文章编号：0253-374X(2022)04-0590-11

DOI：10.11908/j.issn.0253-374x.21477

新闻推荐系统中的边信息融合卷积神经网络

卫　刚，邵　伟，王志成

（同济大学 电子与信息工程学院，上海 201804）

摘要：现有的新闻推荐模型一般由文本特征提取网络和推荐网络两部分组成。新闻相关的边信息（如类别信息）并没有作用在文本特征提取过程中。在未融合边信息的情况下，文本特征提取网络和推荐网络两部分的优化目标是有差异的。提出SIACNN（Side Information Aggregated CNN）的结构，它通过注意力机制的方式，将边信息结合到文本特征提取中，缩小了文本特征提取和推荐网络之间优化目标的差异，有效提升了新闻推荐的效果。将SIACNN替换多个典型新闻推荐网络中的卷积神经网络，并利用MSN（微软新闻）采集的大型新闻数据集MIND（MIcrosoft News Dataset）来进行实验，通过实验证明了SIACNN能提高推荐效果，并同时具有泛化性。

关键词：新闻推荐系统；边信息；文本特征；深度学习

中图分类号：TP399　　**文献标志码**：A

Side Information Aggregated Convolutional Neural Network in News Recommendation

WEI Gang, *SHAO Wei*, *WANG Zhicheng*

(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: Existing news recommendation models generally consist of the text feature extraction network and the recommendation network. News-related side information, such as category, is not fused into the text feature extraction network. Without fusing it, there are differences between the optimization targets of the text feature extraction network and the recommendation network. In this paper, a general SIACNN (side information aggregated CNN) layer is proposed. The SIACNN layer fuses the side information into the text feature through the attention mechanism, which bridges the gap between text feature extraction and recommendation tasks and improves the effectiveness of the recommendation. CNNs are replaced in many state-of-the-art models which used CNNs to extract the text feature with the SIACNN and several experiments are conducted in a large real-world news recommendation dataset MIND (MIcrosoft News Dataset) collected from MSN (MicroSoft News). The recommendation effectiveness and generality of SIACNN are verified by several experiments.

Key words: news recommendation; side information; text feature; deep learning

如今MSN新闻和谷歌新闻等平台吸引了大量用户[1-2]，这些平台从各个渠道获取了海量的新闻。海量的新闻导致了严重的信息过载[3]。新闻的个性化推荐算法能够挖掘每个人的兴趣爱好和新闻的语义，能够缓解信息过载的问题[4-6]。

基于协同过滤的推荐算法往往根据新闻的历史点击情况对用户进行推荐，然而最新的新闻缺乏用户点击记录，因此基于协同过滤的推荐算法只能适用于热门新闻，对最新的新闻（冷启动）效果不佳。缓解新闻冷启动问题的关键在于优化文本信息的利用方式。对文本信息的利用方式直接影响了最新的新闻的曝光机会。这是本文优化文本特征提取的直接动机。此外标题和摘要信息往往是吸引用户的关键因素之一，这是另一个优化文本特征提取的动机。

在推荐系统中有一些边信息，如多级类别信息、图像信息、社交信息等。大多数的新闻推荐模型先提取文本特征，后结合边信息通过推荐网络对新闻打分和推荐。文本特征提取和推荐2个任务在网络结构中是割裂的。NLP（自然语言处理）领域的文本特征提取网络的设计初衷是出于更好的文本特征提取，而推荐网络的设计初衷是为了更好的推荐效果。两部分优化目标是不同的，很难做到两者同时最优。

收稿日期：2021-10-13

第一作者：卫　刚（1973—），男，副研究员，工学博士，主要研究方向为计算机应用、人工智能、计算机辅助设计。E-mail：weigang@tongji.edu.cn

通信作者：邵　伟（1996—），男，硕士生，主要研究方向为计算机推荐系统。E-mail：weis_96@163.com

论文拓展介绍

给出2个案例来说明将边信息融合到文本特征提取过程的必要性。

例如有一个新闻摘要:A公司已召回了所有的可能感染细菌的苹果。如果这是一个健康类别的新闻,“细菌”是用户重点关注的词;如果是投资类别的新闻,“A公司”将是用户重点关注的词。在新闻推荐的任务中,每个词的重要程度随着类别这种边信息而改变。而NLP领域的文本特征提取网络则对这些词一视同仁,它的任务是尽可能把摘要中所有语义放入特征中。而将边信息(类别信息)融合到文本特征提取过程中,对不同词语产生不同的关注度,可以缩小2个任务的目标差异,能将文本特征提取的优化目标对齐到更好的推荐上。

边信息除了影响不同词语的关注度以外,还影响词语的语义。比如“跳水”这个词,在运动类别的新闻中表示一种运动的语义,在投资中则表示了股价迅速大幅度下滑的意思。又比如“唐山大地震”在电影类的新闻里,表示一部灾难片的语义,而在其他新闻中表征的语义可能就是一场地震。同一个词在不同领域会有不同语义。因此类别这种边信息应该融合到短语的特征提取中,使得同一个短语也能产生出不同的特征。

上述2个例子显示了将边信息融合到文本特征提取中的重要性。但是目前绝大多数算法都并没有在边信息的指导下融合文本特征。现在有很多主流的新闻推荐网络都是用了CNN(卷积神经网络)作为文本特征提取器,如NPA[7](Neural News Recommendation with Personalized Attention)、LSTUR[8](Neural News Recommendation with Long- and Short-term User Representations)、NAML[9](Neural News Recommendation with Attentive Multi-View Learning)等。在本文中,SIACNN的设计是为了解决上述2个例子中的问题,将边信息类别信息融合到文本特征中。

1　研究背景

新闻推荐系统[10-15]的技术涉及NLP和数据挖掘领域。随着信息过载问题越发严重,新闻推荐系统也越来越广泛地被研究。早期新闻推荐系统主要是靠繁杂的特征工程[4,16-18]。Liu等[13]使用不同类型新闻的点击分布来构建用户的长短期兴趣特征。Phelan等[17]结合了推特上的内容和RSS(Really Simple Syndication)新闻的投放数据流来提高推荐效果。这些方法都需要基于先验的知识来设计特征。在近些年,深度学习降低了特征工程的难度,被广泛应用于新闻推荐中。目前大多数深度学习模型都基于自动编码器、CNN[19]、RNN(循环神经网络)和transformer来提取文本特征(如标题和摘要)。Okura等[20]使用带有弱监督的自动编码器来获取文本特征。还有用SDAE(stacked denoising auto-encoder)自动编码器来提取文本特征[21]。DKN[22](Deep Knowledge-Aware Network)利用CNN将文本特征和知识图谱信息融合。Wang等[23]将新闻标题、类别和二级类别拼接成单词序列,并用3D卷积来提取文本特征。Lee等[24]使用BIGRU(Bi-directional Gated Recurrent Unit)来提取文本特征。Wu等[25]用transformer对标题和摘要分别进行特征提取,还使用BERT(Bidirectional Encoder Representations from Transformers)提取文本特征,并用模型蒸馏来缩减模型复杂度[26]。

比起NLP的任务,新闻推荐系统有额外的边信息。大多数模型没有很好地利用边信息提取文本特征。比如NPA[7]使用CNN提取单词之间的局部特征,再将局部特征聚合成新闻的文本特征。LSTUR[8]用类似的方法提取文本特征后简单地将类别信息拼接在新闻特征上。在NAML[9]中,分别提取标题、摘要的文本特征还有类别特征,然后将三者用注意力机制聚合成新闻特征。在DAN[27](Deep Attention Neural Network)中,新闻特征由标题特征和实体特征卷积后拼接而成。这些模型的共同点是文本特征提取过程中没有融合边信息,而在推荐网络中才使用到了边信息。文献[28]提出了CSCNN(Category-Specific CNN)的概念,证明了提前将边信息融合到图像特征提取的过程中也有利于推荐系统。受此启发,本文设计SIACNN以替换NPA、NAML、LSTUR等模型的CNN部分。

有少数模型已在文本特征提取过程中融合了边信息,如Wang等[23]将类别作为单词序列拼接在标题后,并用3D卷积将文本信息和边信息融合。但是这种融合方式过于简单直接,并不是所有边信息都可以变成单词序列。DKN使用不同大小的卷积核对文本和知识图谱共同卷积,并用max over time pooling来提取新闻特征。在本文中方法可以视作对DKN文本特征提取模式的扩充和改进。

2 算法与模型

SIACNN (Side Information Aggregated Convolutional Neural Network)由SWANN(Single Word Attention Neural Network)、CNN和MWACNN(Multiple Words Attention Convolutional Neural Network)组成。

2.1 SIACNN的结构

先给出NPA、NAML等模型的共有结构。如图1,第1部分是单词嵌入层。随机初始化一个嵌入矩阵 $\boldsymbol{W}_E$, $\boldsymbol{W}_E \in \mathrm{R}^{V\times D}$,其中 V 为单词总量,D 为单词特征的维度。每个单词 w_i 通过查表的方式从嵌入矩阵中获取到它对应的单词特征向量 $\boldsymbol{e}_i$。第2个部分是卷积层。卷积层共享权重的特点有助于提取相邻单词之间的局部特征。每个单词特征 $\boldsymbol{e}_i$ 通过卷积层得到结合局部特征的 $\boldsymbol{c}_i$。最后一部分是将这些单词特征聚合成新闻的文本特征。如求和聚合,又如NPA使用用户信息生成的权重聚合等。NPA、LSTUR、NAML等模型大致遵从这个结构,也有一些微小的差别。如LSTUR将类别特征拼接在了聚合后的新闻文本特征后,而NAML则是对标题和摘要分别使用这样的结构提取文本特征,最后再将两者合并起来。

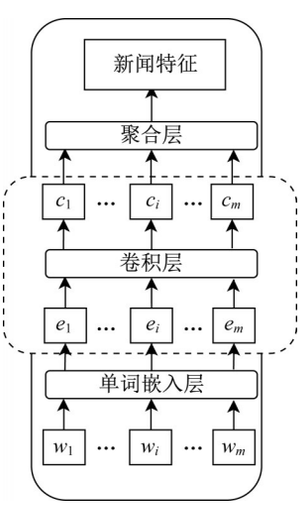

图1 新闻推荐中文本特征提取的通用结构

Fig. 1 General structure of text feature extraction in news recommendation

图1里虚线框中的部分被置换成了SIACNN,如图2。SIACNN比普通的卷积层多了2个模块:SWANN和MWACNN。SWANN结合了边信息和单词特征,根据不同的新闻类型,给予不同单词不同关注度。MWACNN将边信息融合到词组的特征提取中,这使得同一个词组在不同类型的新闻中有不同的语义。SIACNN的2个子结构分别针对上文提及的2个案例做出优化。

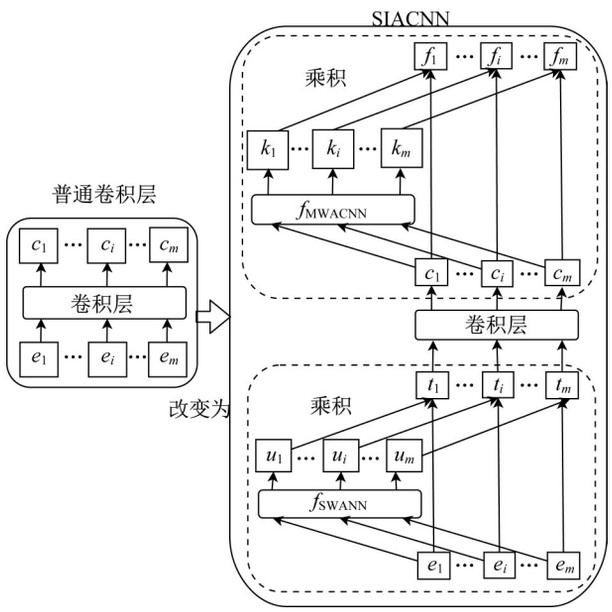

图2 SIACNN网络结构

Fig. 2 Network structure of SIACNN

图2是SIACNN的结构图。每个单词的特征向量 $\boldsymbol{e}_i$ 作为单词特征输入,经过 f_{SWANN} 后,每个单词获得一个关注度权重 u_i,将 u_i 和对应的 $\boldsymbol{e}_i$ 相乘,每个单词 $\boldsymbol{e}_i$ 被映射成结合边信息的单词特征 $\boldsymbol{t}_i$,对应公式见式(1):

$$u_i = f_{\mathrm{SWANN}}\left(\boldsymbol{e}_i,\ v_{\mathrm{side\_information}}\right)$$
$$\boldsymbol{t}_i = u_i \cdot \boldsymbol{e}_i \tag{1}$$

式中:$u_i \in \mathrm{R}$, $\boldsymbol{t}_i$、$\boldsymbol{e}_i \in \mathrm{R}^D$, $i \in [1, m]$; $v_{\mathrm{side_{information}}}$ 为边信息;m 为句子单词数量。

紧接着,$\{\boldsymbol{t}_i \mid i \in [1, m]\}$ 组成了一张图,在这张图上利用CNN来捕获相邻单词的局部特征,见式(2):

$$\boldsymbol{c}_i = f_{\mathrm{Relu}}\left(\boldsymbol{F}_t \times \boldsymbol{t}_{(i-l_t):(i+l_t)} + b_t\right) \tag{2}$$

式中:$\boldsymbol{t}_{(i-l_t):(i+l_t)}$ 是由以第 i 个单词为中心的、连续 $2l_t+1$ 个单词特征拼接成的图,大小为 $(2l_t+1)\times D$; $\boldsymbol{F}_t \in \mathrm{R}^{N_f \times (2l_t+1)D}$ 代表 N_f 个卷积核,每个卷积核的大小为 $(2l_t+1)\times D$; b_t 为每个卷积核的偏置。

最后每个特征向量 $\boldsymbol{c}_i$ 都会根据MWACNN模块得到一个修正向量 $\boldsymbol{k}_i$。最终的每个特征向量与修正向量相乘得到结合边信息的文本特征 $\boldsymbol{f}_i$,见式(3):

$$\boldsymbol{k}_i = \mathrm{f}_{\mathrm{MWACNN}}\Big(\boldsymbol{c}_{(i-K_1):(i+K_1)}, \boldsymbol{c}_{(i-K_2):(i+K_2)}, \boldsymbol{c}_{(i-K_3):(i+K_3)}, \boldsymbol{v}_{\mathrm{side\_information}}\Big)$$

$$\boldsymbol{f}_i = \boldsymbol{k}_i \odot \boldsymbol{c}_i \tag{3}$$

式中：K_j为不同词组长度；$\boldsymbol{c}_{(i-K_j):(i+K_j)}$为以$i$为中心、$2K_j+1$为长度的词组的特征图，一共预设3种词组长度；$\boldsymbol{c}_i$、$\boldsymbol{f}_i$和$\boldsymbol{k}_i$都是$N_f$维度的向量。

2.2　SWANN的结构

在不同类型的新闻中用户对每个单词的关注度是不一样的。例如一个新闻摘要：A公司已召回了所有的可能感染细菌的苹果，在投资类新闻中“A公司”的关注度更大，而在健康类的新闻里“细菌”的关注度更高。SWANN就是为了针对这种情况设计的网络，它将类别信息和文本融合，分配不同的关注度到不同单词。模型结构如图3所示。

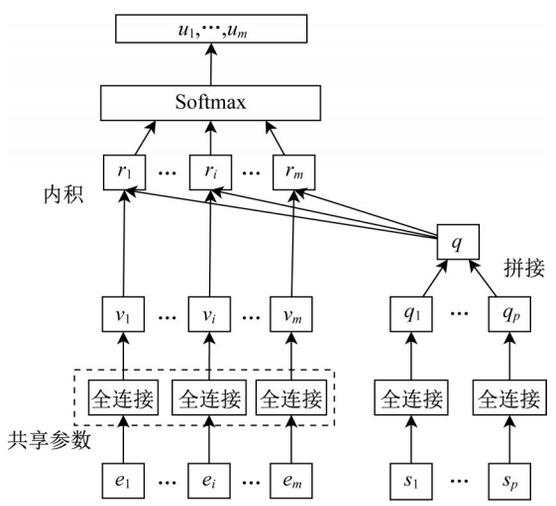

图3　SWANN网络结构

Fig. 3　Network structure of SWANN

图3中，SWANN将所有单词的嵌入向量和所有的边信息作为模型的输入，利用一个共享的全连接层用于对每个单词做映射，最后产出每个单词的特征向量$\boldsymbol{v}_i$。

$$\boldsymbol{v}_i = f_{\tanh}(\boldsymbol{V}_w \times \boldsymbol{e}_i + b_w) \tag{4}$$

式中：$\boldsymbol{V}_w$和b_w分别为所有单词共享全连接层的参数矩阵和偏置；$\boldsymbol{V}_w$的大小是$H \times D$，得到H维向量$\boldsymbol{v}_i$。

类似地，每种边信息也都通过全连接层做映射。最后边信息的特征由所有映射后的向量拼接而成，见式(5)：

$$\boldsymbol{q}_i = f_{\tanh}(\boldsymbol{V}_{s_i} \times \boldsymbol{S}_i + b_{s_i})$$

$$\boldsymbol{q} = f_{\mathrm{concat}}(\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_p) \tag{5}$$

式中：$\boldsymbol{V}_{s_i}$和b_{s_i}分别为全连接层的参数矩阵和偏置，不同的边信息对应的全连接层不共享参数。一共p种边信息，所有边信息的特征拼接后得到边信息特征$\boldsymbol{q}$，$\boldsymbol{q}$也是一个H维向量。

最后将每个单词的特征向量$\boldsymbol{v}_i$和边信息特征向量$\boldsymbol{q}$内积得到单词级别的关注度r_i。通过softmax函数对r_i进行归一化得到归一化后的u_i。

$$r_i = \boldsymbol{v}_i \cdot \boldsymbol{q}$$

$$u_i = \frac{\exp(r_i)}{\sum_{j=1}^{m}\exp(r_j)} \tag{6}$$

2.3　MWACNN的结构

在通过SWANN后，每个单词具有不同的关注度。在通过CNN层后，每个单词的特征$\boldsymbol{c}_i$也获得了相邻单词的局部特征。在结合了单词关注度和局部信息后，单词特征将送入MWACNN的模块中。

前文提及“跳水”在投资类新闻和运动类新闻里含义的差别，以及“唐山大地震”在电影类新闻和其他类新闻里的含义差别。利用MWACNN可以结合边信息丰富词组的语义，使得同一个词组在多个类型的新闻中能获得多种语义。

如图4，MWACNN也和SWANN一样对每个单词使用一个共享的全连接层，对每个边信息使用独立的全连接层。

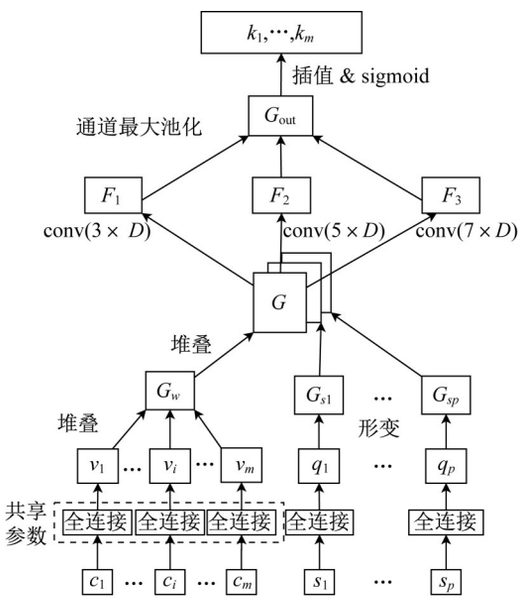

图4　MWACNN网络结构

Fig. 4　Network structure of MWACNN

$$\boldsymbol{v}_i = f_{\tanh}(\boldsymbol{U}_w \times \boldsymbol{c}_i + b_w) \tag{7}$$

$$\boldsymbol{q}_i = f_{\tanh}(\boldsymbol{U}_{s_i} \times \boldsymbol{s}_i + b_{s_i}) \tag{8}$$

式中：$\boldsymbol{U}_w$和b_w分别为单词映射的全连接层参数和偏

置；U_{s_i}和b_{s_i}分别为每个边信息的映射全连接层参数和偏置；c_i为N_f维向量，U_w为$d\times N_f$的参数矩阵，其中$d\ll N_f$。

$$G_w=f_{\text{tile}}(q_1,q_2,\cdots,q_m) \tag{9}$$

将所有v_i拼接得到大小为$m\times d$的图G_w，比起直接将c_i拼接成$m\times N_f$大小的图来说，G_w更小，因为v_i的维度比c_i小很多，对c_i的映射起到了压缩G_w的作用。

每个边信息特征q_i被形变成了同样$m\times d$大小的图G_{s_i}。p个边信息对应的图和单词对应的图按照通道维度堆叠，从而组成了$m\times d\times(p+1)$大小的图G。

$$G_{s_i}=f_{\text{reshape}}(q_i) \tag{10}$$

$$G=f_{\text{tile}}(G_w,G_{s_1},G_{s_2},\cdots,G_{s_p}) \tag{11}$$

在图G上使用CNN卷积可以将多种边信息和文本融合在一起。为了将词组长度的因素考虑在内，使用了3组不同大小的卷积核($3\times d,5\times d,7\times d$)，每种卷积核对应提取不同长度词组的特征。

$$\begin{aligned}F_1&=f_{\text{Conv}_{3\times d}}(G)\\F_2&=f_{\text{Conv}_{5\times d}}(G)\\F_3&=f_{\text{Conv}_{7\times d}}(G)\end{aligned} \tag{12}$$

通过3种尺度的卷积核得到F_1、F_2、F_3大小都是$m\times d$的特征图。同一个单词在3种尺度的卷积核中得到3个特征向量。例如，爱奇艺是当下热门的视频APP，“爱”这个单词在长度为3的卷积核下获得了词组“爱奇艺”的语义F_1，而在长度为5的卷积核下获得了“爱奇艺是当”的语义F_2，在长度为7的卷积核下获得了“爱奇艺是当下热”的语义F_3。由于F_1对应的“爱奇艺”是惯用词组，常出现于数据集中，在实验中表现出更大的激活值，因此对产出的特征图使用最大池化，这样能使得惯用词组对应的特征值主导语义，而屏蔽非惯用词对语义的影响。最大池化的公式见式(13)：

$$G_{\text{out}_{i,j}}=\max(F_{1\,i,j},F_{2\,i,j},F_{3\,i,j}) \tag{13}$$

在式(7)中对单词特征做了降维，导致G_{out}的大小是$m\times d$，小于$m\times N_f$，因此MWACNN在最后对特征图G_{out}做了双线性插值使其大小恢复到$m\times N_f$。为了限制边信息融合对原特征的改动量，增加了sigmoid函数。

$$G_{\text{out}}=f_{\text{bilinear\_interpolation}}(G_{\text{out}}) \tag{14}$$

$$G_{\text{out}}=f_{\text{sigmoid}}(G_{\text{out}}) \tag{15}$$

$$k_i=G_{\text{out}\,i} \tag{16}$$

MWACNN在式(7)中对特征做了降维，使得特征图缩小，后又在式(14)中通过双线性插值做特征升维，使其恢复成原来的特征维度。这样做的原因是为了控制过拟合，降维越多防止过拟合的效果越好。降维后的特征维度d可以根据模型的拟合情况灵活调整。最后G_{out}的大小是$m\times N_f$。将其分裂后得到m个大小为N_f的修正向量k_i，将它分别和原来的m个特征向量c_i做元素积，如式(3)。

最后整合SIACNN的整个流程，得到伪代码如下：

输入：新闻的文本矩阵W，新闻的边信息矩阵S

输出：结合边信息的文本特征矩阵f

变量说明：V为单词总量，D为单词表的特征维度，d为边信息维度，p为边信息个数，m为新闻文本单词个数，$W\in\mathbf{R}^{m\times V}$，$S\in\mathbf{R}^{p\times d}$，$f\in\mathbf{R}^{m\times N_f}$，$N_f$为特征维度，单词嵌入矩阵$W_E\in\mathbf{R}^{V\times D}$。

$e=W*W_E$

#$e\in\mathbf{R}^{m\times D}$，将文本矩阵$W$查单词嵌入矩阵后得到文本特征矩阵$e$

$v=$Fully_connect_layer(e)

#$v\in\mathbf{R}^{m\times H}$，$H$是全连接层输出的维度。

$q=$None

for $i\in[1,p]$do:

    $q_i=$Fully_connect_layer(S_i)

    $q=$concat(q,q_i)

#$q\in\mathbf{R}^{m\times H}$，边信息特征维度之和为$H$。

$r=$sum($q\odot$v)

#$\odot$表示矩阵的元素积，此处sum函数按照第2维度求和，得到$r\in\mathbf{R}^m$，r_i表示对第i个单词的关注度

$u=$softmax(r)

$t=$u$\odot e$

$c=$conv_2d(t,3)

#$t\in\mathbf{R}^{m\times D}$，将$t$视作图，在其上用2维卷积函数，设卷积核个数为$N_f$，则得到$c\in\mathbf{R}^{m\times N_f}$，3为卷积窗口大小

$G_w=$Fully_connect_layer(c)

#$G_w\in\mathbf{R}^{m\times d}$，$d$为全连接层输出的维度。

$G=G_w$

for $i\in[1,p]$do:

    $q_i=$Fully_connect_layer(S_i)

    $q_i=$reshape(q_i)

    #将q_i形变成$m\times d$的特征图

    $G=$tile(G,q_i)

#$G\in\mathbf{R}^{m\times d\times(p+1)}$

$F_1=$conv_2d(G,3)

$F_2=$conv_2d(G,5)

$F_3=$conv_2d(G,7)

$k=$sigmoid(bilinear_interpolation(max(F_1,F_2,F_3)))

#$k\in\mathbf{R}^{m\times N_f}$

$f=k\odot c$

3　实验

3.1　数据集和实验设定

使用MSN新闻收集的大规模新闻数据集MIND[29]来进行实验，它是少数大规模并且拥有丰富文本信息、类别信息、用户点击信息和用户曝光信息的数据集。按照时间排序，将前80%的数据作为训练集，10%作为验证集合，最后的10%作为测试集合。

在实验前，对数据集做了统计学分析。表1详细展现了数据集的基本信息。表1中所有的结果被取整。

表1　数据集基本信息

Tab. 1　Basic information of dataset

用户数/个	新闻数/条	会话数/次	新闻类别数/个	新闻子类别数/个	人均曝光新闻数/个
736 349	104 152	2 609 219	19	286	101
人均点击新闻数/个	标题平均单词数/个	摘要平均单词数/个	平均历史点击数/次	测试集新增用户数/个	测试集新增新闻占比/%
5	11	36	19	37 983	38

实验中，出于显存的考虑，单词嵌入的维度D被设定为200。使用了预训练的Glove嵌入矩阵[30]来初始化本实验的嵌入矩阵$\boldsymbol{W}_E$。根据表1中统计的标题平均单词数，将标题的最大长度设置为15。出于显存的考虑，将摘要的最大长度设置为30，所以每个新闻的文本信息一共使用45个单词。根据用户点击新闻数的分位数，用户点击历史的最大长度设置为20（超过80%分位数），当用户点击历史长度超过20的时候取最新的20次点击。训练时使用训练集中所有的正样本和下采样后的负样本，负样本与正样本数量的比例控制为5∶1。负样本是从用户曝光后未点击的新闻里随机采样，正样本是用户点击的新闻。据表1，类别和子类别的个数分别是19和286，对应的特征维度依照经验设置为10和20。在SWANN的实验中，$\boldsymbol{q}_i$的维度设置为100时，可以在过拟合与欠拟合之间平衡。类似在MWACNN实验中，$\boldsymbol{q}_i$的维度d被设置为12。使用主流的Adam[31]作为模型的优化器，训练批次大小按照惯例设置为256。实验中根据拟合情况将droupout rate[32]设置为0.3。

3.2　基准模型和评价标准

为了证明SIACNN的有效性，选择了几个推荐系统中通用和使用广泛的模型：LibFM（Factorization Machine Library）、Wide&Deep、DeepFM（Deep Factorization Machine）以及DIN（Deep Interest Network）。同时还选择了几个使用CNN的最先进的新闻推荐模型：LSTUR、NAML、NPA。

LibFM[33]：基于因子分解机进行推荐，被广泛使用。测试集中有37.76%的新闻没有出现在训练集中，为了提高LibFM的效果，需要通过手动构造文本特征结合到LibFM中。将文本的TF-IDF（Term Frequency - Inverse Document Frequency）特征作为LibFM的输入。

Wide & Deep[34]：由wide和deep两部分组成，wide部分是线性映射层，deep部分是多层感知机，特征同LibFM。

DeepFM[35]：由FM（Factorization Machine）和deep组成，FM部分是一个因子分解机，deep部分是多层感知机，特征同LibFM。

DIN[36]：使用注意力机制聚合用户点击的历史新闻。将候选新闻、历史新闻、两者的差、两者的元素积通过注意力机制生成不同历史点击的权重。特征仍然同LibFM。

LSTUR[8]：将类别和子类别的特征拼接在新闻的文本特征上，利用GRU（Gated Recurrent Unit）学习历史点击新闻序列的特征。

NPA[7]：将用户id用于生成单词级别和新闻级别的注意力机制。用注意力机制聚合每个新闻特征和历史点击序列特征。

NAML[9]：一个多视角注意力深度神经网络。

LSTUR、NAML和NPA中的CNN被置换为SIACNN，在后续的实验中分别记作LSTUR-SIACNN、NAML-SIACNN和NPA-SIACNN。为了更加公平地比较，实验中所有模型使用相同长度的标题和摘要，并确保在输入的信息量上是对等的。

在实验中，使用的验证指标是推荐系统点击率预估模型常用的验证指标，如AUC（Area Under Curve）、MRR（Mean reciprocal rank）、nDCG@5（Normalized Discounted cumulative gain@5）、nDCG@10（Normalized Discounted cumulative gain@10）。每个实验结果都是重复实验8次后的均值。

3.3　效果评估

实验结果如表2，测试集类型为：全部测试集、老

用户测试集、新用户测试集。老用户测试集表示测试集合中出现在训练集中的用户对应的样本集合。新用户测试集表示测试集合中未出现在训练集中的新用户对应的样本集合。根据对文本特征的处理方式，又将模型分成了3组，A组包括LibFM、Wide&Deep、DeepFM、DIN，A组的模型都利用TF-IDF手动构造的文本特征。B组包括NPA、LSTUR、NAML，是利用神经网络提取文本特征的推荐模型。C组在B组的基础上将边信息利用SIACNN融合到文本特征提取网络中。

表2 MIND数据集中多组实验结果

Tab. 2 Multiple experiment results of MIND dataset

测试集类型	指标	A组				B组			C组		
		LibFM	Wide&Deep	DeepFM	DIN	NPA	LSTUR	NAML	NPA−SIACNN	LSTUR−SIACNN	NAML−SIACNN
全部测试集	AUC	0.593	0.642	0.601	0.646	0.653	0.675	0.693	**0.672**	**0.687**	**0.697**
	MRR	0.308	0.330	0.319	0.347	0.354	0.374	0.391	**0.374**	**0.385**	**0.395**
	nDCG@5	0.272	0.284	0.271	0.299	0.301	0.321	0.336	**0.322**	**0.331**	**0.343**
	nDCG@10	0.334	0.352	0.339	0.366	0.370	0.389	0.403	**0.389**	**0.398**	**0.409**
老用户测试集	AUC	0.599	0.647	0.603	0.648	0.655	0.676	0.694	**0.675**	**0.688**	**0.698**
	MRR	0.308	0.328	0.312	0.346	0.352	0.373	0.394	**0.376**	**0.384**	**0.395**
	nDCG@5	0.271	0.283	0.270	0.293	0.295	0.315	0.334	**0.317**	**0.325**	**0.337**
	nDCG@10	0.333	0.351	0.335	0.361	0.363	0.383	0.402	**0.385**	**0.392**	**0.404**
新用户测试集	AUC	0.591	0.640	0.599	0.645	0.654	0.675	0.693	**0.671**	**0.687**	**0.696**
	MRR	0.309	0.333	0.323	0.347	0.354	0.375	0.395	**0.374**	**0.385**	**0.395**
	nDCG@5	0.272	0.284	0.273	0.300	0.303	0.322	0.342	**0.323**	**0.332**	**0.344**
	nDCG@10	0.335	0.353	0.340	0.367	0.372	0.390	0.409	**0.390**	**0.400**	**0.411**

从表2可以观察到几个现象。B组的表现普遍优于A组。在A组中，先构造TF-IDF特征，再利用模型学习和推荐，这是一个异步的过程。而在B组中，文本特征提取和推荐都交给神经网络来学习，是端到端的学习过程。这个现象表明，网络端到端的学习效果优于手动构造文本特征后推荐的异步学习。

第2个能观察到的是，C组基于B组替换了CNN为SIACNN后效果均有提升。NAML、LSTUR和NPA模型在使用了SIACNN后AUC指标分别提升了0.003 5、0.011 9、0.017 8，在MRR指标上分别提升0.004 6、0.011 0、0.020 5，在nDCG@5指标中分别提升了0.006 6、0.010 2、0.020 4，在nDCG@10指标上分别提升了0.005 8、0.009 6、0.018 7。NPA在使用SIACNN后提升最为显著，这主要因为NPA的网络结构中并没有利用好新闻类别信息，SIACNN的引入增加了类别的信息量。LSTUR和NAML引入SIACNN后提升了效果，验证了SIACNN能更好地将类别信息融合到推荐系统中，对文本的表征方式也更加利于优化推荐效果。在上述经典的几个使用CNN的新闻推荐系统中，SIACNN均产生了正向的效果，展现了它的通用性和扩展性。

第3个发现是，NAML表现比LSTUR好，LSTUR表现比NPA好。NAML在文本提取的最后利用注意力将边信息融合到文本特征中，而LSTUR仅仅将边信息拼接到文本特征后，NPA在提取文本特征时甚至没有利用到边信息。从这个角度看，文本和边信息的融合方式直接会影响推荐的效果。

最后，全部测试集合、老用户测试集和新用户测试集这3个集合中各模型的表现差异不大。老用户的AUC略高于新用户的AUC。这意味着模型已经能很好通过用户点击过的新闻理解用户的兴趣，并找到了新用户和老用户的行为模式共性。

3.4 消融实验

SIACNN由SWANN和MWACNN两部分组成。验证SWANN和MWACNN两者分别带来的贡献是必要的。将SWANN或者MWACNN分别移除，进行了多组实验，来验证两者各自的有效性，避免网络的冗余。

如图5，CNN版本标记为‘＋’，SWANN版本

标记为星号,MWACNN版本标记为三角形,SIACNN版本标记为圆形。图5中又分为4个子图,分别对应指标:AUC、MRR、nDCG@5、nDCG@10。每条曲线下的横轴标注了它对应的基线模型。可以观察到星号和三角形全都高于对应的'+'号,因此SWANN和MWACNN各自在CNN基础上提升了效果。说明2个部分都是对推荐起到正向效果的。另外所有的圆形都是曲线中的最高点,因此SIACNN比单独SWANN或者单独MWACNN表现更好,意味着SWANN和MWACNN的功能是不重合的。因此两者能共同促进模型的推荐效果,两者都是不可或缺的。

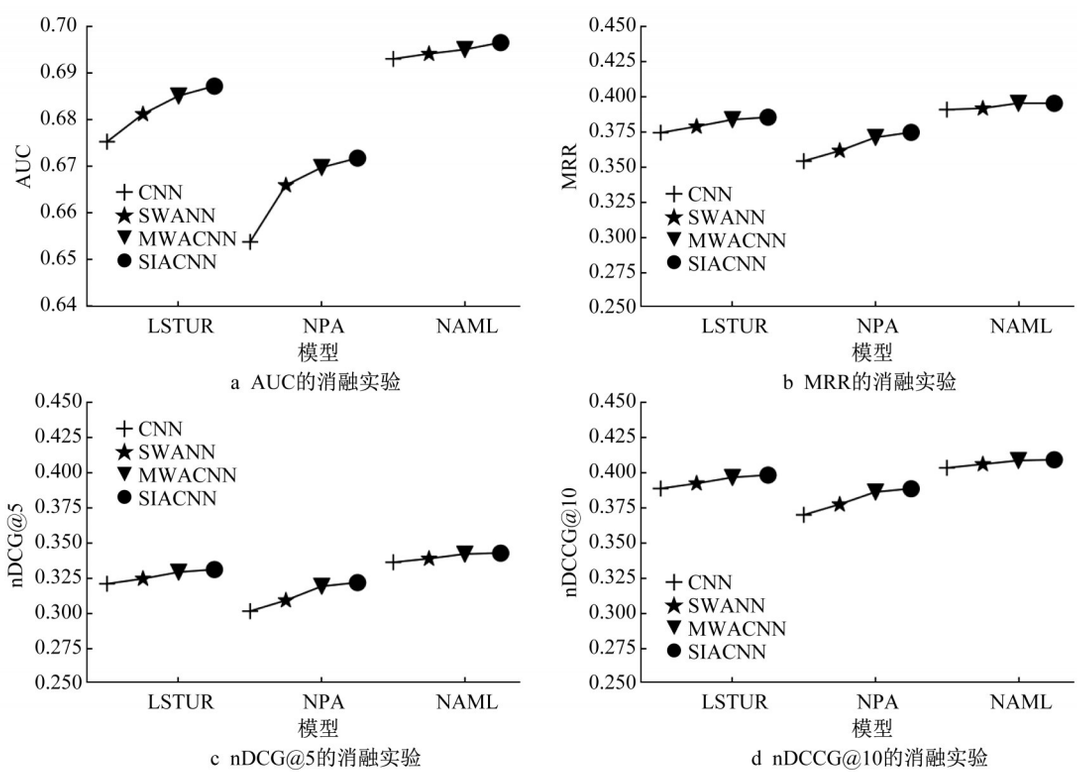

图5　SWANN和MWACNN的消融实验

Fig. 5　Ablation experiments of SWANN and MWACNN

3.5　超参数分析

探索在MWACNN中的一个重要的超参数。MWACNN选择多种不同大小的卷积核来捕获不同长度词组的特征。尝试了以下5组卷积核:1×1&3×3&5×5、3×3&5×5&7×7、5×5&7×7&9×9、1×1&5×5&9×9、3×3&5×5&7×7&9×9。分别实验,得到的结果如表3。

表3　多卷积核组合的实验结果

Tab. 3　Experimental results of convolution kernel combination

卷积核组合	AUC	MRR	nDCG@5	nDCG@10
1×1, 3×3, 5×5	0.686 9	**0.385 2**	0.330 8	0.398 0
3×3, 5×5, 7×7	**0.687 2**	0.385 1	**0.331 0**	**0.398 3**
5×5, 7×7, 9×9	0.684 2	0.381 4	0.327 8	0.396 2
1×1, 5×5, 9×9	0.683 1	0.381 2	0.328 0	0.396 1
3×3, 5×5, 7×7, 9×9	0.679 3	0.379 9	0.323 2	0.390 3

经过实验确定最佳的卷积核组合为3×3、5×5、7×7。本实验基于LSTUR-SIACNN模型。

3.6　案例分析

使用一些案例来可视化SIACNN的注意力权重。测试的案例为"an agriculture company has recalled all infected apples"。在未使用SWANN时,所有单词在卷积时都是一视同仁的。SWANN能学习出不同类别新闻对应单词的影响力,因此将该句子每个单词的权重u_i用柱状图表征出。如图6,当新闻类别是健康时,"infected"、"apples"和"recalled"的

权重最高；当新闻类别是财经时，“company”、“agriculture”和“infected”的权重最高。权重和类别之间的关系是符合预期的。在本案例中，SWANN利用类别这种边信息，使得网络对不同单词产生不同的关注度，从而提高了推荐的效果。

另一个案例是“beyond your dream”这个词组。“beyond your dream”在汽车类别下对应的含义是比亚迪的汽车品牌全称BYD，而在生活方式的类别下是指超越梦想，参照词组设定为“public transport”以及“work hard”。理想的情况下“beyond your dream”应该在汽车类别的新闻里接近“public transport”的语义，而在生活方式的类别下接近“work hard”的语义。用马尔可夫算法抽取最高频的2 000个词组，它们和“beyond your dream”、“public transport”以及“work hard”组成2 003个词组。将这些词组分别作为单独的句子送入模型中，每个句子利用SIACNN得到融合边信息的文本特征向量f。将2 003个f向量利用主成分分析(PCA)算法投射到二维平面内，如图7所示。

图7中，圆形符号表示“public transport”，三角形表示“beyond your dream”，星号表示“work hard”。图7a代表汽车类别，其中的“beyond your dream”更接近“public transport”。图7b代表生活方式类别，其中的“beyond your dream”更接近“work hard”。这也是符合MWACNN的预期的，利用边信息给词组更丰富的表达方式，这也有助于提升推荐效果。

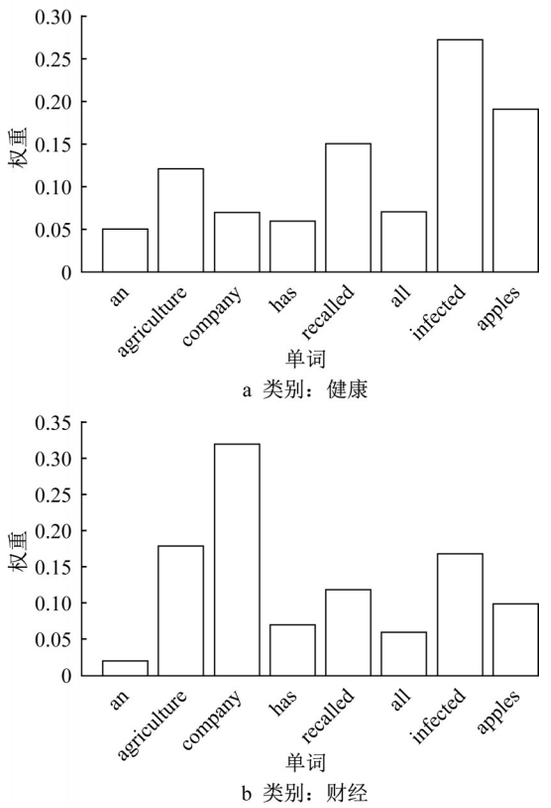

图6 单词注意权重可视化

Fig. 6 Visualization of weight of each word

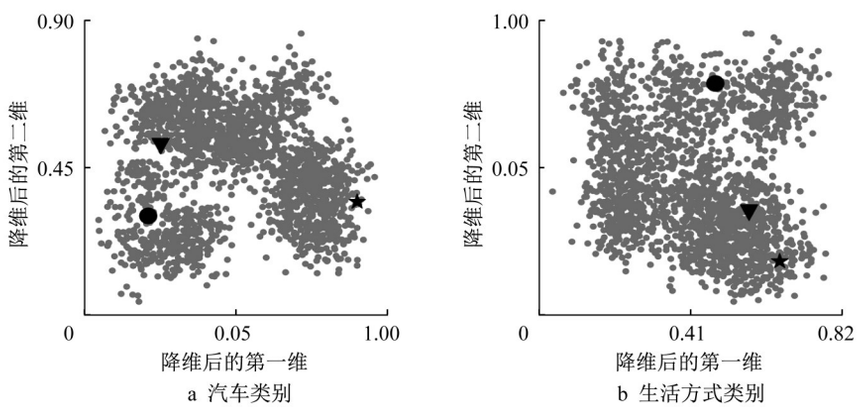

图7 词组特征向量主成分分析投影图

Fig. 7 PCA projection image of feature vectors of phrases

4 结语

设计和呈现了SIACNN的模型结构，它是一种适用于推荐系统的、能将边信息和文本特征提取相融合的网络，由SWANN和MWACNN组成，SWANN是赋予不同单词不同关注度的网络，MWACNN是根据边信息给予词组不同含义的网络。通过微软新闻采集的大型新闻数据集MIND进行大量实验，证明了SIACNN有效提升了推荐效果。通过消融实验证明了SIACNN子结构的有效性。将SIACNN代替3个经典的新闻推荐网络中的CNN均取得了很好的效果，证明了它的通用性和扩展性。

作者贡献声明:

卫　刚:论文撰写、深度神经网络设计。

邵　伟:论文撰写、深度神经网络设计与程序设计。

王志成:深度神经网络设计与数据分析。

参考文献:

[1] DAS A S, DATAR M, GARG A, *et al*. Google news personalization: Scalable online collaborative filtering [C]// Proceedings of the 16th international conference on World Wide Web. New York: Association for Computing Machinery, 2007: 271-280.

[2] LAVIE T, SELA M, OPPENHEIM I, *et al*. User attitudes towards news content personalization [J]. International Journal of Human-Computer Studies, 2010, 68(8): 483.

[3] MORALES G D F, GIONIS A, LUCCHESE C. From chatter to headlines: Harnessing the real-time web for personalized news recommendation [C]//Proceedings of the fifth ACM international conference on Web search and data mining. New York: Association for Computing Machinery, 2012: 153-162.

[4] BANSAL T, DAS M, BHATTACHARYYA C. Content driven user profiling for comment-worthy recommendations of news and blog articles [C]//Proceedings of the 9th ACM Conference on Recommender Systems. New York: Association for Computing Machinery, 2015: 195-202.

[5] LI L, CHU W, LANGFORD J, *et al*. A contextual-bandit approach to personalized news article recommendation [C]// Proceedings of the 19th International Conference on World Wide Web. New York: Association for Computing Machinery, 2010: 661-670.

[6] WU C, WU F, AN M, *et al*. Neural news recommendation with topic-aware news representation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 1154-1159.

[7] WU C, WU F, AN M, *et al*. Npa: Neural news recommendation with personalized attention [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery, 2019: 2576-2584.

[8] AN M, WU F, WU C, *et al*. Neural news recommendation with long-and short-term user representations [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 336-345.

[9] WU C, WU F, AN M, *et al*. Neural news recommendation with attentive multi-view learning [J]. arXiv preprint, 2019: 1907.05576.

[10] CHUANG Y N, CHEN C M, WANG C J, *et al*. TPR: Text-aware preference ranking for recommender systems [C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: Association for Computing Machinery, 2020: 215-224.

[11] GE S, WU C, WU F, *et al*. Graph enhanced representation learning for news recommendation [C]//Proceedings of The Web Conference 2020. New York: Association for Computing Machinery, 2020: 2863-2869.

[12] HU L, LI C, SHI C, *et al*. Graph neural news recommendation with long-term and short-term interest modeling [J]. Information Processing & Management, 2020, 57(2): 102142.

[13] LIU J, DOLAN P, PEDERSEN E R. Personalized news recommendation based on click behavior [C]// International Conference on Intelligent User Interfaces. New York: Association for Computing Machinery, 2010: 31-40.

[14] WANG Y, SHANG W. Personalized news recommendation based on consumers' click behavior [C]//2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Zhangjiajie: IEEE, 2015: 634-638.

[15] LU Z, DOU Z, LIAN J, *et al*. Content-based collaborative filtering for news topic recommendation [C]//In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Menlo Park: Association for the Advancement of Artificial Intelligence, 2015: 217-223.

[16] LEI L A, LI Z A, FAN Y B, *et al*. Modeling and broadening temporal user interest in personalized news recommendation [J]. Expert Systems with Applications, 2014, 41(7): 3168.

[17] PHELAN O, MCCARTHY K, SMYTH B. Using twitter to recommend real-time topical news [C]//Proceedings of the third ACM conference on Recommender Systems. New York: Association for Computing Machinery, 2009: 385-388.

[18] SON J W, KIM A Y, PARK S B. A location-based news article recommendation with explicit localized semantic analysis [C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2013: 293-302.

[19] ZHANG Y, WALLACE B C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing. Taipei: Asian Federation of Natural Language Processing, 2017: 253-263.

[20] OKURA S, TAGAMI Y, ONO S, *et al*. Embedding-based news recommendation for millions of users [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2017: 1933-1942.

[21] ZHANG F, YUAN N J, LIAN D, *et al*. Collaborative knowledge base embedding for recommender systems [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York: Association for Computing Machinery, 2016: 353-362.

[22] WANG H, ZHANG F, XIE X, *et al*. DKN: Deep knowledge-aware network for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference. Switzerland: International World Wide Web Conferences Steering Committee, 2018: 1835-1844.

[23] WANG H, WU F, LIU Z, *et al*. Fine-grained interest matching for neural news recommendation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 836-845.

[24] LEE D, OH B, SEO S, *et al*. News recommendation with topic-enriched knowledge graphs[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: Association for Computing Machinery, 2020: 695-704.

[25] WU C, WU F, HUANG Y, *et al*. Neural news recommendation with negative feedback[J]. CCF Transactions on Pervasive Computing and Interaction, 2020, 2(3): 178.

[26] WU C, WU F, YU Y, *et al*. NewsBERT: Distilling pre-trained language model for intelligent news application [J]. arXiv preprint, 2021: 2102.04887.

[27] ZHU Q, ZHOU X, SONG Z, *et al*. Dan: Deep attention neural network for news recommendation[C]// In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. Stroudsburg: Association for the Advancement of Artificial Intelligence, 2019, 33(1): 5973-5980.

[28] LIU H, LU J, YANG H, *et al*. Category-specific CNN for visual-aware CTR prediction at JD. com[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery, 2020: 2686-2696.

[29] WU F, QIAO Y, CHEN J H, *et al*. Mind: A large-scale dataset for news recommendation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Stroudsburg: Association for Computational Linguistics, 2020: 3597-3606.

[30] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[31] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint, 2014:1412.6980.

[32] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, *et al*. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929.

[33] RENDLE S. Factorization machines with LibFM [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(3): 1.

[34] CHENG H T, KOC L, HARMSEN J, *et al*. Wide & deep learning for recommender systems[C]//Proceedings of the 1st Workshop on Deep learning for Recommender Systems. New York: Association for Computing Machinery, 2016: 7-10.

[35] GUO H, TANG R, YE Y, *et al*. DeepFM: a factorization-machine based neural network for CTR prediction [J]. arXiv preprint, 2017: 1703.04247.

[36] ZHOU G, ZHU X, SONG C, *et al*. Deep interest network for click-through rate prediction [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery, 2018: 1059-1068.