Hans Journal of Data Mining 数据挖掘, 2021, 11(3), 181-186
Published Online July 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.113016

Hans汉斯

基于修正Tanimoto系数的电视节目个性化推荐方法研究

王子涛[1]，伍发珍[1]，江铭海[1]，董云薪[2]，林　耿[1]

[1]闽江学院数学与数据科学学院，福建 福州
[2]福建农林大学计算机与信息学院，福建 福州
Email: 976674498@qq.com

收稿日期：2021年5月25日；录用日期：2021年6月25日；发布日期：2021年7月5日

摘　要

随着互联网产业的发展，电视节目的个性化推荐已经成为了许多电视节目制作者和观众们的共同需求。为达到电视节目个性化推荐的目的，本文通过结合用户共同评分项和用户所有评分项之间的关系，将基于修正Tanimoto系数的推荐算法运用于电视节目个性化推荐的研究中。实验结果表明，本文所运用的算法具有一定的提高电视节目个性化推荐系统的精度的效果。

关键词

Tanimoto系数，相似度，协同过滤算法，电视节目推荐

Research on Personalized Recommendation Method of TV Programs Based on Modified Tanimoto Coefficient

Zitao Wang[1], Fazhen Wu[1], Minghai Jiang[1], Yunxin Dong[2], Geng Lin[1]

[1]College of Mathematics and Data Science, Minjiang University, Fuzhou Fujian
[2]College of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou Fujian
Email: 976674498@qq.com

Received: May 25th, 2021; accepted: Jun. 25th, 2021; published: Jul. 5th, 2021

Abstract

With the development of the Internet industry, the individualization of TV programs has become the common demand of many TV program makers and viewers. In order to achieve the purpose of

文章引用: 王子涛, 伍发珍, 江铭海, 董云薪, 林耿. 基于修正 Tanimoto 系数的电视节目个性化推荐方法研究[J]. 数据挖掘, 2021, 11(3): 181-186. DOI: 10.12677/hjdm.2021.113016

personalized recommendation of TV programs, this paper combined the relationship between user common rating items and user all rating items, and applied the collaborative filtering recommendation algorithm modified Tanimoto coefficient to improve similarity measure in the research of personalized recommendation of TV programs. Experimental results show that the algorithm used in this paper can improve the accuracy of TV program personalized recommendation system to a certain extent.

Keywords

Tanimoto Coefficient, Similarity, Collaborative Filtering Algorithm, TV Program Recommendation

1. 引言

伴随着互联网的迅猛发展，信息化的浪潮逐渐席卷各行各业，报纸新闻、广播、电视节目、电影等传统媒体都面临着向互联网产业转型的时代需求。对于报纸、电影等传统媒介，由于其内容的整体性，很容易作为一个个媒体单元出现在互联网上。然而，以各类电视剧为主的电视节目媒体，却由于整部电视剧的体量庞大，难以在碎片化阅读的时代获得观众们的持续关注。此外，由于信息爆炸的现象，观众也难以从浩如烟海的电视节目中选取自己喜爱的节目。做好电视节目的推荐算法，对观众来说，能够为观众减少大量用以寻找符合个人口味的电视节目的时间，达到增加效率的目的；对于电视节目制作方来说，能够增加节目在对应受众中的曝光度，降低许多无效的宣传费用，获得更好的经济效益。

现有的各种推荐算法大多是基于用户兴趣数据进行推荐，例如以点赞数为主导的短视频推荐算法、搜索记录为主导的热门信息推荐算法、以用户过往评分为主导的评价推荐算法等。这些推荐算法的共同特点是需要大量的用户基础，在用户群体数量足够大的情况下，作出大范围的用户画像，推荐给个人用户的往往是大多数人喜欢的内容。这就导致现有的推荐算法容易引导用户查看更加热门的内容，使得热门内容更加热门，冷门内容无人问津。然而每位用户都是相对独立的个体，其所选择的内容更多是基于个人喜好的，与大众喜好有可能存在一定的差异，这部分用户就难以寻找到与个人爱好相似的内容，从而导致潜在的用户流失风险。

因此，近年来许多文献致力于改进相似性度量方法达到更好的内容推荐效果。Banagans [1]等人选取了收藏列表、观看时长、点播记录三个方面的数据，对基于评分的用户兴趣偏好进行加权处理，进而优化基于用户偏好的相似度算法。宋月亭[2]基于用户隐式评分特征的相似度优化，利用流形学习改进聚类算法的方式改进了协同过滤算法。陈娅昵、苏岐芳[3]等人在相似度计算中加入了相对偏好和绝对偏好，改进了协同过滤推荐算法。虽然这些算法在一定程度上改善了在现有的相似性度量下的协同过滤推荐的效果，但仍存在针对用户个性化、精确化推荐性能不够完善的问题。

就用户行为而言，每个人都有过对自己喜好的内容反复观看，而对自己不喜欢的内容直接忽略或中途中断的经历，这就提示我们从用户对内容的观看时长出发，利用用户对内容的观看时长的相似性，寻找更加精准的用户喜好，进而达到改进推荐算法的目的。本文针对现有的各种推荐算法中对于相似性度量部分的考虑，提出一种修正 Tanimoto 系数改进相似性度量的算法，以用户对内容的观看时长为数据基础，针对用户个性化的观看时长，达到针对用户个人喜好的精准推荐。同时，对于没有任何观看时长的

初始用户，将利用对其提问的方式，令用户事先选择好自己喜爱的领域，作为本文相似度算法的初始值。改进的相似性度量算法，能够有效规避传统相似度度量中的大多数人数据，在用户个性化行为的基础上作出推荐。

2. 算法步骤

本文通过结合用户共同评分项和用户所有评分项之间的关系，利用传统相似度计算方法和加入修正Tanimoto 系数，二者不同权重来计算用户之间相似度，预测评分，得出推荐结果，根据此方法得出的结果准确度较高，效果较好。具体算法步骤如下：

Step 1 获取用户－项目评分矩阵

协同过滤算法的输入数据通常表示为 $a \times b$ 的矩阵 R [4]，其中，矩阵中的每一行代表系统中的一个用户，每一列代表系统中的每一个项目(在本文中表示电视节目)，用 U 表示用户集合，用 I 表示项目集合，a 行表示有 a 个用户，b 列表示有 b 个项目，R_{ij} 表示矩阵中第 i 行第 j 列的数据，代表用户 i 对项目 j 的评分。用 $sim(m,n)$ 表示用户 m 与用户 n 之间的相似度[5]。R_ω 表示用户 m 和 n 共同评分过的项目集合，R_m 、R_n 分别表示用户 m 和用户 n 评分过的项目集合，相似度计算公式为式(2-1)。

$$sim(m,n)=\frac{\sum_{i\in R_\omega}\left(R_{m,i}-\bar{R}_m\right)\times\left(R_{n,i}-\bar{R}_n\right)}{\sqrt{\sum_{i\in R_m}\left(R_{m,i}-\bar{R}_m\right)}\times\sqrt{\sum_{i\in R_n}\left(R_{n,i}-\bar{R}_n\right)}} \tag{2-1}$$

Step 2 找到最近邻居用户集合

通过相似度计算之后，对于目标用户 m，根据其与系统中其他用户相似度排序后，可以找到与该目标用户的邻居集合 $M=\{m_1,m_2,\cdots,m_i \mid m\notin M\}$，相似度由大到小排序。$M$ 表示项目 b 的 i 个邻居集合。

Step 3 计算共同评分项目占用比

假设有两个用户 m_1,m_2，他们感兴趣的电视节目相似，则他们共同评分的项目在二者所有已经评过分的项目中所占的比例较大，记为共同评分占用比。可以有效解决当这两个用户有且只有一个共同评分项目 n_1 且评分值相同时造成相似度为 1 的问题。在本文中，利用 Tanimoto 系数表示目标 m、n 之间的共同评分项目占用比，记为 $T(m,n)$，其表达式为：

$$T(m,n)=\frac{\sum_{i\in I}(m_i\wedge n_i)}{\sum_{i\in I}(m_i\vee n_i)} \tag{2-2}$$

从用户对项目评分的角度来看，共同评分项目占用比越高，则说明用户之间感兴趣的电视节目越相似；相反，共同评分项目占用比越低，说明二者的相似度越低。可以得出结论，T 系数值越大，用户 m 和 n 的整体相似度越高。其中，Tanimoto 系数是一种针对二元数据的计算，m_i 和 n_i 表示用户 m 和 n 是否对项目 i 进行评分，如果用户有进行评分，则 $m_i=n_i=1$，反之，其值为 0。这样得出的 Tanimoto 系数存在不足之处，此计算方法只关注用户是否对项目进行评价，没有考虑到用户对项目的评分值的大小问题。因此，在确保共同评分项目占用比还能够准确衡量用户整体相似度的前提下，本文利用修正的 Tanimoto 系数来计算共同评分项目占用比，从而计算用户之间相似度。修正 Tanimoto 系数表达式如下：

$$T(m,n)^+=\frac{\sum_{t\in R_\omega}\left(1-\sqrt{\left(\frac{m_t-n_t}{N}\right)^2}\right)}{\sum_{t=1}^{k}(m_t\wedge n_t)} \tag{2-3}$$

其中，k 表示评分项目集合的数量，m_t 和 n_t 表示用户对项目 i 的评分，系统评分所允许的最大分值设为 N。在该式中，只有用户之间对相同的项目有进行评分且评分值也相同时，用户此项目的相似度才为 1。相对于传统的 Tanimoto 系数，修正后的系数充分考虑到了用户之间评分值的差异，通过使用此方法能够得到更为准确的共同评分交集数目，从而得到更加准确合理的用户相似度，使产生的推荐结果更加精准。

Step 4 组合修正 Tanimoto 系数相似度

利用修正 Tanimoto 系数与传统修正余弦相似性度量公式结合之后可以得出本文最终采用的相似性度量方法，即组合修正 Tanimoto 系数相似度[6]：

$$sim_{\tau}(m,n)=\alpha\times T(m,n)+(1-\alpha)\times sim(m,n) \tag{2-4}$$

参数 α 为调节因子，主要作用是调节修正 Tanimoto 系数与传统相似度二者之间的权重。

Step 5 根据评分产生推荐结果

利用以上方法计算得出项目之间的相似度从而得到目标项目的最近邻居，在获取了邻居集合后，本文使用权值和法对目标用户进行预测评分[7]。具体公式如下：

$$P_{m,i}=\bar{R}_i+\frac{\sum_{j\in M(i)} sim(i,j)\times\left(R_{m,j}-\bar{R}_j\right)}{\sum_{j\in M(i)}\left|sim(i,j)\right|} \tag{2-5}$$

$P_{m,i}$ 表示对目标用户的预测评分，$M(i)$ 是通过计算相似度得出的最近邻居集合，$sim(i,j)$ 表示项目 i 与其最近邻居 j 之间的相似度，$\bar{R}_i$ 和 $\bar{R}_j$ 分别表示用户对项目 i 和项目 j 的评分。最终对预测评分进行从高到低进行排序产生推荐结果，结果准确度较高。

3. 实验结果及分析

3.1. 实验数据集

本实验选用的数据集来源于第六届“泰迪杯”数据挖掘挑战赛 B 题电视产品的营销推荐，此数据集是由用户收视信息数据、电视产品信息数据和用户基本信息数据等数据组成。我们将数据进行预处理后最终得到数据集包括 1000 个观看用户对 1200 部电视节目的观看时长，定义指标将观看时长转化为观看用户评分，每个观看用户对电视节目的评分数量不少于 15 条，评分是 1~5 的整数，表示观看用户对电视节目的喜好程度。随机地选取 90%的数据进行训练集，10%的数据作为检验和测试集，使得同一个观看用户在训练集和测试集中评论的电视节目不相同，通过训练集中观看用户的评分，确定用户偏好，根据训练集计算出的结果预测测试集中用户对电视节目的评分。

3.2. 评价标准

电视节目个性化推荐任务是为观看用户推荐其最感兴趣的相关的 n 个电视节目，准确率 precision、召回率 recall 和综合评估的指标 f-measure 等都是推荐体系常用的一个评估指标[8]。准确率表示观看节目的用户对推荐系统所推荐的电视节目喜爱的概率，召回率表示观看用户感兴趣的电视节目被推荐的概率，两个概率值越高说明推荐的质量越好。f-measure 主要作用是对 precision 和 recall 的加权调和平均，用作准确率和召回率的综合评定。

$$\text{precision}=\frac{\left|his_m\right|}{\left|recset_m\right|} \tag{3-1}$$

$$\text{recall}=\frac{\left|his_m\right|}{\left|testset_m\right|} \tag{3-2}$$

其中 his_m 表示推荐系统正确地推荐给观看用户 m 的感兴趣电视节目数量，$|recset_m|$ 表示推荐系统推荐给观看用户的电视节目的总数量，$|testset_m|$ 表示用户实际感兴趣的所有电视节目数量。

$$\text{f-measure} = \frac{\left(\alpha^2+1\right)\text{precision}\times\text{recall}}{\alpha^2\left(\text{precision}+\text{recall}\right)} \tag{3-3}$$

当参数 $\alpha=1$ 时，就是最常见的 f_1，也即

$$f_1 = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{3-4}$$

可知 f_1 综合了 precision 和 recall 的结果，当 f_1 较高时则说明推荐系统的个性化推荐效果较好。

3.3. 实验结果分析

该实验将本文研究的改进相似性度量的协同过滤推荐算法和传统的基于用户的协同过滤算法[9]在相同数据集上的推荐效果进行了对比分析。该算法的工作过程采用 Python 编程语言进行实现，两种算法的准确率、召回率和 f_1 值的对比分析结果如图 1、图 2 所示。

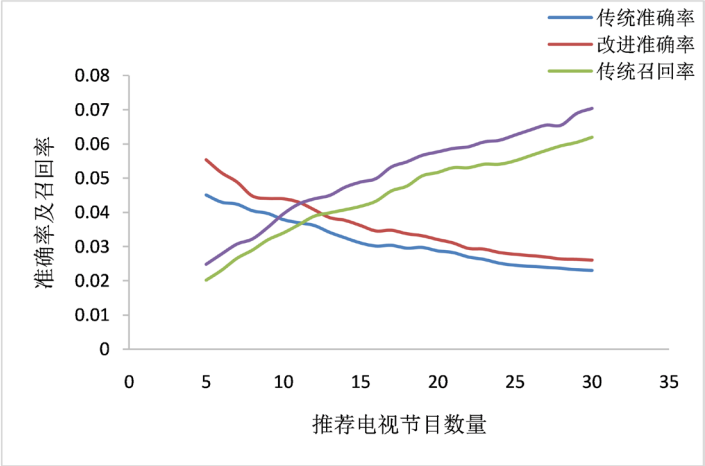

Figure 1. The accuracy and recall rate of the collaborative filtering recommendation algorithm modified Tanimoto coefficient and improved similarity measure and the traditional user-based collaborative filtering algorithm
图 1. 修正 Tanimoto 系数改进相似性度量的协同过滤推荐算法与传统基于用户的协同过滤算法准确率及召回率

从图 1 中可以看出，两种算法的召回率随着推荐电视节目数量的增多而升高。两种算法的准确率随着推荐电视节目数量的增多而降低，这可能是由于随着推荐电视节目数量的增加导致与目标用户兴趣不同的电视节目可能会被推荐，降低了推荐电视节目的准确率。但是本文所研究的改进的协同过滤算法的准确率和召回率都优于传统的基于用户的协同过滤推荐算法的准确率和召回率，个性化推荐电视节目的效果较好。

f_1 值是准确率和召回率的综合，从图 2 可以看出，当推荐电视节目数量 $n\le 10$ 时，随着推荐电视节目数量的增加，两种算法的 f_1 值都呈现上升趋势。当推荐电视节目数量 $n>10$ 时，两种算法的 f_1 值呈现下降趋势。但本文所提的基于修正 Tanimoto 系数改进相似性度量的电视节目个性化推荐算法的 f_1 值始终高于传统的基于用户的协同过滤算法的 f_1 值，个性化推荐电视节目效果较好，并且当推荐电视节目数量 $n=10$ 时，推荐效果最好。

通过以上实验结果我们可以清晰地看出，本文提出的基于修正 Tanimoto 系数改进相似性度量的电视

节目个性化推荐算法和传统的基于用户的协同过滤算法相比，具有了更好的预测准确性，因而可以进一步改善和提高个性化推荐信息的准确性。但是随着推荐的电视节目数量增多的情况下，算法的准确率会降低。综合来看，本文所提的算法具有一定的优势，可以提高电视节目个性化推荐系统的精度。

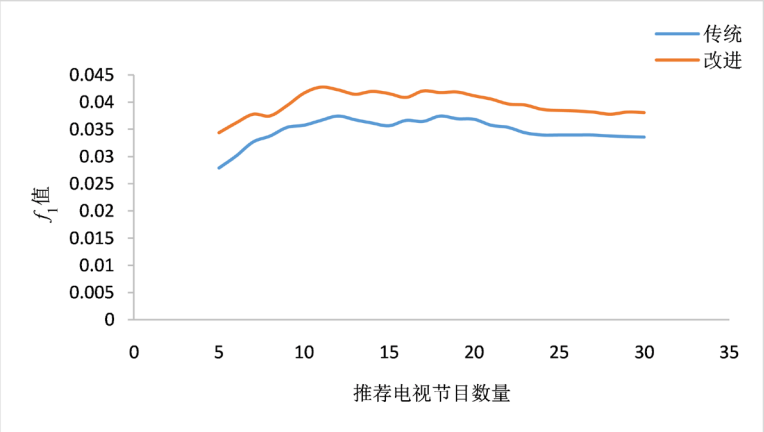

Figure 2. The value of TV program personalized recommendation algorithm based on modified Tanimoto coefficient and improved similarity measure and traditional user-based collaborative filtering algorithm
图 2. 基于修正 Tanimoto 系数改进相似性度量的电视节目个性化推荐算法与传统基于用户的协同过滤算法 f_1 值

4. 总结

本文将共同评分项目占所有评分项目的比例引入到传统的相似性度量方法中，提出了一种改进 Tanimoto 系数的新的相似性度量方法。实验结果表明，通过与传统的基于用户的协调过滤算法的对比，本文所提的基于修正 Tanimoto 系数改进相似性度量的电视节目个性化推荐算法在准确率、召回率和综合 f_1 值上具有一定的优势，能够进一步提高推荐系统的质量，能够对电视节目推荐系统产生较好的个性化推荐效果。

参考文献

[1] Barragans-Martinez, A.B., Costa-Monatenegro, E., Burguillo, J.C., *et al*. (2010) Ahybrid Contene-Based and Item-Based Collaborative Filtering Approach to Recommend TV Programs Enhanced with Singular Value Decomposition. *Information Sciences*, **180**, 4290-4311. https://doi.org/10.1016/j.ins.2010.07.024

[2] 宋月亭. 基于用户特征的电视节目混合推荐算法研究[D]: [硕士学位论文]. 昆明: 昆明理工大学, 2020.

[3] 陈娅昵, 苏岐芳. 基于协同过滤的电视产品营销推荐[J]. 台州学院学报, 2019, 41(3): 5-10+22.

[4] 张洪顺. 推荐系统中矩阵稀疏性问题的研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2018.

[5] 黄传飞, 万剑怡, 王明文, 李茂西. 协同过滤中一种项目综合相似度计算方法[J]. 山西大学学报(自然科学版), 2015, 38(2): 199-205.

[6] 文俊浩, 舒珊. 一种改进相似性度量的协同过滤推荐算法[J]. 计算机科学, 2014, 41(5): 68-71.

[7] 赵永生, 祁云嵩. 基于改进相似度计算方法的协同过滤算法研究[J]. 计算机与数字工程, 2021, 49(3): 447-450+541.

[8] 郭丽莎, 邓棉予, 李秋雨, 冯琪, 郭仲凯. 基于协同过滤算法的电视产品打包推荐[J]. 中南民族大学学报(自然科学版), 2020, 39(6): 655-660.

[9] Wang, H., Shen, Z., Jiang, S., Sun, G. and Zhang, R. (2021) User-Based Collaborative Filtering Algorithm Design and Implementation. *Journal of Physics*: *Conference Series*, **1757**, Article ID: 012168. https://doi.org/10.1088/1742-6596/1757/1/012168