融合项目图像的混合推荐算法

王宇飞 陈 璐
(浙江工业大学管理学院 浙江 杭州 310023)

摘 要 针对传统协同过滤推荐算法存在的数据稀疏问题,提出融合项目图像和项目评分的混合推荐算法。基于卷积神经网络提取的图像特征值计算项目图像之间的相似度,并与评分相似度动态加权后对项目进行推荐。在 MovieLens 数据集上进行的实验结果表明,改进算法有效解决了传统协同过滤算法存在数据稀疏的问题。

关键词 协同过滤 卷积神经网络 项目图像 动态加权 项目相似度

中图分类号 TP391.3 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2020.10.047

HYBRID RECOMMENDATION ALGORITHM FOR FUSED PROJECT IMAGES

Wang Yufei Chen Lu
(*School of Management, Zhejiang University of Technology, Hangzhou 310023, Zhejiang, China*)

Abstract Aiming at the sparse problem of traditional collaborative filtering recommendation algorithm, this paper proposes a hybrid recommendation algorithm for fusion project images and project scores. Based on the image feature values extracted by convolution neural network, the similarity between the project images was calculated, and the project was recommended after dynamically weighted with the similarity between the project and the score. The experimental results on the MovieLens dataset show that the improved algorithm effectively solves the problem of sparse data in the traditional collaborative filtering algorithms.

Keywords Collaborative filtering Convolutional neural network Project image Dynamic weighting Project similarity

0 引 言

随着信息时代的到来,信息过载问题愈发突出,面对海量的数据,传统信息获取工具已无法满足用户个性化的需求。推荐系统使用用户行为产生的数据来探求用户的兴趣、预测用户行为,从大量繁杂的数据中找到符合用户的项目(如信息,服务,物品等),并将其推荐给用户,不仅节约了用户筛选信息的时间,还提高了信息的利用率[1]。

传统的推荐算法[2]必须依赖用户评价信息。以电影推荐为例,由于电影的数量庞大,用户只可能对其中的一部分进行评价,因此会产生用户共同评分项目太少导致评分数据稀疏的问题,这会导致在使用用户对项目的评分计算用户或物品的相似度时,产生较大的误差。这样,就难以保证最近邻居搜索结果的准确性,使得预测值与真实值之间可能误差较大,从而影响推荐的效果。

针对传统算法中存在的问题,相关领域的研究人员使用聚类[3-4]、矩阵分解[5-6]、slope one 算法[7-8]等方法来解决这个问题。朱丽中等[9]将云模型方法引入到协同过滤算法中来,利用项目聚合产生预测结果,从而提高推荐结果的稳定性。韦素云等[10]利用项目属性构造项目相似度矩阵,并与兴趣度相结合,使用改进的公式重新衡量项目的相似度,提升推荐质量。魏甜甜等[11]在项目相似度计算时引入项目流行度,提升了推荐准确度。上述方法都通过分析项目,利用项目的不同特性提高推荐的质量,但并没有很好地将项目的内容信息提取出来加以利用。

伴随着读图时代的到来,快节奏的生活导致时间的碎片化程度加重,许多人没有整块时间去阅读文字,

收稿日期:2019-05-20。王宇飞,硕士生,主研领域:机器学习,协同过滤。陈璐,硕士生。

人们也不愿意将注意力集中在文字上。图像等类似媒体(包括短视频)替代了文字,它们能让人们以最轻松的方式快速获取信息,人们也更加愿意通过图像来了解信息。图像在一定程度上可以反映项目的内容,内容信息相近的项目,其图像往往也具有很高的相似度。本文将项目图像与评分动态组合,得到一种新的推荐算法,从项目图像和项目评分两方面出发,解决传统算法中存在的问题,提高推荐质量。

1 基于项目的协同过滤算法

1.1 数据描述

本文将用户集合定义表示为 $U=\{u_1,u_2,\cdots,u_m\}$,项目集合表示为 $I=\{i_1,i_2,\cdots,i_n\}$,用户-项目的评分矩阵表示为 $\boldsymbol{R}=(r_{ui})_{m\times n}$,得到的评分矩阵 $\boldsymbol{R}$ 如下:

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix}$$

1.2 项目评分相似度计算

使用 1.1 节得到的矩阵 $\boldsymbol{R}$ 进行项目相似度的计算。$U_i\cap U_j$ 为项目 i 和 j 共同用户集合,$\bar{r}_i$ 和 $\bar{r}_j$ 分别表示项目 i、j 的平均分,r_{ui}、r_{uj} 分别指用户 u 对项目 i 和 j 的打分,本文使用皮尔森相关系数计算项目 i 和 j 的评分相似度:

$$sim_1(i,j)=\cos(i,j)=\frac{\sum_{u\in U_i\cap U_j}(r_{ui}-\bar{r}_i)(r_{uj}-\bar{r}_j)}{\sqrt{\sum_{u\in U_i\cap U_j}(r_{ui}-\bar{r}_i)^2}\sqrt{\sum_{u\in U_i\cap U_j}(r_{ui}-\bar{r}_j)^2}} \tag{1}$$

2 融合项目图像的混合推荐算法

针对根据项目进行推荐时并没有完全利用项目图像这一因素,本文利用卷积神经网络提取图像特征,利用提取后的特征值得到项目图像之间的相似度,与项目评分相似度动态结合,进而产生推荐。

2.1 VGG16 卷积神经网络

卷积神经网络(CNN)是近年发展起来的一种图像特征提取方法[12],由于其抽取的图像特征是通过数据学习得来的,克服了传统手工需要通过设计才能提取图像特征的缺陷,节约了大量的人力和时间,因而受到广泛的关注。VGG16 是使用最多、效果显著的卷积神经网络模型[13],其利用卷积计算层、池化层进行图像特征的提取[14]。

卷积层采用卷积的方式,即一种加权的方法,通过一个加权函数(卷积核)对提取特征需要的信息赋予更高的权重,进行特征提取,将卷积核重复作用于图像的每一块区域,即完成一次特征提取,提取出的结果称之为特征图。池化层的主要作用就是将经过卷积层处理后得到的特征图进行一定程度的压缩,防止过拟合现象的出现。

2.2 图像相似度

经过 VGG16 卷积神经网络进行特征提取,每一个电影海报输入到 VGG16 模型中进行特征提取,会形成一个 $512\times7\times7$ 的输出,即提取到的电影海报的特征数据,把这个输出处理为一维的形式,将每个电影海报的特征记录下来,利用余弦相似度公式获得项目图像相似度 sim_2:

$$sim_2(i,j)=\frac{\boldsymbol{i}_a\cdot\boldsymbol{j}_a}{\|\boldsymbol{i}_a\|\|\boldsymbol{j}_a\|}=\frac{\sum_{k=1}^{n}\boldsymbol{p}_{ik}\cdot\boldsymbol{p}_{jk}}{\sqrt{\sum_{k=1}^{n}\boldsymbol{p}_{ik}^2}\sqrt{\sum_{k=1}^{n}\boldsymbol{p}_{jk}^2}} \tag{2}$$

式中:$\boldsymbol{i}_a$、$\boldsymbol{j}_a$ 分别代表项目 i 和 j 的图像经过特征提取后形成的向量;$\boldsymbol{p}_{ik}$、$\boldsymbol{p}_{jk}$ 分别代表向量 $\boldsymbol{i}_a$、向量 $\boldsymbol{j}_a$ 在位置 k 上的取值。

2.3 混合相似度

由于电影海报中会有一些通用的元素,比如人物形象、场景等,这些元素可能使关联性不是很强的项目具有很高的图像相似度,因此引入项目类别属性来对项目图像相似度进行修正,使利用图像相似度找到的都是项目类型、内容与目标项目更加接近的项目。电影有动作、科幻、爱情、冒险等项目类别属性,利用数据集存储的项目信息,提取出项目的类别属性矩阵,如表 1 所示。

表 1 项目-类别属性矩阵

项目	*S*1	*S*2	*S*3	*S*4	*S*5
*I*1	0	1	0	1	1
*I*2	1	0	0	1	0
*I*3	0	0	1	1	0
*I*4	1	1	0	0	1
*I*5	1	0	1	0	0

表 1 中: $I1$–$I5$ 表示项目; $S1$–$S5$ 表示项目类别属性; 1 表示项目有这个类别属性; 0 表示项目没有这个类别属性。通过数据集,获取项目-类别属性信息,可得到项目-类别属性矩阵 $\boldsymbol{CN}$,从中可以得到每个项目的类别属性,并利用式(3)[15]来度量两个项目属性相似度。

$$\lambda = \frac{|CN_i \cap CN_j|}{|CN_i \cup CN_j|} \tag{3}$$

式中: CN_i、CN_j 代表类别属性集; $CN_i \cap CN_j$ 是 i 与 j 都拥有的类别属性数量; $CN_i \cup CN_j$ 为项目 i 与 j 拥有的所有类别属性的量。λ 越大表示项目类别属性相差越大,λ 越小表示项目类别属性相差越小。本文设计类别属性因子 λ,将评分与图像的相似度线性结合,计算方式如下:

$$sim(i,j) = (1-\lambda)\, sim_1(i,j) + \lambda sim_2(i,j) \tag{4}$$

引入 λ 来组合图像和评分的相似度,λ 与项目的属性类别有关,利用项目类别属性可以修正项目图像由于图像共有元素而导致的计算偏差,可以更好地找到与目标项目特性相同的项目。

2.4　预测评分生成

通过式(4),可获得项目间的相似度,构造出相似度矩阵 S。对于任意用户 u,根据用户-项目评分矩阵 $\boldsymbol{R}_{mn}$ 获得用户 u 的已评分项目集合 I_u,可以得到 u 的未评价的项目集 $N_u = I - I_u$。对于集合中的元素 $i(i \in N_u)$,可以根据矩阵 $\boldsymbol{R}$ 和矩阵 S 选取出与项目 i 相似度最高且有评分的一些项目构成其最近邻 $I_{\text{near}} = \{i_{i1}, i_{i2}, \cdots, i_{im}\}$,并用式(5)来预测 u 对 i 的评分[16]。

$$P_{ui} = \bar{r}_i + \frac{\sum_{j \in I_{\text{near}}} sim(i,j)(r_{uj} - \bar{r}_j)}{\sum_{j \in I_{\text{near}}} |sim(i,j)|} \tag{5}$$

式中: r_{uj} 为用户 u 对电影 j 的评分; $\bar{r}_i$ 为对电影 i 评过分的用户的评分均值; $\bar{r}_j$ 为所有对电影 j 评过分的用户的评分均值。得到未评分项目的预测评分后,针对推荐目标将预测评分最高的几个电影推荐给用户。

2.5　算法流程

本文引入图像信息,把项目图像信息和评分信息进行线性结合,得到新的项目相似度计算方法; 通过新的计算方式,得到项目的最近邻集; 通过最近邻集,预计评价值; 使用预计分值进行排序,将预计评分高的项目推荐给用户。融合项目图像的混合推荐算法(PB-CF)的具体流程图如图 1 所示。

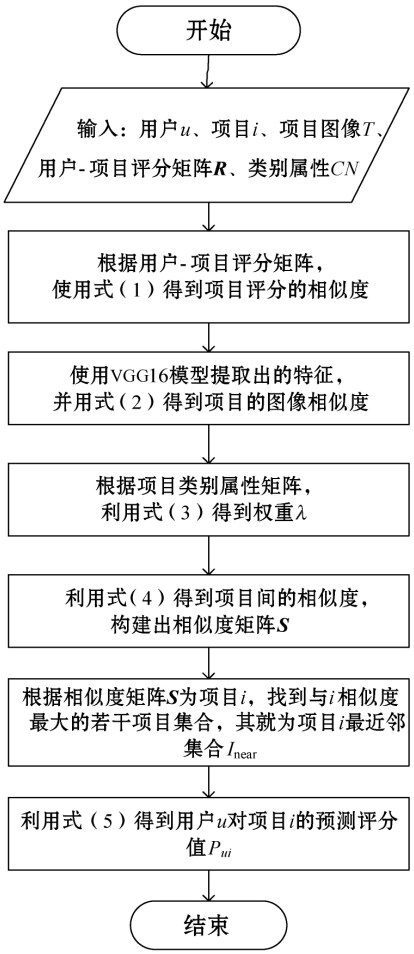

图 1　PB-CF 算法流程图

由于只利用图像信息同样可以得到项目的相似度找到项目的最近邻集进行评分的预测,因此本文同时提出只基于项目图像信息的相似度计算方法,融合项目图像的混合推荐算法(P-CF)的具体流程图如图 2 所示。

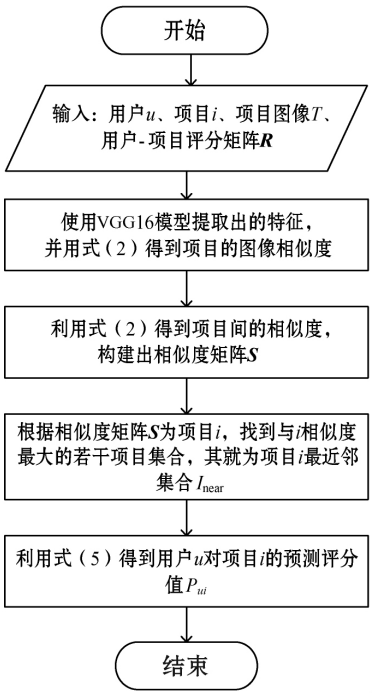

图 2　P-CF 算法流程图

3 实 验

3.1 数据集

本文使用 MovieLens100k[17] 作为实验数据集，MovieLens 是推荐算法研究领域中最被人熟知、最常被使用的数据集，其中有 943 名测试者对 1 683 部影片的 10 万条记录信息，电影海报数据集来自 IMBD，通过 API 接口找到 MovieLens 数据集中的相关电影海报。

实验环境使用 Windows 10 操作系统，8 GB 内存的 Intel i5-7300 处理器，在 Spyder 开发环境下，使用 Python 语言进行算法代码的编写和测试。将 MovieLens 数据集中的数据按照 4∶1 的比例进行随机划分，其中 80% 的数据作为训练集数据，另外 20% 的数据作为测试集数据。

3.2 评价指标

实验使用平均绝对误差(MAE)和均方根误差(RMSE)作为实验的评价指标，其具体的计算公式如下:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (p_i - q_i)^2}{N}} \tag{7}$$

式中: N 为预测评分集合中的项目的个数; p_i 为通过本文算法得到的用户对项目 i 的预测评分; q_i 表示用户对该项目的实际评分。推荐的效果与 MAE 和 RMSE 的值呈负相关关系，MAE、RMSE 值越小，表示推荐的效果越好，推荐质量越高。

3.3 最近邻数量确定实验

为确定本文算法中最近邻数量的最佳值，本次试验选取 10、15、20、25、30、35、40、45、50 的几种最近邻数量进行实验，图 3、图 4 为实验结果。

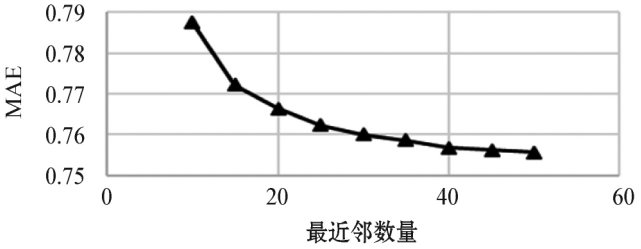

图 3 各最近邻数量下 MAE 误差情况

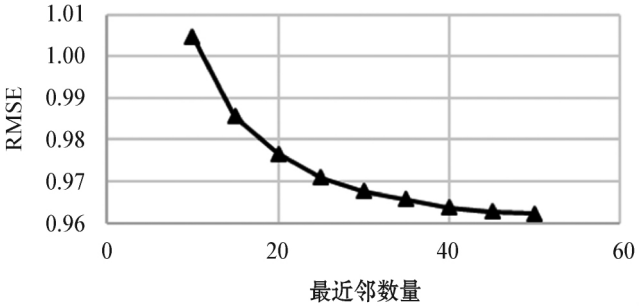

图 4 各最近邻数量下 RMSE 误差情况

由图 3 可知，随着最近邻数量的增加，MAE 的值逐渐变小，最近邻数为 40 时，MAE 值趋于稳定，约为 0.75。由图 4 可知，随着最近邻数量的增加，RMSE 的值逐渐变小，最近邻数为 40 时，RMSE 值趋于稳定，约为 0.96。

3.4 对比实验

为了体现本文 PB-CF 算法和 P-CF 算法的推荐结果准确性与其他算法的差别，本文选取传统的推荐算法进行对比试验，其中包括基于项目的协同过滤算法(IBCF)、基于用户的协同过滤算法(UB)，对比结果如图 5、图 6 所示。可以看出，相较其他推荐方法，本文的 PB-CF 和 P-CF 算法推荐误差值较小。基于图像的 P-CF 算法的 MAE 和 RMSE 值比其他三种算法的准确度更高，但考虑到信息的多样性和单一数据集数据的特殊性问题，还是利用 PB-CF 算法进行推荐。

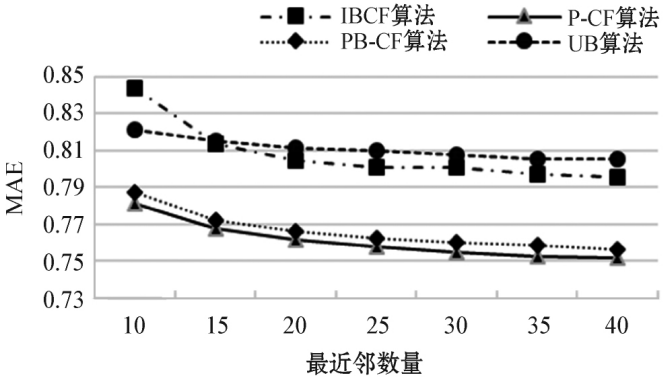

图 5 对比实验结果 1

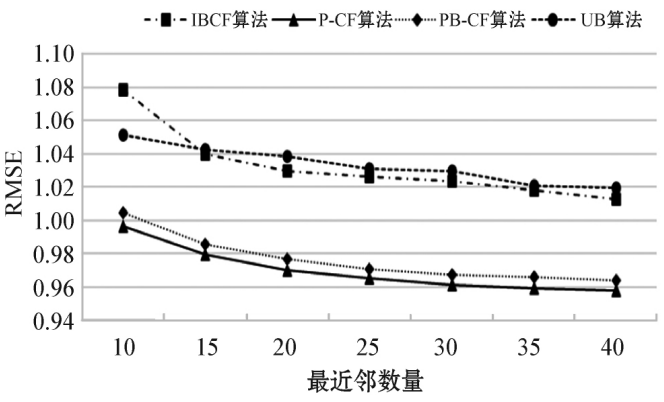

图 6 对比实验结果 2

4 结　语

本文将图像信息引入到项目相似度计算公式中,并设计动态加权方式对其进行结合,获得新的项目相似度衡量方式,通过算法获得的项目最近邻更精准。实验结果表明,本文提出的融合项目图像相似度计算方式,在引入项目图像特征后有效地缓解由于数据稀疏引起的计算偏差,提升推荐的准确度,不足之处在于图像特征提取的准确度还不够高,针对不同的数据集利用图像信息是否能取得同样良好的效果,需要进一步深入研究。

参 考 文 献

[1] Sarwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for e-commerce [C]//ACM Conference on Electronic Commerce, 2000.

[2] 毛明松,张富国. 基于多重图排序的用户冷启动推荐方法[J]. 计算机工程,2019,45(5):175-181.

[3] 段元波,高茂庭. 基于项目评分与类型评分聚类的推荐算法[J]. 计算机工程,2018,44(6):13-17,23.

[4] Frmal S, Lecron F. Weighting strategies for a recommender system using item clustering based on genres [J]. Expert Systems with Applications, 2017, 77: 105-113.

[5] 熊丽荣,刘坚,汤颖. 基于联合概率矩阵分解的移动社会化推荐[J]. 计算机学,2016,43(9):255-260,265.

[6] Zhao Z, Wang C, Wan Y, et al. FTMF: recommendation in social network with feature transfer and probabilistic matrix factorization [C]//2016 International Joint Conference on Neural Networks, 2016.

[7] 孙丽梅,李悦,Ejike Ifeanyi M,等. 简化的 Slope One 在线评分预测算法[J]. 计算机应用,2018,38(2):497-502.

[8] Liu Y, Liu D, Xie H, et al. A research on the improved slope one algorithm for collaborative filtering [J]. International Journalof Computing Science & Mathematics, 2016, 7(3):245-253.

[9] 朱丽中,徐秀娟,刘宇. 基于项目和信任的协同过滤推荐算法[J]. 计算机工程,2013,39(1):58-62,66.

[10] 韦素云,业宁,吉根林,等. 基于项目类别和兴趣度的协同过滤推荐算法[J]. 南京大学学报(自然科学版),2013,49(2):142-149.

[11] 魏甜甜,陈莉,范婷婷,等. 结合项目流行度加权的协同过滤推荐算法[J]. 计算机应用研究,2020,37(3):676-679.

[12] Ciresan D, Giusti A, Gambardella L M, et al. Deep neural networks segment neuronal membranes in elec-tron microscopy images [C]//Advances in Neural Information Processing Systems, 2012.

[13] Razavian A S, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[14] Babenko A, Lempitsky V. Aggregating deep convolutional features for image retrieval [C]//2015 International Conference on Computer Vision, 2015.

[15] 李淑芝,李志军,邓小鸿. 结合评分比例因子及项目属性的协同过滤算法[J]. 计算机应用研究,2020,37(3):680-683.

[16] 孙金刚,艾丽蓉. 基于项目属性和云填充的协同过滤推荐算法[J]. 计算机应用,2012,32(3):658-660,668.

[17] MovieLens100K [EB/OL]. [2017-03-23]. https://grouplens.org/datasets/movielens/.

(上接第 289 页)

参 考 文 献

[1] Schmidt R O. Multiple emitter location and signal parameter estimation [J]. IEEE Transactions on Antennas and Propagation, 1986, 34(3):276-280.

[2] 赵益民,鞠德航. 单通道接收机实施空间谱估计测向[J]. 通信学报,1997,18(2):8-12.

[3] 张玮. 双通道接收机实施空间谱估计测向的研究[D]. 西安: 西安电子科技大学,2013.

[4] 杨洪亮,赵益民. 一种三通道权微扰谱估计测向方法[J]. 电子科技,2014,27(11):14-16.

[5] 王鼎,吴瑛. 一种新的阵列误差有源校正算法[J]. 电子学报,2010,38(3):517-524.

[6] 王珍,段翔,刘周. 基于子空间类法的阵列误差有源校正方法[J]. 雷达科学与技术,2014,12(5):546-550,556.

[7] Friedlander B, Weiss A J. Direction finding in the presence of mutual coupling [J]. IEEE Transactions on Antennas and Propagation, 1991, 39(3):273-284.

[8] 韩芳明,张守宏,潘复平. 阵列误差对 MUSIC 算法性能的影响与校正[J]. 西安电子科技大学学报,2003,30(5):585-589.

[9] 王永良,陈辉,彭应宁,等. 空间谱估计理论与算法[M]. 北京: 清华大学出版社,2004: 430-432.

[10] 刘松,廖勇,谢远举. 一种高效稳健的均匀圆阵互耦校正方法[J]. 电子学报,2017,45(9):2170-2176.