文章编号:1674-2974(2022)08-0208-09 DOI:10.16339/j.cnki.hdxbzkb.2022294

基于策略记忆的深度强化学习序列推荐算法研究

陈卓,姜伟豪†,杜军威
(青岛科技大学 信息科学技术学院,山东 青岛 266061)

摘 要:推荐系统旨在从用户-项目的交互中进行建模,为用户推荐感兴趣的内容,从而提高用户体验.然而大多数用户-项目的序列并不总是顺序相关的,而是有更灵活的顺序甚至存在噪声.为解决这一问题,提出一种基于策略记忆的深度强化学习序列推荐算法,该算法将用户的历史交互存入记忆网络,使用一个策略网络将用户当前的行为模式更细致地划分为短期偏好、长期偏好以及全局偏好,并引入注意力机制,生成相应的用户记忆向量,利用深度强化学习算法识别对未来收益较大的项目.在用户和项目的交互中不断更新、强化学习网络的策略以提高推荐准确性.在两个公共数据集的实验中表明,本文所提出的算法与最先进的基线算法相比,召回率指标在2个数据集上分别提升了8.87%和11.20%.

关键词:推荐系统;强化学习;策略网络;注意力机制
中图分类号:TP181 **文献标志码**:A

Research on Deep Reinforcement Learning Sequential Recommendation Algorithm Based on Policy Memory

CHEN Zhuo, JIANG Weihao†, DU Junwei
(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: The recommender system aims to build a model from the user-item interaction and recommend the content of interest to users, so as to improve the user experience. However, most user-item sequences are not always sequentially related but have more flexible sequences and even noise. In order to solve this problem, a deep reinforcement learning sequence recommender algorithm based on strategy memory is proposed. The algorithm stores the user's historical interaction in the memory network, and then uses a strategy network to divide the user's current behavior pattern into short-term preference, long-term preference, and global preference, and introduces the attention mechanism to generate the corresponding user memory vector. The deep reinforcement learning algorithm is used to identify the projects with great benefits in the future. The strategy of the reinforcement learning network is continuously updated in the interaction between users and items to improve the accuracy of the recommender. Experiments on two public data sets show that the proposed algorithm improves the recall index by 8.87% and 11.20%, respectively, compared with the most advanced baseline algorithm.

Key words: recommender systems; reinforcement learning; policy network; attention mechanism

---

* **收稿日期**:2021-06-29
基金项目:国家自然科学基金资助项目(F030810,61806107),National Natural ScienceFoundation of China (F030810,61806107);山东省重点研发计划资助项目(2018GGX101052), Key Research and Development Program of Shandong Province(2018GGX101052)
作者简介:陈卓(1978—),女,山东青岛人,青岛科技大学副教授,博士
† 通信联系人,E-mail:591614545@qq.com

随着科学技术的发展,信息过载的问题也越来越严重,推荐系统的成功应用可以有效地缓解这一难题.然而,用户兴趣总是随着时间的推移而产生变化,因此,序列推荐系统(Sequential Recommender Systems, SRS)[1]应运而生.序列推荐系统将用户–项目交互视为一个动态序列,捕捉用户当前和最近的偏好,以获得更准确的推荐,该系统在购物以及影音网站等都有着很好的应用.

不同于基于内容的协同过滤[2]以及基于矩阵分解[3]的传统推荐系统,序列推荐系统根据模型的复杂程度可以分为两类,包括传统的序列模型和神经网络模型.传统的序列模型例如基于马尔科夫链的模型对序列中的用户–项目交互进行建模,并计算相互之间的转移概率,该方法只能捕获短期依赖并且忽略用户的总体兴趣.神经网络模型例如基于循环神经网络的模型通过给定的用户–项目交互序列来预测下一交互,但该方法假设任何相邻交互都是有关系的并忽略多个项目的共同作用.

大多数的序列推荐系统只专注于当前收益,对于即时回报较小但有较大未来收益项目的识别度不高.比如用户在观看了有关游戏和天气的短视频后,可能不会再关注今天的天气;但是看了喜欢的游戏之后,更倾向于观看更多与游戏相关的视频.深度强化学习可以统一优化即时收益和长期收益[4],将用户的状态进行动态建模,并学习到最优推荐策略,以提高推荐准确率.现有的深度强化学习状态的表示是将用户之前交互的项目按照一定的顺序进行建模,无法区分用户交互序列中的不同行为模式,因而无法较为准确地预测用户的当前兴趣偏好.深度强化学习做预测时,可选择的动作空间较大且数据较为稀疏,导致算法收敛困难.本文使用深度强化学习中的深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)算法,该算法不再计算每个项目的概率而是基于确定的策略,这更有利于处理连续动作,并且提高了算法的收敛速度.

本文提出了一种将用户策略记忆与DDPG算法结合的方式来解决以上问题.本文的贡献可以总结为以下几点:

1)使用记忆网络存储用户历史交互序列,并训练一个策略网络,通过用户和其历史交互对用户当前行为模式进行划分.引入注意力机制,根据策略网络的输出对记忆网络使用不同的注意力表示方法并生成用户当前的记忆向量表示.

2)提出结合策略记忆的深度强化学习推荐算法(Deep Reinforcement Recommendation with Policy Memory, DRRM).将用户表示、用户当前偏好以及用户记忆作为状态输入,利用DDPG算法预测用户喜好,在交互中不断更新推荐策略,该算法同时考虑了推荐项目对用户的当前收益及其对用户的长期影响.使用探索策略提高推荐多样性.

3)通过在两个公共数据集上进行的实验,验证了所提出算法的有效性,其效果明显强于基线算法.

1　相关工作

1.1　序列推荐系统

序列推荐系统是近年来的研究热点,通过对用户的行为序列进行建模,考虑用户兴趣的依赖性,为其推荐下一时刻所感兴趣的项目.传统的序列推荐算法有:Feng等人[5]提出了基于距离嵌入的个性化排序算法(Personalized Ranking Metric Embedding, PRME),将用户序列建模为马尔可夫链,并嵌入欧氏空间,基于距离计算项目相互之间的转移概率.Liu等人[6]提出了基于上下文感知的序列推荐算法(Context-Aware Recurrent Neural Networks, CA-RNN),该算法使用特定输入矩阵和转移矩阵进行推荐,但该算法不利于捕获交互序列中高度相关的项目信息.Wang等人[7]提出的基于注意力机制的事务嵌入推荐算法(Attention-Based Transaction Embedding Model, ATEM)通过注意力机制来观察和识别与下一个项目相关的上下文信息.Kang等人[8]提出的基于自注意力的序列推荐算法(Self-Attentive Sequential Recommendation, SASRec)将自注意力机制应用于序列推荐系统,用于捕获序列的长期语义,并在训练速度上有一定的提升.

上述算法进行序列推荐时,总是假定相邻两交互是相关的,而对于具有灵活顺序的长序列的识别能力不强.

1.2　基于记忆网络的推荐算法

由于需要长期记忆来存储问答知识或聊天的语境信息,记忆网络(Memory Networks, MN)[9]由Weston首次提出,以这种有效的方式来简单地读写

此类信息,该模型最初被应用于智能问答领域. Chen等人[10]提出了一种基于用户记忆网络的推荐算法(Recommender system with User Memory networks, RUM),该算法首次将记忆网络应用于推荐系统,通过对用户交互项目的读取、写入等操作更好地利用用户的历史记录;但该算法并没有对用户兴趣进行区分. Ebesu等人[11]提出了一种将用户协同过滤与记忆网络相结合的推荐系统(Collaborative Memory Networks, CMN),分别利用潜在因素的全局结构以及邻域的局部结构进行推荐. Ouyang等人[12]提出了一种基于记忆增强的深度神经网络推荐算法(Memory Augmented Deep Neural Networks, MA-DNN),该算法为每一个用户都创建喜欢和不喜欢两个外部记忆向量,以此来对用户点击率进行预测.

上述基于记忆网络的推荐算法仅能够识别记忆向量对当前交互的作用,无法识别未来收益较大的交互项目,并且没有利用更深层的神经网络模型对用户偏好进行预测.

1.3 基于深度强化学习的推荐算法

近几年来,随着强化学习的发展,与推荐系统的结合也越来越多. 深度强化学习的目标是智能体在与环境的不断交互中学习到最佳策略,因此,有利于序列推荐的动态建模. Wang等人[13]提出了基于递归神经网络的监督强化学习算法(Supervised Reinforcement Learning with Recurrent Neural Network, SRL-RNN),该算法使用演员-评论家框架处理多种药物、疾病以及个体之间的关系,以进行个性化药品推荐. Zhao等人[14]提出了一种基于多智能体强化学习的DeepChain算法,该算法使用多个智能体协同优化,捕捉多场景中用户行为顺序相关性,以获得整体的最大回报. Zheng等人[15]将深度强化学习中的深度Q网络(Deep Q Learning, DQN)与推荐系统相结合,提出了基于深度强化学习的新闻推荐框架(Deep Reinforcement Learning Framework for News Recommendation, DRN),该框架同时使用Dueling Bandit梯度下降方法来进行有效的探索. Liu等人[16]利用DDPG算法提出了用于解决电影等推荐的深度强化学习推荐算法(Deep Reinforcement Learning based Recommendation, DRR).

上述基于深度强化学习的推荐算法仅使用单一的方式对用户当前状态进行建模,对用户兴趣的划分存在一定偏差,且无法对用户历史喜好的不同权重进行建模.

2 DRRM模型

2.1 符号定义

本文将深度强化学习应用于序列推荐,推荐代理(Agent)根据时间顺序对用户和项目进行交互,以获得最大化的累计回报. 将此过程建模为一个马尔可夫决策过程,其中包括状态、动作、奖励、衰减因子等四元组,具体定义如下:

状态S:由用户及其之前交互项目的集合组合而成,交互过的项目按照时间顺序依次排列,并将其输入演员-评论家网络中进行预测.

动作A:采用基于策略的DDPG算法,每次交互时根据策略网络生成一个确定的动作A,该动作表示用户当前的喜好,再与项目集进行内积得到为用户推荐的项目.

即时奖励R:推荐代理将一个项目推荐给用户,根据用户是否在当前时刻与该项目进行过交互,并给予一定的奖赏或惩罚.

衰减因子γ:$\gamma\in[0,1]$,是衡量短期收益与累计收益的标准,当γ越接近于零,代表算法越重视推荐的短期回报,反之则更加重视累计回报.

2.2 模型框架

DRRM模型框架如图1所示,该模型分为3部分.

第1部分为图中下半部,即用户记忆网络部分,该部分引入注意力机制用来区分用户历史交互项目的不同权重关系,从而学习用户的兴趣变化;根据不同的行为模式生成不同的用户记忆向量,并将用户的历史记录存入外部的记忆网络中,为状态的更新提供依据. 第2部分为策略网络部分,该部分对用户当前行为模式进行划分. 本文使用基于策略梯度算法的策略网络对其进行划分(详见2.2.2节),从而使记忆网络选择不同的注意力计算方式,得到用户记忆向量,并结合用户向量生成当前状态S. 第3部分为DDPG网络部分,该部分用户预测动作A. 该网络由演员和评论家两个网络组成,演员网络通过输入的状态S,来输出用户当前的喜好;评论家网络通过该部分输出来更新演员网络. 该算法在与用户的交

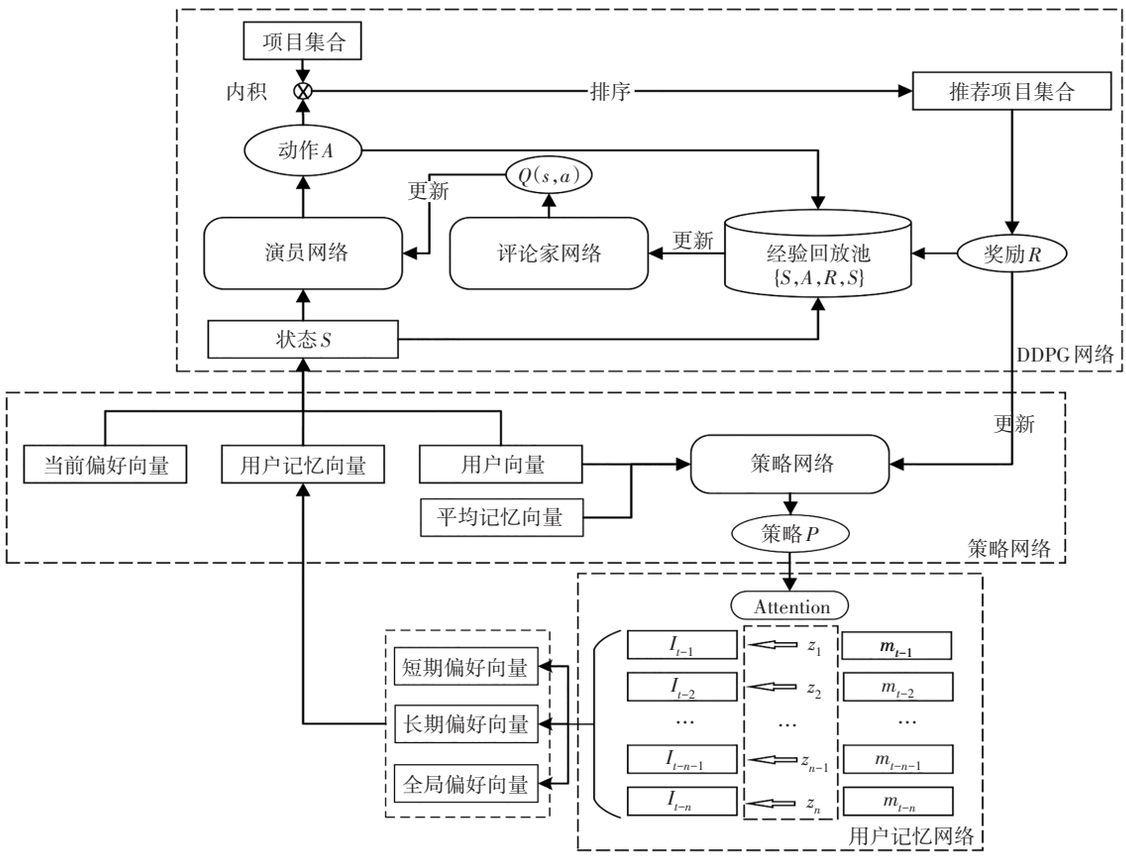

图 1　DRRM 模型框架图

Fig.1　DRRM model framework diagram

互中不断更新,以达到最优策略的输出.

2.2.1　记忆网络模块

本文将用户的历史交互信息存储在一个外部的组件——记忆网络中,记忆网络的具体构建如下:

令 $U=\{u_1,u_2,\cdots,u_{n-1},u_n\}$, $T=\{t_1,t_2,\cdots,t_{m-1},t_m\}$. 分别表示用户和项目的集合,而 n 和 m 分别表示数据集中用户以及项目的个数. 令 $\boldsymbol{S}=\{s_{11},s_{12},\cdots,s_{ij}\}$且 $\boldsymbol{S}\in R^{n\times m}$,表示用户和项目的交互矩阵,其中 s_{ij} 表示用户 u_i 对项目 t_j 的评分,矩阵中不同得分表示用户对该项目的喜好程度,若用户没有对该项目进行过评分,则该评分为 0. 对于每个用户 u 都有该用户自己评过分的项目集合 $T^{\mathrm{u}}=\{t_1^{\mathrm{u}},t_2^{\mathrm{u}},\cdots,t_{k-1}^{\mathrm{u}},t_k^{\mathrm{u}}\}$,其中 k 为该用户所有评分项目的个数,该集合按用户对项目评分的时间序列依次排列.

在每个时间节点 t 时刻都将每个用户 u 在此时刻之前的历史交互 T_t^{u} 存入其对应的记忆网络中,即为 $M_t^{\mathrm{u}}=\{m_1^{\mathrm{u}},m_2^{\mathrm{u}},\cdots,m_{t-2}^{\mathrm{u}},m_{t-1}^{\mathrm{u}}\}$. 根据本文 3.5 节实验研究表明,记忆网络中每个项目对于用户下一次行为的影响并不相同. 因此,本文将用户行为分为短期偏好、长期偏好和全局偏好.

1)当用户行为受上一次影响即为短期偏好时,对用户交互集合 M_t^{u} 中的所有项目与最近一次交互 m_{t-1}^{u} 计算注意力权重,权重的计算如式(1)所示.

$$w_n=m_{t-1}^{\mathrm{u}}\times(m_n^{\mathrm{u}})^{\mathrm{T}},\ z_n=\frac{\exp(w_n)}{\sum_j\exp(w_j)} \tag{1}$$

式中:n、$j\in[1,t-2]$;w_n 为最近一次交互的项目向量与记忆网络中每个向量的乘积,输出为一个一维的数值;exp()表示以 e 为底的指数函数;z_n 为第 n 个项目向量在此记忆网络中的权重值. 在得到权重值之后,计算当前状态下的权重向量. 经过注意力机制的权重向量的计算如式(2)所示.

$$\boldsymbol{A}_t=\sum_j(z_j\times m_j^{\mathrm{u}}) \tag{2}$$

式中:$\boldsymbol{A}_t$ 为所求的前 t-2 个项目的注意力向量,以此来表示该用户的记忆向量.

2)当用户行为受之前某一行为影响即为长期偏好时,首先使用式(1)得出与最近一次行为注意力权重最高的项目 m_q^{u};再使用 m_q^{u} 和用户的历史交互向量

计算注意力权重,再通过式(2)计算用户记忆向量.

3)当用户行为与历史交互无直接关系即为全局偏好时,对使用用户向量与该用户的历史交互向量进行注意力权重的计算,并计算用户的记忆向量.

对于用户记忆网络的写入操作,本文采用先入先出的策略进行更新.将每个用户的记忆网络设置为固定的长度,先将每个用户的前两次交互放入网络中,从第3次交互开始训练.若用户的记忆数小于记忆网络可存放的记忆个数,则直接将最近一次交互写入记忆网络,否则将最开始的用户交互记忆删除并写入最近交互.

2.2.2 基于策略梯度的策略网络

本文将训练一个策略网络来对用户当前行为的3种模式进行划分,记忆网络将根据该网络的输出结果选择如2.2.1节所示的不同的注意力计算方式,生成用户记忆向量.

该策略网络基于策略梯度(Policy Gradient)算法,通过用户向量和用户历史交互向量的平均值作为输入,计算得到用户3种行为模式的概率$\pi_\theta(a_t|s_t)$,并输出该交互的行为模式.在经过记忆网络以及DDPG网络预测出的结果得到奖励值$R(a_t)$,通过该奖励值计算策略网络的损失,该损失函数如式(3)所示.

$$\text{Loss\_p} = -R(a_t)\ln\pi_\theta(a_t|s_t) \tag{3}$$

在得到损失函数后,使用梯度下降来优化损失,更新网络参数.

2.2.3 基于DDPG算法的训练框架

用户当前状态的表示模型由用户表示、用户记忆网络以及用户当前偏好3部分组成,以此来表示用户与项目交互的序列关系.用户u在t时刻的偏好即当前状态的表示如式(4)所示.

$$S_t = \text{Concat}(u, A_t, m^{\text{u}}_{t-1}) \tag{4}$$

式中:Concat()为拼接函数;u代表用户的向量表示;A_t代表用户的记忆向量表示;m^{u}_{t-1}代表用户的当前偏好表示,若为短期偏好则使用用户最近一次交互,若为长期偏好则使用权重最高的交互向量,若为全局偏好则使用历史交互的平均值.

使用DDPG算法对数据进行训练,DDPG算法是演员-评论家网络的一种,它输出的是一个确定的动作,结合DQN以及策略梯度算法的优势,加速网络的收敛,可以更好地解决连续动作空间上的求解问题.该算法由两个网络组成,即演员网络和评论家网络.同时,为了提高网络的稳定性和收敛性,设置两个更新较慢的目标网络来提高两个预测网络的更新效率.两个预测网络的功能及其更新过程如下所述:

演员网络将输入的用户当前状态S_t经过几层神经网络后输出一个确定的用户喜好向量$\boldsymbol{a}_t$.为了增加推荐的多样性,防止算法得到局部最优解,在训练过程中,使用高斯噪声进行探索,探索过程计算如式(5)所示.

$$\boldsymbol{a}_t \sim N(\mu, \sigma^2) \times \beta \tag{5}$$

式中:μ为数学期望;σ^2为方差;β为探索时的衰减系数.使用经过探索得到的喜好向量跟项目集合T相乘,将得到的值经过sigmoid函数变换为每个项目的得分G_i,并将此得分进行排序为用户推荐前n个项目.具体计算如式(6)所示.

$$G_i = \text{sigmoid}(a_t \cdot t_i^{\text{T}}), t_i \in T \tag{6}$$

每轮预测都将向经验回放池传入$B=\{S_t, A, R, S_{t+1}\}$四元组,其中S_{t+1}为下一时刻的状态.演员网络更新的损失函数梯度如式(7)所示.

$$\text{Loss\_a} = -\frac{1}{N}\sum\nolimits_i Q(s, a, \theta_{\text{a}}) \tag{7}$$

式中:θ_{a}为演员网络中的参数;N为经验回放池中的batch大小;$Q(s,a,\theta_{\text{a}})$为评论家网络的输出.

评论家网络用来评估演员网络.通过从经验回放池中获得一个batch的数据,利用现实Q值和估计Q值的均方差来更新其网络参数.更新的损失函数如式(8)所示.

$$\text{Loss\_c} = \frac{1}{N}\sum_i [y_i - Q(s_i, a_i, \theta_{\text{c}})]^2 \tag{8}$$

式中:$Q(s_i, a_i, \theta_{\text{c}})$为估计的$Q$值网络得到的值.$y_i$的计算如式(9)所示.

$$y_i = r_i + \gamma Q'(s_{i+1}, a_{i+1}, \theta_{\text{c}}') \tag{9}$$

式中:γ为衰减因子,用于权衡即时收益与未来总收益之间的关系;$Q'(s_{i+1}, a_{i+1}, \theta_{\text{c}}')$为使用现实$Q$网络对下一步交互所计算的值;$r_i$为当前状态的奖励值.奖励值$R$的计算如式(10)所示.

$$R = \begin{cases} 1 & \text{if} \quad t_i = t_i^{\text{u}} \\ 1/2 & \text{if} \quad t_i \in T^{\text{u}} \\ -1 & \text{其他} \end{cases} \tag{10}$$

若推荐的项目集合中存在用户当前的喜好项目

则奖励值为1;若推荐的项目集合中没有当前喜好,但出现在用户的交互序列中则奖励值为1/2;其他情况奖励值为-1. 以此来区分不同动作所获得的回报.

3　实验与结果分析

3.1　数据集与评价指标

本文使用Amazon Instant Video以及Amazon Automotive(http://jmcauley.ucsd.edu/data/amazon/)两个公开的数据集来进行实验分析. 为保证序列的长度,将交互个数小于10个项目的用户删除,经过预处理后的可用数据量如表1所示. 两个数据集都具有时间戳信息,因此,可对用户序列按时间进行排序并进行序列推荐. 使用每个用户前80%的数据进行训练,后20%进行测试,以证明本文所提出观点的有效性.

表1　数据集统计表

Tab.1　Statistics of data sets

数据集	用户数	项目数	记录数
Amazon Instant Video	1 533	7 786	22 893
Amazon Automotive	1 144	24 590	28 733

本文所使用的评价指标[10]由精准度(Precision)、召回率(Recall)、F_1值(F_1-score)以及HR值(Hit-ratio)组成,从多个方面评估模型的好坏.

3.2　实验环境

本实验采用的软硬件环境如表2所示,本算法所使用的Python版本为3.7.3,并基于Tensorflow深度学习框架实现本算法的深度强化学习.

表2　实验环境表

Tab.2　Experimental environment table

配置名称	配置参数
CPU	Intel Xeon E5-2650 2.20 GHz
内存	32 G×6
显卡	NVIDIA GeForce GTX 1080Ti
操作系统	Ubuntu 20.04.2 x86_64
Python版本	3.7.3
Tensorflow版本	1.9.0

3.3　实验参数设定

DRRM模型所使用的超参数有学习率、batch_size以及衰减因子.

为了研究超参数对模型性能的影响,首先对DDPG网络使用不同学习率、batch_size进行实验. 学习率和batch_size是两个最重要的模型超参数,合适的学习率和batch_size,不仅可以加速模型收敛,防止陷入局部最优,还可以提高模型的性能. 在Amazon Instant Video数据集上进行参数设置,不同学习率、batch_size的对比实验如表3所示.

表3　学习率、batch_size对比实验表

Tab.3　Learning rate, batch_ size of comparative experiments

batch_size	评价指标	学习率			
		0.005	0.002	0.001	0.000 5
4	Precision	0.883	0.979	1.378	1.203
	F_1-score	0.986	1.079	1.499	1.157
8	Precision	1.302	1.375	1.530	1.364
	F_1-score	1.130	1.424	1.761	1.438
16	Precision	1.223	0.990	1.186	0.989
	F_1-score	1.347	1.102	1.259	1.005

从表3可以看出,当学习率为0.001且batch_size为8时,所得到的Precision以及F_1-score指标最高,模型达到最佳性能.

衰减因子是深度强化学习中最重要的参数之一,是衡量当前收益与未来总收益的标准,通过设置不同的大小来表明当前动作对未来的影响. 当学习率和batch_size分别为0.001和8时,进行衰减因子的设定对比实验,实验结果如图2所示.

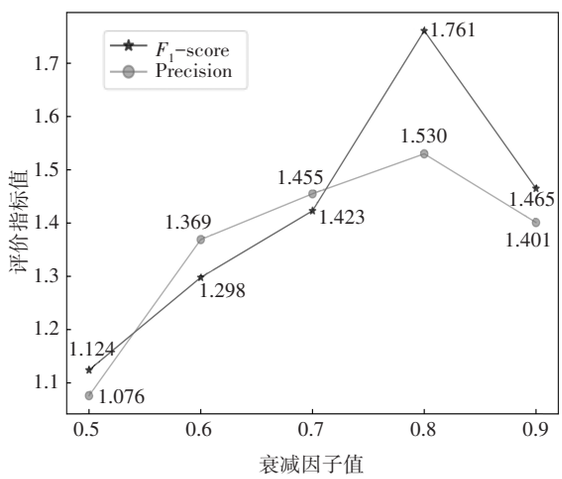

图2　不同衰减因子的指标对比

Fig.2　Comparison of indicators of different attenuation factors

3.4 对比实验

为了证明所提出的DRRM算法的有效性,本文从传统的序列推荐模型、基于神经网络的模型、基于强化学习的模型以及记忆网络模型4个方面,使用以下5种具有代表性的基线算法进行对比实验.

贝叶斯个性化排序算法[17](Bayesian Personalized Ranking, BPR):该算法基于贝叶斯的个性化Top-N推荐方法,提出了一个通用的准则BPR-Opt来优化推荐排序.

个性化马尔科夫链算法[18](Factorizing Personalized Markov Chains, FPMC):该算法将矩阵分解和马尔科夫链模型相结合来学习用户对应的转移矩阵,并引入BPR算法来处理数据进行序列推荐.

动态递归推荐算法[19](Dynamic Recurrent Basket Model, DREAM):该算法的主要思想为基于递归神经网络学习用户的动态兴趣表示,并捕捉用户的全局序列特征.

深度强化学习推荐算法[16](Recommendation Based on Deep Reinforcement Learning, DRR):该算法基于显式用户-项目交互的深度强化学习推荐系统,仅使用用户和项目的交互向量作为输入项,并使用DDPG算法进行预测.

用户记忆网络推荐算法[10](Recommender with User Memory Networks, RUM):该模型的主要思想是在序列推荐算法中引入用户记忆网络,存放用户的历史交互并进行Top-N推荐.

Amazon Instant Video和Amazon Automotive数据集模型性能比较如表4所示.相较于最优基线算法,DRRM算法的精准度在Amazon Instant Video数据集上有8.89%的提升,在Amazon Automotive数据集上略有下降;召回率在2个数据集上分别有8.87%和11.20%的提升;F_1值在2个数据集上分别有18.10%和7.23%的提升;HR在2个数据集上分别有8.89%和1.07%的提升.由此证明了本文所提算法的有效性.

表4 Amazon Instant Video和Amazon Automotive数据集模型性能比较

Tab.4 Performance comparison of Amazon Instant Video and Amazon Automotive dataset models %

评价指标@5	Amazon Instant Video						Amazon Automotive					
	BPR	FPMC	DREAM	DRR	RUM	DRRM	BPR	FPMC	DREAM	DRR	RUM	DRRM
Precision	1.198	1.301	1.312	1.348	1.405	1.530	0.822	0.812	0.792	0.825	0.842	0.831
Recall	1.624	1.612	1.652	1.791	1.905	2.074	0.481	0.483	0.463	0.493	0.501	0.557
F_1-score	1.300	1.322	1.342	1.538	1.491	1.761	0.593	0.594	0.576	0.617	0.622	0.667
Hit-ratio	5.917	5.917	6.097	6.120	6.287	6.846	3.917	3.921	3.812	4.009	4.111	4.155

3.5 消融实验

为了研究本文所提出的策略网络以及基于优先记忆模型在用户记忆网络和DRRM模型中的作用,在Amazon Instant Video数据集上进行多组消融实验,并使用精准度以及召回率指标进行对比.

具体实验设置如下:①将用于生成用户行为模式的策略网络去除,仅使用用户最近交互对用户的记忆向量表示进行计算的DRRM_s模型;②将策略网络去除,仅使用与用户最近交互项目权重最高的记忆项目对用户的记忆向量进行计算的DRRM_l模型;③将策略网络去除,仅使用用户向量计算注意力权重并生成用户记忆向量表示的DRRM_h模型.实验对比图如图3所示.

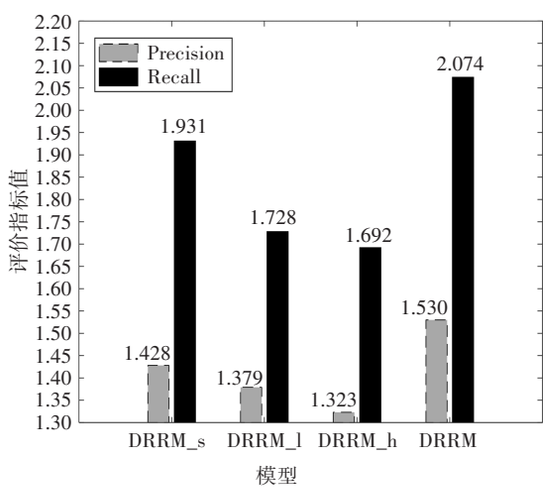

图3 消融实验对比图

Fig.3 Comparison diagram of ablation experiment

由图3可知,没有加入策略网络的3种单一记忆向量表示方法的精准度与召回率均低于DRRM算法,从而说明用户策略网络对用户当前行为模式的预测起着重要的作用.

3.6 用户行为模式

为了证明所提出的3种用户行为模式,即短期偏好、长期偏好以及全局偏好的可解释性,使用Amazon Instant Video数据集的实例进行实验验证.如图4~图6所示,实验使用的记忆网络内存长度为5,x轴和y轴均表示用户的交互序列,每一个小格代表对应两个向量之间的相关度,颜色越深则说明两向量越相关,对角线元素均为1.

由图4可知,在用户的第6~9次交互时,策略网络预测为短期偏好,最近一次交互对当前交互的影响最大,对应的颜色也越深.对应于Amazon Instant Video数据集中的实例是该用户在看了一集某电视剧后又接连观看了后面的3集.

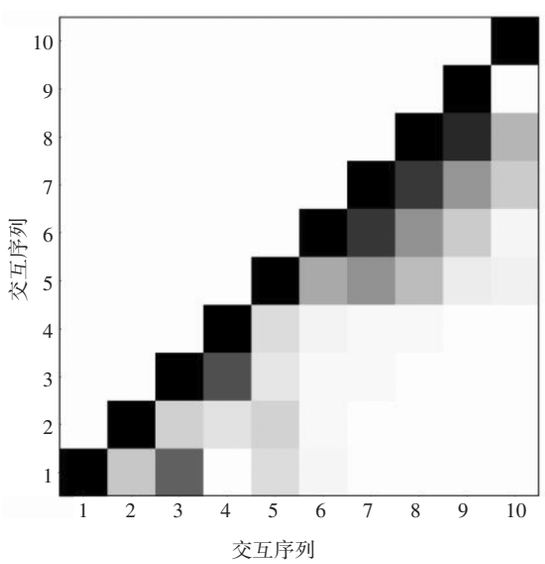

图4　短期偏好图

Fig.4　Short-term preference diagram

由图5可知,在用户的第8~11次交互时,策略网络预测为长期偏好.用户的第6次交互对其影响最大,对应颜色也越深,即为用户的长期偏好.对应于Amazon Instant Video数据集中的实例是该用户当看了一部之前没看过的喜剧类型电影(第6次交互)后,又连续看了几部该类型(第8~11次交互)但互相关联不大的电影.

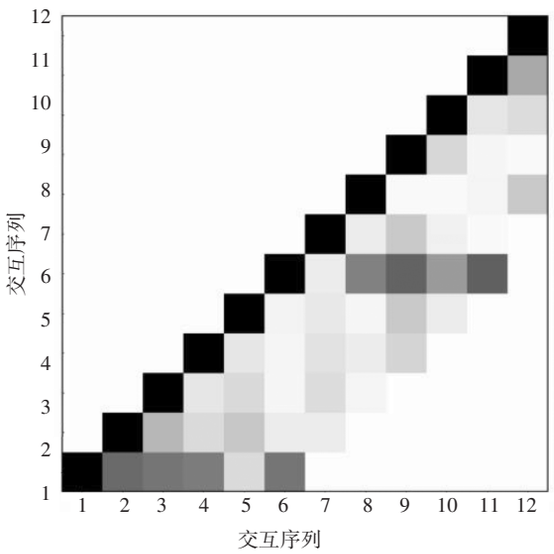

图5　长期偏好图

Fig.5　Long-term preference diagram

由图6可知,在用户的第6~9次交互时,策略网络预测为全局偏好.其之前的几次交互权重值相差不大且颜色近似,说明此次交互为用户的全局偏好.对应于Amazon Instant Video数据集中的实例是该用户当看了一部惊悚类型(第6次交互)的电影之后,又看了喜剧、爱情、传记类型的电影.

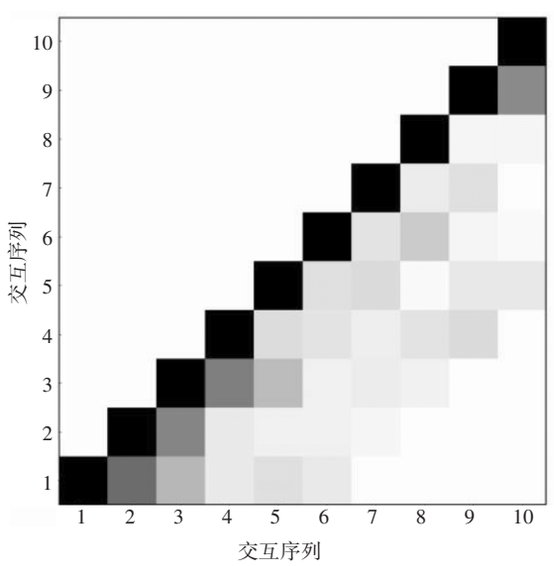

图6　全局偏好图

Fig.6　Global preference diagram

4 结　论

本文研究了结合用户策略记忆和深度强化学习

的序列推荐算法,提出一个新的DRRM模型框架.该算法通过策略网络对用户与项目交互的行为模式进行更加细致的划分,以解决用户-项目交互序列并不总是顺序相关甚至存在噪声的问题.通过对衰减因子的设定,证明深度强化学习对DRRM的影响;在消融实验中,验证了用户策略网络以及注意力机制在记忆网络中的重要性.通过在两个数据集上对比先进序列推荐模型的大量实验,证明了本文所提算法的有效性.

本文只是对数据集中的用户和项目进行矩阵分解得到相应的向量,此外还有许多可以利用的信息比如用户的身份信息、社交信息等,来更新用户、项目表示,以提高模型的可解释性.

参考文献

[1] WANG S, HU L, WANG Y, *et al*. Sequential recommender systems: challenges, progress and prospects[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Amsterdam: Elsevier, 2019: 6332-6338.

[2] MOONEY R J, ROY L. Content-based book recommending using learning for text categorization[C]// Proceedings of the 5th ACM Conference on Digital Libraries. New York: ACM, 2000: 195-204.

[3] 刘胜宗,樊晓平,廖志芳,等.基于PMF进行潜在特征因子分解的标签推荐[J].湖南大学学报(自然科学版),2015,42(10):107-113.
LIU S Z, FAN X P, LIAO Z F, *et al*. A tag recommending algorithm with latent feature factor jointly factorizing based on PMF[J]. Journal of Hunan University (Natural Sciences), 2015, 42(10): 107-113. (In Chinese)

[4] 刘朝阳, 穆朝絜, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 314-326.
LIU Z Y, MU C X, SUN C Y. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 314-326. (In Chinese)

[5] FENG S, LI X, ZENG Y, *et al*. Personalized ranking metric embedding for next new POI recommendation[C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Amsterdam: Elsevier, 2015: 2069-2075.

[6] LIU Q, WU S, WANG D Y, *et al*. Context-aware sequential recommendation[C]//Proceedings of the IEEE 16th International Conference on Data Mining. Stroudsburg: IEEE, 2016: 1053-1058.

[7] WANG S, HU L, CAO L, *et al*. Attention-based transactional context embedding for next-item recommendation[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 2532-2539.

[8] KANG W C, MCAULEY J. Self-attentive sequential recommendation[C]//Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM). Piscataway, NJ: IEEE, 2018: 197-206.

[9] WESTON J. Memory networks for recommendation[C]//Proceedings of the 11th ACM Conference on Recommender Systems. New York: ACM, 2017: 4.

[10] CHEN X, XU H T, ZHANG Y F, *et al*. Sequential recommendation with user memory networks[C]// Proceedings of the 11th ACM International Conference on Web Search and Data Mining. New York: ACM, 2018: 108-116.

[11] EBESU T, SHEN B, FANG Y. Collaborative memory network for recommendation systems[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2018: 515-524.

[12] OUYANG W T, ZHANG X W, REN S K, *et al*. Click-through rate prediction with the user memory network[C]//Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data with KDD 2019. New York: ACM, 2019: 1-4.

[13] WANG L, ZHANG W, HE X F, *et al*. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 2447-2456.

[14] ZHAO X, XIA L, ZOU L, *et al*. Model-based reinforcement learning for whole-chain recommendations[C]//Proceedings of the 13th ACM International Conference on Web Search and Data Mining. New York: ACM, 2019: 4-8.

[15] ZHENG G J, ZHANG F Z, ZHENG Z H, *et al*. DRN: a deep reinforcement learning framework for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference. New York: ACM, 2018: 167-176.

[16] LIU F, TANG R, LI X, *et al*. State representation modeling for deep reinforcement learning based recommendation [J]. Knowledge-Based Systems, 2020, 205(1): 106170.

[17] RENDLE S, FREUDENTHALER C, GANTNER Z, *et al*. BPR: bayesian personalized ranking from implicit feedback [C]//Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal: AUAI Press, 2009: 452-461.

[18] RENDLE S, FREUDENTHALER C, SCHMIDT-THIEME L. Factorizing personalized Markov chains for next-basket recommendation [C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 811-820.

[19] YU F, LIU Q, WU S, *et al*. A dynamic recurrent model for next basket recommendation [C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2016: 729-732.