解决推荐系统中数据量庞大、难以捕捉深层非线性关系的推荐问题。但是,在深度神经网络中,高效地捕捉用户和项目隐向量间的交互信息,是以不断提升网络层数为代价的,深度神经网络层数的增加会导致新的参数数量膨胀问

题[14]。此外,DNN 无法对时间序列上的变化进行建模,不能反映用户兴趣的动态变化,而样本出现的时间顺序对推荐任务又有着非常重要的意义。

3.2 基于注意力机制的 CNN 推荐方法

CNN 即卷积神经网络,具有限制参数个数和挖掘模型局部结构的特点。为了解决 DNN 训练数据时带来的参数数量膨胀问题,有学者将注意力机制和 CNN 结合用在推荐系统研究中。

针对微博的话题标签推荐任务,经常需要大量人工进行分类这一问题,Gong 等人[15]提出了一种基于注意力机制的 CNN 微博标签推荐模型。该模型使用全局和局部注意力 2 个通道,有效提高了推荐性能;但推荐数据仅使用了微博文本标签,未考虑使用图像等其它形式数据提取微博特征。针对这一问题,Zhang 等人[16]加入协同注意力机制对标签与图像、文本中的局部关联性进行建模,相较于仅使用文本信息的模型,推荐效果更好。不足之处是作者仅验证了 1 层和 2 层的协同注意力机制对推荐结果的影响,没有在层数上做更多的尝试。针对在线新闻网站中,平台编辑手动挑选推荐候选文章的耗时问题,Wang 等人[17]构建了一种动态注意力深度模型 DADM(Dynamic Attention Deep Model),DADM 将专业与时间 2 个潜在因素加入注意力机制,自适应地为编辑分配偏好权重,使模型在处理动态数据和编辑行为方面拥有很小的方差。但是,文章中的文字和图像对编辑选择行为的影响应该是不同的,此模型未加以区分。

针对传统推荐算法对评论文本信息提取能力有限的问题,文献[18]提出了一种融合注意力机制对评论文本深度建模的推荐模型 ACoNN(deep Cooperative Neural Networks based on Attention),通过注意力机制设计一层权值更新层对文本矩阵进行重新赋权,再使用一组并行的 CNN,充分挖掘用户和项目的隐含特征。推荐流程如图 3 所示。

ACoNN 推荐模型的主要实现步骤:

(1)输入层:利用词嵌入模型,将用户与项目的评论文本表示成词嵌入矩阵 $\boldsymbol{M}_u$ 和 $\boldsymbol{M}_i$。

(2)注意力机制层:以更新用户评论文本的注意力权重为例,先计算 $\boldsymbol{M}_u$ 中每个词向量 $\boldsymbol{W}_k^u$ 与所有用户评论文本词向量 $\boldsymbol{W}_{1:n}^u$ 之间的相似度系数 sim_k,如式(6)所示:

$$sim_k = \sum_{k=1}^{d} \cos((\boldsymbol{W}_k^u),(\boldsymbol{W}_{1:n}^u)) \tag{6}$$

然后对 sim_k 进行归一化处理,获取用户 u 第 k 个词汇的注意力权值 a_k^u,如式(7)所示:

$$a_k^u = \frac{e^{sim_u}}{\sum_{k=1}^{d} e^{sim_u}} \tag{7}$$

最后对目标用户词向量矩阵进行加权,得到更新权值后的矩阵 $\boldsymbol{S}_u$,如式(8)所示:

$$\boldsymbol{S}_u = \boldsymbol{A}_{(u)} \times \boldsymbol{M}_u \tag{8}$$

其中,d 是用户评论文本中词向量的个数,$\boldsymbol{A}(u)$ 是注意力权值矩阵,且 $A_{(u)} = (a_1^u, a_2^u, \cdots, a_k^u, \cdots, a_d^u)$。

(3)CNN 层:利用 CNN 对词向量矩阵 $\boldsymbol{S}_u$ 进行卷积、池化和全连接操作得到用户向量 $\boldsymbol{output}_u$,同理可得项目向量 $\boldsymbol{output}_i$。

(4)推荐:连接 $\boldsymbol{output}_u$、$\boldsymbol{output}_i$,构建用户-项目特征向量 $\boldsymbol{z}$;向向量 $\boldsymbol{z}$ 加入因子分解机,根据最小化损失函数进行训练,完成参数更新,如式(9)所

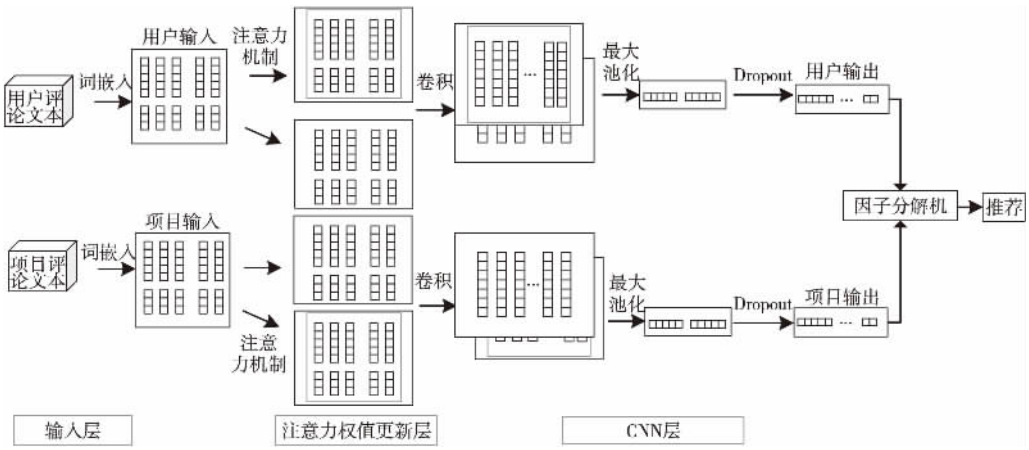

Figure 3 ACoNN recommendation model

图 3 ACoNN 推荐模型

示：

$$J(w_i)=y_{\text{real}}-\left(w_0+\sum_{i=1}^{|z|}w_i z_i+\sum_{i=1}^{|z|}\sum_{j=i+1}^{|z|}w_{ij}z_i z_j\right) \tag{9}$$

其中，y_{real} 为用户对项目的真实评分值，w_0 为全局偏置量，w_i 表示向量 z 中第 i 个分量的权重值，z_i 和 z_j 分别表示向量 z 的第 i 和第 j 个分量，w_{ij} 表示 z 中第 i 个与第 j 个特征向量的交互值。

相比深度神经网络，该模型训练阶段参数较少、复杂度较低。此外，注意力权值更新层的设计有助于捕捉文本中的重点信息，结合 CNN 具有共享权值和局部连接的特性，更加易于模型的优化[19]。此方法适用于解决图像视觉领域的图像分类和文本处理等问题，运用注意力机制能使 CNN 在每一步关注图像或者文本上的不同位置，提高对重点特征的提取效率。虽然基于注意力机制的 CNN 推荐方法能从输入中获取最有效的信息[20]，但是这种方法也不能表示动态变化的用户兴趣。

3.3 基于注意力机制的 RNN 推荐方法

RNN 即循环神经网络，是一类用以处理序列数据的神经网络。针对 DNN 和 CNN 不能解决时序数据的问题，一些研究者将注意力机制和 RNN 结合应用于推荐任务中，刻画用户兴趣的动态变化。LSTM(Long Short-Term Memory)和 GRU(Gated Recurrent Unit)是 RNN 的 2 种改进版本，它们在简化 RNN 内部循环结构的同时，缓解了 RNN 无法检测长序列的问题[21]。

针对微博的话题标签推荐没有考虑文本的时序特征问题，Li 等人[22]构建了一种基于主题注意力机制的 LSTM 模型，该模型与文献[15]中的 CNN 推荐模型相比，加入了时序特征的影响，有效提升了推荐性能。不足之处是忽略了用户信息、时间信息等数据对标签推荐的影响。Xing 等人[23]提出了基于词级与语句级注意力机制的用户-项目推荐模型，在 Yelp 和 Amazon 数据集上的实验中，推荐性能皆提升了近 2%，验证了考虑语义层面的推荐是有效的。但是，这种方法只有当目标用户为目标项目编写的评论可用时，才表现出最佳性能，数据量较少时会降低推荐效果。冯兴杰等人[24]提出了深度协同模型 DeepCLFM (Deep Collaborative Latent Factor Model)，解决了用户与项目的深层抽象特征挖掘不充分问题，通过对评论文本信息作全局偏倚项的补充，有效缓解了冷启动问题。但是，DeepCLFM 学习到的用户偏好向量是静态的，而同一用户对不同项目的偏好向量是不同的，此模型未加以区分。

为了解决标签推荐中存在的微博噪声问题，文献[25]提出了基于 LSTM 的时态增强语句级注意力模型。通过在语句级注意力层引入时间信息，减少了噪声数据对分类器的影响。其推荐模型如图 4 所示。其中，$M_i(i=1,2,\cdots,N)$表示第 i 条微博的词向量矩阵。

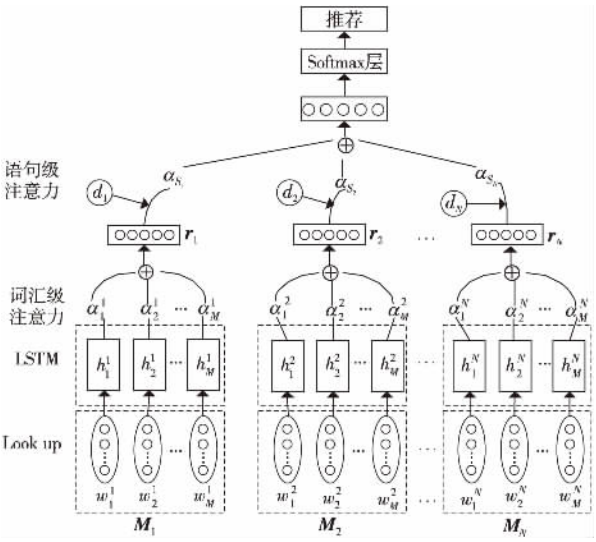

Figure 4 Temporal enhanced sentence-level attention model based on LSTM

图 4 基于 LSTM 的时态增强语句级注意力模型

基于 LSTM 的时态增强语句级注意力模型的主要实现步骤如下所示：

(1)Lookup 层：将微博中的单词 w_i 映射到一个低维向量中，得到嵌入向量 e_i 。

(2)LSTM 层：将实值嵌入向量序列 $b_N=\{e_1,e_i,\cdots,e_N\}$ 输入 LSTM，获得微博的高级语义表示 H，且 $H=\{h_1,h_2,\cdots,h_M\}$ 。其中，N 和 M 分别表示微博条数和最大长度。

(3)词汇级注意力层：通过更新每个隐状态 h_j 的注意力分数，得到词汇级注意力矩阵 α_w，然后求解隐状态的加权和，得到语句向量 r，如式(10)～式(11)所示：

$$\alpha_w=Softmax(\omega^T\tanh(H)) \tag{10}$$

$$r=H\alpha_w^T \tag{11}$$

其中，ω 是一个训练好的参数向量，ω^T 是它的转置，通过预训练得到。

(4)语句级注意力层：将词汇级注意力层输出的句子向量集合 $S=\{r_1,r_2,\cdots,r_N\}$ 输入语句级注意力层，先计算语句向量 r_i 与标签查询向量 t 的匹配分数 m_i；然后加入时间信息 d_i，得到每个语

句向量 $\boldsymbol{r}_i$ 的注意力权重 $\alpha_{\boldsymbol{M}_i}$;最后求解集合 S 中语句向量的加权和,记为 $\boldsymbol{R}$,如式(12)~式(14)所示:

$$m_i = \boldsymbol{r}_i \boldsymbol{A} \boldsymbol{t} \tag{12}$$

$$\alpha_{\boldsymbol{M}_i} = \frac{\exp(m_i \times d_i)}{\sum_{k=1}^{N} \exp(m_k \times d_k)} \tag{13}$$

$$\boldsymbol{R} = \sum_{i=1}^{N} \alpha_{\boldsymbol{M}_i} \boldsymbol{r}_i \tag{14}$$

其中,d_i 表示时间元素,当给定一个〈microblog $\boldsymbol{M}_i$,hashtag h〉的元组时,根据微博词向量矩阵 $\boldsymbol{M}_i$ 和标签,可以从一个需要训练的二维矩阵 $\boldsymbol{B} \in \mathbf{R}^{|time| \times |hashtag|}$ 中查找对应的 d_i。$|time|$ 是时间节点的个数,$|hashtag|$ 是标签的个数,$\boldsymbol{A}$ 是一个加权对角矩阵。

(5)推荐:将 $\boldsymbol{R}$ 处理成与标签相关的输出向量,送入 Softmax 层,得到候选标签的概率分布 $\hat{p}(\boldsymbol{t} \mid \boldsymbol{M}, \theta)$。根据最小化交叉熵误差,定义目标函数,如式(15)所示。由目标函数的结果进行标签推荐。

$$J(\theta) = \sum_{i=1}^{N} \lg \hat{p}(\boldsymbol{t}_i \mid \boldsymbol{M}_i, \theta) \tag{15}$$

其中,θ 是模型的所有参数,$\boldsymbol{M}_i$ 和 $\boldsymbol{t}_i$ 分别表示第 i 个微博向量和标签向量。

该模型不仅从词汇和语句 2 个级别对微博特征进行分层刻画和关联,还将时间信息引入注意力机制模型,弥补了文献[22]未考虑时间信息的不足,更形象地刻画了微博数据的动态性。因此,适用于解决文本翻译、语言识别和推荐中的序列预测问题,应用注意力机制使 RNN 能够将输出序列中的每一项与输入序列相关项对应,克服传统循环神经网络在学习超长序列上的限制问题[26]。但是,LSTM 和 GRU 等作为 RNN 的衍生,只可以处理欧几里得空间数据,对非欧空间数据的处理存在一定局限性,也无法解决非欧空间的推荐问题。

3.4 基于注意力机制的 GNN 推荐方法

GNN 即图神经网络,不仅对数据具有强大的特征提取和表示能力,还可以表示非欧几里得结构数据,可用于解决非欧空间的推荐问题[27]。针对传统协同过滤方法的稀疏性问题,Wu 等人[28]提出了一种双图注意力网络协作学习双重社会效应的推荐方法。该方法一方面由用户特定的注意力权重建模,另一方面由动态的、上下文感知的注意力权重建模,通过将用户领域的社会效应扩展到项目领域,缓解了数据稀疏性问题。模型可学习多方面社会影响的有效表示,具有良好的表达性,但社会图网络的构建相应增加了模型的时间复杂度。考虑当前网络社区推荐未充分考虑用户会受朋友偏好影响的问题,Song 等人[29]提出了一种基于动态图注意力神经网络的社区推荐模型,图注意力网络用来捕获朋友的短期与长期偏好对用户的影响。其模型图如图 5 所示。详细步骤主要有:

(1)用户动态偏好建模:通过 RNN 对用户近期的浏览内容进行建模,得到用户的偏好 $\boldsymbol{h}_n$。

(2)朋友偏好表示:利用 RNN 对朋友最近消费的项目进行建模,输出向量 $\boldsymbol{s}_k^s$ 表示朋友 k 的短期偏好。朋友的长期偏好 $\boldsymbol{s}_k^l$ 由用户的长期行为向量化后得到。连接 $\boldsymbol{s}_k^s$ 和 $\boldsymbol{s}_k^l$ 得到朋友 k 的整体偏好 $\boldsymbol{s}_k$。

(3)图注意力网络建模:对每个用户构建一个由节点表示用户及其朋友的图网络;然后计算目标

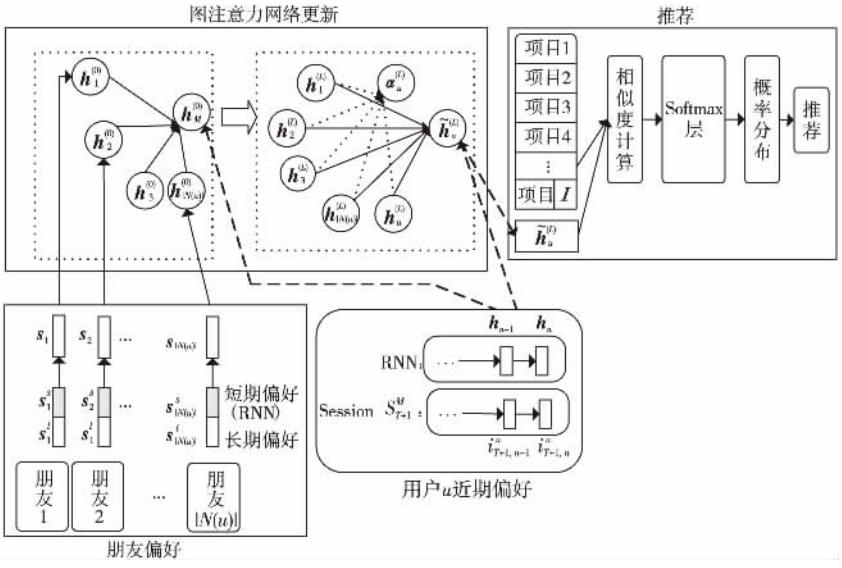

Figure 5 Dynamic graph attention network social recommendation model

图 5 动态图注意力网络社会推荐模型

用户 u 与朋友 k 的注意力分数 $a_{uk}^{(l)}$，如式(16)所示。

$$a_{uk}^{(l)}=\frac{\exp(f(\boldsymbol{h}_u^{(l)},\boldsymbol{h}_k^{(l)}))}{\sum_{j\in N(u)\cup\{u\}}\exp(f(\boldsymbol{h}_u^{(l)},\boldsymbol{h}_k^{(l)}))} \tag{16}$$

将用户 u 所有朋友的偏好向量联合注意力分数进行加权求和，得到朋友偏好的混合向量 $\tilde{\boldsymbol{h}}_u^{(l)}$，如式(17)所示：

$$\tilde{\boldsymbol{h}}_u^{(l)}=\sum_{k\in N(u)\cup\{u\}}a_{uk}^{(l)}\boldsymbol{h}_k^{(l)} \tag{17}$$

其中，$\boldsymbol{h}_u^{(l)}$ 和 $\boldsymbol{h}_k^{(l)}$ 表示目标用户节点和朋友节点的偏好向量，$N(u)$ 是用户 u 的好友集合，l 是当前注意力更新的次数。

(4)推荐：连接 $\boldsymbol{h}_n$ 和 $\boldsymbol{h}_u^{(L)}$ 得到融合了朋友偏好的用户偏好表示 $\hat{\boldsymbol{h}}_n$，如式(18)所示：

$$\hat{\boldsymbol{h}}_n=\boldsymbol{W}_2[\boldsymbol{h}_n:\boldsymbol{h}_u^{(L)}] \tag{18}$$

其中，$\boldsymbol{W}_2$ 是一个线性变换矩阵，L 表示注意力叠加的次数，$\boldsymbol{h}_u^{(L)}$ 是合并后的朋友偏好向量，由 $\tilde{\boldsymbol{h}}_u^{(l)}$ 经过 L 次注意力叠加得到。

之后由 Softmax 函数得到项目 y 的概率，表示用户对项目 y 可能感兴趣的程度，如式(19)所示。最终根据这个概率的大小，向用户进行推荐。

$$p(y\mid i_{T+1,1}^{u},\cdots,i_{T+1,n}^{u};$$

$$\{\vec{\boldsymbol{S}}_T^k,k\in N(u)\})=\frac{\exp(\hat{\boldsymbol{h}}_n^{\mathrm{T}}\boldsymbol{z}_y)}{\sum_{j=1}^{I}\exp(\hat{\boldsymbol{h}}_n^{\mathrm{T}}\boldsymbol{z}_j)} \tag{19}$$

其中，$\boldsymbol{z}_y$ 为项目 y 的嵌入，I 为项目的总数量，$\vec{\boldsymbol{S}}_T^k$ 是第 k 个朋友在 T 时刻的浏览记录，且 $\vec{\boldsymbol{S}}_T^k=\{i_{T,1}^k,\cdots,i_{T,N_{k,T}}^k\}$，$N_{k,T}$ 表示用户 k 在会话 T 中消费的项目的总数量，$i_{T,N_{k,T}}^k$ 表示用户朋友 k 在会话 T 中消费的第 $N_{k,T}$ 个项目。

该模型能充分利用朋友的短期与长期偏好，获取社会关系对用户偏好的影响，但项目的特征提取过程过于粗糙，忽略了用户和项目之间的互动关系。将注意力机制应用到 GNN 邻近节点上，能够学习每个邻近节点与该节点之间的影响[30]。此外，基于图结构的广义神经网络能够表示除语言、视频和图像之外的非欧几里得结构数据，通过对图数据进行处理，可深入挖掘其内部的特征和规律，解决如社交网络、信息网络和基础设施网络等领域中的推荐问题。

4 基于注意力机制的深度学习推荐的难点

4.1 提取注意力方法的选择问题

在一些场景下，可选择的注意力方法可能有多种。如文献[6]中，引入多头注意力与单层自注意力皆可提升分类任务的性能，但较使用自注意力而言，多头注意力更能提升模型在语句层面的特征表达能力，在 SemEval-2010 数据集上的实验中，多头注意力模型的 $F1$ 值相对自注意力模型的提高了 2.0%左右，说明不同的注意力方法对提升模型性能的贡献是不同的。近年来，许多研究者在不同任务场景下又提出了不同注意力机制的新变体，如双注意力[31]、双向分块自注意力[32]等，如何结合这些新变体，选择适合当前推荐任务的注意力方法仍具有一定的复杂性。

4.2 注意力融入时机的选择问题

在注意力机制与 CNN 相结合的工作中，Yin 等人[33]和 Santos 等人[34]通过实验证实了注意力机制用于池化层的效果比卷积层好。在此基础上，文献[35]将注意力与 CNN 池化层、项目潜在向量层及 MLP 输入层相结合进行对比实验，发现在稠密数据集上，注意力与池化层相结合的模型表现得更加稳定；而在稀疏数据集上，注意力与隐藏层相结合模型预测效果更佳，说明注意力引入时机的差异、数据集稠密度差别，都会影响最终的推荐结果。CNN 相对神经网络，结构较简单，而在更加复杂的任务场景下，使用的神经网络也更加复杂，增加了注意力机制融入深度神经网络中的时机的难度。

4.3 融入注意力机制引起推荐模型复杂度增加问题

虽然注意力机制可以改善传统编码器-解码器的部分问题，但引入注意力机制获得注意力分配权重时，需要计算源语言句子中所有词语的权重，该过程计算资源耗费大，增大了推荐模型复杂度，还会导致模型的训练速度和推断速度下降。同时，引入注意力机制可能需要更多的存储资源，例如对于自注意来说，需要很大的存储空间来保存元素的对齐分数，需要的存储空间随序列长度呈二次方增长，因此在保证效率的前提下降低推荐模型的复杂度存在一定的难度。

4.4 融入注意力机制的推荐效果评价问题

注意力机制应用范围广，但并不是对所有模型引入注意力机制都可以提高性能。例如，因子分解机 FM(Factorization Machine)利用同一特征向量表示某个特征和其它特征间的交互显然是不合理的。于是 Juan 等人[36]和 Xiao 等人[37]分别提出了 FFM(Field-aware Factorization Machine)和 AFM(Attentional FM)2 种新的方法。FFM 通过引入“域”的概念，对不同域使用不同的向量来解决这一

问题。而 AFM 通过引入注意力机制对不同的交互项计算注意力权重,区分特征的重要程度。比较来看,AFM 虽然和 FFM 效果相当,但是 AFM 通过引入新参数来弥补某方面的拟合能力,可能会造成过拟合现象。所以,对模型引入注意力机制后的推荐效果进行多方面的评价,也是当前基于注意力机制的深度学习推荐的一个难点。

5 基于注意力机制的深度学习推荐未来研究方向

5.1 多特征交互的注意力机制深度学习推荐

当涉及多特征交互时,通常采用矩阵分解模型来实现,如文献[37]利用一个神经注意力网络对不同交互特征的重要程度进行区分,改善了因子分解机的性能,并在真实数据集上将回归任务的性能提高了 8.6%。但是,基于矩阵分解的协同过滤方法仅使用评分信息,不能捕捉更深层的特征信息。而文献[38]利用多层交互的非线性网络结构获取不同层次的交互结果,将 RMSE 指标的值降低了 2%左右。但是,这种基于深度学习的推荐模型在提升推荐效果的同时,难以对推荐效果做出合理的解释。所以,考虑在多特征交互的推荐模型中加入注意力机制,以提高模型的可解释性,是值得研究的重要课题之一。

5.2 多模态注意力机制的深度学习推荐

信息的媒介有音频、文字、语音和图像等多种模态,目前对多模态信息的使用仍不够广泛,在多模态注意力机制中,主要使用语音和图像信息。文献[39]认为不同模态对于情感状态的影响是不同的,作者通过多模态注意力机制,将视频特征和音频特征进行融合,相比一些采用主流深度学习方法进行情感分析的任务,在性能上提高了 2%左右。在深度学习的推荐研究中,除了利用文本、评分等信息外,还可以从视频和它模态信息中提取用户的偏好特征。所以,将多模态注意力机制结合深度学习技术,用于推荐系统也是未来的一个研究方向。

5.3 注意力机制的 GNN 推荐和其他推荐方法融合

由于 GNN 可以用来表示其它神经网络无法表示的非欧几里得结构数据,将其作为辅助工具应用在推荐系统领域,可有效缓解数据稀疏性问题[40]。文献[28]引入双图注意力网络来协作学习用户的静态和动态双重社会效应,同时考虑到用户领域和项目领域中不同的社会效应会相互作用,提出了基于多臂赌博机的一种新的融合策略来衡量这种交互作用,在真实数据集上的实验表明,其推荐精度最高提高了 9.33%。因此,将注意力机制的 GNN 推荐融合其它推荐算法或深度学习技术,有利于提高推荐的效果。

5.4 基于注意力机制的深度学习群组推荐

大多数推荐技术应用于个性化推荐,但在很多情况下,推荐的产品或服务被一群用户所消费[41]。文献[42]提出了一种 AGR(Attention-based Group Recommendation)模型,利用注意力机制学习群体中每个用户的影响权重,相较于基准模型其推荐性能提高了 3%以上。但是,作者只在模型中使用了项目的 ID 信息,得到的信息非常有限,对模型性能的提升也有一定的限制。而李振新[43]提出的基于 Phrase-LDA 模型从评论中提取用户主题,更细致地从语义层面描述了用户的偏好,在群组推荐领域中具有一定的新颖性。考虑在 AGR 模型的基础上,将诸如社交关系、文本信息(例如事件描述)或时间等上下文信息用来学习群组推荐中的注意力模型,也是未来的一个研究方向。

5.5 基于注意力机制和深度学习的跨领域推荐

单领域个性化推荐中容易出现数据稀疏性和冷启动问题,使得推荐效果不够理想。而在跨领域推荐中,其它辅助域信息可以为目标域推荐提供帮助,从而解决传统单域推荐中数据稀疏和冷启动问题,因此逐渐成为学术界的研究热点。文献[44]构建了一个基于注意力机制和知识迁移方法的卷积-双向长短期记忆 AC-BiLSTM(Convolution-Bi-directional Long Short-Term Memory based on Attention mechanism)模型,向 BiLSTM 中引入注意力机制得到不同词汇对文本的贡献程度,并且在目标函数中加入了正则约束项,避免在迁移过程中出现负迁移现象,使跨领域情感分类的平均准确率在 2 个数据集上分别提高了 6.5%和 2.2%。结合相关情感分类模型,将注意力机制应用到跨领域推荐研究中也是未来的一个研究方向。

6 结束语

注意力机制的特点是能主动从海量输入信息中选择对当前目标任务更重要的信息,在提高推荐模型性能的同时,提升深度学习可解释性。将注意力机制应用到深度学习推荐研究中,扩展了推荐模型中神经网络的能力。本文围绕注意力机制的结

构、分类以及注意力机制在深度学习推荐中的研究等角度展开,并针对深度学习推荐模型中存在的注意力机制的选择、阶段融入、评价和模型复杂度增加等难点与挑战进行了分析,最后指出了基于注意力机制的深度学习推荐未来的研究方向。

参考文献:

[1] Li J W, Will M, Dan J. Understanding neural networks through representation erasure[J]. arXiv: 1612.08220, 2016.

[2] Treisman A M, Gelade G. A feature-integration theory of attention[J]. Cognitive Psychology, 1980, 12(1): 97-136.

[3] Zhu Zhang-li, Rao Yuan, Wu Yuan, et al. Research progress of attention mechanism in deep learning[J]. Journal of Chinese Information Processing, 2019, 33(6): 1-11. (in Chinese)

[4] Bahdanau D, Cho K, Be Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409, 0473, 2014.

[5] He X D, Golub D. Character-level question answering with attention[C]// Proc of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1598-1607.

[6] Liu Feng, Gao Sai, Yu Bi-hui, et al. Relation classification based on multi-head attention and bidirectional long short-term memory networks[J]. Computer Systems & Applications, 2019, 28(6): 118-124. (in Chinese)

[7] Li Yu-yu, Lang Cong-yan, Feng Song-he. Rating prediction recommendation model based on dual-level attention mechanism[J]. China Sciencepaper, 2018, 13(18): 2076-2081. (in Chinese)

[8] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[C]// Proc of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1412-1421.

[9] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// Proc of the 32nd International Conference on Machine Learning, 2015: 2048-2057.

[10] Zhang Quan-gui, Zhang Xin-xin, Li Zhi-qiang. Music recommendation algorithm based on attention mechanism[J]. Application Research of Computers, 2019, 36(8): 2297-2299. (in Chinese)

[11] Shen Dong-dong, Wang Hai-tao, Jiang Ying, et al. Item similarity recommendation model based on attention mechanism [J]. Electronic Measurement Technology, 2019, 42(15): 150-154. (in Chinese)

[12] Guo Xu, Zhu Jing-hua. Deep neural network recommendation model based on user vectorized representation and attention mechanism[J]. Computer Science, 2019, 46(8): 111-115. (in Chinese)

[13] Xie En-ning, He Ling-min, Wang Xiu-hui. Deep collaborative filtering model based on attention mechanisms[J]. Journal of China Jiliang University, 2019, 30(2): 219-225. (in Chinese)

[14] Liu Jing. Research on collaborative filtering algorithm based on deep learning [D]. Beijing: Beijing University of Posts and Telecommunications, 2019. (in Chinese)

[15] Gong Y Y, Zhang Q. Hashtag recommendation using attention-based convolutional neural network[C]// Proc of the 25th International Joint Conference on Artificial Intelligence, 2016: 2782-2788.

[16] Zhang Q, Wang J W, Huang H R, et al. Hashtag recommendation for multimodal microblog using co-attention network[C]// Proc of the 26th International Joint Conference on Artificial Intelligence, 2017: 3420-3426.

[17] Wang X J, Yu L J, Ren K, et al. Dynamic attention deep model for article recommendation by learning human editors demonstration[C]// Proc of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 2051-2059.

[18] Huang Wen-ming, Wei Wan-cheng, Zhang Jian, et al. Recommendation method based on attention mechanism and review text depth model[J]. Computer Engineering, 2019, 45(9): 176-182. (in Chinese)

[19] Zhou Fei-yan, Jin Lin-peng, Dong Jun. Review of convolutional neural networks [J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251. (in Chinese)

[20] Xiao Qing-xiu, Tang Kun. A deep learning recommendation system of movie based on dual-attention model[J]. Computer and Modernization, 2018(11): 109-114. (in Chinese)

[21] Wang Hong, Shi Jin-chuan, Zhang Zhi-wei. Text semantic relation extraction of LSTM based on attention mechanism [J]. Application Research of Computers, 2018, 35(5): 1417-1420. (in Chinese)

[22] Li Y, Liu T, Jiang J, et al. Hashtag recommendation with topical attention-based LSTM[C]// Proc of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 3019-3029.

[23] Xing S N, Liu F A, Wang Q Q, et al. A hierarchical attention model for rating prediction by leveraging user and product reviews[J]. Neurocomputing, 2019, 332: 417-427.

[24] Feng Xing-jie, Zeng Yun-ze. Joint deep modeling of rating matrix and reviews for recommendation [J]. Chinese Journal of Computers, 2020, 43(5): 884-900. (in Chinese)

[25] Jun M, Chong F, Ge S, et al. Temporal enhanced sentence-level attention model for hashtag recommendation [J]. CAAI Transactions on Intelligence Technology, 2018, 3(2): 95-100.

[26] Li Mei, Ning De-jun, Guo Jia-cheng. Attention mechanism-based CNN-LSTM model and its application [J]. Computer Engineering and Applications, 2019, 55(13): 20-27. (in Chinese)

[27] Wu Guo-dong, Zha Zhi-kang, Tu Li-jing, et al. Research advances in graph neural network recommendation [J]. CAAI Transactions on Intelligent Systems, 2020, 15(1): 14-24. (in Chinese)

[28] Wu Q T,Zhang H R,Gao X F,et al. Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems[C]// Proc of the 2019 World Wide Web Conference,2019:2091-2102.

[29] Song W P,Xiao Z P,Wang Y F,et al. Session-based social recommendation via dynamic graph attention networks[C]// Proc of the 12th ACM International Conference on Web Search and Data Mining,2019:555-563.

[30] Qu M,Tang J,Shang J B,et al. An attention-based collaboration framework for multi-view network representation learning[C]// Proc of the 2017 ACM Conference on Information and Knowledge Management,2017:1767-1776.

[31] Liang Tian-an. Research on personalized conversation recommendation model of double attention [D]. Wuhan:Huazhong University of Science and Technology,2019. (in Chinese)

[32] Shen T,Zhou T Y,Long G D,et al. Bi-directional block self-attention for fast and memory-efficient sequence modeling[J]. arXiv:1804.00857,2018.

[33] Yin W P, Schütze H. MultiGranCNN:An architecture for general matching of text chunks on multiple levels of granularity[C]// Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,2015:63-73.

[34] Santos C,Tan M,Xiang B,et al. Attentive pooling networks [J]. arXiv:1602.03609,2016.

[35] Zhang Hao-bo. Research on collaborative filtering prediction score integrating attention mechanism and text context information [D]. Changchun:Northeast Normal University,2019. (in Chinese)

[36] Juan Y C, Zhuang Y, Chin W S, et al. Field-aware factorization machines for CTR prediction[C]// Proc of the 10th ACM Conference on Recommender Systems,2016:43-50.

[37] Xiao J,Ye H,He X N,et al. Attentional factorization machines:Learning the weight of feature interactions via attention networks[C]// Proc of the 26th International Joint Conference on Artificial Intelligence,2017:3119-3125.

[38] Li Tong-huan,Tang Yan,Liu Bing. Multi-interaction hybrid recommendation model based on deep learning [J]. Computer Engineering and Applications,2019,55 (1):135-141. (in Chinese)

[39] Tang Yu-hao,Mao Qi-rong,Gao Li-jian. Dimensional emotional recognition based on hierarchical attention mechanism [J/OL]. Computer Engineering[2020-05-15]. https://doi.org/10.19678/j.issn.1000-3428.0054127. (in Chinese)

[40] Zhou J,Cui G Q,Zhang Z Y,et al. Graph neural networks:A review of methods and applications[J]. arXiv:1812.08434,2018.

[41] McCarthy J F. Pocket restaurant finder:A situated recommender system for groups[C]// Proc of Workshop on Mobile Ad-Hoc Communication at the 2002 ACM Conference on Human Factors in Computer Systems, 2002:2-5.

[42] Vinh T D Q,Pham T A N,Cong G,et al. Attention-based group recommendation[J]. arXiv:1804.04327,2018.

[43] Li Zhen-xin. Research on tea product group recommendation based on the phrase-LDA theme model [D]. Hefei:Anhui Agricultural University,2016. (in Chinese)

[44] Gong Qin, Lei Man, Wang Ji-chao, et al. Cross-domain sentiment classification method of convolution-bi-directional long short-term memory based on attention mechanism [J]. Journal of Computer Applications,2019, 39(8):2186-2191. (in Chinese)

附中文参考文献:

[3] 朱张莉,饶元,吴渊,等.注意力机制在深度学习中的研究进展[J].中文信息学报,2019,33(6):1-11.

[6] 刘峰,高赛,于碧辉,等.基于 Multi-head Attention 和 Bi-LSTM 的实体关系分类[J].计算机系统应用,2019,28(6):118-124.

[7] 李钰钰,郎丛妍,冯松鹤.基于双层注意力机制的评分预测推荐模型[J].中国科技论文,2018,13(18):2076-2081.

[10] 张全贵,张新新,李志强.基于注意力机制的音乐深度推荐算法[J].计算机应用研究,2019,36(8):2297-2299.

[11] 沈冬东,汪海涛,姜瑛,等.基于注意力机制的项目相似性推荐模型[J].电子测量技术,2019,42(15):150-154.

[12] 郭旭,朱敬华.基于用户向量化表示和注意力机制的深度神经网络推荐模型[J].计算机科学,2019,46(8):111-115.

[13] 谢恩宁,何灵敏,王修晖.基于注意力机制的深度协同过滤模型[J].中国计量大学学报,2019,30(2):219-225.

[14] 刘晶.基于深度学习的协同过滤算法的研究[D].北京:北京邮电大学,2019.

[18] 黄文明,卫万成,张健,等.基于注意力机制与评论文本深度模型的推荐方法[J].计算机工程,2019,45(9):176-182.

[19] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(6):1229-1251.

[20] 肖青秀,汤鲲.基于双层注意力机制的深度学习电影推荐系统[J].计算机与现代化,2018(11):109-114.

[21] 王红,史金钏,张志伟.基于注意力机制的 LSTM 的语义关系抽取[J].计算机应用研究,2018,35(5):1417-1420.

[24] 冯兴杰,曾云泽.基于评分矩阵与评论文本的深度推荐模型[J].计算机学报,2020,43(5):884-900.

[26] 李梅,宁德军,郭佳程.基于注意力机制的 CNN-LSTM 模型及其应用[J].计算机工程与应用,2019,55(13):20-27.

[27] 吴国栋,查志康,涂立静,等.图神经网络推荐研究进展[J].智能系统学报,2020,15(1):14-24.

[31] 梁天安.双注意力个性化会话推荐模型研究[D].武汉:华中科技大学,2019.

[35] 张昊博.集成注意力机制和文本上下文信息的协同过滤预测评分研究[D].长春:东北师范大学,2019.

[38] 李同欢,唐雁,刘冰.基于深度学习的多交互混合推荐模型[J].计算机工程与应用,2019,55(1):135-141.

[39] 汤宇豪,毛启容,高利剑.基于层次注意力机制的维度情感识别方法[J/OL].计算机工程[2020-05-15]. https://doi.org/10.19678/j.issn.1000-3428.0054127.

[43] 李振新.基于 Phrase-LDA 主题模型的茶产品群组推荐研究[D].合肥:安徽农业大学,2016.

[44] 龚琴,雷曼,王纪超,等.基于注意力机制的卷积双向长短期记忆模型跨领域情感分类方法[J].计算机应用,2019,39(8):2186-2191.

作者简介:

陈海涵(1994-),女,安徽蚌埠人,硕士生,研究方向为人工智能及其应用。**E-mail**:15665593987@163.com

CHEN Hai-han, born in 1994, MS candidate, her research interest includes artificial intelligence and its applications.

吴国栋(1972-),男,安徽安庆人,博士,副教授,CCF 会员(08063M),研究方向为智能决策和推荐系统。**E-mail**:wugd@ahau.edu.cn

WU Guo-dong, born in 1972, PhD, associate professor, CCF member(08063M), his research interests include intelligent decision making, and recommendation system.

李景霞(1976-),女,安徽巢湖人,博士,讲师,CCF 会员(40303M),研究方向为分布式计算和云计算。**E-mail**:jxiali@163.com

LI Jing-xia, born in 1976, PhD, lecturer, CCF member(40303M), her research interests include distributed computing, and cloud computing.

王静雅(1996-),女,安徽合肥人,硕士生,研究方向为农业电子商务和智能信息处理。**E-mail**:912718123@qq.com

WANG Jing-ya, born in 1996, MS candidate, her research interests include agricultural E-commerce, and intelligent information processing.

陶鸿(1995-),男,安徽马鞍山人,硕士生,研究方向为人工智能及其应用。**E-mail**:869546564@qq.com

TAO Hong, born in 1995, MS candidate, his research interest includes artificial intelligence and its applications.