

基于会话的双层注意力机制新闻推荐方法

王海艳^{1,2,3}, 胡 阳¹

(1.南京邮电大学 计算机学院,江苏 南京 210023
2.南京邮电大学 江苏省无线传感网高技术研究重点实验室,江苏 南京 210023
3.南京邮电大学 江苏省大数据安全与智能处理重点实验室,江苏 南京 210023)

摘要:随着新闻信息的爆炸式增长,个性化新闻推荐对用户快速找到感兴趣的新闻变得非常重要。基于会话的推荐旨在根据用户的行为序列向用户推荐下一个可能感兴趣的项目。但是,现有方法大都忽略了会话内用户阅读行为的随机性和偶然性,难以捕捉用户的主要兴趣,或者将用户会话视为单个序列,忽视了用户兴趣在不同会话之间的演变和关联。文中提出了一种基于会话的双层注意力机制新闻推荐方法,该方法将用户行为序列划分成多个会话:首先使用卷积神经网络对用户点击的新闻特征进行提取;其次在兴趣感知注意力层结合门控循环单元和自注意力机制获取用户在每个会话内的主要兴趣;最后,在会话感知注意力层使用注意力机制建模当前会话和历史会话之间用户兴趣的关联程度以形成最终的用户兴趣表示。通过在真实世界的数据集上与基准方法进行对比实验,结果表明了所提方法在新闻推荐上的有效性。

关键词:新闻推荐;基于会话的推荐;注意力机制;神经网络

中图分类号:TP311 **文献标志码:**A **文章编号:**1673-5439(2021)01-0099-08

News recommendation method for session-based neural neural network with two-layer attention mechanism

WANG Haiyan^{1,2,3}, HU Yang¹

(1.School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
2.Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
3.Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: With the explosive growth of news information, personalized news recommendations have become very important for users to specially push news they are really interested in. The task of session-based recommendations is to recommend the next item that may be of interest to the user based on the user history behavior sequence. However, most of the existing methods ignore the randomness and the contingency of the user reading in a session, thus it is difficult to capture the user main interest. Some other methods treat the user sessions as a single sequence, ignoring the evolution and the association of user interests between different sessions. A news recommendation method for session-based neural network with two-layer attention mechanism is proposed, by dividing the user behavior sequence into multiple sessions. Firstly, it uses a convolutional neural network to extract the news features clicked by the user. Then, the gate recurrent unit and the self-attention mechanism are combined to learn the user main interest in the

interest-aware attention layer. Finally, the attention mechanism is used in the session-aware attention layer to model the correlation of user interests between tcurrent sessions and historical sessions to form the final representation of user interests. The simulation experiment is conducted on real-world news datasets. Experimental results demonstrate the effectiveness of the proposed method for news recommendations.

Keywords: news recommendations; session-based recommendation; attention mechanism; neural network

在线新闻平台(如头条新闻和 Google 新闻)提供的新闻数量众多容易导致信息过载。为了减轻信息过载问题,针对用户的阅读兴趣向用户推荐感兴趣的新闻是很有必要的^[1-4]。与其他项目(电影、音乐和书籍)不同,新闻通常具有高度的时间敏感性,过时的新闻经常被新发布的新闻所取代^[3]。此外,用户的兴趣变化通常不如其他领域稳定,用户的某些兴趣可能持续很长时间,也可以由特定的上下文或者临时的需求触发。而传统的个性化推荐方法(如基于协同过滤的方法)由于忽略了用户浏览行为的顺序信息,难以捕捉用户不断变化的兴趣。

因此,有研究提出序列化新闻推荐方法,这些方法通常利用循环神经网络^[2,5-6](Recurrent Neural Network, RNN)和注意力机制^[4,7-8]对用户的历史交互行为进行建模,以捕获用户的顺序阅读模式,如文献[5]将用户浏览历史中的会话信息作为循环神经网络中的输入序列生成用户表示,文献[6]通过门控循环单元(Gate Recurrent Unit, GRU)捕获用户行为序列中的顺序信息,从而学习用户的长期兴趣,并将 GRU 网络的最后一个隐藏状态作为用户的短期兴趣,最终结合长短期兴趣进行推荐。文献[7]利用了注意力机制对用户近期交互的顺序行为进行建模从而学习当前的用户兴趣。然而,这类方法仍然面临一些问题:

(1) 忽略会话内用户行为具有随机性和偶然性,难以捕捉会话内用户的主要阅读兴趣。

(2) 将用户的行为视为单个序列,忽略各个会话之间用户兴趣的相关性。

为了解决上述问题,本文提出了基于会话的双层注意力机制新闻推荐模型(Session-based Neural Network with two-layer Attention Mechanism for News Recommendation, TLAM-NR)。在 TLAM-NR 模型中,用户阅读行为被划分成多个会话,并通过兴趣感知注意力层细粒度的学习用户在各个会话内的主要阅读兴趣。此外,在会话感知注意力层通过对当前会话和历史会话之间用户兴趣的相关程度进行建模,从而学习最终的用户表示,实现新闻推荐。本文

的主要工作如下:

(1) 本文提出了一种双层注意力机制新闻推荐模型(TLAM-NR),该模型结合了会话内和会话间的用户兴趣。

(2) 在 TLAM-NR 模型中,首先使用卷积神经网络完成对新闻特征的提取;其次,在兴趣感知注意力层利用 GRU 网络 and 自注意力机制提取每个会话内的用户兴趣;最后,会话感知注意力层通过注意力机制来建模当前会话和历史会话之间用户兴趣的相关程度,并融合各个会话中的用户兴趣完成对候选项的预测从而实现新闻推荐。

(3) 本文在公共的数据集上进行了实验,与现有的方法对比验证了所提模型的有效性,同时分析了影响 TLAM-NR 模型推荐性能的参数。

1 相关工作

1.1 传统新闻推荐方法

传统的方法包括基于协同过滤^[9-11]的方法,基于内容的方法^[12-13]以及混合方法^[14]。基于协同过滤的方法利用用户过去对项目的评分或者与给定用户相似的用户来预测用户的偏好。如文献[11]利用了用户的显式评分和隐式反馈来进行推荐。但是基于协同过滤的方法通常会遇到冷启动问题,因为新闻项经常被替换。基于内容的方法可以缓解冷启动问题,它通过对用户浏览新闻的内容进行分析从而向用户推荐类似的新闻。文献[13]是一个基于内容的深度神经网络模型,它通过余弦相似度计算文档之间的相关性从而对一组文档进行排名。混合方法通常结合不同的推荐方法进行推荐,文献[14]提出了一个两阶段的推荐框架,将基于协同过滤的方法和基于内容的方法相结合。然而,这些方法忽略了用户浏览历史中的序列信息,难以学习用户不断变化的兴趣。

1.2 序列化新闻推荐方法

会话的概念通常会在序列化推荐中提及,序列由会话组成。最近很多研究基于循环神经网络建模。文献[5]使用基于张量的卷积神经网络来提取会话内的新闻特征表示,并利用 RNN 捕

获用户点击新闻的顺序信息。文献[6]提出通过 GRU 捕获用户行为序列中的顺序信息,从而学习用户的长期兴趣,并将 GRU 网络的最后一个隐藏状态作为用户的短期兴趣。随着注意力机制在各种机器学习任务(如图像/视频处理^[15]和自然语言处理^[16])中成功应用,近年来,注意力机制也经常被用于推荐任务中。文献[17]首次利用注意力机制来获取当前会话内的用户兴趣。文献[7]通过设计了一种基于注意力机制的 RNN 来捕获用户点击新闻的顺序信息。自注意力机制最早由 Google 提出并用于机器翻译^[18]。自注意力是注意力机制的一种特例,它只需要一个序列即可计算其表示形式。与传统的注意力机制不同,自注意力机制侧重于句子中多个单词

之间的交互学习^[19],它可以捕捉单词之间的长距离依赖关系。文献[8]提出了一种考虑时间因素的自注意力机制从用户最近的浏览序列中学习短期兴趣,并将其与用户的长期兴趣结合进行新闻推荐。文献[20]提出了一种顺序知识感知推荐模型,它使用自注意力机制来发现用户交互序列中的顺序模式。

2 TLAM-NR 模型

本文提出了基于会话的双层注意力机制新闻推荐模型 TLAM-NR,如图 1 所示。该模型由两个重要部分组成:(1) 结合 GRU 网络和自注意力机制的兴趣感知注意力层。(2) 基于注意力机制的会话感知注意力层。

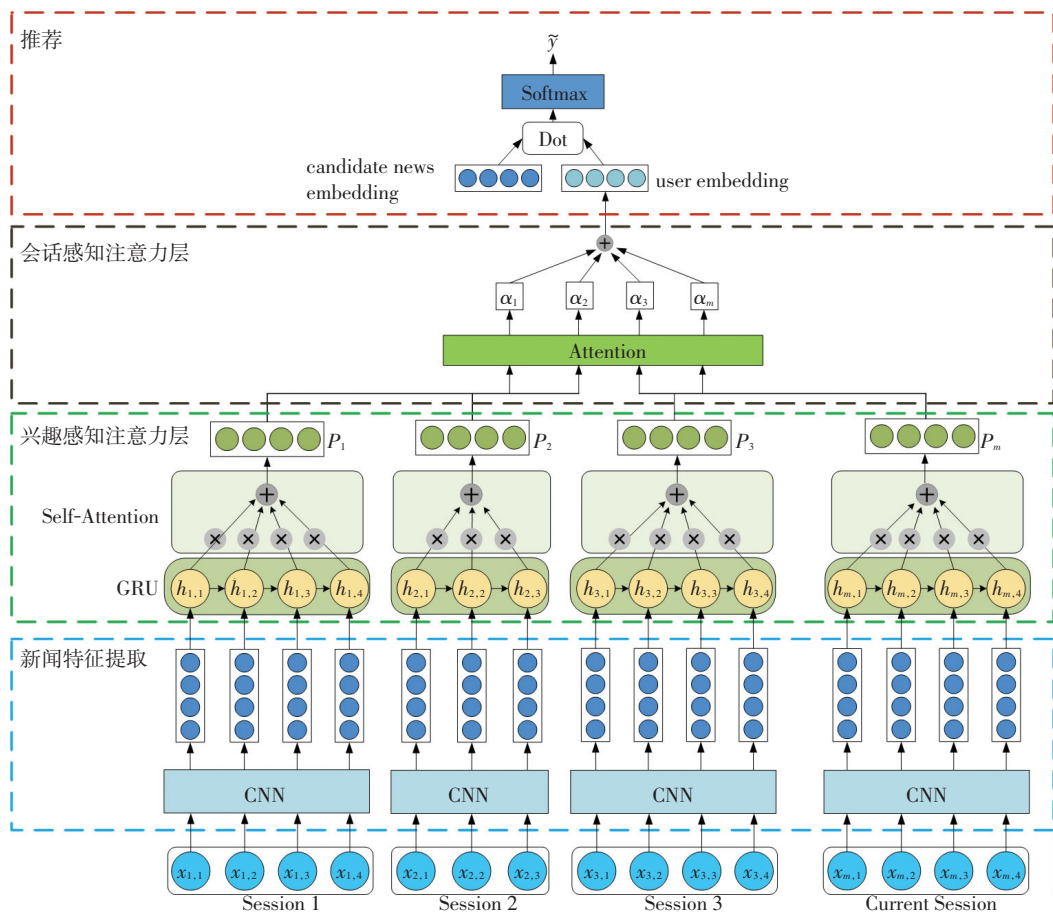


图 1 TLAM-NR 模型

2.1 问题描述

用户集为 $U = \{u_1, u_2, \dots, u_{|U|}\}$, 新闻集 $X = \{x_1, x_2, \dots, x_{|X|}\}$, 每个用户 u 对应其会话交互集表示为 $S^u = \{s_1^u, s_2^u, \dots, s_m^u\}$, 其中历史会话为 $S_h^u = \{s_1^u, s_2^u, \dots, s_{m-1}^u\}$, 当前会话为 s_m^u 。用户 u 对应每

个会话中的交互序列表示为 $S_j^u = \{x_{j,1}^u, x_{j,2}^u, \dots, x_{j,t}^u\}$ 。给定一个用户 u 的所有会话, 推荐模型旨在预测用户 u 在当前会话 s_m^u 下是否对其未见过的新闻有潜在的兴趣。表 1 定义了本文使用的主要符号。

表 1 相关符号定义

符号	定义
U	用户集
X	新闻集
S_u	用户 u 的所有会话集
$W_{1:n}$	新闻标题原始输入
$\mathbf{v}$	新闻特征表示向量
S_j^u	用户 u 对应的第 j 个会话
$h_{j,k}$	第 j 个会话中的第 k 个新闻项的隐藏状态
$\alpha_{j,k}$	第 j 个会话中的第 k 个新闻项的注意力值
$\mathbf{p}_j$	第 j 个会话用户主要兴趣特征表示向量
P	所有会话的兴趣集
α_j	第 j 个会话的注意力值

2.2 新闻特征提取

受卷积神经网络成功应用在计算机视觉领域的启发^[21],研究人员提出了许多基于卷积神经网络的语义嵌入模型^[22-23]并在 NLP 领域得到了广泛使用。在本文中,使用卷积神经网络作为新闻特征提取器,将新闻的标题作为输入,学习每条新闻的特征表示。

令 $\mathbf{W}_{1:n}$ 是长度为 n 的新闻标题的原始输入, $\mathbf{W}_{1:n} = [w_1 w_2 \cdots w_n] \in \mathbf{R}^{d \times n}$ 为输入标题的词嵌入矩阵,其中 $w_i \in \mathbf{R}^{d \times 1}$ 为第 i 个单词在标题中的嵌入, d 为单词嵌入的维数。然后将卷积核应用于单词嵌入矩阵中进行卷积运算,其中 $l(l \leq n)$ 为卷积核的窗口大小。具体来说,新闻的一个特征表示向量 $\mathbf{c}_i$ 由子矩阵 $\mathbf{W}_{i:i+l-1}$ 产生。

$$\mathbf{c}_i = f(\mathbf{h} \odot \mathbf{W}_{i:i+l-1} + \mathbf{b}) \tag{1}$$

其中, f 为非线性函数, $\odot$ 为卷积操作, $\mathbf{b}$ 为偏置向量。将卷积应用于词嵌入矩阵中的每个可能位置后得到特征图。

$$\mathbf{c} = [c_1, c_2, \cdots, c_{n-l+1}] \tag{2}$$

然后在特征图上使用最大池化操作来提取标题中最重要的特征。

$$\tilde{\mathbf{v}} = \max \{ \mathbf{c} \} = \max \{ c_1, c_2, \cdots, c_{n-l+1} \} \tag{3}$$

最后将这些特征进行拼接形成新闻特征表示向量 $\mathbf{v}$ 。

2.3 兴趣感知注意力层

兴趣感知注意力层根据用户在会话内点击的新闻记录的顺序信息并结合自注意力机制来学习用户在该会话内的主要阅读兴趣。由于用户点击新闻的顺序信息可以很好地反映一段时间内用户兴趣的变化和多样性,因此,使用 GRU 网络对用户顺序行为进行建模。给定一个会话中用户点击的新闻序列 $S_j = \{x_{j,1}, x_{j,2}, \cdots, x_{j,t}\}$, 则它们的嵌入表示为 $\{v_{j,1},$

$v_{j,2}, \cdots, v_{j,t}\}$ 。将新闻特征嵌入表示向量作为 GRU 的输入,在每个时间步长它都会生成一个隐藏状态 $h_{j,k}$, 它由先前的隐藏状态 $h_{j,k-1}$ 和候选状态 $\tilde{h}_{j,k}$ 线性组成,其中 $1 \leq k \leq t$ 。使用每个隐藏状态作为用户的兴趣表示

$$h_{j,k} = (1 - z_{j,k}) \cdot h_{j,k-1} + z_{j,k} \cdot \tilde{h}_{j,k} \tag{4}$$

更新门为 $z_{j,k}$

$$z_{j,k} = \sigma(W_z v_{j,k} + U_z h_{j,k-1}) \tag{5}$$

其中, $W_z, U_z \in \mathbf{R}^{d \times d}$ 为更新参数, $\sigma(\cdot)$ 为 sigmoid 函数。候选隐藏状态 $\tilde{h}_{j,k}$ 的计算公式为

$$\tilde{h}_{j,k} = \tanh(W_{\tilde{h}} v_{j,k} + U_{\tilde{h}}(r_{j,k} \odot h_{j,k-1})) \tag{6}$$

重置门的隐藏状态 $r_{j,k}$ 为

$$r_{j,k} = \sigma(W_r v_{j,k} + U_r h_{j,k-1}) \tag{7}$$

其中, $W_r, U_r \in \mathbf{R}^{d \times d}$ 为重置参数。

在获得了用户每个时间步长的隐藏状态后,使用自注意力机制去捕捉序列中用户点击项之间的长距离依赖关系。自注意力是注意力机制的一种特例,它以序列本身作为查询,键和值向量,通过单个序列与其自身匹配来捕获序列内部用户对于不同新闻项的注意力。将 GRU 输出的隐藏特征 $\mathbf{H}_j = \{h_{j,1}, h_{j,2}, \cdots, h_{j,t}\}$ 通过非线性变换投影到相同的特征空间

$$\mathbf{Q}_j = \text{ReLU}(W_Q^j \mathbf{H}_j) \tag{8}$$

$$\mathbf{K}_j = \text{ReLU}(W_K^j \mathbf{H}_j) \tag{9}$$

$$\mathbf{V}_j = \text{ReLU}(W_V^j \mathbf{H}_j) \tag{10}$$

其中, $W_Q^j, W_K^j, W_V^j \in \mathbf{R}^{d \times d}$ 为投影矩阵。ReLU 为激活函数。因此,注意力矩阵计算为

$$\alpha_{j,j} = \text{softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_j^T}{\sqrt{d}}\right) \tag{11}$$

其中, $\alpha_{j,j} \in \mathbf{R}^{d \times d}$ 。比例因子 $\sqrt{d}$ 是为了避免点积过大并加快收敛。因此用户在第 j 个会话中的主要兴趣特征 $\mathbf{p}_j$ 可以表示为:

$$\mathbf{p}_j = \sum_{k=1}^t \alpha_{j,k} v_{j,k} \tag{12}$$

最终通过兴趣感知注意力层能够得到各个会话下的用户兴趣集 $P = \{p_1, p_2, \cdots, p_m\}$ 。

2.4 会话感知注意力层

由于会话内的行为有限,缺乏用户的长期行为交互,不足以进行准确的预测。考虑到用户不同时期的会话体现了用户的阅读兴趣,并且会话之间并非彼此绝对独立,因此使用注意力机制来建模当前会话和历史会话之间用户兴趣的关联程度。

将兴趣感知注意层提取的各个会话下用户主要兴趣 $P = \{p_1, p_2, \dots, p_m\}$ 作为会话感知注意力层的输入。因此各个会话用户兴趣的注意力权重为

$$\alpha_j = \frac{\exp(f(p_m, p_j))}{\sum_{j=1}^m \exp(f(p_m, p_j))} \quad (13)$$

式中

$$f(p_m, p_j) = W_0 \tanh(W_f p_m + U_f p_j) \quad (14)$$

其中, $W_0 \in \mathbf{R}^d$, $W_f \in \mathbf{R}^{d \times d}$, $U_f \in \mathbf{R}^{d \times d}$ 。将各个会话下的用户兴趣融合生成最终的用户兴趣表示

$$u_p = \sum_{j=1}^m \alpha_j p_j \quad (15)$$

2.5 推荐

设 v_c 为候选新闻嵌入表示,在获取用户 u 的嵌入表示 u_p 后,通过点积计算候选项的得分

$$\tilde{z}_c = v_c^T u_p \quad (16)$$

最后通过函数 softmax 来获取模型的预测值 $\tilde{y}$

$$\tilde{y} = \text{softmax}(\tilde{z}_c) \quad (17)$$

其中, $\tilde{y}$ 表示用户 u 点击候选新闻项的概率。

为了训练提出的模型,从现有观察到的点击阅读历史中选择正样本,从未观察到的阅读中选择负样本。训练样本表示为 $X = (s^u, x, y)$, 其中 x 是预测是否点击的候选新闻项。输入正样本时, y 值为 1, 负样本 y 值为 0。使用交叉熵损失作为损失函数, 则有

$$L = - \sum_{x=1}^X y \log(\tilde{y}) + (1 - y) \log(1 - \tilde{y}) \quad (18)$$

3 实验结果与分析

为了验证 TLAM-NR 模型的有效性和准确性, 本节首先介绍整个实验的实验准备, 然后展示实验结果并对其进行分析。

3.1 实验准备

主要对数据集、实验环境与具体设置、基准方法与评价指标这 3 个方面进行叙述。

(1) 数据集

为了验证提出模型的有效性, 选取了真实世界的数据集 Adressa^[24] 进行实验。使用两种版本的数据集 Adressa-1week 和 Adressa-1month, 分别对应一周和一个月的新闻点击日志。选择具有订阅服务的非匿名用户, 移除了少于 3 个交互的会话, 并且保留了 4 个以上会话以提供足够的历史会话信息。对于每个用户, 80% 的会话划分在训练集中, 剩下的用于

构建测试集。将每个用户的会话集分成一系列的序列和标签。例如, 对于会话集 $S_u = \{s_1^u, s_2^u, \dots, s_m^u\}$, 生成历史会话 $S_h = \{s_1^u, s_2^u, \dots, s_{m-1}^u\}$ 和当前会话以及标签 $s_m = [x_{c,1}^u, x_{c,2}^u, x_{c,3}^u, \text{label}; x_{c,4}^u]$ 。统计数据如表 2 所示。

表 2 Adressa 数据集

数据集	Adressa-1week	Adressa-1month
# of users	5 503	20 907
# of news	2 820	12 057
# of events	170 517	1 075 133
# of sessions	37 132	245 485
# of train session	31 683	22 4621
# of test session	5 449	20 864

(2) 实验环境和具体设置

实验环境是 CPU 为 Intel (R) Xeon (R) silver 4210 @ 2.20 GHz, 内存 64 GB, 系统为 Ubuntu16.04LTS。为了实现 TLAM-NR, 主要使用 python 及其深度学习框架 TensorFlow 与科学运算库 Numpy 实现。使用 Adam 方法来优化损失函数, 设置卷积核窗口大小为 $\{2, 3, 4\}$, GRU 隐藏单元设置为 100, 隐藏层设置为 1, 学习率设置为 0.001, 批处理大小设置为 256。

(3) 基准方法与评估指标

在 Adressa 两个版本的数据集上与 4 个基准方法进行了对比:

DeepJoNN^[5] 使用基于张量的卷积神经网络提取基于会话的新闻特征表示, 并使用 RNN 来捕获点击和相关特征流中的顺序模式。

LSTUR^[6] 利用门控循环单元 (GRU) 捕获用户行为序列中的顺序信息, 从而学习用户的长期兴趣, 并将 GRU 网络的最后一个隐藏状态作为用户的短期兴趣, 最后结合长短期兴趣形成统一的用户表示。

DAN^[7] 设计了一种基于注意力机制的 RNN 来捕获用户近期点击新闻的顺序信息, 实现新闻推荐。

TANN^[8] 使用自注意力机制来发现用户交互序列中的顺序模式, 以学习更准确的用户短期兴趣, 并将通过潜在因子模型获得的长期兴趣和短期兴趣拼接进行预测。

为了评估推荐效果, 采用了 3 种广泛采用的指标:

① MRR@K。平均倒数排名, 即被正确推荐项目的倒数排名的平均值, 当排名位置超过 K 时, 倒数排名设置为 0。MRR 评估指标考虑了推荐排名的顺序, 其中 MRR 值大的表示正确推荐的项目位

于排名列表的顶部,表达形式为

$$\text{MRR@K} = \frac{1}{K} \sum_i^K \frac{1}{\text{rank}(i)} \tag{19}$$

② nDCG@K。归一化折损累计增益,其表达形式为

$$\text{nDCG@K} = \frac{\text{DCG}}{\text{IDCG}}$$
$$\text{DCG} = \sum_{i=1}^K \frac{\text{rel}_i}{\log_2(i+1)} \tag{20}$$

在计算折扣累积收益(DCG)的式(20)中,rel_i表示第*i*个推荐的相关程度。当rel_i=1时表明该项目实际被选中,否则rel_i=0。IDCG是归一化的DCG,即根据理想排序,获得的最大DCG值。

③ AUC。ROC曲线下面积,表示从样本中取正样本的预测值大于负样本的概率,定义为

$$\text{AUC} = \frac{S - M(M+1)/2}{M \times N} \tag{21}$$

其中,*M*,*N*分别代表正样本和负样本的个数,*S*为所有样本的排名。

3.2 实验结果与分析

(1) 推荐模型结果比较

将TLAM-NR模型与基准方法进行比较,表3和表4分别显示了在Adressa-1week和Adressa-1month两个版本数据集上各个方法的结果。

表 3 在 Adressa-1week 数据集上不同方法的性能比较

方法	MRR@5	MRR@20	nDCG@5	nDCG@20	AUC
DeepJoNN	0.103 1	0.141 7	0.392 1	0.493 5	0.672 2
DAN	0.112 3	0.153 2	0.423 3	0.524 8	0.702 5
LSTUR	0.119 6	0.161 5	0.445 6	0.541 7	0.723 4
TANN	0.128 5	0.167 8	0.469 3	0.561 8	0.741 1
TLMA-NR	0.136 2	0.178 6	0.486 2	0.584 2	0.758 3

表 4 在 Adressa-1month 数据集上不同方法的性能比较

方法	MRR@5	MRR@20	nDCG@5	nDCG@20	AUC
DeepJoNN	0.091 4	0.130 6	0.358 6	0.461 3	0.597 4
DAN	0.108 2	0.145 1	0.438 9	0.516 2	0.642 5
LSTUR	0.113 6	0.151 3	0.440 3	0.521 6	0.667 4
TANN	0.116 4	0.156 2	0.443 7	0.534 9	0.681 3
TLAM-NR	0.123 3	0.165 9	0.454 2	0.549 1	0.706 9

从表3和4可以看出,TLAM-NR的指标在所有基准方法中总体性能表现最佳,尤其是MRR@5和nDCG@20分别提高了6.43%和3.98%,这验证了TLAM-NR模型的有效性。这是因为TLAM-NR模型不仅利用门控循环单元捕捉用户的顺序信息,而且结合了自注意力机制更好地学习会话内用户的主要兴趣。此外,相较于其他基准方法,所提的模型不

仅考虑了会话内的用户交互行为,而且还考虑了会话之间用户兴趣关联性,从而进一步提升推荐的性能。

(2) 模型变体分析

对模型的3种变体进行了比较:(1)变体TLAM-NR-H不考虑使用自注意力机制学习各个会话内用户的主要兴趣。(2)变体TLAM-NR-S不考虑使用注意力机制学习用户兴趣在当前会话与历史会话之间的相关性。(3)变体TLAM-NR则同时考虑了各个会话内用户的主要兴趣以及用户兴趣在会话之间的关系。图2和图3分别显示在两个版本数据集上的实验结果。

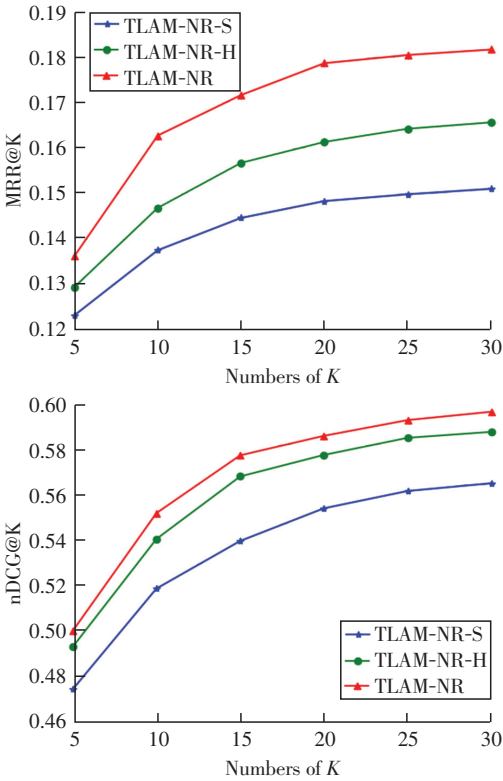


图 2 Adressa-1week

从图2和图3可以看出,当不使用自注意力机制学习会话内的用户兴趣时模型性能会下降明显,这验证了兴趣感知注意力层设计的有效性,这是因为自注意力机制能够学习会话内各个点击新闻项所占的不同权重,可以有效捕捉单个会话下用户的主要兴趣。此外,当移除会话感知注意力层时,实验结果稍差,这说明对不同会话之间的用户兴趣相关性进行建模从而学习最终的用户表示是合理的。

(3) 模型参数分析

主要分析新闻特征嵌入维度和会话数量对于推荐模型性能的影响。

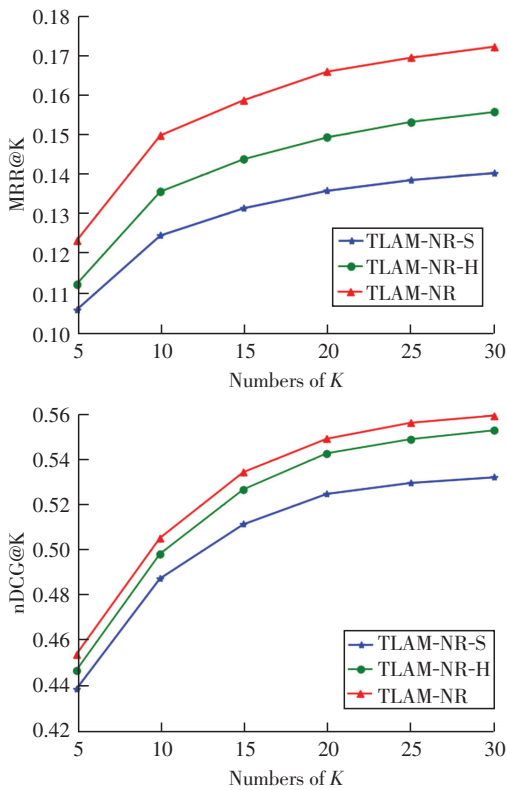


图 3 Adressa-1month

特征维度的影响。分别设置新闻特征嵌入向量的维度 d 为 $\{50, 100, 150, 200\}$ 。观察特征维度对指标 $MRR@20$ 和 $nDCG@20$ 值的影响如图 4 所示。

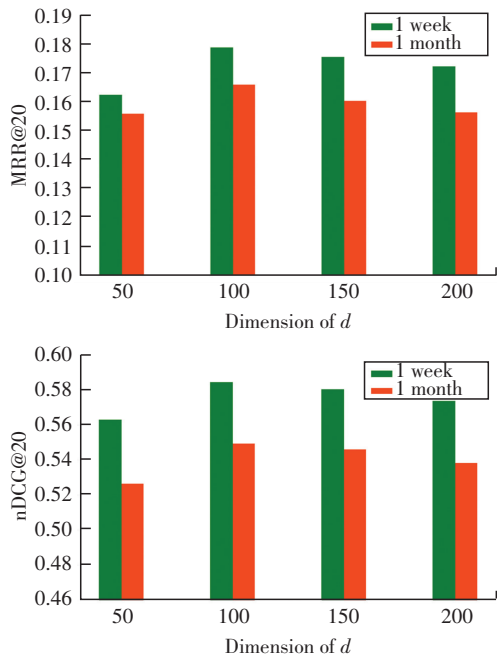


图 4 特征维度 d 对于推荐性能的影响

从图 4 可以看出模型性能随着 d 的增加先增加,后降低。当 d 的维度等于 100 时推荐性能最佳,这表明该维度最能表达新闻的语义信息。综合以上

情况分析,过低的维数不足以捕获有效的信息,而过高的维数会导致模型的过拟合,这些情况都会导致推荐性能的下降。

会话数量的影响。图 5 给出了随着会话数量增加,评估指标值的变化趋势。

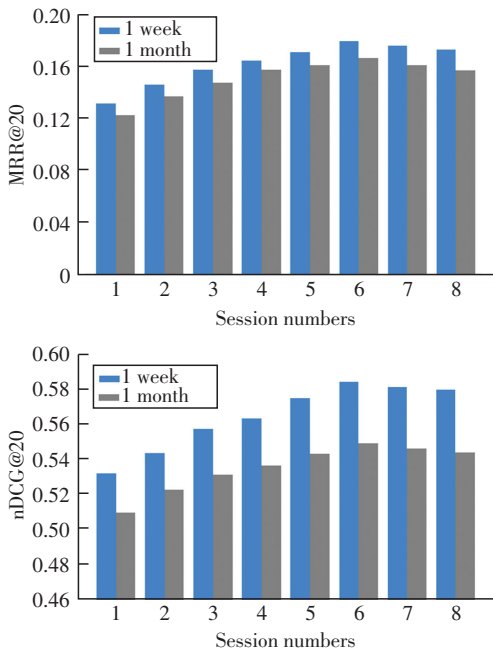


图 5 会话数量对于推荐性能的影响

从图 5 可以看到,随着会话数量的增加,模型性能逐渐提高,这说明引入历史会话信息可以丰富用户的兴趣表示,提高推荐的准确性。但是当会话数量大于 6 时,模型性能有所下降,可能的原因是用户兴趣随着时间而发生偏移,这意味着结合过多的会话信息会给融合后的用户表示带来一定程度的噪声。

4 结束语

本文提出了一种基于会话的双层注意力机制新闻推荐模型 (TLAM-NR)。该模型综合分析了会话内和会话间的用户兴趣,它不仅利用兴趣感知注意力层捕捉会话内用户的主要兴趣,而且还使用会话感知注意力层建模当前会话与历史会话之间用户兴趣的相关度,有效地捕捉不断变化的用户兴趣并实现新闻推荐,进而提高新闻推荐的准确度。

在后续工作中,将利用更多的新闻属性信息 (如主题、类别以及实体关系等) 来挖掘深层次的新闻语义特征,进一步提高新闻推荐的性能。

参考文献:

[1] 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.2015;195–202.
- [2] 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.2017; 1933–1942.
 - [3] WANG H, ZHANG F, XIE X, et al. DKN: deep knowledge-aware network for news recommendation[C] // Proceedings of the International Conference on World Wide Web.2018; 1835–1844.
 - [4] HU L M, 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.
 - [5] ZHANG L M, LIU P, GULLA J A. A deep joint network for session-based news recommendations with contextual augmentation[C] // Proceedings of the 29th ACM Conference on Hypertext and Social Media.2018;201–209.
 - [6] AN M X, WU F Z, WU C H, 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.2019;336–345.
 - [7] ZHU Q N, ZHOU X F, SONG Z L, et al. DAN: deep attention neural network for news recommendation[C] // Proceedings of the AAAI Conference on Artificial Intelligence.2019;5973–5980.
 - [8] PANG Y T, ZHANG Y M, TONG J N, et al. Time-aware attentive neural network for news recommendation with long-and short-term user representation[C] // International Conference on Knowledge Science, Engineering and Management.2020;76–87.
 - [9] 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.2007;271–280.
 - [10] WANG C, BLEI D M. Collaborative topic modeling for recommending scientific articles[C] // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.2011;448–456.
 - [11] XUE H J, DAI X, ZHANG J, et al. Deep matrix factorization models for recommender systems [C] // IJCAI.2017;3203–3209.
 - [12] KOMPAN M, BIELIKOVÁ M. Content-based news recommendation[C] // International Conference on Electronic Commerce and Web Technologies.2010;61–72.
 - [13] HUANG P S, HE X D, GAO J F, et al. Learning deep structured semantic models for web search using click-through data[C] // Proceedings of the 22nd ACM International Conference on Information & Knowledge Management.2013;2333–2338.
 - [14] LI L, WANG D D, LI T, et al. SCENE: a scalable two-stage personalized news recommendation system[C] // Proceedings of the 34th ACM SIGIR International Conference on Research and Development in Information. 2011; 125–134.
 - [15] LU J S, XIONG C M, PARIKH D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning[C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2017;3242–3250.
 - [16] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL].[2020-08-27].<https://arxiv.org/abs/1409.0473>.
<https://arxiv.org/abs/1409.0473>.
 - [17] LI J, REN P J, CHEN Z M, et al. Neural attentive session-based recommendation[C] // Proceedings of the ACM Conference on Information and Knowledge Management. 2017;1419–1428.
 - [18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Advances in Neural Information Processing Systems.2017;5998–6008.
 - [19] ZHANG S, TAY Y, YAO L, et al. Next item recommendation with self-attentive metric learning[C] // Proceedings of the 33rd AAAI Conference on Artificial Intelligence.2019;9.
 - [20] CHU Q F, LIU G S, SUN H R, et al. Next news recommendation via knowledge-aware sequential model[C] // China National Conference on Chinese Computational Linguistics.2019;221–232.
 - [21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90.
 - [22] KIM Y. Convolutional neural networks for sentence classification[C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014;1746–1751.
 - [23] ZHANG X, ZHAO J B, LECUN Y. Character-level convolutional networks for text classification[EB/OL].[2020-08-27].<https://arxiv.org/abs/1509.01626>.
 - [24] GULLA J A, ZHANG L, LIU P, et al. The adressa dataset for news recommendation[C] // Proceedings of the International Conference on Web Intelligence. 2017; 1042–1048.