

基于二分网络社团检测的协同过滤推荐算法研究

高梦瑶¹, 刘晨晨², 邓治文^{1*}

(1.新疆科技学院, 新疆 库尔勒 841000; 2.新疆移动伊犁分公司, 新疆 伊宁 835000)

摘要: 文章提出了一种基于二分网络社团检测的协同过滤推荐算法 (Bipartite Network Community Detection-Based Collaborative Filtering Recommendation Algorithm, BCD-CF), 以优化“数据过载”问题: 首先, 通过用户评分数据挖掘用户兴趣, 构建用户-项目兴趣二分网络; 其次, 运用谱聚类思想对用户-项目兴趣二分网络进行社团检测; 最后, 在每个社团内借助修正的 Pearson 相关系数寻找邻居用户进行推荐。

关键词: 二分网络; 社团检测; 协同过滤; 修正 Pearson 相关系数

中图分类号: TP391

文献标识码: A

文章编号: 1003-9767 (2025) 08-040-03

Research on Collaborative Filtering Recommendation Algorithm Based on Bipartite Network Community Detection

GAO Mengyao¹, LIU Chenchen², DENG Zhiwen^{1*}

(1.Xinjiang College of Science & Technology, Korla Xinjiang 841000, China;

2.Yili Branch of Xinjiang Mobile,Yining Xinjiang 835000, China)

Abstract: The paper proposes a bipartite network community detection-based collaborative filtering recommendation algorithm (BCD-CF) to optimize the “data overload” problem. Firstly, user interests are mined through user rating data, and a user-project interest bipartite network is constructed. Secondly, the idea of spectral clustering is used to perform community detection on the user-project interest bipartite network. Finally, within each community, the modified Pearson correlation coefficient is used to find neighboring users for recommendation.

Keywords: bipartite network; community detection; collaborative filtering; modified Pearson correlation coefficient

0 引言

如今“信息过载”问题日益加剧,这使得用户在短时间内难以筛选出所需信息。同时,各个网站在准确推荐用户感兴趣的内容方面也面临挑战。因此,推荐系统至今都受到许多学者的关注^[1]。

推荐系统的核心是推荐算法,推荐算法主要分为基于协同过滤的推荐算法^[2]、基于内容的推荐算法^[3]和基于关联规则的推荐算法^[4],其中应用最广泛的是基于协同过滤的推荐算法。传统的协同过滤推荐算法是通过用户-项目的评分数据实现推荐的,但随着用户项目数量

呈指数形式的增长,用户-项目矩阵越来越稀疏,不仅会导致推荐速率的降低,推荐质量也会受到影响。

针对上述推荐速率降低的问题,许多学者将聚类算法与传统的协同过滤推荐算法相结合,或通过对传统聚类算法改进后与协同过滤算法共同实现推荐,将用户分为不同的类别,当目标用户需要推荐时,只需将其分配至相似度最大的类别中进行推荐^[5-7]。

当前,为提升推荐速率并解决相似度计算方式绝对值不敏感所导致的推荐算法质量问题,文章提出了一种基于二分网络社团检测的协同过滤推荐算法。

收稿日期: 2025-02-22

作者简介: 高梦瑶,女,硕士,助教。研究方向: 调和分析。

通信作者: 邓治文,男,硕士,助教。研究方向: 数据智能分析。

1 BCD-CF 推荐模型的构建

首先,研究通过用户的历史评分数据挖掘用户兴趣并构建用户-项目兴趣二分网络;通过奇异值分解-多尺度社团检测算法(Singular Value Decomposition-MultiScale, SVD-MS)实现社团检测^[8],以期同时将用户和项目划分到同一个社团减少推荐运行时间;通过用户评分相似性和用户评分项目数量对 Pearson 相关系数修正,以期提高邻居用户质量从而提高推荐质量;最后,在每一个小社团内实现基于用户的协同过滤推荐。

1.1 构建用户-项目兴趣二分网络

用户和项目是两类节点,边只存在于不同类型的节点之间,表示用户喜欢该项目。用户对项目的评分高低反映了他们对项目的兴趣程度。本文采用每个用户所给出评分的平均值作为阈值,以此来判断每个用户对兴趣项目的偏好。设用户 u_i 对 n 个项目进行了评分 $R_i = \{R_{i1}, R_{i2}, \dots, R_{in}\}$, 则用户 u_i 的阈值如式(1)所示:

$$\bar{R}_i = \frac{R_{i1} + R_{i2} + \dots + R_{in}}{n} \quad (1)$$

当 $R_{ij} \geq \bar{R}_i$, 则说明用户 u_i 喜欢项目 j , 用户 u_i 和项目 j 就存在一条连边;反之,则不存在。这样,就构成了一个用户-项目兴趣二分网络。

1.2 用户的相似度计算

Pearson 相关系数是协同过滤算法中最常用的相似度计算方法之一,但在实际应用中忽略了用户评分习惯和评分项目数量的影响。为此,将评分相似因子作为权重,并考虑评分项目的数量,共同对 Pearson 相关系数进行修正,计算步骤如下。

(1) 假设用户 u_1 和 u_2 共同评分的项目集合为 $I = \{I_1, I_2, \dots, I_t\}$, 用户 u_1 和 u_2 共同评分的项目分别为 $u_1 = \{I_{u_1, I_1}, I_{u_1, I_2}, \dots, I_{u_1, I_t}\}$ 和 $u_2 = \{I_{u_2, I_1}, I_{u_2, I_2}, \dots, I_{u_2, I_t}\}$, 将用户 u_1 和 u_2 的评分差异度 $d(u_1, u_2)$ 定义为式(2):

$$d(u_1, u_2) = (R_{u_1, I_1} - R_{u_2, I_1}, R_{u_1, I_2} - R_{u_2, I_2}, \dots, R_{u_1, I_t} - R_{u_2, I_t}) \quad (2)$$

$$= (d_1, d_2, \dots, d_t)$$

利用欧氏距离得出两用户之间的距离,为了在两个用户评分完全相同时, $d_{ag}(u_1, u_2)$ 仍然有意义,则用户 u_1 和 u_2 评分相似度 $d_{ag}(u_1, u_2)$, 如式(3)所示:

$$d_{ag}(u_1, u_2) = \frac{1}{1 + d_{ed}(u_1, u_2)} \quad (3)$$

其中, $d_{ed}(u_1, u_2)$ 为用户 u_1 和 u_2 之间的欧几里得距离。两个用户共同评分项目的数量越多,用户的相似度越大。

以杰卡德距离作为权重,则新的用户评分相似度计算公式为式(4):

$$d_{ag}(u_1, u_2) = \frac{J(u_1, u_2)}{1 + d_{ed}(u_1, u_2)} \quad (4)$$

其中, $J(u_1, u_2)$ 为杰卡德距离,取值范围为 0 到 1。

1.3 社团内推荐

实现二分网络社团检测后,每个用户和项目都有所属社团,每个社团是一个小型的用户-项目兴趣二分网络,可将其直接转换为用户-项目评分矩阵。在寻找目标用户的邻居用户时,首先计算两个用户之间的相似度,再对社团内未评分的项目进行评分预测,最后将评分最高的前 n 个项目推荐给目标用户。

2 实验结果及分析

2.1 实验数据来源

文章使用了 MovieLens 100k 数据集^[9]。此次实验选取的数据集涵盖了 943 位用户对 1682 部电影的十万条评分记录,数据稀疏度高达 98.367%,每位用户至少评价了 20 部电影。评分范围为 1~5,数值由低到高代表用户的喜好程度。实验选择 80% 的数据作为训练集,20% 的数据作为测试集。

2.2 评价标准

实验采用平均绝对偏差 (Mean Absolute Error, MAE) 和均方根误差 (Root Mean Square Error, RMSE) 衡量 BCD-CF 算法的质量。MAE (RMSE) 值越大,说明预测值和真实值之间的差别越大,推荐精度越低;反之,则推荐精度越高。除此之外,还采用了召回率 (Recall) 和准确率 (Precision) 进一步说明文中所提出算法的有效性。

2.3 实验结果及分析

2.3.1 社团数量的确定

社团数目对推荐算法有重要影响,社团数目过小会导致每个社团内部的用户和项目数量过多,没有起到缓解“数据过载”的问题,过多则影响推荐效果。对该数据进行多次迭代实验发现,当设定划分为 4 个社团时,可以得到最大模块度 $Q_B = 0.471$ 。第一、二、三、四个社团中分别包括 243、193、251、255 个用户和 423、538、441、278 个项目。

2.3.2 与其他推荐算法精度的对比

对数据进行社团划分后,通过比较 BCD-CF 算法、CF 算法和 KM-CF 算法的 MAE 和 RMSE 值,研究了这三种推荐算法的推荐精度。在实验中,最近邻居数目区间为 [5,50],并以 5 为增量进行调整。实验结果分别如图 1 和图 2 所示。

结果表明,随着最近邻居数目的增加,所有推荐算法的 MAE 值都呈先减后增的趋势,且在邻居用户数目为 30 时达到最低值。RMSE 值逐渐下降,并趋于稳定。在最近邻居数目为任意值时,BCD-CF 推荐算法都比其他推荐算法值低,即改进后的推荐算法比传统推荐算法的推荐效果好。

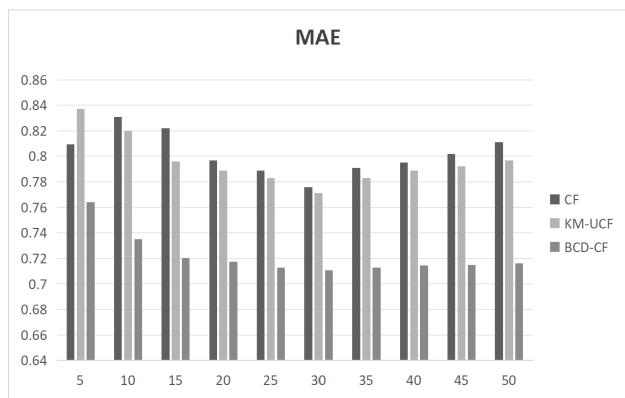


图1 协同过滤推荐算法 MAE 值的比较

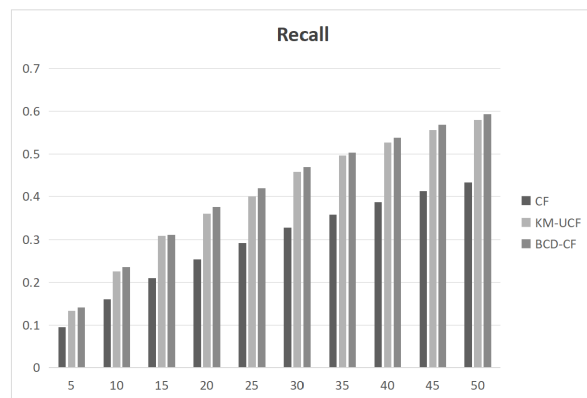


图4 各推荐算法召回率对比

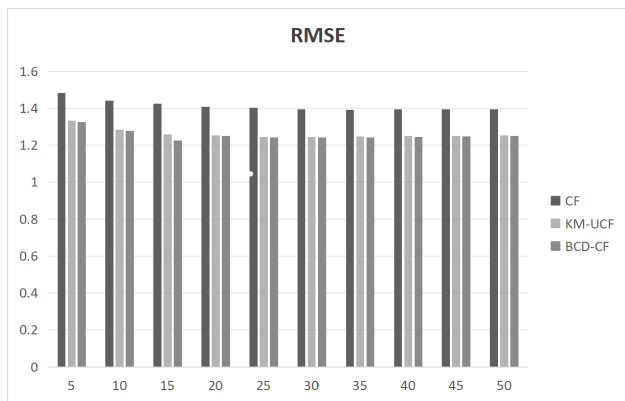


图2 协同过滤推荐算法 RMSE 值的比较

进一步实验研究推荐列表数量对推荐效果的影响,根据前面的结果将最近邻居设置为 30,将推荐列表的区间设置为 [5,50],间隔为 10,观察推荐列表长度对各推荐算法的召回率、准确率和 F_1 值的影响,实验结果分别如图 3、图 4 所示。

如图所示,各推荐算法的准确率随着推荐列表长度的增加而呈下降趋势,召回率呈上升趋势,且 BCD-CF 推荐算法在推荐列表为任何长度时的准确率和召回率都高于其他算法。召回率和准确率两个度量标准相互矛盾,当召回率增加时,准确率会下降;反之,召回率减少时,准确率会上升。

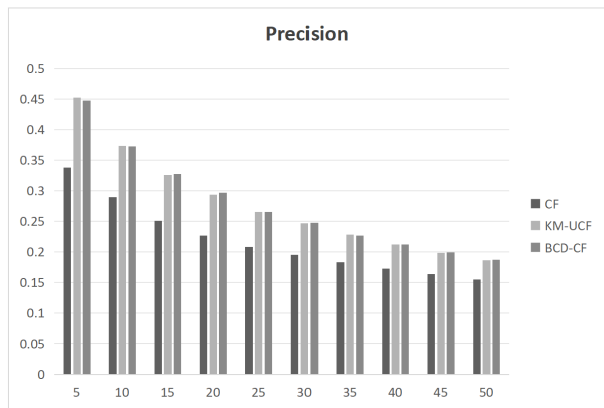


图3 各推荐算法准确率对比

3 结 语

为了缓解“数据过载”问题对传统推荐算法效率和质量的影响,本文从二分网络社团检测的角度出发,提出了一种基于二分网络社团检测的协同过滤推荐算法。该方法能够更有效地根据用户兴趣划分用户和项目,从而减少推荐运行时间。实验结果表明,所提算法不仅缩小了邻居用户的搜索范围,提高了邻居用户的质量,还显著提升了推荐效率,并有效缓解了数据稀疏性对推荐质量的影响。

参考文献

- [1] 林啸轩,季一木,刘尚东,等.融合数据挖掘和评分预测的推荐算法[J].南京邮电大学学报(自然科学版),2024,44(01):101-108.
- [2] 刘宇,朱文浩.基于内容和标签权重的混合推荐算法[J].计算机与数字工程,2020,48(04):773-777.
- [3] 向程冠,周东波,李雷,等.基于关联规则与知识追踪的编程题目推荐算法[J].计算机工程与设计,2022,43(11):3135-3142.
- [4] 林建辉,王茜冉,詹可强.基于聚类与差异协调的协同过滤推荐算法[J].兰州文理学院学报(自然科学版),2023,37(06):50-54.
- [5] 李希文.密度聚类研究及其在电影推荐算法中的应用[D].兰州:西北师范大学,2021.
- [6] 施天虎,徐洪珍.基于改进K-means和优化评分的协同过滤推荐算法[J].江苏科技大学学报(自然科学版),2021,35(06):72-77.
- [7] 孙海岗,李玲娟.融合隐性社交网络社团划分和协同过滤的推荐算法[J].南京邮电大学学报(自然科学版),2023,43(04):93-100.
- [8] 刘晨晨,许英.基于谱聚类的二分网络社团检测算法[J].吉首大学学报(自然科学版),2023,44(06):9-13+19.
- [9] 刘晓蒙.基于协同过滤的学习资源推荐算法[J].信息与电脑(理论版),2023,35(01):63-65.