基于动态动作覆盖的深度强化学习新闻推荐

董相宏，安俊秀
成都信息工程大学软件工程学院，四川 成都 610000

摘要
新闻推荐系统对新媒体新闻传播有着重要作用。提出了一种以深度强化学习为基础的推荐系统，旨在结合神经网络的表征能力和强化学习的策略选择能力来提升新闻推荐效果。使用动态动作掩码加强对用户短期兴趣的判断能力，使用优化缓存机制提升经验缓存的使用效率，通过区域遮蔽性质的奖励设计加快模型训练，从而提高推荐系统在新闻推荐领域的表现。实验表明，所提模型在新闻数据集上的推荐准确率与主流的神经网络推荐方法相当，且在排序性能上优于当前先进的推荐算法。

关键词
新闻推荐；强化学习；动态掩码；优势缓存；内在奖励

中图分类号：TP311.5 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023069

Deep reinforcement learning news recommendation based on dynamic action coverage

DONG Xianghong, AN Junxiu
School of Software Engineering, Chengdu University of Information Technology, Chengdu 610000, China

Abstract
News recommendation system plays an important role in news dissemination of new media. This paper proposed a recommendation system based on deep reinforcement learning, which aimed to combine the representation ability of neural network and the strategy selection ability of reinforcement learning to improve the effect of news recommendation. This paper used dynamic action masks to enhance the ability of judging the short-term interests of users, used the optimization cache mechanism to improve the efficiency of experience cache use, and accelerated model training through the reward design of regional masking nature to improve the performance of the recommendation system in the field of news recommendation. Experimental results show that the accuracy of the proposed model in news data sets is comparable to the current mainstream neural network recommendation methods,

and its ranking performance is better than others.

Key words
news recommendation, reinforcement learning, dynamic mask, advantage cache, internal reward

0 引言

新闻推荐通常被建模为序列推荐，但在实际推荐场景中，用户对新闻主题的选择通常呈现不规则的多样性，用户一般不会连续点击相似的新闻，这与商务类推荐等序列推荐场景不同[1]。因此，对于新闻的推荐而言，通过监督学习建立一个以历史会话数据为基础的模型，是无法适应多变的新闻主题推荐的，无法很好地识别用户短期内以会话为单元的新闻倾向性变化，也就是监督学习建立的推荐模型更适应静态的、关联性强的推荐领域，而不适应动态的、随时间快速变化的领域。新闻推荐模型需要一定的序列决策能力以增加对新闻推荐的适用性。为解决新闻推荐的问题，文献[2]首次将深度强化学习应用于新闻推荐任务，提高了新闻推荐的工作效率，但是仍然存在模型收敛速度过慢、使用贪婪算法导致重复率过高的问题。为解决以上的问题，本文在深度Q网络（deep Q network，DQN）方法的基础上，提出动态动作掩码深度强化学习（dynamic action masking deep reinforcement learning，DAMDRL）对新闻推荐系统进行建模。本文的主要贡献如下。①提出在深度强化学习的推荐方法中采用动态掩码表示动作空间，并使用优先缓存机制强化部分缓存经验，从而加快模型收敛；②使用避免回路的探索方式改进推荐系统对新项目的探索效率，很好地平衡了利用率和探索性；③提出了基于深度强化学习的DAMDRL框架，结合用户的多重反馈和动态动作进行新闻推荐。

1 相关研究

1.1 新闻推荐

作为推荐系统的一个重要研究方向，新闻推荐既有一般物品推荐的特点，如项目的关联性等，也有自身的特点。新闻推荐与其他推荐任务的区别如下。①新闻的时效性高，大多数新闻文章的有效期很短，通常仅有几天时间。因此，在设计新闻推荐算法时要充分考虑时效性。②新闻文章具有丰富的内容和上下文信息，捕捉用户行为和候选新闻之间的相关性对于理解用户对特定候选新闻文章的兴趣非常重要。因此，对用户行为进行多维度建模是改进新闻推荐效果的有效方法。③用户对新闻的选择存在强烈的时间多样性偏好。对于其他推荐任务，用户可能更喜欢点击非常相似的项目，而在新闻推荐中，通常用户更倾向于点击与之前有关联但有所不同的新闻。文献[1]指出，将新闻推荐建模为标准的顺序推荐任务可能不太合适，在用户建模中考虑时间多样性偏好是很重要的。如GRU-4Rec[3]使用的循环神经网络模型和SLi-Rec[4]使用的注意力机制模型，更多地关注全局上下文，而不是顺序依赖性。文献[5]提出一种在BERT框架的基础上结合了时间信息的重排推荐模型，该方法较好地关注了全局历史信息，本文借鉴了其对时间信息的处理。

在新闻建模发展的初期，通常使用TF-IDF（term frequency-inverse document frequency）方法抽取新闻的重要特征[6]。随着深度学习技术的发展，使用神经网络进行新闻表示成为主流。例如，文献[7]提出了一种基于词嵌入的新闻推荐方法，该方法使用去噪编码器表征新闻特征；文献[8]使用doc2vec提取新闻表征，然后使用神经网络学习隐藏特征；文献[9]通过word2vec嵌入来表征新闻。单纯基于特征工程的新闻推荐模型在文本质量有保证的小规模系统上表现较好，但在实际新闻推荐中通常需要其他方法的协同。一种方法是利用新闻的其他信息，例如，以新闻的类别进行聚类；将新闻的一些动态特征（如流行度和建立时间等）应用于基于特征的新闻建模方法中。此外，还可以利用一些环境因素，如新闻事件发生地与用户地理位置的关联性因素[10]。

此外，有研究对不同类型的用户兴趣进行建模，如GRU4Rec和SLi-Rec考虑用户的长期和短期兴趣，从而更好地捕捉用户的兴趣动态。这两种方法都综合了不同类型的用户兴趣，提高了对用户兴趣的理解。然而，用户兴趣存在多样性和演化性，这些方法仍难以全面、准确地对用户兴趣进行建模。

1.2 深度强化学习新闻推荐

近年来，随着机器学习研究的深入，深度学习和强化学习在推荐系统构建中起到了很好的效果。深度学习方法可以通过神经网络发现用户和物品之间复杂的非线性关系[11]。虽然神经网络构建推荐系统的方式解决了在原有模型上处理信息能力较弱和模型召回率、准确度不高的问题，但是基于神经网络的模型需要大量的标签化数据用于训练，而推荐系统的关系矩阵是稀疏的，尤其在新用户或物品进入系统时，仅凭借少量数据不足以实现对用户和物品的精确建模。借助强化学习在少样本渐近式构建模型方面的优势，优化深度学习模型，成为解决上述问题的一个重要方法。

使用强化学习进行推荐，主要思想是建立一个可优化的候选新闻文章目标，并在此基础上通过优化长期的总奖励[12]得到一个符合用户习惯的决策序列。文献[13]提出将个性化新闻推荐问题建模为上下文相关的多臂老虎机问题，这种方法通过混合线性模型计算收益，从而得到推荐序列。文献[2]使用DQN估计真实奖励，根据推荐后返回的待优化列表进行计算，并且采用博弈老虎机梯度下降（dueling bandit gradient descent）[14]算法，解决了随机探索造成的性能下降问题。文献[15]描述了在线推荐中的稀疏奖励问题，并提出了一种新的状态感知体验重放模型，该方法以重放优化为基础，提升了重放中稀疏奖励的下限。

强化学习的基本思想是使用马尔可夫决策过程（Markov decision process）对用户的浏览行为进行模拟和预测。从模型上看，强化学习可以分为无模型（model free）和基于模型（model-based）[16]两类。无模型方法尝试理解环境，环境给什么信息就使用什么信息，如Q-learning方法。传统的Q-learning方法使用表格来表示和计算动作空间与状态空间的关系，而在复杂的拟真环境中，表格法存在空间维度爆炸、难以表示的问题。因此，在神经网络的基础上提出了DQN以及相关的改进方法，用于计算和表征每个动作的分数。

DQN算法使用Q网络进行策略选择。Q网络的输入是用户状态和智能体的动作空间，Q网络的输出是智能体在某一状态下执行动作得到的累计奖励值。文献[17]提出了一种基于值的强化学习模型，使用

交叉熵捕获用户的兴趣变化，并在重放经验中使用优势重放。文献[18]使用图卷积神经网络得到商品的嵌入表征，并使用循环神经网络融合用户行为得到用户的状态，然后采用改进的DQN算法选择推荐策略。通过对上述文献的实验复现和研究对比，本文最终选定以优化重放缓存的方式加快模型收敛，以优化探索策略的方式减少探索步骤和增加准确率。

2 动态动作的深度强化学习框架

本节首先介绍DAMDRL如何定义新闻推荐任务，然后从顶层架构的视角描述DAMDRL的系统结构，最后详细说明DAMDRL各个构件的组成。

2.1 新闻推荐的任务定义

本文将推荐系统的推荐过程定义为在由用户历史数据构造的环境中，通过向用户推荐项目这一动作得到该动作的奖励，并尝试改变用户历史数据以进行下一步动作。整个过程构建序列 (s_t,a_t,r_t,s_{t+1})，序列中元素的含义如下。

状态空间S由用户的正面反馈、负面反馈、隐性反馈和待推荐项目组成，定义状态$s_t \in S$，s_t表示t时刻下用户历史数据的状态[19]。

动作空间A由召回后得到的小规模待排序项目序列组成，定义动作 $a_t \in A$，a_t 表示在 s_t 状态下向用户推荐某一物品的行为。

基础奖励 $r_t \in \{0,1\}$，若用户接受推荐项目，则得到奖励1；若用户不接受推荐项目，则得到奖励0。总的奖励表示为 $R = r_t + r^{in}$。除了基础奖励，使用内在奖励 r^{in} 来减少智能体的旧区域探索[20]，关于内在奖励将在后文中讨论。

状态 $s_{t+1} \in S$，s_{t+1} 表示用户选择动作 a_t 后，根据用户是否接受推荐项目更新正负反馈队列，并与 s_t 状态下的待推荐列表重新结合，得到新的状态 s_{t+1}。

2.2 DQN算法和双层深度Q网络算法的改进

DQN算法通过学习动作值函数 $Q(s,a)$ 估计推荐系统在状态 s 下向用户推荐项目 a 的奖励。时序差分（temporal difference，TD）是一种用来估计一个策略的价值函数的方法，传统的基于价值的强化学习方法（如Q-learning）使用表格的方式对TD误差进行估计，但在推荐系统中，无法在巨大的动作空间中使用表格法。针对这一问题，DQN算法使用神经网络进行函数拟合来表述动作价值函数。DQN网络存在两个问题：一是强化学习交互经验之间存在关联性，不满足建立神经网络所需的数据独立同分布假设；二是DQN算法通常会过高估计探索值，使部分探索值无效。为解决这两个问题，双层深度Q网络（double DQN，DDQN）使用了经验回放（experience replay）机制和目标网络两种方式。

经验回放：为了将Q学习和深度神经网络相结合，DQN算法采用了经验回放方法，具体做法为维护一个回放缓冲区，将每次从环境中采样得到的四元组数据 (s_t,a_t,r_t,s_{t+1}) 存储到回放缓冲区中，训练Q网络时再从回放缓冲区中随机采样若干数据。

目标网络：由于TD误差目标由神经网络的输出组成，在更新网络参数时目标也在不断改变，这导致神经网络训练不稳定。为了解决这一问题，DQN构建与Q网络一致的目标网络，在目标网络中固定TD目标，在一定的迭代间隔后更新目标网络。

2.3 DAMDRL整体架构

DAMDRL模型的体系结构主要由Embedding词嵌入模块、用户历史行为序列存储模块、经验采集模块以及双层Q网络输出模块组成，如**图1**所示。首先，通过Embedding词嵌入模块对所有新闻候选推荐项目进行初始化表征并构建环境；然后，由经验采集模块与环境进行交互，采集所需的经验数据；最后，使用DDQN进行探索学习，并存储每一次探索的行为序列数据。

(1) Embedding词嵌入模块

以往的推荐算法主要关注用户的正反馈信息，而忽略了用户行为的负反馈和隐性反馈。但在新闻推荐场景下，单独的正反馈并不能确定用户对此新闻有兴趣，并且正反馈一般会存在数据稀疏问题，进而造成强化学习的奖励稀疏[21]。因此，本文提出通过正面反馈、负面反馈以及隐性反馈来建模用户行为。隐性反馈包括新闻的时间特征，用户是否有深度阅读、转发和评论等行为，最终混合多种反馈与待推荐项来构造推荐环境[22]。本文使用GRU网络提取3个部分的特征，即用户行为序列向量 $U_t=\{u_p,u_o,u_n\}$ ，其中 u_p 为用户正反馈向量， u_o 为用户负反馈向量， u_n 为隐性反馈向量。3个向量的长度仅为 k，特征提取首先使用item2vec构建新闻向量表示，然后分别使用3个GRU编码器对用户的反馈序列进行编码，最终将推荐项目编码与用户行为序列编码组成强化学习所需的环境反馈[23]，整体流程如**图2**所示。

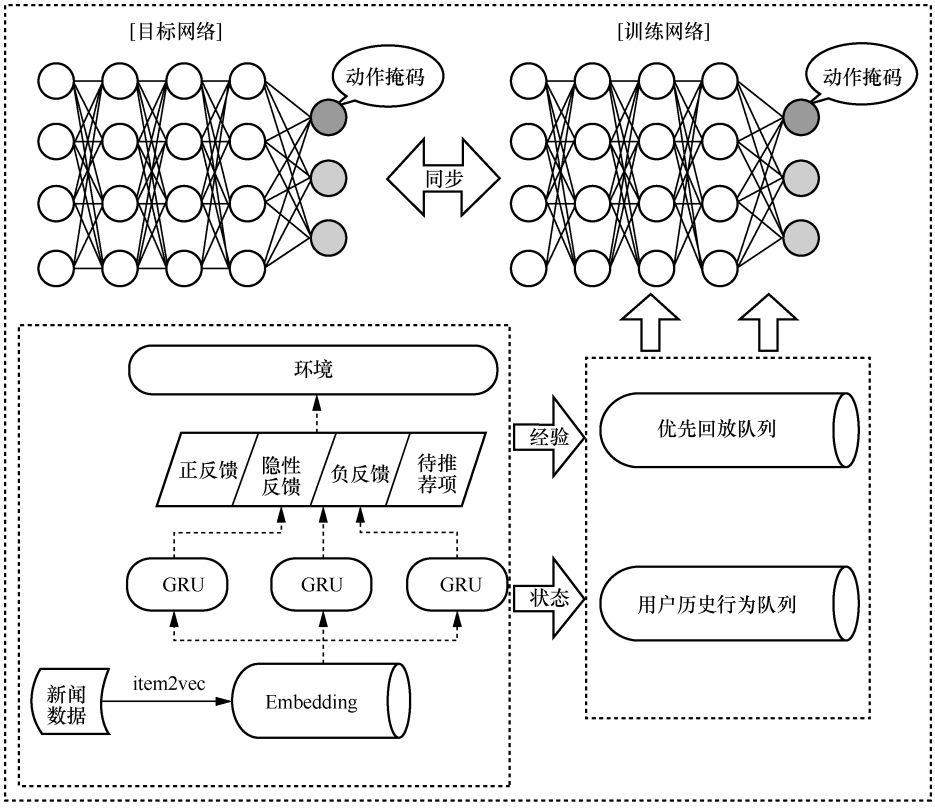


图1 DAMDRL模型的体系结构

（2）用户历史行为序列存储模块

在深度强化学习过程中，为了得到独立同分布的数据集，采用“先采集经验池再进行计算”的方式切割连续相关的用户行为。本文采用独立的存储池保存用户行为数据的分布情况。

（3）经验采集模块

深度强化学习使用经验池机制将经验数据以四元组 (s_t,a_t,r_t,s_{t+1}) 的形式存储起来，然后按照一定的策略抽取样本，并更新神经网络参数 θ 以对 Q 函数进行近似：$Q(s,a;\theta)\approx Q^*(s,a;\theta)$。本文采用了基于优选缓存的经验池构建方法来实现非随机抽样数据[24]。优先缓存可以评价每条经验数据源的重要性，在数据进入队列时可根据评估的重要性指标计算采样概率，计算式为：

$$P(i)=\frac{p_i}{\sum_k p_k} \tag{1}$$

基于值的强化学习算法每一步都将计算TD误差，这种度量方式非常适合在连续性推荐预测的场景中使用。在经验更新的过程中，具备最大误差的经验数据有更大概率进行回放。δ_i 表示第 i 条经验的TD误差。$p_i=|\delta_i|$，也就是说经验池中每一个样本的变化都是由TD误差的正则化权重决定的。这种更新方式使经验数据不再独立，经过若干次迭代后，具有高权重的数据将被多次使用，而具有较低权重的数据将逐渐消失。为了解决经验回放中缓存更新不平衡问题，本文引入系数 γ 和偏移 β，在Q网络训练过程中保证所有的经验数据都能够被更新，并且保证低优先级的数据不被遗弃。定义优势缓存的概率为：

$$P(i)=\gamma\frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}}+\beta \tag{2}$$

（4）双层Q网络输出模块

双层Q网络模块利用优势经验进行训练，计算当前用户多重反馈序列和待推荐物品之间的值函数 $Q^*(s,a;\theta)$。在该模块中，强化学习智能体分别从优势经验回放缓存池和用户反馈历史数据缓存中取出属于该用户的经验数据并进行预测，根据预测结果更新参数 θ。在这个过程中，由于推荐系统的推荐项目不可重复，本文在推荐选择的位置使用了掩码机制，即通过初始化掩码向量 $M=\{m_1,m_2,m_3,\cdots,m_k\,|\,m_i\in\{0,1\}\}$，掩码长度 k 等于动作空间大小，在动作Q网络计算完成后将动作空间与 M 做并集，计算得到最终可用的动作并按 Q 值进行选择。

此外，为了保证探索的稳定性和新颖性的平衡，本文使用旧区域回避策略的 $\varepsilon-$贪婪算法进行探索。该方法鼓励智能体远离当前最近访问过的一片区域。区域是由状态组成的集合，通过周期性地更换最近访问过的区域，从区域中找出转移状态和最近状态，再根据它们的距离计算内在奖励，智能体能够最大程度地远离最近访问过的区域。若智能体在第 i 个周期访问了 $s_{i,0},s_{i,1},\cdots,s_{i,j}$ 状态，r^{in} 指智能体在第 i 个周期 j 时刻的内

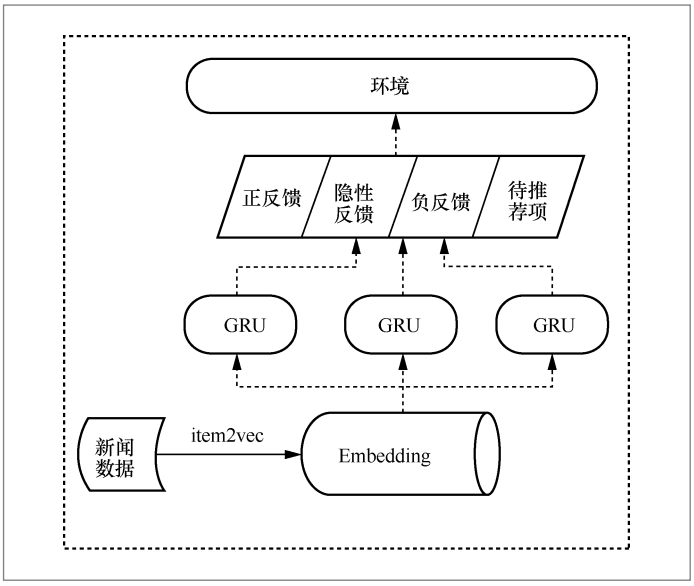


图 2 用户行为序列编码

在奖励，内在奖励如下。

$$r^{\text{in}} = -\frac{1}{\min\left\|s_{i,j+1} - s_k\right\|} \tag{3}$$

3 实验分析

3.1 基线方法

为验证提出的DAMDRL模型的性能，本文选取了5组方法作为基线方法。

- SVD++[25]使用概率矩阵分解方法，在预测函数中考虑用户所有隐式交互行为。
- NCF[26]是一种神经协同过滤方法，采用多层感知机学习用户行为信息。
- GRU4Rec[3]是一种使用循环神经网络的推荐方法，使用GRU从短期的会话数据中抽取用户行为信息。
- SLi-Rec[4]是一种用注意力网络模拟长期兴趣、用LSTM模拟短期兴趣的推荐算法。
- DQN[14]是结合深度神经网络的基于价值优化的强化学习方法，使用神经网络近似价值函数。

3.2 评价指标

为了评价推荐列表质量，选择了3个评价指标，分别是命中率（hit ratio，HR）、平均倒数排名（mean reciprocal rank，MRR）与归一化折损累计增益（normalized discounted cumulative gain，NDCG）。

HR反映了在推荐序列中是否包含用户真正点击的项目，定义如式（4）所示，k表示推荐列表的长度，hit(i)函数表示是否命中，即用户选择的项目是否在推荐序列中，存在为1，否则为0。

$$\text{HR}@k = \frac{1}{k}\sum_{i=1}^{k}\text{hit}(i) \tag{4}$$

MRR指标反映了推荐的项目是否处于用户推荐列表中的明显位置，强调位置关系。定义如式（5）所示，k表示推荐列表的长度，表示用户真实访问的项目在推荐列表中的位置，如果不在推荐序列中，则p_i为无穷大，$\frac{1}{p_i}=0$。

$$\text{MRR}@k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{p_i} \tag{5}$$

折扣累计收益（discounted cumulative gain，DCG）反映了用户喜欢的商品是否在推荐列表中靠前的位置。DCG无法直接比较，需要进行归一化处理。把所有待推荐项目置放在期望的次序下，选取前K项并计算它们的DCG。然后，将DCG除以期望状态下的DCG就可以得到NDCG，它是一个0～1的数，具体定义为：

$$\text{NDCG}@k = Z_k\sum_{i=1}^{k}\frac{2^{r_i}-1}{\log_2(i+1)} \tag{6}$$

r_i表示第i个位置的“等级关联性”，如果该位置的物品在测试集合中，则$r_i=1$，否则为0。Z_k为归一化系数，表示式（6）中的累加求和计算式在最好情况（满足$r_i=1$）下的倒数。

3.3 对比实验

表1展示了DAMDRL模型与其他所有基线方法在新闻数据集上的实验结果，分别在HR@10、MRR@10和NDCG@10 3种评价指标下进行对比。

通过**表1**发现，以HR@10为指标的情况下，GRU4Rec和SLi-Rec两种用神经网络进行监督学习的模型表现较好，而未进

表1 对比实验结果

方法	HR@10	MRR@10	NDCG@10
SVD++	0.38105	0.19401	0.11597
NCF	0.59171	0.38729	0.39812
GRU4Rec	0.73882	0.48101	0.41101
SLi-Rec	0.68491	0.42635	0.47539
DQN	0.31023	0.43916	0.44379
DAMDRL	0.60177	0.49348	0.50181

行改进的DQN模型的命中率较低，本文改进后的DAMDRL模型相比DQN模型有较大的提升。而从MRR@10和NDCG@10两个更专注顺序的指标来看，DAMDRL的结果，相对于NCF、GRU4Rec和SLi-Rec这3种神经网络模型以及未改进的DQN强化学习模型，均有较为明显的提升。实验表明,本文提出的DAMDRL模型在推荐准确率和排序性上均表现良好。

3.4 消融实验

为确定各个模块对于推荐效果的影响，本文在DAMDRL模型的基础上提出了该模型的变种DAMDRL-1与DAMDRL-2。DAMDRL-1模型在经验采集上使用平均采样方式，以验证优势采样对模型的影响。DAMDRL-2模型不使用避免回路的探索策略，仅使用贪婪算法，以验证推理探索策略对模型的影响。消融实验使用与对比实验一致的数据集和评价指标。**表2**给出了DAMDRL、DAMDRL-1和DAMDRL-2这3个模型的实验结果。在HR@10和MRR@10这两项指标中，使用

表2 消融实验结果

方法	HR@10	MRR@10	NDCG@10
DAMDRL-1	0.59021	0.47834	0.41032
DAMDRL-2	0.58911	0.48921	0.37079
DAMDRL	0.60177	0.49348	0.50181

基础的采样方式和探索策略并未大幅度改变推荐精度，但在以NDCG@10为指标的消融实验中，可发现精度变化幅度较大。该实验说明，本文所提的DAMDRL框架采用强化学习的推荐方法提高了推荐命中率，而对缓存策略和探索策略的改进的提升点主要集中在推荐后的排序表现上。

4 结束语

为了提升在新闻推荐场景下推荐系统应对动态的用户期望的能力，本文在深度强化学习DQN方法的基础上，提出动态动作掩码深度强化学习来加强对用户短期兴趣的判断能力。通过动态掩码和区域遮蔽性质的奖励设计加快模型训练，提高推荐系统在新闻推荐领域的表现。实验结果表明，DAMDRL在新闻数据集上的推荐准确率与当前的神经网络推荐方法相当，且在排序性能上优于当前先进的推荐算法。下一步的工作将研究在会话新闻推荐的环境下，强化学习应用于会话新闻推荐的可能性。

参考文献:

[1] LIN C, XIE R Q, GUAN X J, et al. Personalized news recommendation via implicit social experts[J]. Information Sciences, 2014, 254: 1-18.

[2] ZHENG G, ZHANG F, ZHENG Z, et al. DRN: a deep reinforcement learning framework for news recommendation[C]// Proceedings of the 2018 World Wide Web Conference. Republic and Canton of Geneva: IW3C2, 2018: 167-176.

[3] HIDASI B, KARATZOGLOU A, BALTRUNAS L, et al. Session-based recommendations with recurrent neural

networks[EB]. arXiv preprint, 2015, arXiv: 1511.06939.

[4] LIN G Y, GAO C, LI Y F, et al. Dual contrastive network for sequential recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 2686–2691.

[5] ZHAO Q H. RESETBERT4Rec: a pretraining model integrating time and user historical behavior for sequential recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 1812–1816.

[6] IJNTEMA W, GOOSSEN F, FRASINCAR F, et al. Ontology–based news recommendation[C]// Proceedings of the 2010 EDBT/ICDT Workshops. New York: ACM, 2010: 1–6.

[7] 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: ACM, 2017: 1933–1942.

[8] KARVELIS P, GAVRILIS D, GEORGOULAS G, et al. Topic recommendation using Doc2Vec[C]//Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2018: 1–6.

[9] CASELLES–DUPRÉ H, LESAINT F, ROYO–LETELIER J. Word2Vec applied to recommendation: hyperparameters matter[C]//Proceedings of the 12th ACM Conference on Recommender Systems. New York: ACM, 2018: 352–356.

[10] ZHANG J D, CHOW C Y, LI Y H. iGeoRec: a personalized and efficient geographical location recommendation framework[J]. IEEE Transactions on Services Computing, 2015, 8(5): 701–714.

[11] KARATZOGLOU A, HIDASI B. Deep learning for recommender systems[C]// Proceedings of the Eleventh ACM Conference on Recommender Systems. New York: ACM, 2017: 396–397.

[12] DEVLIN S M, KUDENKO D. Dynamic potential–based reward shaping[C]// Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. Richland: IFAAMAS, 2012: 433–440.

[13] LI L H, 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: ACM, 2010: 661–670.

[14] YUE Y S, JOACHIMS T. Interactively optimizing information retrieval systems as a dueling bandits problem[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 1201–1208.

[15] XIAOCONG C, LINA Y,et al. Locality–sensitive state–guided experience replay optimization for sparse rewards in online recommendation[C]// Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2022: 1316–1325.

[16] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1–27.
LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1–27.

[17] ZHANG Y Y, SU X Y, LIU Y. A novel movie recommendation system based on deep reinforcement learning with prioritized experience replay[C]//Proceedings of 2019 IEEE 19th International Conference on Communication Technology (ICCT). Piscataway: IEEE Press, 2020: 1496–1500.

[18] LI Y Q, CHEN W Z, YAN H F. Learning graph–based embedding for time–aware product recommendation[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 2163–2166.

[19] LIU Q, ZENG Y F, MOKHOSI R, et al. STAMP: short–term attention/

memory priority model for session-based recommendation[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1831-1839.

[20] 蔡丽娇, 秦进, 陈双. 远离旧区域和避免回路的强化探索方法[J]. 计算机工程, 2023, 49(7): 118-124,134.
CAI L J, QIN J, CHEN S. Reinforcement exploration method to keep away from old areas and avoid loops[J]. Computer Engineering, 2023, 49(7): 118-124,134.

[21] ZHAO X Y, ZHANG L, DING Z Y, et al. Recommendations with negative feedback via pairwise deep reinforcement learning[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1040-1048.

[22] GONG S, ZHU K Q. Positive, negative and neutral: modeling implicit feedback in session-based news recommendation[EB]. arXiv prprint, 2022, arXiv: 2205.06058.

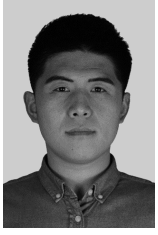
[23] 刘树栋, 张可, 陈旭. 基于多维度兴趣注意力和用户长短期偏好的新闻推荐[J]. 中文信息学报, 2022, 36(9): 102-111.
LIU S D, ZHANG K, CHEN X. Multi-dimensional interest-attention-based news recommendation with long and short-term user preferences[J]. Journal of Chinese Information Processing, 2022, 36(9): 102-111.

[24] 陈希亮, 曹雷, 李晨溪, 等. 基于重抽样优选缓存经验回放机制的深度强化学习方法[J]. 控制与决策, 2018, 33(4): 600-606.
CHEN X L, CAO L, LI C X, et al. Deep reinforcement learning via good choice resampling experience replay memory[J]. Control and Decision, 2018, 33(4): 600-606.

[25] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2008: 426-434.

[26] HE X N, LIAO L Z, ZHANG H W, et al. Neural collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2017: 173-182.

作者简介



董相宏(1993-),男,成都信息工程大学软件工程学院硕士生,主要研究方向为云计算技术、推荐系统。

安俊秀(1970-),女,成都信息工程大学软件工程学院教授,主要研究方向为云计算与大数据技术、大数据分析与服务、云计算技术及应用等。

收稿日期: 2023-02-21
基金项目: 国家社会科学基金项目(No.22BXW048)
Foundation Item: The National Social Science Foundation of China (No.22BXW048)