文章编号：2095-2163(2020)08-0100-05 中图分类号：TP391 文献标志码：A

基于深度学习的协同过滤推荐算法

刘 航，李锡祚
（大连民族大学 计算机科学与工程学院，辽宁 大连 116000）

摘 要：利用深度学习在特征提取方面的优势，挖掘嵌入用户和项目信息中的隐藏信息，改善传统协同过滤算法中存在的数据稀疏性及冷启动问题，将提取到的特征信息运用协同过滤算法评分预测，并且考虑用户的兴趣漂移情况和物品流行度情况，增加用户时间偏置和项目时间偏置，使算法具有实时性。最后与多种算法进行对比实验，通过计算 RMSE，评估算法的可行性与有效性。实验结果表明，基于深度学习的协同过滤推荐算法可行有效，能缓解传统协同过滤算法中存在的数据稀疏性、冷启动问题，具有实时性，提高推荐准确率，具有良好的推荐效果。

关键词：深度学习；协同过滤；数据稀疏性；冷启动；实时性

Collaborative filtering recommendation algorithm based on deep learning

LIU Hang, LI Xizuo
(College of Computer Science and Engineering, Dalian Minzu University, Dalian 116000, Liaoning, China)

【Abstract】 The advantages of deep learning in feature extraction is utilized in mining hidden information embedded in user and item information. The data sparsity and cold start problems in traditional collaborative filtering algorithms is alleviated by using the extracted feature information in score predictions and considering the user´s interest drift situation and item popularity situation to increase the user time offset and project time offset, so that the algorithm can run in real time. Finally, a comparison experiment with various algorithms is carried out, and the feasibility and effectiveness of the algorithm are evaluated by calculating the RMSE. The experimental results show that the collaborative filtering recommendation algorithm based on deep learning is feasible and effective, and can alleviate the data sparsity and cold start problems in the traditional collaborative filtering algorithm with real-time performance, which improves the recommendation accuracy, and has a good recommendation effect.

【Key words】 deep learning; collaborative filtering; data-sparsity; cold-start; real-time

0 引 言

近年来，人工智能成为家喻户晓的名词，其中推荐系统更是深入到人们衣食住行方方面面当中。如：优酷、爱奇艺的影视剧推荐，QQ 音乐、网易云音乐的歌曲推荐、美团外卖的美食推荐，抖音、快手的短视频推荐，今日头条、QQ 看点、百度的文章推荐，淘宝、亚马逊的商品推荐等等，为大数据时代提供便利。

协同过滤推荐算法是推荐系统中应用最早和最为成功的技术之一。通过分析用户的行为数据，找出近邻用户并依靠近邻用户的行为数据为目标用户奉送推荐，能够对难以进行内容分析的物品进行过滤，可以发现用户的潜在喜好。协同过滤算法虽然应用最为广泛，但是也存在缺陷，如：数据稀疏问题、冷启动问题和实时性问题。

数据稀疏问题是由于待处理的数据规模大（用户和物品的数量多），而用户评价过的物品及用户间重叠评价过的物品数量较少，导致很难准确找到用户的相似用户，推荐效果不理想；冷启动问题是对于新注册的用户或者新上市的软件，没有新用户的行为信息，无法找到他的相似用户为其推荐；用户的兴趣和物品的流行度都会随时间的推移发生变化，使算法具有实时性会大大提高推荐效果。

深度学习在特征挖掘和特征表示上有着强大的功能，已经成功应用于计算机视觉、语音识别、自然语言处理等。本文提出一种混合深度学习和协同过滤的推荐算法，缓解传统协同过滤算法存在的数据稀疏、冷启动问题，并使算法具有实时性，提高推荐的准确率。

1 国内外相关研究概况

1998 年，Amazon 上线了协同过滤推荐算法，使销售额提高了 35%；2016 年，YouTube 将深度学习应用于视频推荐，取得良好效果。目前，推荐算法在学术界及商业界均获得了广泛的关注与发展。有文

作者简介：刘 航（1993-），女，硕士研究生，主要研究方向：机器学习和推荐系统；李锡祚（1963-），男，博士，教授，硕士生导师，主要研究方向：机器学习和推荐系统。
通讯作者：李锡祚 Email：lixizuo@163.com

收稿日期：2020-05-20

献针对用户冷启动和扩展性问题，提出一种融合用户特征优化聚类的协同过滤算法[1]；有文献提出协同过滤混合填充算法，分别从用户和物品角度出发，两次填充稀疏矩阵，缓解数据稀疏问题[2]。

在基于深度学习的推荐系统方面，有文献对近几年基于深度学习的推荐系统研究进行综述，分析其与传统推荐系统的区别以及优势，并对其主要的研究方向、应用进展等进行概括、比较和分析[3]；有文献使用真实的电影数据进行实验，与另外四种优秀算法对比，证明深度学习可以真实有效得解决由于数据稀疏使得性能降低的问题，并提高推荐的准确度[4]；有文献集成了长短期记忆网络 LSTM 和概率矩阵分解 PMF，基于用户评分学习用户特征，深度挖掘辅助信息，学习更精确的物品特征[5]。有文献提出一种新式混合神经网络模型，该模型由栈式降噪自编码器和深度神经网络构成，学习得到用户和项目的潜在特征向量以及用户-项目之间的交互行为模型，有效解决数据稀疏问题从而提高系统推荐质量[6]。

有文献从真实的用户行为数据中提取时间规律，通过导入时间因素改进矩阵分解的算法[7]；有文献提出了融合动态协同过滤和深度学习的推荐算法，通过时间段划分方法考虑用户的兴趣偏好以及商品的受欢迎程度随时间变化情况[8]；有文献采用 CFDP算法对项目集合聚类，并采用 Slope-One 算法数据填充，有效的缓解了数据稀疏以及冷启动的问题，并添加时间因子，使模型具有实时性[9]。

2　基于卷积神经网络的协同过滤推荐算法

本文提出的基于深度学习的协同过滤推荐算法结构体如图 1 所示，模型大体分为 CNN 网络，CNN 与协同过滤融合，动态时间偏置三部分。

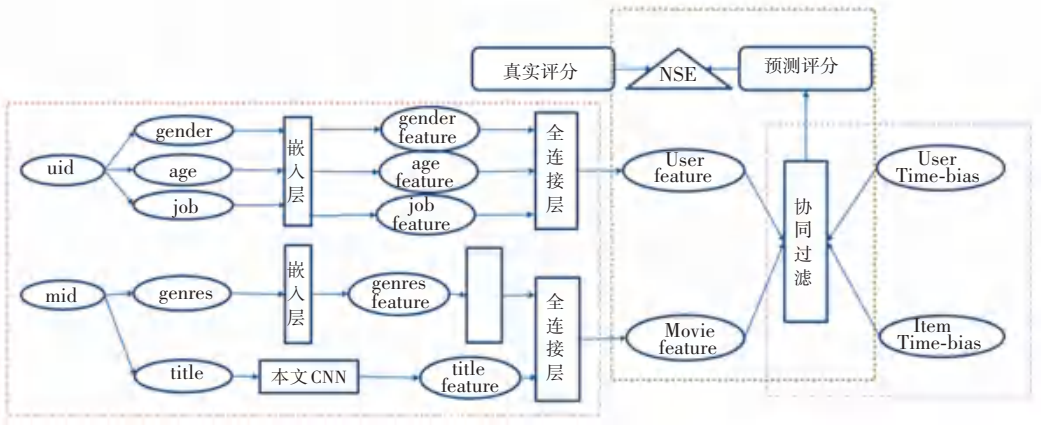


图 1　基于深度学习的协同过滤推荐算法结构图

Fig. 1　Structure chart of collaborative filtering recommendation algorithm based on deep learning

2.1　构建卷积神经网络模型

本文使用 MovieLens 1M 数据集，包含 6000 个用户在近 4000 部电影上的 1 亿条评论。数据集分为三个文件：用户数据 users.dat，电影数据 movies.dat 和评分数据 ratings.dat。

数据预处理：

(1)用户数据 users.dat 预处理，数据变化情况如图 2 所示。

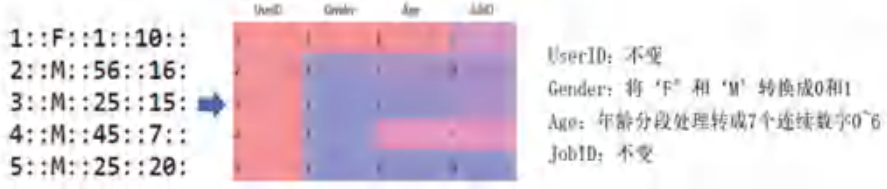
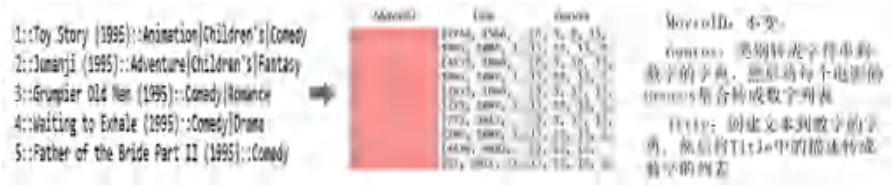


图 2　用户数据处理变化

Fig. 2　User data processing changes

(2)电影数据 movies.dat 预处理，数据变化情况如图 3 所示。

将用户各数据向量输入到卷积网络的第一层嵌入层引出特征，将各特征传入全连接层，使用 Relu 函数激活，生成用户特征矩阵。如公式(1)所示，其中 f_u 表示训练得到的用户特征矩阵，f_{relu} 表示激活函数，combine( )表示全连接：



处理前电影数据 处理后电影数据

图 3 项目数据处理变化

Fig. 3 Project data processing changes

$$f_u = combine(f_{relu}(embedding(uid), embedding(gender), embedding(age), embedding(job))). \quad (1)$$

将 MovieID、Genres 向量输入到卷积网络的第一层嵌入层引出特征;对于 Title,先使用卷积网络进行文本处理[10],输入到嵌入层得到电影名对应的各个单词的嵌入向量,对文本嵌入层使用不同尺寸的卷积核进行文本特征学习,采用最大池化生成 Title 特征。最后将各特征传入全连接层,使用 Tanh 函数激活,生成电影特征矩阵。如公式(2)(3)所示,其中 f_i 表示训练得到的项目特征矩阵,f_{tanh} 表示 Tanh 激活函数,f_{relu} 表示 Relu 激活函数,$combine()$ 表示全连接,t_i 表示 Title 文本特征,f_{pool_max} 表示最大池化,K_n 表示卷积核:

$$t_i = f_{pool\_max}(f_{relu(K_n * title)}), \quad (2)$$

$$f_i = combine(f_{tanh}(embedding(mid), embedding(genres)), t_i). \quad (3)$$

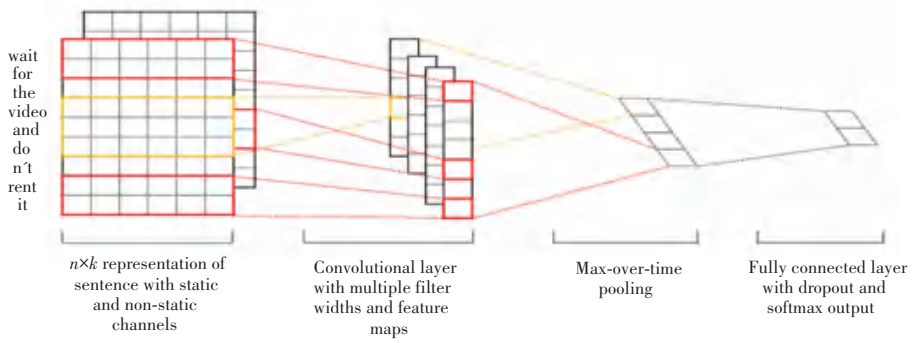


图 4 Title 卷积网络文本处理结构图

Fig. 4 Structure chart of convolution network text processing

2.2 融合深度学习的协同过滤推荐算法

通过构建卷积神经网络模型,训练得到用户特征矩阵 f_u 和项目特征矩阵 f_i,可以使用多种算法拟合评分,操作最简单的一个就是矩阵分解模型。该模型的基本思想是将用户和项目映射到一个矩阵 R 中,用矩阵的内积来表示用户对项目的预测评分。如公式(4)所示,通过计算 f_i^T 和 f_u 的内积就可以得到用户 u 对项目 i 的预测评分 $\hat{r}_{ui}$:

$$\hat{r}_{ui} = f_i^T f_u. \quad (4)$$

使用 MSE 优化损失,如公式(5)所示:

$$loss = MSE = \frac{1}{m}\sum_{i=1}^{m}(r_{ui} - \hat{r}_{ui})^2. \quad (5)$$

优化融合模型,通过增加偏置项的方法,增加推荐准确度。如公式(6)所示,增加了偏置项 b_u 和 b_i,b_u 表示用户因素对预测评分的偏置,b_i 表示项目因素对预测评分的偏置:

$$\hat{r}_{ui} = f_i^T f_u + b_i + b_u. \quad (6)$$

2.3 动态推荐算法模型

用户兴趣度和物品流行度都会随时间变化,本文通过增加动态时间权重函数,使模型具有实时性。如公式(7)所示,对于用户兴趣漂移,借助艾宾浩斯遗忘曲线,使用 Scipy 拟合时间权重函数 w_u,对于物品流行度变化,统计若干真实电影不同时期的弹幕数,得到流行度随时间的变化规律权重函数 w_i:

$$\hat{r}_{ui} = f_i^T f_u + b_i w_i + b_u w_u. \quad (7)$$

3 实验及分析

实验使用的是 MovieLens 1M 数据集,分为 3 个文件:用户数据 users.dat,电影数据 movies.dat 和评分数据 ratings.dat,包含6 040个用户在 3 652 部电影项目上的1 000 209个评分,还包含用户性别、年龄、职业、评分时间戳,电影名、电影类型等信息。

3.1　模型训练

实验经过 CNN 学习获得用户特征和项目特征并采用矩阵分解方式预测评分。使用 MSE 优化损失将计算值回归到评分，调整模型参数进行反向传播训练，如图 5 所示，多次迭代直至训练损失和测试损失收敛，获得稳定模型：

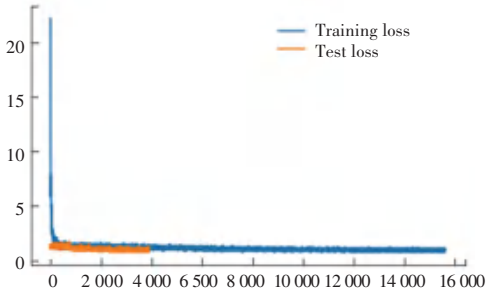


图 5　模型训练 Loss 曲线

Fig. 5　Model training loss curve

3.2　评价标准

实验采用十折交叉验证，将评分数据以 2∶8 分为测试集与训练集，使用 *RMSE* 作为评价指标。*RMSE* 是推荐系统中一种性能评价标准，衡量观测值与真实值之间的偏差，是真实值与预测值的差值的平方然后求和平均再开方。*RMSE* 值的计算如公式(8) 所示，$\hat{r}_{ui}$ 表示预测评分，r_{ui} 表示用户实际评分，m 表示 $\hat{r}_{ui}$ 或者 r_{ui} 的数量：

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(r_{ui} - \hat{r}_{ui})^2}. \tag{8}$$

RMSE 值越小说明该算法准确度越高。

3.3　实验对比

将本文提出的模型与几种具有代表性的模型在同一实验数据中进行结果对比，验证本文提出的基于深度学习的协同过滤推荐算法模型的有效度。模型列举见表 1。

表 1　几种推荐算法模型

Tab. 1　Several recommended algorithm models

模型简称	模型含义	模型功能
SVD	协同过滤推荐	缓解数据稀疏性
CNN	深度学习推荐	缓解数据稀疏性、冷启动
TimeSVD	考虑时间因素的协同过滤推荐	具有实时性
CNN-TimeSVD(本文)	基于深度学习并考虑时间因素的协同过滤推荐	缓解数据稀疏性、冷启动、具有实时性

实验采用 Python3 语言，Spyder 编译器，TensorFlow 架构，验证各模型 RMSE 结果，见表 2。

表 2　各模型 RMSE 计算结果

Tab. 2　RMSE calculation results of each model

模型	SVD	CNN	TimeSVD	CNN-TimeSVD (本文)
RMSE 值	0.928	0.885	0.916	0.872

为更直观展示，将结果制作成柱状图，如图 6 所示。

将参考文献中功能相近的模型与本文提出的模型在同一实验数据中进行结果对比，模型列举见表 3。

各模型 RMSE 结果见表 4。

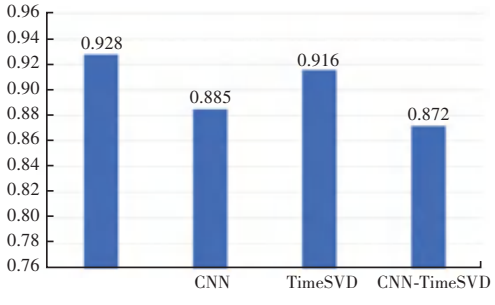


图 6　各模型 RMSE 值比较

Fig. 6　RMSE value comparison of each model

表 3　参考文献模型

Tab. 3　Reference model

模型简称	模型含义及功能
文献[9]CFDP 算法	通过项目聚类缓解数据稀疏性、冷启动，添加时间因子迎合用户兴趣漂移
文献[5]LSTM+PMF 算法	PMF 模型缓解数据稀疏性、冷启动，LSTM 模型增添实时性
本文 CNN-TimeSVD 算法	CNN 模型缓解数据稀疏性、冷启动，融合时间偏置考虑用户兴趣漂移问题和物品流行度问题，使模型具有实时性

表 4 各模型 RMSE 计算结果

Tab. 4 RMSE calculation results of each model

模型	CFDP	LSTM+PM	CNN-TimeSVD(本文)
RMSE 值	0.986	0.905	0.872

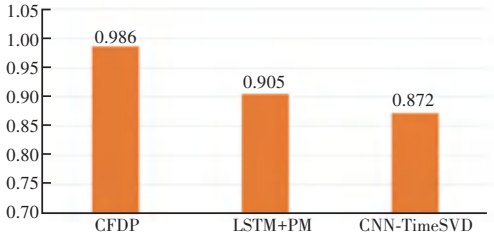


图 7 各模型 RMSE 值比较

Fig. 7 RMSE value comparison of each model

4 结束语

本文提出了一种基于深度学习的协同过滤推荐算法,能缓解传统协同过滤算法存在的数据稀疏性、冷启动问题;考虑用户兴趣漂移和物品的流行度,增加时间偏置使模型具有实时性。在 MovieLens 1M 数据集进行对比实验 RMSE 值均低于传统协同过滤算法、单纯深度学习算法、单纯实时性协同过滤算法,证明了将深度学习与协同过滤算法融合并增添时间信息的可行性和有效性。将本文算法与参考文献中功能相近算法进行实验对比,本文算法的 RMSE 值均低于其他算法,表明本文提出的基于深度学习的协同过滤推荐算法能提高推荐准确率,提升推荐效果。

参考文献

[1] 梁丽君,李业刚,张娜娜,张晓,王栋.融合用户特征优化聚类的协同过滤算法[J/OL]. 智能系统学报,2020,3:1-6.
[2] 任永功,王思雨,张志鹏.缓解数据稀疏问题的协同过滤混合填充算法[J]. 模式识别与人工智能,2020,33(2):166-175.
[3] 黄立威,江碧涛,吕守业,刘艳博,李德毅.基于深度学习的推荐系统研究综述[J]. 计算机学报,2018,41(7):1619-1647.
[4] 冯楚滢,司徒国强,倪玮隆.协同深度学习推荐算法研究[J]. 计算机系统应用,2019,28(1):169-175.
[5] 曾安,赵恢真.融合了 LSTM 和 PMF 的推荐算法[J]. 计算机工程与应用,2020,56(19):68-75.
[6] 张杰,付立军,刘俊明.基于混合自编码器的协同过滤推荐算法优化[J]. 计算机系统应用,2019,28(5):161-166.
[7] 赵蝶祥.基于时间因素影响的矩阵分解算法的研究[D]. 内蒙古大学,2017.
[8] 邓存彬,虞慧群,范贵生.融合动态协同过滤和深度学习的推荐算法[J]. 计算机科学,2019,46(8):28-34.
[9] 张凯辉,周志平,赵卫东.结合 CFDP 与时间因子的协同过滤推荐算法[J]. 计算机工程与应用,2020,56(15):80-85.
[10] Yoon Kim. Convolutional Neural Networks for Sentence Classification[J]. Computation and Language,2014,8:85-91.

(上接第 99 页)

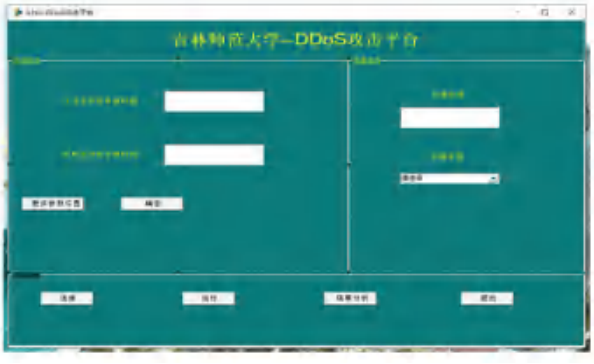


图 2 基于 IPv6 的模拟集成 DDoS 攻击平台 GNS3 模型

Fig. 2 The GNS3 model of simulate integrated DDoS attack platform depend on IPv6

5 结束语

本文分析了 IPv6 的特点(包括 IPv6 的现状及先进性)及 DDoS 的特点(包括 DDoS 的分类、现象和原理),对 IPv6 面临的 DDoS 威胁进行了研究,利用基于树状结构的动态分布式网络模型,并结合插件技术,设计了基于 IPv6 下的模拟集成 DDoS 攻击平台,展示了平台的 GNS3 模型。在 IPv6 逐渐普及的今天,此平台的设计具有很强的研究意义。

参考文献

[1] 薛晓敏. 基于 IPv6 的协议解析和 DoS/DDoS 攻击检测[D]. 暨南大学,2007:1-67.
[2] 中国 IPv6 发展状况 https://network.51cto.com/art/201907/599898.htm
[3] 雷振洲. 全球 IPv6 的发展状况[J]. 信息技术与标准化, 2003(7):19-21.
[4] Keith Barker. The security implications of IPv6[J]. Network Security,2013,2013(6).
[5] 张连成,郭毅. IPv6 网络安全威胁分析[J]. 信息通信技术,2019,13(6):7-14.
[6] 王麦玲. 分布式拒绝服务攻击与防范措施[J]. 办公自动化:综合版(9):42-43.
[7] 曹洪英. 基于 IPv6 的集成 DDoS 攻击平台设计与部分实现[D]. 北京邮电大学,2008:1-67.
[8] Seo Jung Woo, Lee Sang Jin. A study on efficient detection of network-based IP spoofing DDoS and malware-infected Systems.[J]. SpringerPlus,2016,5(1).
[9] 张永铮,肖军,云晓春,等. DDoS 攻击检测和控制方法[J]. 软件学报,2012,23(8):2058-2072.
[10] 张佳欣. DDoS 攻击检测与跟踪方法研究[D]. 哈尔滨理工大学,2019.
[11] 杨明明. IPv6 下的 DDoS 防御研究[D]. 重庆大学,2010:1-56.
[12] 姜川.计算机软件中的插件技术及应用研究[J]. 数字技术与应用,2013(1):94.
[13] 李帷笳.计算机软件技术中插件技术的应用研究[J]. 电脑知识与技术,2019,15(28):60-61.