基于情感分析的个性化电影推荐

黄剑波,陈方灵,丁友东,吴利杰
(上海大学,上海 200072)

摘 要:随着互联网技术的飞速发展,越来越多的信息和服务充斥着网络,如何实现精准高效的推荐,已成为亟待解决的问题之一。现有个性化电影推荐方法,将用户的历史评分作为推荐的重要依据,然而用户评分标准不一,很难挖掘出用户真正的喜好,难以形成精准推送。因此,为了实现高质量的电影个性化推荐,挖掘用户评论的情感就变得尤为重要。文中提出一种基于影评情感分析的个性化推荐方法,运用自然语言处理技术,挖掘用户影评情感倾向,将影评情感值与用户评分结合,共同计量用户喜好倾向。并利用点击率预估模型预测点击率,为用户提供个性化的推荐服务。实验结果表明,这种方法不仅有效解决了用户评分尺度不一等问题,且充分展现其个性化推荐的优越性。

关键词:电影推荐;情感分析;数据挖掘;点击率预估;个性化
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2020)09-0132-05
doi:10.3969/j.issn.1673-629X.2020.09.024

Personalized Movie Recommendation Based on Sentiment Analysis

HUANG Jian-bo, CHEN Fang-ling, DING You-dong, WU Li-jie
(Shanghai University, Shanghai 200072, China)

Abstract: With the rapid development of Internet technology, more and more information and services are flooding the network. How to achieve accurate and efficient recommendation has become one of the urgent problems to be solved. The existing personalized movie recommendation method takes the user's historical score as an important basis for recommendation. However, the user rating standards are different, so it is difficult to find out the real preference of the users and form a precise push. Therefore, in order to achieve accurate movie personalized recommendation, it is especially important to tap the emotion of user comments. We propose a personalized recommendation method based on sentiment analysis of emotions. The natural language processing technology is used to explore the emotional tendency of user's film reviews, and the emotional value of the film reviews is combined with user rating to jointly measure the user's preference. In addition, the click-through rate is predicted by click-through rate estimation model to provide personalized recommendation services for users. The experiment shows that the proposed method not only effectively solves the problem of different user ratings, but also fully demonstrates the superiority of its personalized recommendation.

Key words: movie recommendation; sentiment analysis; data mining; click-through rate estimation; personalized

0 引 言

近年来,移动互联网飞速发展,网络数据过载,生活节奏加快,如何实现精准高效推荐成为亟待解决的问题。传统的推荐方法,将用户评分作为评判用户倾向性的重要指标。其假设相似评分的用户具有类似喜好,而近邻用户并不能完全客观、真实地反映用户自身的偏好[1]。用户评分数据在一定程度可以代表用户对商品的态度,但用户评分产生差异的原因得不到合理解释,而评论为心中所想,更能反映用户心理。且心理学研究表明大多数人都有从众心理,人们对物品的喜好或情感状态会受多数人的情感影响[2]。因此,为了实现精准推荐,挖掘用户评论的情感就变得尤为重要。

文中运用自然语言处理等技术,分析电影评论文本,将其应用到个性化电影推荐中,挖掘用户情感信息,提高推荐的准确性。具体方法如下:首先抓取网络公开电影基本信息和影评数据,然后使用多人人工交叉标注部分影评数据集,为影评情感倾向性打分,训练情感分析模型。将情感值与用户评分结合,作为用户的喜好程度,消除用户评分标准不一的影响,能更加真实地反映用户的偏好。最后,使用点击率预测模型,对用户观影历史行为进行训练,并预测每个用户对未观看的电影的点击率,排序选取前 N 个数据为用户推

收稿日期:2019-10-21 **修回日期**:2020-02-25
基金项目:国家自然科学基金(61303093);上海市科委工程技术研究中心建设专项(16dz2251300)
作者简介:黄剑波(1980-),男,博士,实验师,硕导,研究方向为图像处理技术、电影特效技术等。

荐。实验结果表明此方法有更好的性能。

1　相关工作

Resnick 等[3]在 20 世纪 90 年代首次提出个性化推荐的概念,经过了二十多年的积累和沉淀,推荐系统逐渐成为一门独立学科在学术研究和业界应用中取得了很多成果。其背后的技术大致可以划分为三类:基于内容的模型、基于协同过滤的模型,以及混合模型[4]。

基于内容的推荐模型主要在于分别建立用户和物品的档案资料,计算用户或物品之间的相似度[5]。物品的档案通常由它的各种属性资料构成,以服装领域为例,包括价格、品牌、类别、颜色、风格、款式、尺寸等。内容推荐虽然是推荐系统的孩童时代,但依然适用于各个领域,主要原因在于,首先只要得到物品或者用户的档案,就可以处理冷启动问题,其次,档案都是显式特征,模型有很好的可解释性。

协同过滤的提出,极大地推动了推荐系统的研究和发展[6]。基于协同过滤的推荐模型不需构建任何档案资料,只收集用户的历史行为记录,就可挖掘用户与用户、物品与物品之间潜在的相似性,并基于这种群组相似性完成推荐。其包括基于邻居的方法和基于模型的方法。基于邻居的方法核心在于根据历史行为记录,构建用户与用户,或者物品与物品的相似度矩阵,能在广泛的兴趣范围中推荐出热门物品,但缺少个性化。基于模型的推荐最常用的是隐因子模型,典型的是 Koren Y 等[7]提出的矩阵分解。在这类模型中,用户和物品都被嵌入到一个低维向量表示,用户和物品的相关性体现于它们对应隐向量的点积关系。这种方法效率高,一旦训练出模型,用户和物品的关系就能很方便地通过点积计算出来,同时准确度也好于邻居模型。但缺点也很明显,不能解决冷启动问题,同时学习出的隐向量不方便解释。

综上,不同推荐算法在应用中有不同效果。因此,工业界常用的是混合模型,结合多种推荐模型,取长补短,能得到更好的推荐效果。

近年来,随着众多学者对点击率(click-through rate estimation,CTR)预估模型的研究,CTR 模型在推荐系统中得到广泛应用,解决了矩阵分解技术在高度稀疏的数据场景下不适用的问题。2011 年 Steffen Rendl 等[8]提出的因子分解机(factorization machine,FM)模型,采用特征组合的方式,解决了推荐数据稀疏的问题。2016 年 Juan Yu-Chin 等[9]提出场感知分解机(field-aware factorization machine,FFM)模型,在 FM 的基础上引入场的概念,将具有相同性质的特征归为同一个场。2017 年 Guo Huifeng 等[10]提出了深度因子分解机(a Factorization-Machine based neural network,DeepFM)模型,可同时学习低阶和高阶特征,提高排序能力。

2　基于情感分析的个性化电影推荐

2.1　整体流程

基于情感分析的个性化电影推荐主要分为以下 3 个步骤:数据采集与预处理,模型训练,预测及电影推荐,流程如图 1 所示。

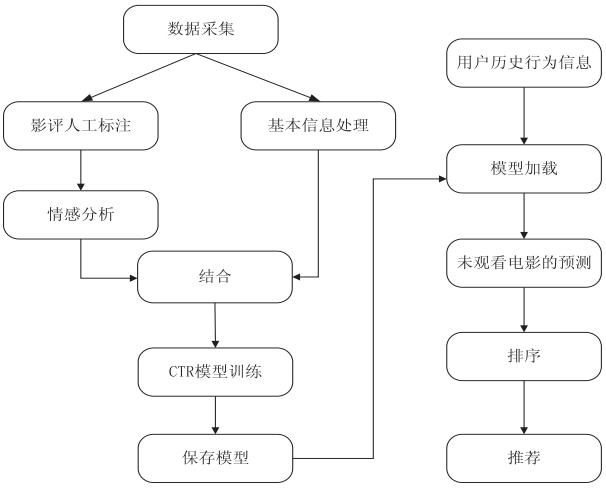

图 1　基于影评情感分析的电影推荐流程

(1)数据采集与预处理。从网络公开数据收集足够的电影相关数据,然后进行数据的清洗处理。对于影评数据,还需采用多人人工交叉精确标注,影评根据情感倾向程度进行标注为 1 ~ 5,其中 1 ~ 5 喜爱度依次递增。文中采用 BERT[11]对影评数据有监督多分类训练,并保存训练模型,对新的情感预测时,载入保存好的训练模型,直接进行预测。

(2)模型训练。使用 CTR 预估模型 DeepFM,对处理好的数据进行训练,并保存相应模型。

(3)电影推荐。加载模型,根据用户历史行为信息,预测用户对未观看电影的 CTR、排序,选取前 N 个数据,实现个性化推荐。

2.2　数据采集与预处理

文中数据集主要来源于网络公开数据的抓取,包括电影数据集、用户数据集以及影评数据集。其中电影数据集包含电影 ID、电影名、演员、导演、类型、编剧、时长等字段。用户数据集包含用户 ID、用户常居地、用户名等字段。影评数据集含有电影 ID、电影名、用户 ID、用户评论、用户评分等字段。对抓取数据首先进行清洗操作,影评数据集还需要多人人工交叉标注,然后进行情感模型的训练。

文中情感分析采用 BERT 模型。BERT 采用的是 Transformer[12]的双向编码器结构,Transformer 不需要循环,而是并行处理序列中的所有单词或符号,同时利

用自注意力机制将上下文与较远的单词结合起来。BERT 的双向为深度双向,与传统双向有所不同。传统双向是从左到右与从右到左的结合,但是两个方向的损失计算相互独立,其本质还是单向的,只是一种简单融合,而 BERT 的深度双向充分结合了上下文信息。BERT 模型可同时用作预训练模型和下游任务模型,且不需要做任何的修改就能实现文本分类。对影评做情感分析时,上下文的语义尤为重要,要充分考虑到上下文的关系,因此采用 BERT 模型能更精确的分类。

传统的情感分析方法分为正向和负向两类,文中的情感分类分为 5 个等级。采用 BERT 模型进行中文文本情感分析的过程如图 2 所示。

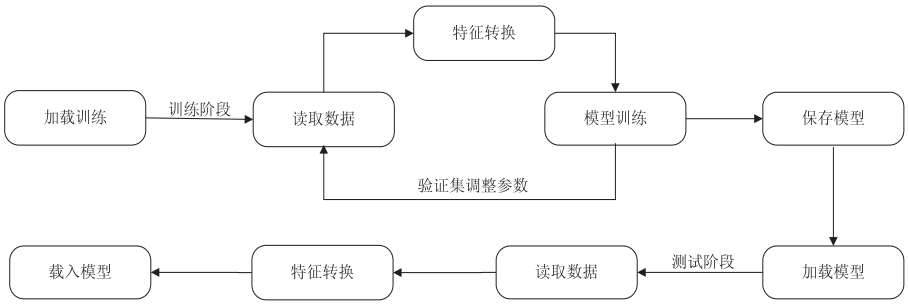

图 2 中文文本情感分析的流程

利用训练好的 BERT 模型预测影评情感值,将影评情感值和用户评分相结合,表示用户的整体倾向性,可以写成:

$$y = w_1 y_1 + w_2 y_2 \tag{1}$$

其中,y_1 和 y_2 分别为用户影评情感值和用户评分,w_1,w_2 为各自的权重,$w_1 + w_2 = 1$,文中 w_1 和 w_2 的取值设为 0.5。

2.3 模型训练

对于 CTR 预估模型,低阶组合特征和高阶组合特征都会影响最终的结果,学习用户行为背后隐含特征组合极其重要。而 DeepFM 模型可以从原始数据中同时学习低维与高维特征。因此,文中推荐模型采用 CTR 预估模型 DeepFM。

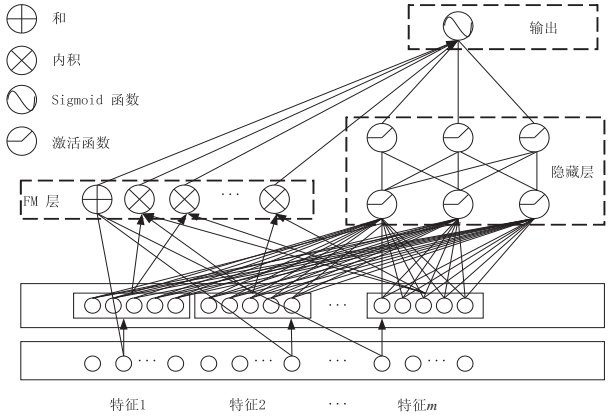

图 3 DeepFM 模型结构

DeepFM 分为神经网络部分和因子分解机部分。DeepFM 模型将 DNN 和 FM 并行组合,同时具有 FM 在推荐中的优势和深度学习在特征学习的优势。模型结构如图 3 所示,因子分解机部分和神经网络部分分别负责提取低阶特征和高阶特征,共享权重矩阵,即共享嵌入层。这样可以从原始数据中同时学习到低维与高维特征,不再需要人为设计特征工程,训练效率更高效。DeepFM 模型的预测结果可以写成

$$\hat{y} = \text{sigmoid}(y_{\text{FM}} + y_{\text{DNN}}) \tag{2}$$

其中,$\hat{y} \in (0,1)$ 是预测的 CTR,y_{FM} 为 FM 部分输出,y_{DNN} 为神经网络输出。

模型训练采用对数似然损失(logarithmic loss function,LogLoss)进行参数更新。LogLoss 采用 KL(Kullback-Leibler)[13] 散度来计算,预测分布越接近真实分布,其值越小。假设样本的真实分布为 P,预测分布为 Q,则 KL 散度定义[14] 如下:

$$D(P \parallel Q) = \sum_{x \in X} P(X) \log \frac{P(x)}{Q(x)} \tag{3}$$

在 CTR 预估中,概率分布为二项分布。设真实的点击率为 tctr,预测的点击率为 pctr。因此真实的二项分布 P(tctr,1 - tctr),预测的二项分布 Q 为(pctr,1 - pctr)。因此损失函数可以写成如下形式:

$$\text{KL}(\text{tctr} \parallel \text{pctr}) = \text{tctr} * \log \frac{\text{tctr}}{\text{pctr}} + (1 - \text{tctr}) * \log \frac{1 - \text{tctr}}{1 - \text{pctr}} \tag{4}$$

为了适应电影推荐这类稀疏数据,参数优化方法采用 Adagrad 优化方法[15]。Adagrad 算法在训练中自动更新学习率,采用较大的学习率调整出现次数较少的参数。

DeepFM 模型输入数据为特征经过独热编码横向拼接而成的高维稀疏向量。首先,各个特征加权求和得到一次项。然后,将原始输入的特征经过嵌入层,一方面两两内积,求和得二次项,另一方面作为输入全连接到 DNN,实现低维和高维特征的结合。

2.4 电影推荐

文中加载训练好的模型,对于给定的用户及其历史行为,对其未观看电影预测 CTR,按照从大到小排序选取前 N 个数据,实现个性化的电影推荐。

3　实验结果

实验采用的操作系统为 Ubuntu 16.04,64 位,基于 TensorFlow 框架,编程语言为 Python3.5。所有的训练均采用 NVIDIA-GTX-TitanX 显卡。

根据 Zhou Guorui 等[16]提出的数据处理方法,为了适用 CTR 预测任务,将数据转换为二分类数据。用户的偏好值是从 0 到 5 的连续值。将偏好值为 4 和 5 的样本标记为 1,其余为 0。将 227 424 个样本划分为训练集,其余 46 036 个样本为测试集。目标是根据历史行为预测用户是否对给定电影的偏好值高于 3(为 1)。

文中采用 LogLoss、AUC、MAP 作为模型评价指标。LogLoss 更关注和观察数据的吻合程度,AUC 更关注排序。MAP 是反映系统在全部相关文档上性能的单值指标,系统检索出来的相关文档越靠前,MAP 就可能越高[17]。图 4 对比了 DNN,FFM,DeepFM 三个模型在测试集上的 AUC 表现。三个模型的预测结果如表 1 所示。

实验结果表明,基于情感分析的个性化电影推荐是可行的,由图 4 和表 1 可知深度因子分解机具有较好的预测结果。

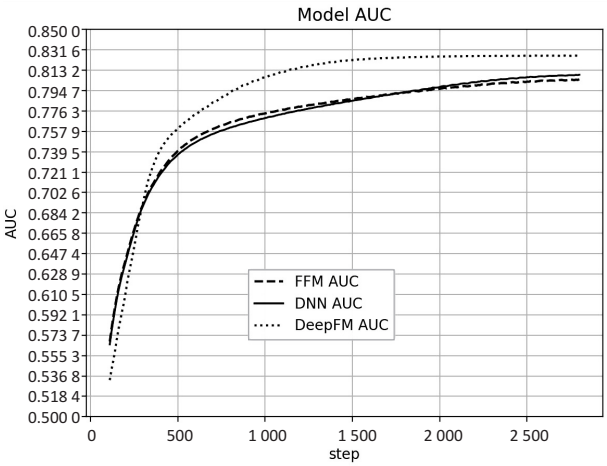

图 4　不同模型 AUC

表 1　不同模型预测结果对比

模型	LogLoss	AUC	MAP@10	MAP@20
FFM	0.473	0.796	0.640	0.630
DNN	0.492	0.800	0.659	0.648
DeepFM	0.485	0.824	0.660	0.652

选取两名用户分别为其推荐 10 部电影,如表 2 所示。从表中看出,用户 B 的历史评分最高分为 3 分,没有表现出明显的倾向性,而评论却表现出明显的喜好

表 2　不同用户电影推荐对比

用户	历史观看记录				基于用户评分的推荐	基于情感分析的推荐
	电影名	影评	评分	情感值		
A	阿童木	这个偷了阿童木名字的人是谁!!!	1	3	1、速度与激情 8 2、红海行动 3、七武士 4、隐秘而伟大 5、指环王 3 6、记忆大师 7、宝贝计划 8、金刚 9、热血高校 2 10、华尔街之狼	1、星河战队 2、速度与激情 8 3、指环王 3 4、冰雪奇缘 5、记忆大师 6、七武士 7、金刚 8、华尔街之狼 9、红海行动 10、被解救的姜戈
	侏罗纪公园 2:失落的世界	科幻电影的里程碑式作品。	3	5		
	丁丁历险记	一场动作捕捉和视觉特效的盛宴。还有萌翻天的白雪和阿道克船长。	3	5		
	孤胆特工	紧张的节奏。绝佳的剪辑。深沉的色調。还有一个平头男人元彬!	4	5		
	催眠大师	结构完整悬疑够味制作精良。结局冗长得好可惜。	4	4		
	魔发奇缘	美得惊人的场景。以及经典版的童话故事。	3	4		
B	古墓丽影:源起之战	女主很有力量感啊! 墓穴那一幕好看,喜欢这样的电影。	3	4	1、乱世佳人 2、指环王 3 3、末代皇帝 4、神奇女侠 5、无间道风云 6、发条橙 A 7、美国往事 8、巴尔干边界 9、云水谣 10、海王	1、金刚 2、狂暴巨兽 3、羞羞的铁拳 4、 指环王 3 5、末代皇帝 6、无间道风云 7、发条橙 A 8、摩天营救 9、神奇女侠 10、巴尔干边界
	速度与激情 8	情节很燃,只不过剧情看看就好。	3	3		
	睡在我上铺的兄弟	很尴尬的剧情场景切换的拼凑,是真的烂。	2	1		
	唐人街探案	看完 2 又特意回顾了一下,佟丽娅真的美啊。张子枫最后的笑容至今依然让人毛骨悚然! 比 2 好看。	3	4		
	撒娇女人最好命	周迅演过这种剧,可怕。	2	2		

倾向。对比两种推荐方法的推荐结果,可以看出基于情感分析的推荐更符合用户的心理。

对于三部不同类型的电影,将 CTR 作为推荐指数,用户的推荐值数如表 3 所示。可知对于同一部电影,不同的用户表现出了明显的差异,表明了提出的个性化电影推荐方法的可行性。

表 3 用户电影推荐指标

用户	电影名称	推荐指数
A	九层妖塔	0.353 999 7
	哪吒之魔童降世	0.714 026 3
	变形金刚 2	0.625 131 1
B	九层妖塔	0.170 344 7
	哪吒之魔童降世	0.578 783 4
	变形金刚 2	0.639 750 2

文中还采集了 25 个不同年龄阶段的历史的信息记录,参与模型训练,为其推荐 5 部电影,并调研反馈信息。如表 4 所示。可以看出只有 8% 的人不喜欢推荐的电影,表明了该推荐方法的有效性。

表 4 用户反馈调研

用户(年龄分布)	喜欢	部分喜欢	不喜欢	总人数
15 ~ 20	1	4	1	6
20 ~ 25	2	5	0	7
25 ~ 30	1	6	1	8
30 ~ 35	2	2	0	4
统计	24%	68%	8%	

4 结束语

提出了一种基于影评情感分析的个性化电影推荐方法。首先,爬取网络公开电影、用户数据、影评数据,然后将影评数据集进行人工交叉标注,使用 BERT 模型进行情感分析,情感分析结果和用户评分相结合,采用 DeepFM 点击率预估模型进行预测。最后,根据 DeepFM 预测的结果,按照 CTR 排序,选取前 N 个数据实现线下推荐,提高了推荐的质量。然而在研究中还发现了一些问题:(1)对中文影评情感多分类难度较大,因此采用强大的 BERT 模型,依然不能有很高的准确率;(2)BERT 模型训练时间消耗很大,对于影评数据句子稍长,需要耗费大量的时间。在下一步的工作中,可以采用混合推荐的方法,评估整体的效果。

参考文献:

[1] 侯银秀,李伟卿,王伟军,等. 基于用户偏好与商品属性情感匹配的图书个性化推荐研究[J]. 数据分析与知识发现,2017,1(8):9-17.

[2] 夏明星. 基于情感分析的评论极性分类和电影推荐系统的设计与实现[D]. 合肥:安徽大学,2016.

[3] RESNICK P, VARIAN H R. Recommender systems[J]. Communications of the ACM,1997,40(3):56-59.

[4] 刘建国,周 涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.

[5] 高建煌. 个性化推荐系统技术与应用[D]. 合肥:中国科学技术大学,2010.

[6] 刘青文. 基于协同过滤的推荐算法研究[D]. 合肥:中国科学技术大学,2013.

[7] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer,2009,42(8):30-37.

[8] RENDLE S. Factorization machines[C]//2010 IEEE international conference on data mining. Sydney:IEEE,2010:995-1000.

[9] JUAN Y,ZHUANG Y,CHIN W S,et al. Field-aware factorization machines for CTR prediction[C]//Proceedings of the 10th ACM conference on recommender systems. New York:ACM,2016:43-50.

[10] GUO H,TANG R,YE Y,et al. DeepFM:a factorization-machine based neural network for CTR prediction[C]//Proceedings of the 26th international joint conference on artificial intelligence. Melbourne, Australia: AAAI Press, 2017: 1725-1731.

[11] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: human language technologies. Minneapolis: NAACL-HLT,2019:4171-4186.

[12] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Advances in neural information processing systems. Long Beach:NIPS,2017:5998-6008.

[13] KULLBACK S,LEIBLER R A. On information and sufficiency[J]. The Annals of Mathematical Statistics,1951,22(1):79-86.

[14] 陈巧红,余仕敏,贾宇波. 广告点击率预估技术综述[J]. 浙江理工大学学报:自然科学版,2015,33(6):851-857.

[15] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12: 2121-2159.

[16] ZHOU G,ZHU X,SONG C,et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. New York:ACM,2018:1059-1068.

[17] 张 文,姜祎盼,张思光,等. 基于经验分布和 KL 散度的协同过滤推荐质量评价研究[J]. 计算机应用研究,2019,36(9):2625-2630.