Computer Science and Application 计算机科学与应用, 2023, 13(5), 1037-1043
Published Online May 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.135101

Hans汉斯

基于长短期兴趣的深度强化学习推荐模型

王世罡，牛连强

沈阳工业大学软件学院，辽宁 沈阳

收稿日期：2023年4月19日；录用日期：2023年5月18日；发布日期：2023年5月25日

摘　要

现有的基于深度学习的推荐模型将推荐过程视为静态过程，在一段时间内使用固定策略进行推荐，难以动态捕捉用户兴趣变化，影响推荐结果的准确性。本文提出了一个利用深度强化学习动态地对推荐过程进行建模的推荐模型，模型以最大化长远收益为目标，通过分别提取长短期序列中的特征信息对用户兴趣进行描述，根据兴趣变化不断改变推荐策略。在Movielens-1m数据集上的实验结果表明，相较于其他基线模型，本文模型可在precision@10和recall@10上分别提升1.7%~7.6%和1.5%~3.8%。

关键词

推荐模型，长短期兴趣，深度强化学习，深度因子分解机，自注意力模型

A Deep Reinforcement Learning Recommendation Model Based on Long and Short Term Interest

Shigang Wang, Lianqiang Niu

School of Software, Shenyang University of Technology, Shenyang Liaoning

Received: Apr. 19th, 2023; accepted: May 18th, 2023; published: May 25th, 2023

Abstract

The existing deep learning based recommendation models treat the recommendation process as a static process and use fixed strategies for recommendation over a period of time, which makes it difficult to dynamically capture changes in user interests and affects the accuracy of recommendation results. This article proposes a recommendation model that utilizes deep reinforcement learning to dynamically model the recommendation process. The model aims to maximize long-term benefits and describes user interests by extracting feature information from long and short term

文章引用: 王世罡, 牛连强. 基于长短期兴趣的深度强化学习推荐模型[J]. 计算机科学与应用, 2023, 13(5): 1037-1043.
DOI: 10.12677/csa.2023.135101

sequences. The recommendation strategy is constantly changed according to changes in interest. The experimental results on the Movielens-1m dataset indicate that compared to other baseline models, our model can be applied in precision@10 and recall@10 Increased by 1.7%~7.6% and 1.5%~3.8% respectively.

Keywords

Recommendation Model, Short and Long-Term Interests, Deep reinforcement Learning, Deep Factor Decomposition Machine, Self Attention Model

1. 引言

个性化推荐模型通过分析用户的过往行为记录，对海量信息进行过滤来达到“千人千面”的推荐效果，可以有效缓解大数据时代下的信息过载问题[1]。为了提高推荐的精准度，目前的深度学习模型主要采取了两种策略，部分推荐模型侧重于将用户历史交互记录视为一个集合，通过用户特征与物品特征提取用户长期兴趣[2]-[7]，处理简单，但忽略了用户短期内的兴趣变化对推荐结果造成的影响，如用户近期可能面临搬家或旅游等特殊情况。此时，短期兴趣与长期兴趣产生较大差别，导致长期兴趣被短期兴趣掩盖，产生的推荐结果并不可靠。

基于序列的深度学习推荐模型将用户的历史交互记录视为序列，侧重于根据用户近期交互记录提取包含在其中的序列信息，以捕捉用户的短期兴趣[8] [9] [10] [11] [12]。此类方法容易陷入局部最优状态，跳出用户当前的兴趣点困难，易导致严重的有偏推荐。

除了不能准确描述用户兴趣，上述两种模型都将推荐过程视为静态过程，而推荐过程是一个用户与推荐系统不断交互的过程，推荐系统应该根据用户的兴趣变化改变自身策略，而基于深度学习的推荐模型根据训练好的模型进行推荐，无法对用户的兴趣变化做出调整。

本文构建一种对长短期兴趣分别建模、提取的深度强化学习推荐模型 LSRWRL (Long-Short Recommendation with Reinforcement Learning)。该模型将拼接后的长短期兴趣作为深度 Q 网络[13]的输入，进而模拟用户与推荐系统的交互过程，使模型能关注到长远收益。

2. 模型梗概

2.1. 问题定义

定义在一个推荐系统中，记用户在 t 时刻访问过的物品集合为 s_t，推荐系统根据 s_t 为用户给出一个推荐结果 a_t，这个推荐结果 a_t 可以使推荐系统可以获得最大的累积长远收益 $Reward_t$，定义为式(1)。

$$Reward_t = \sum_{i=1}^{n+1} \gamma^{i-1} r_{t+i} \tag{1}$$

其中，r_t 为 t 时刻用户对推荐结果 a_t 作出的反馈，即收益值，γ 为衰减因子，每经过一个时刻，收益值就衰减到原来的 γ 倍。

2.2. 模型结构

本文将推荐系统视作智能体，将推荐过程视为马尔科夫决策过程，将用户视为强化学习中的环境，

进而由深度因子分解机模块、自注意力模块以及深度 Q 网络构成一个整体模型。

首先，将用户的历史交互序列的最近 n 个物品从序列中分割出来作为短期序列，这里将 n 定义为 20，长期序列为序列中用户所有喜欢的物品。对于不足长度 n 的序列，在前端使用 padding 操作进行填充。将长短期序列分别进行 one-hot 编码和 embedding 嵌入。其次，将长期序列输入到深度因子分解机模型提取长期兴趣特征，短期序列输入到自注意力模型中提取短期兴趣特征。最后，将两部分兴趣拼接送入深度 Q 网络，得到每个物品的 Q 值，以此进行推荐并得到反馈。如此往复直到智能体到达最终状态。图 1 说明了模型的整体结构。

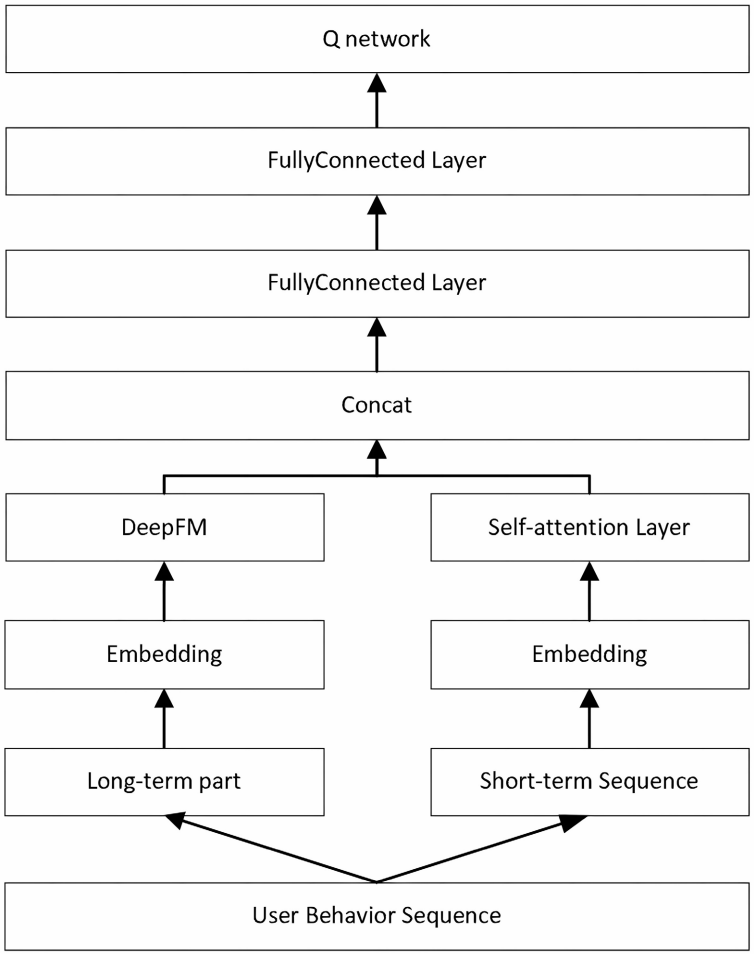

Figure 1. Overall structure of the model
图 1. 模型整体结构

3. 任务与模型

3.1. 短期兴趣提取

短期兴趣提取的自注意力模型以最近 n 个物品的特征嵌入及对应的位置编码作为输入，其输出向量考虑了整个输入序列的信息。

位置编码 p_{pos} 以式(2)的方式嵌入到特征向量 $\boldsymbol{x}$ 中。

$$\boldsymbol{x} = \left(x_1 + p_1, x_2 + p_2, \cdots, x_n + p_n\right) \tag{2}$$

p_{pos} 的计算方式如式(3)所示：

$$p_{pos}=\begin{cases}\sin\left(\dfrac{pos}{10000^{\frac{2i}{d}}}\right), & \text{if } pos \text{ is odd}\\ \cos\left(\dfrac{pos}{10000^{\frac{2i+1}{d}}}\right), & \text{if } pos \text{ is even}\end{cases} \tag{3}$$

其中，d 为输出向量的维度，i 为输入序列中当前位置的下标。

位置编码嵌入后，为每个位置上的特征向量 $\boldsymbol{X}$ 乘上三个可训练的权重矩阵 $\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V$，生成式(4)所示的查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。

$$\boldsymbol{Q}=\boldsymbol{W}_Q\boldsymbol{X},\boldsymbol{K}=\boldsymbol{W}_K\boldsymbol{X},\boldsymbol{V}=\boldsymbol{W}_V\boldsymbol{X} \tag{4}$$

由式(5)得到自注意力模型的输出，其中 d_k 为查询向量的维度。

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

对自注意力模型输出结果进行平均池化操作，使之转换成能被全连接层接受的形式，作为用户的短期兴趣。

3.2. 长期兴趣提取

长期兴趣由两部分决定，用户本身的特征和长期序列中物品的特征。采用平均池化的方式对物品特征进行提取，与用户特征拼接后输入到 DeepFM (深度因子分解机模型)中进行特征组合。长期兴趣提取模型如图 2 所示：

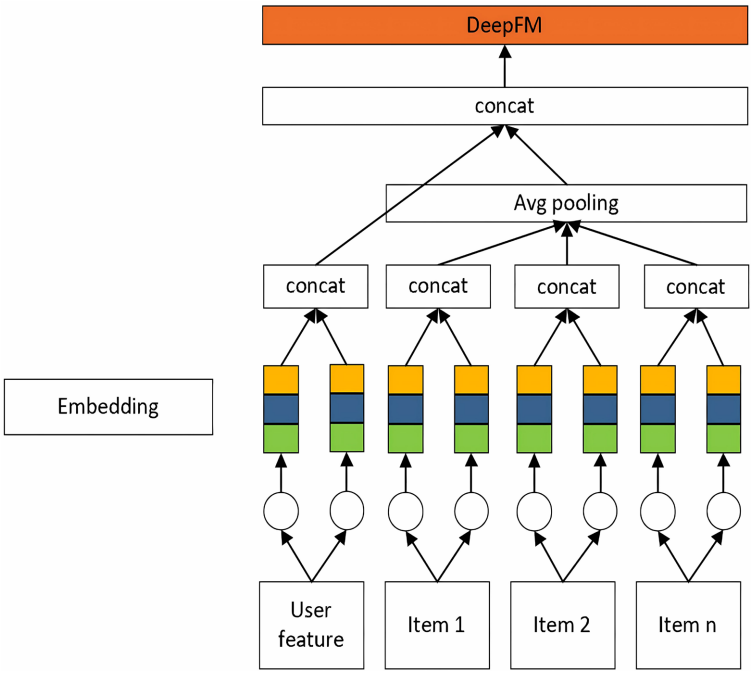

Figure 2. Long term interest extraction
图 2. 长期兴趣提取

DeepFM 的输出由 FM [14]和 DNN 两部分组成。FM 部分的输出表示为式(6)。

$$y_{\mathrm{FM}} = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \left\langle v_i, v_j \right\rangle x_i x_j \tag{6}$$

其中，$\boldsymbol{w}$ 为权重向量，$\boldsymbol{x}$ 为特征向量，d 为输入的维度，$\left\langle v_i, v_j \right\rangle$ 是特征隐向量的内积，x_i 和 x_j 为第 i 和第 j 个一阶特征，点乘后得到二阶特征组合。

DNN 部分的输出 y_{DNN} 表示为式(7)。

$$y_{\mathrm{DNN}} = \sigma\left(\boldsymbol{w}^{l+1}\boldsymbol{a}^{l} + \boldsymbol{b}^{l+1}\right) \tag{7}$$

其中，σ 为激活函数，l 为隐藏层的层数，$\boldsymbol{a}^{l}$ 为第 l 层的输出，$\boldsymbol{w}^{l}$ 和 $\boldsymbol{b}^{l}$ 为模型的权重和偏置项。

两部分的相加结果为式(8)，作为用户的长期兴趣。

$$y_{\mathrm{DeepFM}} = y_{\mathrm{FM}} + y_{\mathrm{DNN}} \tag{8}$$

3.3. 基于深度 Q 网络的最优推荐策略学习

基于深度强化学习的推荐系统将推荐系统看作智能体，将用户看作环境。Q 网络以用户长期兴趣与短期兴趣的拼接结果作为输入，在得到每个物品的 Q 值并获得用户反馈后，不断改变推荐策略，直到到达最终状态。

深度 Q 网络分为当前 Q 网络和目标 Q 网络。当前 Q 网络负责在前向传播过程中计算 Q 值，而目标 Q 网络只参与计算反向传播过程中下一个状态的 Q 值。迭代遵循如下流程：

1) 初始化经验回放池，用于提供独立同分布的数据分批次训练 Q 网络。

2) 初始化当前 Q 网络和目标 Q 网络的参数。每一回合中，对每位用户重复多次推荐，根据当前的状态以 ε-greedy 策略选择物品进行推荐并获得奖励。如果用户喜欢推荐结果，奖励值根据相应的评分设置，并更新到下一个状态 s_{t+1}。否则，奖励值为 0 且不更新状态。

将四元组 $\left\langle s_t, a_t, r_t, s_{t+1} \right\rangle$ 存储到经验回放池。

3) 每隔一定时间步从经验回放池中批量采样 m 组样本用来学习并更新参数，使用损失函数更新当前 Q 网络的参数。如果当前状态不是终止状态，由式(9)计算目标 Q 网络值。否则目标 Q 网络的值由当前奖励 r_t 决定。

$$y = r_t + \gamma \max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}\right) \tag{9}$$

损失函数采用式(10)所示的均方差损失函数。

$$loss = \frac{1}{m} \sum_{j=1}^{m} \left(y - Q\left(s_j, a_j\right)\right)^2 \tag{10}$$

其中 y 为目标 Q 网络计算得到的最优 Q 值，用于参数更新时的目标值。$Q\left(s_j, a_j\right)$ 为当前 Q 网络的 Q 值。由于每次都使用目标 Q 网络中下一个状态的最优 Q 值作为目标值，通过不断缩小当前 Q 值与目标 Q 值的差距，可确保当前 Q 网络参数朝着获取最大奖励值的方向改变，从而学习到最优的推荐策略。

4. 实验与分析

4.1. 实验准备

本文的实验环境为 Windows11、Intel i5-11320H @3.20GHZ4 核 8 线程 16 G 内存。采用 Movielens-1m 和 Movielens-100k 电影推荐数据集，按照 80%~20%比例对训练集和测试集进行划分。

实验比较的模型包括：

UserCF：基于用户的协同过滤模型，根据用户历史行为记录找到相似用户并进行推荐。

ItemCF：基于物品的协同过滤模型，根据用户历史行为记录计算物品之间的相似度并进行推荐。

SASREC [9]：采用多层自注意力的长期兴趣的序列推荐。

STAMP [10]：采用注意力机制的短期兴趣优先的序列推荐。

GRU4REC [12]：采用 GRU 模型捕捉短期兴趣的序列推荐。

本模型设置 1000 回合训练，参数设置包括 batch_size 为 128，e-greedy 参数为 0.9，奖励衰减因子为 0.9，学习率为 0.01，每隔 50 个时间步更新一次目标 Q 网络参数。

实验采用准确率 precision 和召回率 recall [15]作为实验指标。

4.2. 实验结果及分析

实验结果如表1所示。

Table 1. Experiment result
表 1. 实验结果

Model	Movielens-1m		Movielens-100k	
	precision@10	recall@10	precision@10	recall@10
ItemCF	0.176	0.098	0.131	0.072
UserCF	0.167	0.108	0.142	0.087
SASREC	0.226	0.125	0.203	0.119
STAMP	0.220	0.116	0.191	0.110
GRU4REC	0.214	0.121	0.178	0.115
LSRWRL	0.243	0.136	0.207	0.121

实验结果显示，本文模型的 precision 和 recall 指标均有提升。其中，基于内容和基于用户的协同过滤模型不涉及用户和物品的特征，在两项指标上均表现较差，说明考虑特征信息对推荐有所帮助。SASREC 模型 precision@10 和 recall@10 两项指标上表现较 STAMP 模型和 GRU4REC 模型更好，说明自注意力模型较注意力模型和 GRU 模型可以更有效地提取序列特征，但这些基于深度学习的推荐模型不能关注到用户的兴趣变化。相比之下，本文模型综合考虑了长短期兴趣，不仅准确捕捉了用户兴趣，还通过强化学习动态捕捉了用户的兴趣变化，因此取得了更好的效果。

5. 结论

为了消除基于深度学习的推荐模型不能捕捉用户兴趣变化的劣势，引入了强化学习方法，动态地对推荐过程进行建模。同时，为了能够充分体现用户短期兴趣的作用，分别对长短期序列进行处理并融合。实验表明，二者的分离可以有效利用长短期兴趣特征之间的数据分布的差异性，更好地体现用户的兴趣变化。同时，通过考虑用户的长远受益更有助于得到精准的推荐策略。

参考文献

[1] 于蒙, 何文涛, 周绪川, 等. 推荐模型综述[J]. 计算机应用, 2022, 42(6): 1898-1913.

[2] Covington, P., Adams, J. and Sargin, E. (2016) Deep Neural Networks for Youtube Recommendations. *Proceedings of the* 10*th ACM Conference on Recommender Systems*, Boston, 15-19 September 2016, 191-198. https://doi.org/10.1145/2959100.2959190

[3] He, X., Liao, L., Zhang, H., *et al*. (2017) Neural Collaborative Filtering. *Proceedings of the* 26*th International Conference on World Wide Web*, Perth, 3-7 May 2017, 173-182. https://doi.org/10.1145/3038912.3052569

[4] Zhou, G., Zhu, X., Song, C., *et al*. (2018) Deep Interest Network for Click-Through Rate Prediction. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Cambridge, July 2018, 1059-1068. https://doi.org/10.1145/3219819.3219823

[5] Chen, H.T., Koc, L., Harmsen, J., *et al*. (2016) Wide & Deep Learning for Recommender Systems. *Proceedings of the* 1*st Workshop on Deep Learning for Recommender Systems*, Boston, 15 September 2016, 7-10. https://doi.org/10.1145/2988450.2988454

[6] Guo, H., Tang, R., Ye, Y., *et al*. (2017) DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction. *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 1725-1731. https://doi.org/10.24963/ijcai.2017/239

[7] Zheng, L., Noroozi, V. and Yu, P.S. (2017) Joint Deep Modeling of Users and Items Using Reviews for Recommendation. *Proceedings of the* 10*th ACM International Conference on Web Search and Data Mining*, Cambridge, 6-10 February 2017, 425-434. https://doi.org/10.1145/3018661.3018665

[8] Li, J., Wang, Y. and McAuley, J. (2020) Time Interval Aware Self-Attention for Sequential Recommendation. *Proceedings of the* 13*th International Conference on Web Search and Data Mining*, Houston, 3-7 February 2020, 322-330. https://doi.org/10.1145/3336191.3371786

[9] Kang, W.C. and McAuley, J. (2018) Self-Attentive Sequential Recommendation. 2018 *IEEE International Conference on Data Mining* (*ICDM*), Singapore, 17-20 November 2018, 197-206. https://doi.org/10.1109/ICDM.2018.00035

[10] Liu, Q., Zeng, Y., Mokhosi, R., *et al*. (2018) STAMP: Short-Term Attention/Memory Priority Model for Session-Based Recommendation. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1831-1839. https://doi.org/10.1145/3219819.3219950

[11] Sun, F., Liu, J., Wu, J., *et al*. (2019) BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, 3-7 November 2019, 1441-1450. https://doi.org/10.1145/3357384.3357895

[12] Hidasi, B., Karatzoglou, A., Baltrunas, L., *et al*. (2015) Session-Based Recommendations with Recurrent Neural Networks.

[13] Mnih, V., Kavukcuoglu, K., Silver, D., *et al*. (2013) Playing Atari with Deep Reinforcement Learning. *Computer Science*, **10**, 431-439.

[14] Rendle, S. (2010) Factorization Machines. 2010 *IEEE International Conference on Data Mining*, Sydney, 13-17 December 2010, 995-1000. https://doi.org/10.1109/ICDM.2010.127

[15] 王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7): 66-76.