文章编号：2095-2163(2020)12-0209-05 中图分类号：TP391 文献标志码：A

基于多模态特征融合的个性化视频推荐方法

谭 晓[1,2]，孙全明[2]，曲志坚[2]
（1 山东水利技师学院 信息技术系，山东 淄博 255130；2 山东理工大学 计算机科学与技术学院，山东 淄博 255049）

摘 要：为了充分利用用户历史行为数据的结构化特征，提高视频个性化推荐效果，本文提出了一种基于多模态特征融合的视频个性化推荐方法。通过 Word2Vec 提取视频的词向量特征，并将视频数据从高维空间映射到低维稠密空间；提取视频图像特征以及文本特征并与结构化特征进行融合，共同完成视频推荐任务。通过融合 LightGBM 和 DeepFM 构建推荐模型，该融合模型既具有在连续特征上的学习能力，也拥有高阶特征组合的泛化能力。该方法能够更好的挖掘用户偏好，提高模型推荐的准确性。

关键词：多模态特征融合；视频推荐；词向量；用户偏好

Personalized Video Recommendation Method Based on Multimodal Representation Learning

TAN Xiao[1,2], SUN Quanming[2], QU Zhijian[2]
(1 Department of Information Technology, Water Conservancy of Shandong Technician College, Zibo Shandong 255130, China; 2 School of Computer Science and Technology, Shandong University of Technology, Zibo Shandong 255049, China)

【Abstract】 To make full use of the structural features of user historical behavior data and improve the effect of personalized video recommendation, a personalized video recommendation method based on multi-modal feature fusion is proposed. The word vector features of the video were extracted by applying the Word-to-Vector, so that the video data was mapped from high-dimensional space to low-dimensional dense space. Then the video image features and text features are extracted. Then the video image features and the text features were extracted and merged with structural features. The model is trained by these merged features. The recommendation model has both the learning ability in continuous features and the generalization ability in combination of high-order features by fusing the LightGBM and DeepFM model. The proposed method can better excavate the user preferences by learning the hidden features of user sequences and improve the accuracy of recommendation.

【Key words】 Multi-model feature fusion; Video recommendation; Word vector; User preference

0 引 言

随着移动互联网的快速发展以及智能终端的广泛使用，用户可获取的视频数据和访问视频资源的终端设备越来越多，在线视频平台的用户规模和视频资源与日俱增。为了给用户提供个性化的视频服务，解决信息过载问题，视频推荐系统应运而生。视频推荐旨在为用户提供个性化、精准化的视频推荐服务。一个有效的推荐系统不仅能提高平台用户的忠诚度，而且用户的转化和留存所带来的流量还可以显著提高视频平台的收益。

目前的推荐系统能实现文字、图像、音频和视频之间的推荐，预先对不同媒体文件进行标记，产生不同的标签，然后通过搜索引擎搜索用户需求的内容。但是随着视频平台的发展，用户个性化需求越来越强烈，推荐场景也越来越复杂，对视频推荐的精准性要求也越来越高。传统的视频推荐方法主要包括基于用户的协同过滤（User-based Collaborative Filtering，UCF）、基于项目的协同过滤（Item-based Collaborative Filtering，ICF）和基于模型的协同过滤（Model-based Collaborative Filtering，MCF）等。但是，由于协同过滤方法本身采用浅层模型，无法学习到用户与视频之间的深层次特征，而且容易遇到冷启动、数据稀疏和可扩展性难等问题。

为了解决协同过滤算法存在的这些问题，学者们从多个方面进行了广泛且深入的研究，并呈现出了诸多研究成果。Sang-Min 等提出一种仅依靠电影分类信息进行个性化推荐的方法[1]，该方法首先通过计算得到电影类别间的相关性矩阵，然后根据

基金项目：山东省自然科学基金（ZR2016FM18，ZR2017LF004）；山东省高等学校青年创新团队发展计划项目（2019KJN048）。

作者简介：谭 晓（1982-），女，硕士，讲师，主要研究方向：数据分析与可视化；孙全明（1995-），男，硕士研究生，主要研究方向：机器学习与数据分析；曲志坚（1980-），男，博士，副教授，主要研究方向：机器学习。

通讯作者：谭 晓 Email：handuhandu@163.com

收稿日期：2020-10-20

用户的历史行为计算用户对所有电影类别的偏好程度,最后利用用户的类别偏好和目标电影类别评分进行推荐;Elkahky 等通过提取用户的浏览记录和搜索记录来丰富用户的特征表示,并提出了一种多视角深度神经网络模型[2],该模型试图通过用户和视频两种信息实体的语义匹配来实现用户的视频推荐;Zarzour 等提出了一种基于降维和聚类技术的协同过滤算法,利用均值聚类(K-means cluster)算法对相似用户进行聚类,利用奇异值分解(Singular Value Decomposition, SVD)对相似用户进行降维[3]。这些方法分别从特征和模型两个方面对推荐效果进行改进,但是忽略了用户随时间变化的行为特征,难以捕捉用户的兴趣变化,无法实现个性化推荐的目的。

针对个性化推荐的问题,赵楠等研究人员提出了一种面向多维特征分析过滤的视频推荐方法[4],该方法从用户行为和视频标签等多个维度对视频进行特征提取,然后进行相似性分析,利用改进的协同过滤算法实现用户个性化视频推荐;王娜等人提出了一种基于用户播放行为序列的个性化视频推荐方法[5],该方法通过提取用户观看视频的历史行为数据,构造用户行为序列,利用词向量模型将高维序列特征映射成低维稠密向量,提取视频的语义特征,并通过聚类建模用户兴趣分布矩阵,结合用户兴趣偏好和用户历史行为生成推荐列表。这类方法能够充分考虑用户的兴趣变化,从多个角度捕捉用户行为与视频之间的内在关系,并利用改进的协同过滤算法或深度学习算法实现个性化推荐。

在大数据背景下,在线视频平台可利用的数据具有多模态、规模大、数据异构等特点,而且视频是集文本、图像、音频为一体的综合体,用户点击视频与否与视频的图像、标题、简介等非结构化数据有很大关系,传统推荐方法很难从这种非结构化数据中学习到有用的知识,进而影响推荐的准确性。基于此本文提出了一种多模态特征融合的个性化视频推荐方法,该方法不仅考虑用户行为特征与视频标签特征,还利用 Word2Vec 分析用户点击行为数据,将视频从高维空间映射到低维稠密空间,提取视频的词向量特征。同时,考虑用户点击视频的行为与视频封面图像和简介之间有直接关系,故提取了视频图像特征以及文本特征,与结构化特征进行融合,共同训练完成视频推荐任务。该方法充分利用不同模态之间的差异性和相似性,增加模型对用户行为的理解能力,提高了视频推荐的效果。

1 数据描述与分析

1.1 数据描述

本文采用由芒果 TV 提供的 30 天线上 feed 流产品用户行为数据集。数据集包括:用户侧历史行为数据、视频侧结构化数据、交互上下文数据以及媒资原始封面和标题描述原始特征数据,总量接近一亿条样本。

用户侧历史行为数据是由用户编号和用户观看记录构成,每条样本表示该用户最近观看的视频及观看时间的集合,反映了用户的历史偏好。

视频侧结构化数据是由视频编号、视频属性以及交互当天视频的相关统计信息构成。

交互上下文数据是由用户与平台交互的时间信息、设备信息、地点以及是否点击视频等信息构成。

媒资原始封面和标题数据是由视频编号、视频封面和标题信息构成。用户在点击视频之前,首先关注的是视频封面和视频标题,符合用户兴趣的封面和标题可以大大提高该视频的点击率,所以如何正确利用这部分数据是提高平台推荐效果的关键。

1.2 探索性数据分析

在推荐系统中,用户是最关键的因素,新老用户的比例对推荐效果会有一定的影响。如果老用户占比大,则容易捕捉用户的兴趣偏好;反之,如果老用户占比较小,则难以学习用户兴趣偏好,造成较大的冷启动问题。

本文数据中用户访问次数分布如图 1 所示。从图 1 可以看出,用户 30 天内平均访问次数较少,呈现长尾分布。且通过进一步分析发现,数据中包含大量单次访问用户或访问次数很少的用户,对于这类用户可通过其历史观看记录或视频实际属性确定是否点击当前曝光的视频;而对于多次访问的用户,可以通过统计学方法提取用户的实际偏好。

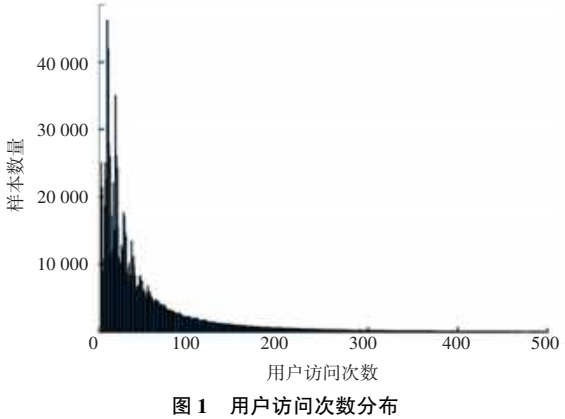

图 1 用户访问次数分布

Fig. 1 Distribution of user visits

视频曝光率与点击率分布如图 2、图 3 所示。图 2、图 3 分别描述了视频每天和每小时的曝光率与点击率分布情况。可以看出视频每天的曝光率差异较大，点击率在第 10 天和第 20 天出现加大起伏，可能是由于特殊日期造成的这种现象。视频在各个小时的曝光率具有明显的潮汐现象，凌晨 5 点～晚上 22 点呈现整体上升趋势，晚上 23 点～凌晨 4 点呈现明显的下降趋势，这符合人们的正常作息。特别是在晚上 19 点～22 点上升趋势明显，可能是由于晚间黄金档的原因造成曝光量剧增。而点击率整体趋于稳定，凌晨 5 点和晚上 20 点出现上升趋势。根据视频曝光率和点击率在时间上的差异，可分别提取用户和视频在时间维度上的统计特征表示其在时间维度上的分布。

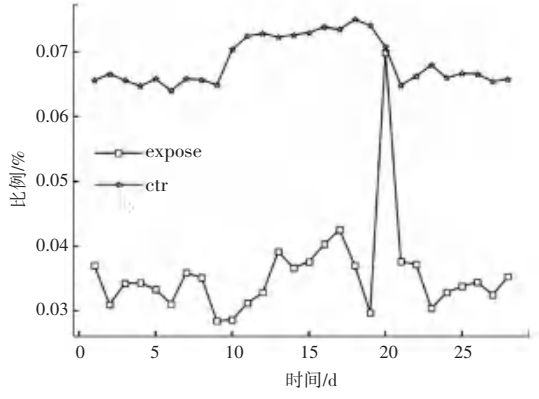

图 2　视频每天的曝光率与点击率分布

Fig. 2　Video daily exposure and click-through rate distribution

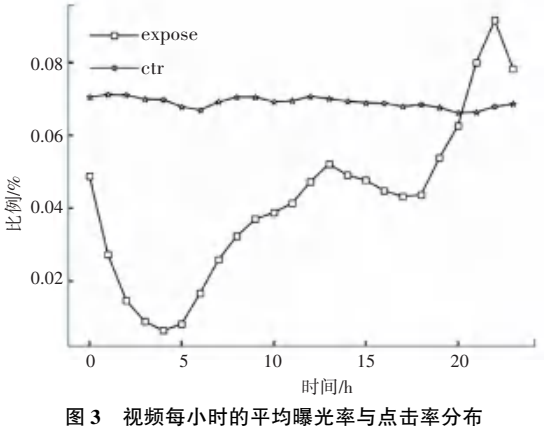

图 3　视频每小时的平均曝光率与点击率分布

Fig. 3　Video hourly average exposure and click－through rate distribution

2　方法描述

2.1　词频-逆文本频率

词频－逆文本频率（Term Frequency Inverse Document Frequency，TFIDF）是一种用于咨讯检索与咨讯探勘的加权技术，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度[6]。词频（Term frequency，TF）指对于给定词语在该文章中出现的次数，反映该词语对于文章主题的重要程度。但是一些通用词语对于文章主题预测并没有太大作用，反而一些出现频率较少的词语更能够表达文章主题，故提出利用逆文本频率（Inverse document frequency，IDF）对词语权重进行修正。逆文本频率的主要思想是如果给定词语在该文章中次数较多，但是在整个语料库中出现次数较少，则说明该词语具有很好的类别区分能力，逆文本频率越大。TFIDF 公式如式（1）～（3）所示。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1}$$

$$IDF_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|}, \tag{2}$$

$$TFIDF = TF \times IDF. \tag{3}$$

其中，$n_{i,j}$ 表示词语 i 在文章 j 中出现的次数；分母表示文章 j 中所有字词出现的次数之和；$|D|$ 表示语料库中所有文件的总数；分母表示包含词语 t_i 的文件总数；式（3）表示给定词语在该语料库中的权重。利用该方法，提取视频标题和简介的文本特征。

2.2　词向量

Word2Vec 是 Google 公司在 2013 年推出的一种用于训练词向量的工具，因为其简单、高效而受到广泛关注。Word2Vec 包含二种词向量学习模型：CBOW（Continues Bag of Words）模型和 Skip-gram（Continues Skip gram）模型[7]。模型结构如图 4、图 5 所示。

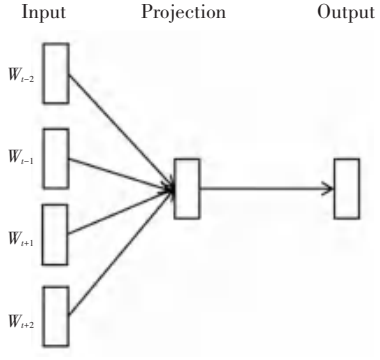

图 4　CBOW 模型

Fig. 4　CBOW model

为增加文本特征的表达能力，本文利用 Word2Vec 提取视频标题和简介的文本特征。同时，为了能更好的捕捉用户历史行为的隐式特征，提取用户历史观看记录，并将其转化为文本序列，利用 Word2Vec 方法将高维视频特征映射成低维稠密向量，与其它特征进行融合并嵌入到推荐模型中[8]。

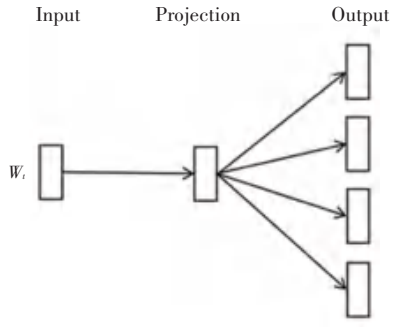

图 5 Skip-gram 模型

Fig. 5 Skip gram model

2.3 视频推荐模型

本文采用 LightGBM 和 DeepFM 融合的方法,实现视频 Top One 推荐。近年来,GBDT 在数据挖掘中得到了广泛应用,通常被用于多分类、点击率预估、搜索排序等任务。主要思想是利用弱分类器进行迭代训练以得到最优模型。LightGBM 是对 GBDT 的高效实现框架,解决了 GBDT 在面对海量数据时出现训练速度慢,内存消耗大等缺点。

DeepFM 也是一个被广泛应用在点击率预估中的深度学习模型,由浅层模型 FM 和深层模型 DNN 组成,二者分别用来提取低阶和高阶特征组合。从本文前面部分的数据分析可以发现,本文采用的数据较为稀疏且维度较大,不同用户对视频有着不同的偏好程度。利用 FM 模型可以对稀疏特征进行稠密嵌入,并使得每个特征的 Embedding 非零,增加了模型的泛化能力。本文提出模型中的 DNN 部分弥补了 FM 无法构造高阶特征的缺点,使模型具备挖掘高阶组合特征的能力,模型的具体结构如图 6 所示。

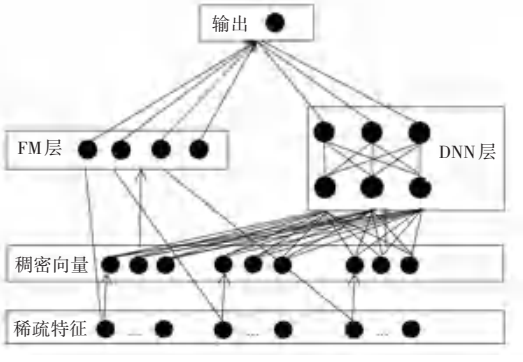

图 6 DeepFM 模型

Fig. 6 DeepFM model

2.4 多模态特征构造

多模态特征构造主要包括结构化特征、文本特征以及图像特征。

结构化特征:利用用户与平台的交互记录统计每个类别特征出现的频率作为该特征的分布。例如,用户观看过的视频类别数,视频所属集合类别数等统计特征。用户的历史观看记录能够隐式表达用户观看视频的喜好,通过统计用户观看视频的平均时长、观看视频的类别众数以及视频的点击率等特征表征用户偏好。在观看视频前,用户除了比较关注视频的类别之外,视频由那些演员参演也是用户比较关注的焦点。如果一个用户比较关注某演员,则用户大概率会点击有关该演员的相关视频。根据该业务特点,本文挖掘了有关视频演员列表的相关特征。例如,用户点击最多的演员,演员之间的相似度以及演员的 TFIDF 等统计特征。同时,由于用户行为与时间具有强相关性,所以提取用户观看视频的时间特征表征用户行为与时间的上下文关系,例如当前时间所对应的周几、几点、是否为周末、是否为节假日等特征。

文本特征:主要对视频标题、视频简介以及用户行为序列提取相应的文本特征。在结构化特征中,利用统计学方法提取视频演员列表特征。但是,由于文本数据的非结构化特点,统计特征无法表征语料的隐式特征,故采用 Word2Vec 将文本转化为词向量,将词向量嵌入到原始特征中进行训练。另外,对于视频标题特征同样采用 TFIDF 和 Word2Vec 方法进行特征提取。为了更好地捕捉用户行为特征,首先提取了用户行为记录,并将其转化为文本序列,采用 Word2Vec 方法将高维视频特征映射成低维稠密向量,嵌入到原始特征中进行训练。

图像特征:媒资原始封面和标题数据中提供了视频封面的原始特征向量,由于其维度较高,采用奇异值分解方法(Singular Value Decomposition, SVD)和主成分分析方法(Principal Component Analysis, PCA)进行降维,生成最终的图像特征。

3 实验结果分析

3.1 实验数据

实验采用芒果 TV 提供的用户侧历史行为数据、视频侧结构化数据、交互上下文数据以及媒资原始封面和标题描述原始特征数据共九千万条样本,利用前 29 天的数据为训练集,第 30 天的数据为测试集。训练集与测试集的统计信息见表 1。

表 1 训练集与测试集统计信息

Tab. 1 Statistics of the test set and train set

数据集	样本数	特征数
训练集	88270998	257
测试集	3324513	257

3.2 评价指标

在机器学习算法中,AUC 是最为常用的评价指

标，反映了模型在整体样本间的排序能力，但是在某些情况下，AUC 并不能真正反映模型的好坏。例如本文实验中，视频 Top1 推荐属于点击率预估问题，不同用户和视频之间的排序是个性化的，无法很好的进行比较，实际需要衡量的是模型在不同用户对不同视频的排序能力，这时全局 AUC 并不能反映模型的真实情况，故采用 Group AUC 作为模型的评价指标。Group AUC 首先计算的每个用户的 AUC，然后进行加权平均，这样能减少不同用户对不同视频排序结果的影响。评价指标如下所示：

$$GroupAUC = \frac{\sum_{(u,p)} w_{(u,p)} * AUC_{(u,p)}}{\sum_{(u,p)} w_{(u,p)}}. \quad (4)$$

式中，$AUC_{(u,p)}$ 表示用户 u 的 AUC 值，$w(u, p)$ 表示用户 u 的权重。在实际计算过程中，用户权重一般为用户的点击次数，并且会忽略掉单个用户全是正样本或负样本的情况。

3.3　实验结果分析

为了验证本文方法的推荐精度，选取 Logistics Regression、Random Forest、XGBoost、CatBoos、单模 LightGBM、单模 DeepFM 与本文融合模型进行比较。表 2 给出了不同模型之间的推荐效果，表 3 给出了本文多模态特征融合方法对推荐效果的影响。

表 2　不同模型的推荐效果

Tab. 2　Recommendation effect of different models

算法	AUC	GAUC
Logistic Regression	0.603	0.600
Random Forest	0.710	0.674
XGBoost	0.721	0.683
CatBoost	0.799	0.683
LightGBM	0.760	0.685
DeepFM	0.753	0.684
LightGBM+DeepFM	0.761	0.686

由表 2 可知，单模 LightGBM 和单模 DeepFM 的 AUC 指标虽然不如 CatBoost，但是在 GAUC 指标下要优于其他单一模型。与集成模型相比，Logistic Regression 效果较差，这符合预期，即集成学习模型在高维数据中更具优势。DeepFM 虽然集成了 FM 和 DNN 的优势，但是与 LightGBM 相比结果稍差。两模型融合之后，AUC 和 GAUC 都分别提升了 0.2%，表明差异化模型之间进行融合有助于提高整体的推荐效果，见表 3。

由表 3 可知，通过构造用户行为序列提取视频词向量特征与纯结构化特征相比提升了 2.6%，融合多模态特征后，模型提升了 0.5%，这表明，本文提出的方法可以有效提高推荐的准确性。

表 3　多模态特征融合对推荐效果的影响

Tab. 3　The influence of multi-modal feature fusion on recommendation effect

方法	AUC	GAUC
结构化特征	0.702	0.655
结构化特征+行为序列特征	0.752	0.681
结构化特征+行为序列特征+文本特征	0.760	0.685
结构化特征+行为序列特征+文本特征+图像特征	0.761	0.686

4　结束语

针对当前在线视频平台推荐场景复杂化，用户个性化需求越来越高的特点，本文提出了一种基于多模态特征融合的个性化视频推荐方法。为了能真实反映实际业务中多模态数据的特点，该方法不仅考虑用户历史行为的结构化特征，而且利用 Word Embedding 捕捉用户行为的隐式特征，将视频从高维空间映射到低维稠密空间，提取视频的词向量特征。同时，考虑到用户点击视频的行为与视频封面图像和简介之间有直接关系，故提取了视频图像特征以及文本特征，从图像和文本中挖掘用户兴趣偏好，并与结构化特征进行融合，共同训练完成视频推荐任务。实验证明，本文方法较其它方法提升了推荐准确度。

参考文献

[1] CHOI S M, KO S K, HAN Y S. A movie recommendation algorithm based on genre correlations [J]. Expert Systems with Applications, 2012, 39(9): 8079-8085.

[2] ELKAHKY A M, SONG Y, HE X. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems [C]// the 24th International Conference. International World Wide Web Conferences Steering Committee, 2015.

[3] ZARZOUR H, AL-SHARIF Z, AL-AYYOUB M, et al. A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques [C]// 2018 9th International Conference on Information and Communication Systems (ICICS). IEEE, 2018.

[4] 赵楠，皮文超，许长桥.一种面向多维特征分析过滤的视频推荐算法[J].计算机科学,2020,47(4):103-107.

[5] 王娜,何晓明,刘志强，等.一种基于用户播放行为序列的个性化视频推荐策略[J].计算机学报,2020,43(1):123-135.

[6] GU Y, WANG Y, HUAN J, et al. An Improved TFIDF Algorithm Based on Dual Parallel Adaptive Computing Model [C]. // 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2018.

[7] SIWEI L, KANG L, SHIZHU H, et al. How to Generate a Good Word Embedding [J]. IEEE Intelligent Systems, 2016, 31(6): 5-14.

[8] ZHUANG Z, KONG X, ELKE R, et al. Attributed Sequence Embedding[C]// 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2020.