基于长尾理论的物品协同过滤Top-N推荐算法

刘向举，袁煦聪，刘鹏程

（安徽理工大学 计算机科学与工程学院，安徽 淮南 232001）

摘要：传统的基于物品的协同过滤算法在计算物品相似度时，热门商品难以与冷门商品相似，从而冷门商品的推广就更为困难。针对上述问题，在传统的物品协同过滤算法基础上，提出了一种改进的类 TF-IDF 物品相似度算法，同时考虑并排除刷分现象对物品推荐产生的负面影响，使冷门商品在基于长尾理论的推荐系统中有更高的覆盖率和准确率。以 MovieLens 上的数据集作为实验数据进行实验，实验结果表明，改进后的算法在保持甚至提高准确率的前提下，有效地提高了在推荐冷门商品时系统的覆盖率。

关键词：推荐系统；长尾理论；协同过滤；冷门商品；物品相似度；权重

中图分类号：TP391.3 **文献标志码：**A **文章编号：**1007-984X(2019)02-0001-04

当今社会正处在从信息时代到数据时代的转变时期。数据时代最显著的特点就是信息过载，如何在海量的信息中找到用户所需要的信息呢，通过搜索引擎，用户可以用关键词去搜索有用的信息，如果用户无法准确描述自己所需要的信息，那么搜索引擎也无法提供有价值的结果，而推荐系统则可以通过分析用户的历史行为记录，向用户推荐他们可能感兴趣的物品和内容。

在现有的商品推荐系统中，使用最广的依然是基于项目的协同过滤算法[1]，该算法主要根据用户的历史浏览记录来分析并比较项目之间的相似度，找出与当前项目相似度最高，与用户历史兴趣度最接近的 N 个项目推荐给用户。相比基于用户的协同过滤算法，该算法在长尾推荐中有着更好的覆盖率，文献[2]通过提取用户浏览特征发现，访问特定浏览页面的前十位用户在信息传播中起着关键作用，并因此加入了新的机制向用户进行长尾推荐。文献[3]通过检索用户的相对不受欢迎的受到的关注度逐渐提高的物品，采用的新的策略进行推广，构造了一个新的用户兴趣综合得分，通过加权用户信息和相关性，调整用户对物品的兴趣度权重，使用户偏向那些不太受欢迎的物品。文献[4]针对冷门商品推荐中，用户信息矩阵稀疏的问题，使用 K 均值聚类对用户先进行了分类，并对用户隐形的物品信息反馈时采用不同的比重。文献[5]考虑长尾推荐时准确度不足的问题，引入长尾物品的群组推荐算法，以群组用户满意度和物品流行度为目标函数，提高了推荐结果的多样性和新颖性。这些改进算法在进行长尾推荐时有了进一步的拓展，但是依然存在着问题：在长尾推荐中，热门物品依然更倾向与热门物品相似，如果一个用户喜欢热门物品，那么在对用户进行推荐时，很难推荐尾部冷门的物品。

本文提出了一种类 TF-IDF 的相似度算法，改进算法对热门物品的评论数进行惩罚，提高冷门物品评论数的权重，在提高冷门物品评论权重的同时，也考虑到了刷分现象对项目相似度的影响。

1 长尾理论

长尾理论是由美国杂志《连线》主编 Chris Anderson 在分析并研究了 Google、亚马逊、Netflix 等互联网零售商的销售数据之后，提出了“长尾”概念[6]，该理论认为只要物品的存储和流通渠道足够大，那些分布在尾部的 80%零碎商品市场聚集起来的商业规模就能超过图 1 中曲线前端 20%的热销商品的商业规模，传统的零售商因为存储的限制不可能把所有商品呈现在用户面前，考虑到成本的因素，往往只能对那些最受用户欢迎的 20%进行销售，互联网的出现，克服了传统零售商的各种困难，在信息传播成本为零，存储空间足够大，商品的边际成本趋于零的情况下，尽管那些尾部的商品销售量没有畅销商品那么惊人，但这些商品累计销售的规模却足以让人重视。对于新用户，用户因为没有对任何物品有过历史记录，推荐系统

收稿日期：2018-10-31
基金项目：国家自然科学基金资助项目（51504010），安徽省高校省级质量工程重点教研项目（2017jyxm0185）
作者简介：刘向举（1978-），男，黑龙江双城人，副教授，硕士，主要从事数据挖掘和物联网方向的研究，xjliu@aust.edu.cn。

往往推荐最热门的物品，而对于经常购买物品的用户来说，更关注的是冷门物品[7]，在新的长尾互联网经济模式下，本文在重点挖掘对冷门物品推荐的同时，也考虑到在长尾推荐中可能会出现的弊端，减少那些因用户刷分现象而进入长尾推荐列表的商品。

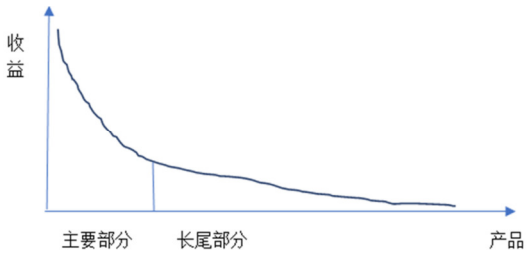

图1 长尾理论示意图

2 算法

2.1 传统的物品相似度计算方法

算法思想：对物品有过行为记录的用户作为物品的特征向量，计算物品之间的的相似度，从用户已经有过历史行为的物品中找到与其相似的 K 个物品，并推荐 Top–N 个相关物品作为该物品的推荐集。计算物品相似度常见的算法有如下3种。

（1）余弦相似度。普通余弦相似度，R_{ui} 是用户 u 对物品 i 的评分，R_{uj} 是用户 u 对物品 j 的评分，物品 i 和物品 j 的相似度 W_{ij} 则可以表示为[8]式（1）

$$W_{ij}=\frac{\sum_{u\in U}R_{ui}*R_{uj}}{\sqrt{\sum_{u\in U}R^2{}_{ui}}*\sqrt{\sum_{u\in U}R^2{}_{uj}}} \tag{1}$$

（2）基于关联的相似度计算。 基于皮尔森逊相关系数，计算两个物品之前的相似度，R_{ui} 表示用户 u 对物品 i 的打分，R_i 表示第 i 个物品打分的平均值，式（2）在式（1）的基础上进行了去中心化处理，减去了物品 i 和物品 j 的评分均值。则物品 i 和物品 j 的相似度 W_{ij} 可表示为[9]式（2）

$$W_{ij}=\frac{\sum_{u\in U}(R_{ui}-R_u)(R_{uj}-R_j)}{\sqrt{\sum_{u\in U}(R^2{}_{ui}-R_i)^2}\sqrt{\sum_{u\in U}(R_{uj}-R_j)^2}} \tag{2}$$

（3）Jaccard 相似度。W_{ij} 表示物品 i 和物品 j 之间的相似度，$N(i)$表示喜欢物品 i 的用户集合，$N(j)$表示喜欢物品 j 的用户集合，则物品 i 和物品 j 的相似度可表示为式（3）

$$W_{ij}=\frac{|N(i)\cap N(j)|}{\sqrt{|N(i)||N(j)|}} \tag{3}$$

2.2 本文算法描述

研究表明，计算用户相似度时使用皮尔逊相关系数法比其他的相似度计算方法更好一些，而从算法的行为相关性度量的比较和计算的角度，本文使用 jaccard 相似度算法计算物品之间的相似度[10]。从 jaccard 相似度算法可以看出，两个物品相似是因为它们共同被很多用户喜欢，但是也存在着一个问题[11]，假设物品 j 太过热门，购买过物品 i 的用户大部分都会购买物品 j，那么式（3）分子中 $|N(i)\cap N(j)|$ 就会越来越接近 $|N(j)|$，热门物品 j 就会获得一个比较大的相似度，这对挖掘长尾信息的推荐系统来说则会出现这样的现象：热门物品倾向和热门物品相似，冷门物品倾向和冷门物品相似[6]，也就是说，如果一个用户喜欢一个热门商品，很难推荐一个冷门商品给他，因此，如何给用户更好的推荐冷门商品，成为一个急需解决的问题。本文引入物品评论数的权重因子，如式（4）所示，其中 $I1$ 为用户总评论数，$N(j)$为物品 j 的评论数。

$$\alpha(j)=\log\frac{|I1|}{(N(j))} \tag{4}$$

权重因子α在惩罚热门物品的同时，提高了冷门物品在计算物品相似度时的权重比，改进的类 TF–IDF 物品相似度算法式（5）所示，另外用式（5）计算物品相似度时，考虑当向用户推荐长尾物品时，防止部分物品因为刷分而进入用户推荐列表，所以加入一个阈值 x，当用户对物品的评价均分为 x 时，判定该用户群体有很大程度上的刷分嫌疑，在计算新的物品相似度前，需对这部分用户进行排除，从而提高推荐结果的准确度[12]。

$$w_{ij}=\frac{|N(i)\cap N(j)|}{\sqrt{|N(i)||N(j)|}}*\alpha(j) \tag{5}$$

2.3　算法实现描述

基于以上分析，本文在新的物品相似度改进算法基础上进行实验，基本步骤如下：

（1）输入：movielens 数据集。

（2）输出：推荐项目，召回率、覆盖率、准确率及平均流行度。

算法设计：

（1）加载并读取 movielens 数据集，将数据集随机分取 10 份，其中 2 份作为测试集，8 份作为训练集。

（2）获取用户–物品的倒排表。

（3）计算项目相似度时，采用新的物品相似度计算方法，并考虑刷分现象对物品推荐产生的影响，设计阈值 x（用户对所有物品评分的平均值）。

（4）找出与当前项目相似度最高，兴趣度最接近的 N 个项目推荐给用户。

（5）计算覆盖率，准确率，召回率，平均流行度。

3　算法实验分析

3.1　实验环境与评价标准

本文所使用的实验语言为 Python，实验平台为 Pycharm，测试数据集使用 Movielens 数据集，包括 943 个用户对 1 682 项目的 10 万条评论[13]，数据集的稀疏度为 93.7%，评分范围为 1 ~ 5。本文希望在保持甚至提高推荐结果准确度的情况下，更有效地挖掘数据集中的长尾信息并进行推荐，且覆盖率为本次实验的重点评价标准[14]。覆盖率测量的是系统推荐给用户的物品占所有物品的比例，覆盖率越高，越能说明推荐算法能将尾部中的物品推荐给用户。准确率和召回率是测量推荐系统准确性的重要指标，准确率测量的是用户喜欢的物品占被推荐列表中的比例，平均流行度是测量推荐系统新颖程度最简单的方法，用户接触到的物品越不热门，则新颖度越高。召回率描述的是用户对物品评分记录在最终推荐列表所占比例。一个好的长尾推荐系统，应该具有较高的准确率和覆盖率，且平均流行度应低于传统的推荐系统[15]。本实验所采用的 4 个指标为准确率(*Precision*)、覆盖率(*Coverage*)、平均流行度(*PopularityAVG*)和召回率(*Recall*)[16]，分别为：

$$Precision=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|};\ Coverage=\frac{\left|\sum_{u\in U}R(u)\right|}{|I|};\ PopularityAVG=\frac{\sum_{i\in I(u)}item\_pop(i)}{N};\ \mathrm{Re}call=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|}$$

其中，U 为用户集合，$R(u)$表示为对每个用户推荐长度为 N 的物品列表，$T(u)$为每个用户在测试集中有过行为记录的物品集合，$I(u)$为用户 u 所得到的物品推荐集合，$|I|$ 表示数据集 I 中所包含的物品数量，N 表示长度为 N 的推荐列表，$item_pop(i)$为物品 i 的流行度。

3.2　实验结果分析

在所有实验中，推荐数目 N=10，邻居数目 k 的范围为 5 ~ 100 之间。传统的基于物品的协同过滤算法在不同近邻 K 值下的表现如表 1 所示。

改进算法在采用新的物品相似度计算时，在不同近邻 K 值下的表现如表 2 所示。

表 1　传统基于物品的协同过滤算法在 N=10，K=5~100 的实验结果

参数 K	准确率/%	覆盖率/%	平均流行度/%
5	12.683	16.300	5.401
10	12.725	11.600	5.517
20	12.927	9.585	5.529
40	12.884	9.158	5.525
80	13.022	9.890	5.531
100	12.895	10.012.	5.535

表 2　改进的算法在 N=10，K=5~100 的实验结果

参数 K	准确率	覆盖率	平均流行度
5	12.439%	26.984%	5.095%
10	13.234%	18.315%	5.349%
20	13.415%	14.835%	5.436%
40	13.616%	12.943%	5.462%
80	13.531%	12.821%	5.474%
100	13.595%	12.149%.	5.487%

由表 1 和表 2 可知，新算法在准确率、覆盖率、平均流行度上均有明显的改善，且在 K=80 时有着良

好表现，在此改进算法基础上考虑刷分现象对推荐效果的影响，设定用户对所有物品的平均评分 x，变化阈值 x 的变化范围为 4.5～4.9，当 K=80，x 发生变化时，算法性能变化如图 2 所示。

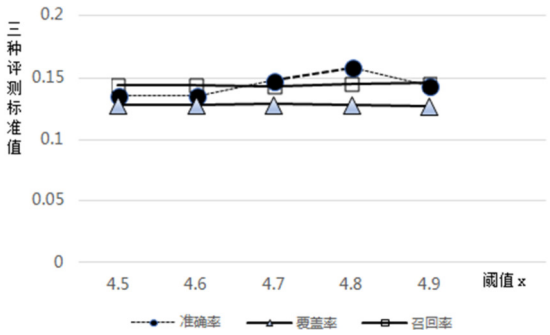

图 2　用户评论均分对推荐结果影响

由图 2 可以看出，在改进算法的基础上，当排除评价均分 x=4.8 的用户群体时，召回率和覆盖率几乎没有变动的情况下，准确率达到了 15.832%，说明该用户群体在所有可能刷分的用户群体中所占比例最高，对推荐结果影响最大。与传统算法相比，最终改进算法在降低流行度的同时，覆盖率、召回率和准确率都有明显的提升，说明改进的物品相似度计算方法和考虑用户评分行为对提高尾部冷门商品的挖掘有着一定的改善作用，以下是近邻 K=80，最终改进算法和传统算法的对比结果如表 3 所示。

表 3　最终改进算法和传统算法对比

算法名称	准确率/%	召回率/%	覆盖率/%	流行度
传统算法	13.022	13.907	9.890	5.531
最终改进算法	15.832	14.451	12.759	5.423

4　结束语

本文提出的物品相似度改进算法和考虑用户评分对推荐结果影响的两个因素，通过离线实验，改进后的算法在一定程度上提升了尾部商品的推荐能力，改善尾部推荐不仅能扩大经济效益，更能给用户带来差异化体验，提高用户的体验度。后续工作在冷门物品随时间函数变化与热门物品是否会进行转变是有意义的下一步的研究方向。

参考文献：

[1] 刘冠军.基于流行性预测的推荐算法研究[D].北京：电子科技大学，2013

[2] Masayuki Ishikawa, Peter Geczy, Noriaki Izumi, Takahira Yamaguchi,Long Tail Recommender Utilizing Information Diffusion Theory[C]. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008(1): 785–788

[3] Mi Zhang,NeiHurley,WeiLi,XiangyangXue, A Double–Ranking Strategy for Long–Tail Product Recommendation[C]. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.2012(1):282–286

[4] 陈联平.关于电商平台冷门商品的推荐系统研究[D].昆明：云南财经大学，2018

[5] 韩亚敏，柴争义，李亚伦，等.长尾群组推荐的免疫多目标优化实现[J].西安：西安电子科技大学学报，2018, 45(3):110–116

[6] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. “Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes”[C]. ACM WSDM, 2012,10:201–210

[7] Yehuda Koren.Factor in the Neighbors: Scalable and Accurate Collaborative Filteriing[C].ACM Transactions on Knowledge Discovery from Data, 2010,4(1):1–24

[8] Badrul Sarwar, George Karypis, Joseph Karypis, John Ried, Item–based Collaborative Filtering Recommendation Algorithms[R].2001, 4(1):285–295

[9] 项亮.推荐系统实践[M].北京：人民邮电出版社，2012:45–50

[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):17–24

[11] 杨博，赵鹏飞.推荐算法综述[J].山西大学学报，2011,34(3):337–350

[12] 赫立燕，王靖.基于项目流行度的协同过滤 TopN 推荐算法[J].计算机工程与设计，2013,34(10): 3497–3501

[13] 邓爱林，朱扬勇，施伯乐.基于项目评分预测的协同过滤推荐算法[J].软件学报，2003,14(9):1621–1628

[14] 汪静，印鉴.一种优化的 Item–based 协同过滤推荐算法[J].计算机技术与发展，2010,31(12):2237–2342

[15] 郑苏洋，姜久雷，王晓峰.基于用户项目体验度的协同过滤推荐算法[J].计算机工程，2017,43(8):215–218,224

[16] 孙竹.基于商品关系改进的协同过滤推荐算法[D].秦皇岛：燕山大学，2016

（下转第 9 页）

[10] 邱立达，刘天键，傅平.基于深度学习的无线传感器网络数据融合[J].计算机应用研究，2016,33(01):185-188
[11] 王丽红，于光华，刘平.无线传感器网络 LEACH 算法的改进研究[J].齐齐哈尔大学学报：自然科学版，2018(03):2-4

Routing protocol algorithm for wireless sensor networks based on deep learning model

WANG Li-hong，SHAO Hui

(School of Computer and Information Engineering, Heihe University, Heilongjiang Heihe 164399,China)

Abstract：To reduce the energy consumption and prolong the lifetime of wireless sensor network (WSN), a WSN routing protocol algorithm based on deep learning model is proposed. Firstly, the algorithm completes training and clustering at the sink node, transfers the trained parameters to each cluster node, and then transfers the collected data to the sink node after feature classification, extraction and fusion. In order to make the distribution of cluster heads more uniform, the clustering method is improved on the basis of estimating the optimal number of cluster heads, which reduces the number of clusters and saves the energy consumption of the network. The simulation results show that the WSN routing protocol algorithm based on cascade automatic encoder reduces the network energy consumption, prolongs the network life cycle, and is more suitable for large-scale long-distance communication.

Key words: wireless sensor network；deep learning；autoencoder；routing protocol

（上接第 14 页）

Top-N recommendation algorithm for item collaborative filtering based on long tail theory

LIU Xiang-ju，YUAN Xu-cong，LIU Peng-cheng

(School of Computer Science and Engineering, AnhuiUniversity of Science and Technology, Anhui Huainan 232001, China)

Abstract:In the traditional item-based collaborative filtering algorithm, unpopular items is difficult to compare with popular products when calculating the similarity of items. To solve the above problems, an improved TF-IDF similarity algorithm is proposed on the traditional item collaborative filtering algorithm. At the same time, consider and eliminate the negative impact of the score cheating phenomenon on the recommendation system, so that the unpopular goods have higher coverage and accuracy in the recommendation system based on the long tail theory. Experiments on the dataset on the MovieLens show that the improved algorithm effectively improves the coverage of the system when recommending unpopular products while maintaining or even improving the accuracy.

Key words: recommend system；long tail theory；collaborative filtering；unpopular item；weight