基于大数据的 UserBased 推荐算法的研究*

任　悦　闫仁武
（江苏科技大学计算机学院　镇江　212000）

摘　要　在科技迅速发展的时代，人们不断地接触新事物和新技术来提高自己，因此，网购成为了最受大家追捧的生活方式之一。论文以电商平台的商品为研究对象，提出了基于用户推荐算法的研究与改进。论文主要研究的是UserBased协同过滤推荐算法，此算法的优点是可以实现跨领域、惊喜度较高；同时也存在着不足之处，比如，推荐结果的个性化比较弱、较为宽泛。通过对物品比用户多、物品时效性较强这种情景进行实验，多方面进行分析，实验结果表明，与传统算法进行比较，论文所研究的UserBased推荐算法，在个性化推荐以及推荐的准确度方面有了较好的改善。

关键词　推荐算法；UserBased；冷启动；协同过滤
中图分类号　TP391.3　　**DOI**：10. 3969/j. issn. 1672-9722. 2022. 01. 015

Research on UserBased Recommendation Algorithm Based on Big Data

REN Yue　YAN Renwu
（School of Computer Science and Engineering，Jiangsu University of Science and Technology，Zhenjiang　212000）

Abstract　In the era of rapid development of science and technology，people are constantly exposed to new things and technologies to improve themselves，so online shopping has become one of the most popular lifestyles. This paper takes the commodities of the e-commerce platform as the research object and puts forward the research and improvement based on the user recommendation algorithm. This paper mainly studies the UserBased collaborative filtering recommendation algorithm，which has the advantages of cross-domain and high degree of surprise. At the same time，there are also shortcomings，for example，the recommendation results of personalization is weak，relatively broad. Through the experiment on the situation that there are more items than users and the items are more time-efficient，the experiment results show that compared with the traditional algorithm，the UserBased recommendation algorithm studied in this paper has a better improvement in personalized recommendation and the accuracy of recommendation.

Key Words　recommendation algorithm，UserBased，cold-star，collaborative filtering
Class Number　TP391.3

1　引言

随着现代科技与网络的飞速发展，计算机技术及其相关知识已经广泛应用到各行各业中。科技的发展敲开了新世纪的大门，更多的购买方式渐渐被人们挖掘出来，例如亚马逊、淘票票、天猫以及京东等，推荐系统的推荐效果对电商的发展有着巨大的影响。为了让人们在电商平台上更有效地购买商品，大数据时代下的基于用户的推荐算法[1]发挥了至关重要的作用。

数据挖掘[2]是通过采集大量的数据，分析数据并进行对比，并将隐含在其中且又潜在有用的知识挖掘出来的过程。通过数据挖掘，我们可以减少对比较隐蔽性的知识的忽略，从而提高一些搜索的精准性。个性化推荐系统是推荐系统最为重要的研究方向，它可以通过对用户的需求及爱好等的分析，利用推荐算法[3]从海量数据中挖掘出用户感兴趣的项目，并将结果推荐给用户。推荐算法成为了

* **收稿日期**：2021年6月11日，**修回日期**：2021年7月22日
基金项目：国家自然科学基金项目“基于鲁棒表现建模的目标跟踪方法研究”（编号：61772244）资助。
作者简介：任悦，女，硕士研究生，研究方向：数据挖掘。闫仁武，男，副教授，研究方向：知识发现（数据挖掘）理论与技术、信息融合理论与技术、模式识别理论与应用等。

个性化推荐系统中最为重要的部分,其一些属性会在推荐系统的性能方面有着直接或者间接的影响。

当今社会拥有着庞大的信息体,我们需要在有效的时间内,将信息进行筛选、过滤,提取出对用户需求有益的数据进行反馈。本文通过对协同过滤算法[4](Collaborative filtering algorithm)的介绍与研究,着重于对基于用户推荐算法(UserBased)的研究与改进,对原有的UserBased推荐算法进行提升,有效改善个性化推荐以及冷启动方面的问题。

2 相关介绍

2.1 协同过滤算法

传统的协同过滤主要任务是找到与目标对象相似的相邻用户,然后将相邻用户所喜欢的项目推荐给该用户。该过程不关注商品的具体内容,可以实现跨类别推荐,提高用户的惊喜度和用户满意度,能应用于复杂的非结构化的推荐系统中[5]。主要过程是构建用户项目评价数据模型;邻居相似性计算与最近邻形成;预测评分与产生推荐,即算法的输入、算法处理和算法的输出。

协同过滤推荐技术是推荐系统中最为常见的方法,类型主要有两种:基于用户的协同过滤和基于项目的协同过滤[6]。基于用户的算法是将和目标用户有相同兴趣爱好的用户所心仪的物品且目标用户没有购买的物品推荐给目标用户。基于项目的算法则是通过目标项目的相似项目集合预测用户对相似物品的喜欢程度。基于用户的协同过滤推荐算法的步骤有建立用户评分表,寻找相似用户,推荐物品,其所存在的问题如下。

1)冷启动问题

当一个项目第一次出现的时候,肯定没有用户对它做过详细评价,因此无法对该项目进行评分预测和推荐。同时,因为新物品出现时,用户评价少,所以准确性也较差[7]。

2)稀疏性问题

在庞大的数据量并且数据稀疏的状态下,首先最近邻用户集的存在比较难发现,其次计算相似性所损耗的费用也会很大。同时,信息常常会丢失,导致推荐效果降低[8]。

3)可扩展性问题

随着推荐系统用户和项目数量的不断增长,协同过滤推荐算法的计算量也会随之增长,导致系统的性能逐步下降,从而影响用户体验[9]。

2.2 数据预处理

数据预处理技术[10]是对数据信息进行提前处理,以此来提升数据挖掘的精准度,比如,在进行关键词检索时,数据预处理能够对数据库内的信息资源进行相关的排序处理,来提升检索精度和效率等。该技术一般经过数据审核、数据筛选、数据排序等,达到数据信息处理效率加强的效果。

预处理技术的工作原理一般包括对数据进行清理、集成、变换、归约等方面的技术处理,来提升后期数据检索的精准性。

1)数据清理通过填充缺失值,识别离群点,纠正数据中的不一致等技术进行[11]。

2)数据集成需要考虑许多问题,如冗余,常用的冗余分析法有皮尔逊积距系数、卡方检验、数值属性的协方差等[12]。

3)数据转换将数据转换为适合学习的形式,包括数据光滑、聚集、泛化、规范化等[13]。

4)数据归约技术是用来得到数据集的归约表示,在接近原始数据完整性的同时将数据集规模从维度到数量大大减小[14]。

3 基于UserBased推荐算法

3.1 传统UserBased推荐算法

基于用户的协同过滤算法(UserBased)是最早出现的协同过滤算法的基本形式,其工作流程分两步[15]。第一步是求出用户之间的相似度,第二步是根据用户之间的相似度找出与待推荐的用户最为相似的几个用户并根据他们的兴趣爱好向待推荐用户推荐其可能会感兴趣的商品。用户u和v之间的相似度的计算主要可以通过Jaccard公式如式(1)和余弦相似度公式如式(2)得到。

$$W_{uv}=\frac{|N_{(u)}\bigcap N_{(V)}|}{|N_{(u)}\bigcup N_{(V)}|} \tag{1}$$

$$W_{uv}=\frac{|N_{(u)}\bigcap N_{(v)}|}{\sqrt{|N_{(u)}|\cdot|N_{(v)}|}} \tag{2}$$

其中,$N_{(k)}$为用户k感兴趣的商品集,$\cdot$为普通乘法。计算用户u对商品i的兴趣度加权打分公式如式(3):

$$p(u,i)=\sum_{v\in S(u,k)\bigcap I(i)}w_{uv}r_{vi} \tag{3}$$

$S(u,k)$包含了和用户u兴趣最接近的k个用户集合,$I(i)$表示对商品i有过打分行为的用户集合,W_{uv}表示计算出的用户u和用户v的兴趣相似度,r_{vi}表示用户v对商品i打的分数。得到了用户u对所有商品的感兴趣程度分数后,将分数最高的几个商品作为用户u最有可能感兴趣的商品推荐给用户u。

3.2　UserBased 推荐算法改进

本文采用余弦相似度算法对两两用户进行相似度的计算。首先要建立商品-用户倒排序如图1，图1中 a、b、c、d 为用户，I_1、I_2、I_3、I_4、I_5为商品，左边部分表示用户喜欢的商品，例如用户 a 喜欢的商品是 I_1、I_2、I_4，右边部分表示喜爱每个物品的用户，例如喜爱商品 I_1 的用户有 a 和 b；然后建立用户相似度矩阵 W，如表1，表1中行和列代表用户，内部表示两两用户共同喜欢的商品数量；最后计算用户相似度。

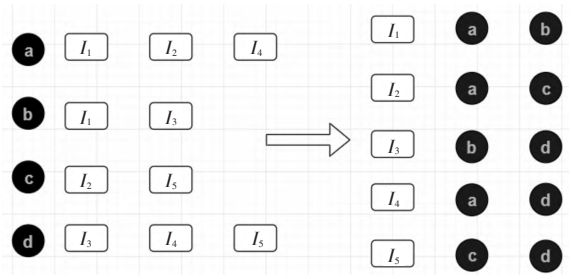

图1　物品-用户图

表1　用户相似矩阵 W

	a	b	c	d
a	3	1	1	1
b	1	2	0	1
c	1	0	2	1
d	1	1	1	3

遍历用户相似度矩阵中所有的两两用户，根据共同喜欢的商品数量，计算相似度，用到的公式为式(2)。比如 a 和 b 这两个用户，根据式(2)计算如下。

基于相似矩阵的基本运算：

$$W_{ab}=\frac{W_{[a][b]}}{(W_{[a][b]}*W_{[a][b]})^{\frac{1}{2}}}=\frac{1}{\sqrt{6}}$$

基于物品-用户图的解释：

$$W_{ab}=\frac{|\{I1、I2、I4\}\cap\{I1、I3\}|}{\sqrt{|\{I1、I2、I4\}|\cdot|\{I1、I3\}|}}=\frac{1}{\sqrt{6}}$$

为了改善冷启动问题的实时性和提高个性化推荐，对相似度的计算公式修改为如式(4)：

$$W_{uv}=\frac{\sum_{i\in N_{(u)}\cap N_{(v)}}\frac{1}{\log\ (1+|I_{(i)}|)}}{\sqrt{|N_{(u)}||N_{(v)}|}}\tag{4}$$

其中，i 表示用户 u 和 v 都有过正反馈的商品集合，$I_{(i)}$ 表示对商品 i 有过正反馈的用户数。该公式降低了用户 u 和 v 共同喜欢的物品中热门物品对他们相似度的影响。简言之，如果不同用户对冷门项目采取过同样的行为，则更能说明他们兴趣的相似度是比较高的。

3.3　UserBased 改进推荐算法

伪代码是一种非正式的，类似于英语结构的，介于自然语言和计算机语言之间的文字和符号(包括数学符号)来描述算法。本文 UserBased 改进算法的伪代码如下。

```
Algorithm：UserBased改进推荐算法
Input：u，items，v
Output：W
1)初始化 item_users=dict()，C=dict()，N=dict()，W=dict()
2)for u，items to train.items() do
3)  for i to items.keys() do
4)    if i ∉ item_user  do
5)      item_users[i]←set()
6)    end if
7)    item_users[i].add(u)
8)  end for
9)end for
10)for i，users to item_users.items() do
11)  for u to users do
12)    N[u]←N[u]+1
13)    for v to users do
14)      if u==v do
15)        continue
16)      end if
17)    end for
18)    C[u][v]←C[u][v]+1/math.log(1+len(users))
19)  end for
20)end for
21)for u，related_users to C.items() do
22)  for v，cuv to related_users.items() do
23)    W[u][v]←cuv/math.sqrt(N[u]*N[v])
24)  end for
25)end for
26)return W
```

4　实验结果与分析

4.1　实验环境及数据

本文涉及的实验所在环境是一台配置为 Intel(R) Core(TM) i5-8250U 处理器 2.6GHz CP 和 1.8GHz 的笔记本。本文实验选取的数据是使用 Movielens 电影评分数据集合，利用其中 1482 个用户对 943 个物品的评分记录，每个用户至少评价过 20 部电影，评分的取值位于整数 1~5 之间，通过数值的高低来判断用户对该电影的偏爱程度。

4.2　实验结果评估标准

本文的实验结果评估方式主要是两方面,一个是精准度Precision,一个是召回率Recall。

精准度Precision描述的最终推荐列表中有多少比例是发生过的用户-物品评分记录,如式(5):

$$Precision=\frac{\sum_{u}|R(u)\cap T(u)|}{\sum_{u}|R(u)|} \tag{5}$$

召回率Recall反映了有多少比例的用户-物品评分记录包含在最终的推荐列表中,如式(6):

$$Recall=\frac{\sum_{u}|R(u)\cap T(u)|}{\sum_{u}|T(u)|} \tag{6}$$

对于这两方面的评估标准,其中,对用户u推荐的物品集合定为$R(u)$,用户u喜欢的物品集合为$T(u)$。

4.3　实验结果分析

本文选取精准度和召回率作为推荐算法质量的衡量标准,那么,精准度和召回率的数值越高,说明推荐结果效果越好。从Movielens电影评分数据集合中选取五组实验数据分别是50、200、400、800和943。五组数据中选四组为训练集,另一组为测试集,数据信息如表2。

表2　五组数据信息

组号	用户数目	电影数目	评价数目
1	50	1084	4797
2	200	1420	18234
3	400	1542	39286
4	800	1671	79265
5	943	1682	89326

在五组实验数据下将本文提出的算法分别进行推荐,并计算精准度和召回率,最终结果如图2、图3所示。

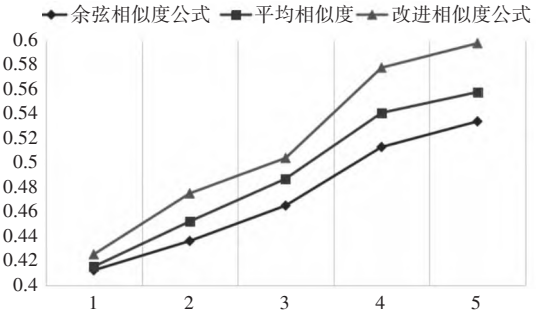

图2　精准度分布图

图2中,纵轴为精准度值,横轴为五组实验数据的组号,根据折线图,我们很明显地会发现,本文提出对余弦公式的优化是可取的,精准度越高说明,算法的效果越好,也就是,为用户推荐的商品越精确。

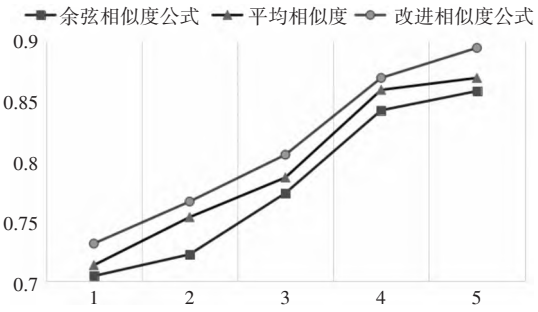

图3　召回率分布图

图3中,纵轴为召回率,横轴同样为五组实验数据的组号,根据折线图,很明显发现,召回率得到很好的提高,说明为用户推荐的商品集合$R(u)$与用户喜欢的商品集合$T(u)$的公共商品,在用户喜欢的集合$T(u)$中覆盖更广。

5　结语

传统的协同过滤推荐算法中存在着冷启动和个性化推荐两个问题,本文针对这两个问题进行深入的研究,并通过相关实验得到了改进方案后的优化实验结果。本文将UserBased推荐算法稍作优化,增加了推荐结果的精准度,并根据数据集分组进行相关实验。通过实验得到,本文的改进算法可以优化推荐算法的冷启动和个性化推荐问题,从而提高大数据环境下的数据处理能力,给用户得到更好的使用体验。

参 考 文 献

[1] Badsha S, Yi X, Khalil I, et al. Privacy Preserving User-Based Recommender System[C]// IEEE International Conference on Distributed Computing Systems,2017.

[2] Chen D, Sain S L, Guo K. Data miningfor the online retail industry: A case study of RFM model-based customer segmentation using data mining[J]. Journal of Database Marketing & Customer Strategy Management, 2012, 19(3):197-208.

[3] 王茜,王均波.一种改进的协同过滤推荐算法[J].计算机科学,2010,37(6):226-228,243.
WANG Xi, WANG Junbo. An Improved Collaborative Filtering Recommendation Algorithm [J]. Computer Science, 2010,37(6):226-228,243.

[4] Yan Jiang, Meihua Dong. Collaborative Filtering Recommendation Algorithm Based on Xml Fuzzy Data[J]. Journal of Physics Conference Series,2019,1288(1):12047.

[5] 王俊淑, 张国明, 胡斌. 基于深度学习的推荐算法研究

综述[J]. 南京师范大学学报(工程技术版), 2018, 18(04):39-49.
WANG Junshu, ZHANG Guoming, HU Bin. Review of Research on Recommendation Algorithm Based on Deep Learning [J]. Journal of Nanjing Normal University (Engineering Technology Edition), 2018, 18 (04): 39-49.

[6] 鲁鹏. 个性化推荐系统中协同过滤技术的研究[D]. 天津:河北工业大学,2013.
LU Peng. Research on Collaborative Filtering Technology in Personalized Recommendation System [D]. Tianjing: HebeiUniversity of technology, 2013.

[7] ZHANG Zike, LIU Chuang, ZHANG Yicheng, et al. Solving the cold-start problem in recommender systems with socialtags[J]. EPL, 2010, 2(2): 1-6.

[8] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 6(6): 734-749.

[9] 张麒增,戴翰波. 基于数据预处理技术的学生成绩预测模型研究[J]. 湖北大学学报(自然科学版),2019,41(1):101-108.
ZHANG Qizeng, DAI Hanbo. Research on the Prediction Model of Student Achievement Based on Data Preprocessing Technology [J]. Journal of Hubei University (Natural Science Edition), 2019, 41(1): 101-108.

[10] Salvador García, Julián Luengo, Herrera F. Data Preprocessing in Data Mining[J]. Computer Science, 2000: 72.

[11] 徐彬,杜卫锋,滕姿. 基于用户的协同过滤推荐系统的数据清洗研究[J]. 福建电脑,2017,33(8):32-34,70.
XU Bin, DU Weifeng, TENG Zi. Research onData Cleaning Based on User Collaborative Filtering Recommendation System [J]. Fujian Computer, 2017, 33 (8): 32-34, 70.

[12] 张尧,冯玉强. 数据稀疏环境下基于用户主题偏好的协同过滤算法[J]. 运筹与管理,2014(2):145-152.
ZHANG Yao, FENG Yuqiang. Collaborative Filtering Algorithm Based on User Theme Preference in Data Sparse Environment[J]. Operations Research and Management, 2014(2): 145-152.

[13] 荣辉桂,火生旭,胡春华, 等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报,2014(2):16-24.
RONG Huigui, HUO Shengxu, HU Chunhua, et al. Collaborative Filtering Recommendation Algorithm Based on User Similarity [J]. Journal of Communications, 2014 (2): 16-24.

[14] 袁利. 基于聚类的协同过滤个性化推荐算法研究[D]. 武汉:华中师范大学,2014.
YUAN Li. Research on Clustering Based Collaborative Filtering Personalized Recommendation Algorithm [D]. Wuhan: Central China Normal University, 2014.

[15] Morgan R, Gatell E, Junyent R, et al. An Improved User-Based Beach Climate Index [J]. Journal of Coastal Conservation, 2000, 6(1): 41-50.

(上接第 44 页)

LI Peng, YU Liang. An Improved Wavelet Threshold Denoising Method [J]. Modern Computer, 2016 (7): 72-75.

[18] 邱毅. 基于正交小波变换的信号降噪算法研究[J]. 电脑与电信,2008(6):67-68.
QIU Yi. Research on Signals Denoising Algorithm Based on Orthogonal Wavelet Transform[J]. Computer & Telecommunication, 2008(6): 67-68.

[19] 曹京京,胡辽林,赵瑞. 一种改进小波阈值函数的光纤光栅传感信号去噪方法[J]. 传感技术学报,2015,28(4):521-525.
CAO Jingjing, HU Liaolin, ZHAO Rui. Improved Threshold De-Noising Method of Fiber Bragg Grating Sensor Signal Based on Wavelet Transform[J]. Chinese Journal of Sensors and Actuators, 2015, 28(4): 521-525.

[20] 周帅,左东广. 基于改进阈值函数和自适应阈值的小波去噪方法[J]. 电子科技,2012,25(11):31-34.
ZHOU Shuai, ZUO Dongguang. Wavelet Denoising Basaed on Improved Threshold Function and Adaptive Threshold[J]. Electronic Science and Technology, 2012, 25(11): 31-34.

[21] 王佐成,刘晓冬,薛丽霞. 双曲线函数在灰度图像小波阈值去噪中的应用[J]. 计算机工程与应用,2010,46(35):177-179.
WANG Zuocheng, LIU Xiaodong, XUE Lixia. Application of hyperbolic functions in gray image wavelet threshold de-noising method [J]. Computer Engineering and Applications, 2010, 46(35): 177-179.

[22] 李智,张根耀,王蓓. 基于一种新的阈值函数的小波图像去噪[J]. 计算机技术与发展, 2014, 24(11): 100-102.
LI Zhi, ZHANG Genyao, WANG Bei. Wavelet Image Denoising Based on a New Threshold Function[J]. Computer Technology and Development, 2014, 24 (11): 100-102.

[23] 王瑞,张友纯. 新阈值函数下的小波阈值去噪[J]. 计算机工程与应用,2013,49(15):215-218.
WANG Rui, ZHANG Youchun. New threshold function in wavelet threshold de-noising[J]. Computer Engineering and Applications, 2013, 49(15): 215-218.