基于协同过滤的图书推荐算法研究

张大伟

（北海职业学院，广西 北海 536000）

摘 要：传统的在线图书阅览以用户主动搜索为主，虽然有系统推荐书目，但是却无法满足大众的个性化需求。为解决目前在线图书馆推荐系统无法满足读者个性化需要的问题，笔者设计实现了以协同过滤算法为核心的智能推荐系统。实验结果表明，基于协同过滤算法核心的智能图书推荐系统不仅能够有效节约用户的时间，还能提高图书借阅效率，具有更好的应用效果。

关键词：协同过滤；智能推荐；相似度

中图分类号：TP391.3 **文献标识码：**A **文章编号：**1003-9767（2021）11-060-03

Research on Book Recommendation Algorithm Based on Collaborative Filtering

ZHANG Dawei

(Beihai Vocational College, Beihai Guangxi 536000, China)

Abstract: Traditional online book reading is dominated by users' active search. Although there is a system recommended bibliography, it cannot meet the individual needs of thousands of people. In order to solve the problem that the current online library recommendation system cannot meet the individual needs of readers, an intelligent recommendation system with a collaborative filtering algorithm as the core is designed and implemented. The experimental results show that the intelligent book recommendation system based on the collaborative filtering algorithm can effectively save users' time, Improve the efficiency of book borrowing and have a better application effect.

Keywords: collaborative filtering; intelligent recommendation; data mining

0 引言

信息和互联网技术的进步使网络数据空前增长，大数据和人工智能技术已与众多行业呈现出交叉融合趋势。在社交平台、电商以及在线服务领域，个性化推荐已经成为基础功能，智能推荐的时代已经到来。在数以万计的书籍中找到需要借阅的书目耗时耗力，协同过滤算法是机器学习的入门算法，其利用相似度分析智能为用户推荐图书列表，可有效节约用户时间，提高图书借阅效率。

1 协同过滤推荐算法

1.1 协同过滤算法

协同过滤算法是一种经典的被广泛应用在推荐和预测场景中的个性化推荐算法，其原理是对用户历史性的行为数据进行挖掘处理，预测用户偏好，根据用户偏好进行分类，进而推荐相似性的内容，是一种人工智能相关的算法[1]。协同过滤算法是最早用于智能推荐的算法，市面上许多流行的算法均基于其改进而来，其稳定、高效的特点非常适合推荐系统。协同过滤算法可以分为两种：基于用户的协同过滤算法和基于物品的协同过滤算法。

1.2 基于用户与基于物品的协同过滤算法

基于用户的协同过滤算法其原理是与当前用户相似的人一定具有相似的爱好，也可以理解成根据某个用户的喜好推测相关用户可能喜好的一种计算方法。核心的思想是查找相似的用户，这一过程涉及比较当前对象和目标对象的相似程度，也就是相似度。基于用户的协同过滤算法的对象需要在同一个域中（获取的数据和计算的范围均在图书推荐系统

基金项目：2020 年度广西高校中青年教师科研基础能力提升项目“数据挖掘技术在高校图书馆服务中的应用研究”（项目编号：2020KY65009）。

作者简介：张大伟（1981—），男，山东青岛人，本科，副教授。研究方向：人工智能。

内），利用数据库中的用户行为数据进行分析、统计，进而查询到相似度接近的用户 [2]。

基于物品的协同过滤算法与基于用户的协同过滤算法类似。简单的说就是若 *UserA* 借阅了 *Book*1 和 *Book*2，并且对两本书的评分都不错；而 *UserB* 也借阅了 *Book*1 这时就可以向其推荐 *Book*2。

1.3 基于用户的协同过滤算法的两种解决方案

1.3.1 非个性化解决方案

将用户对图书的评分均值作为当前项的评分。这种方法太过粗糙，为了让结果更准确，通常在计算均值时会减去用户本身的评分均值，其计算公式如式（1）所示：

$$P_{a,i}=\frac{\sum_{u=1}^{n} r_{u,i}}{n} \tag{1}$$

式中，$P_{a,i}$ 为当前项的评分，$r_{u,i}$ 为用户对当前项的评分，n 为参与评分的人数。

1.3.2 个性化解决方案

在计算平均分时只考虑与当前用户相似的用户对图书的评分，这样每一个评分均有一个权重，如式（2）所示：

$$P_{a,i}=\bar{r_a}+\frac{\sum_{u=1}^{n}(r_{u,i}-\bar{r_u})\ W_{a,u}}{\sum_{u=1}^{n} W_{a,u}} \tag{2}$$

式中，$P_{a,i}$ 为当前项的评分，$\bar{r}_a$ 为相似用户对当前项的评分的均值；$W_{a,u}$ 为用户的相似度；$r_{u,i}$-$\bar{r}_u$ 为修正的评分。相对来说，个性化的解决方案要优于非个性化的解决方案。

2 相似度度量

在协同过滤中，相似度是常用的概念，用于判断目标用户与已经存在的某用户的相似程度，根据已经存在的用户所喜好的图书推断目标用户的喜好图书。因此，相似度度量是为了寻找目标用户的邻居（与其最相似的那个用户）[3]。用户 - 图书矩阵如表 1 所示，根据平均值很容易计算用户对某一本书的喜好。

计算相似度的算法有很多，常见的有余弦相似度、皮尔逊相似度、Jaccard 相似度等，每种算法有其适应的场景。余弦相似度常用于度量文本相似度、用户相似度等，但在计算的时候余弦相似度需要对向量长度归一化 [4]；皮尔逊相关度是计算两个变量的变化趋势是否一致，不适合布尔值向量的计算；Jaccard 相似度是两个集合的交集元素个数在并集内所占的比例，其非常适合布尔向量表示，用于隐式反馈数据。

2.1 余弦相似度

余弦相似度是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。余弦相似度适用于二维或 N 维向量，其根本方法是计算两个向量的夹角余弦值，计算公式如下：

$$\cos\theta=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{3}$$

由式（3）知，cosθ 取值在 [-1,1] 之间，若两个向量的方向完全相反，此时余弦相似度为 -1，二者完全不相关；若两个向量方向相同时，余弦相似度为 1，说明二者完全相关。两个向量之间的夹角余弦越大，表明二者越相似。因此余弦相似度可以用来度量两个变量之间的相似程度。

以表 1 所示矩阵，若直接采用对书籍喜好的量化方式，以 *UserA* 与 *UserB* 为例，则可以量化成（6,0,0,7,1,0,0）、（7,7,7,5,0,0,0），根据式（3）得出 cosθ=0.633，也就是 *UserA* 和 *UserB* 两个人的相似度是 0.633。在实际的应用中，达到一定的阈值即认为两人或多人存在较高的相似性，即认为他们有相似的爱好，很明显这种计算过于粗糙。

若两个用户都借阅了《新华字典》很难确定他们兴趣相似，因为大多数人都会借阅过这类工具书 [5]。但如果两个人都借阅《现代网络技术》，那么可以认为他们的兴趣相似，只有相关专业的人才会借阅这本书。JOHN S 提出一个称为用户活跃度对数的倒数的参数，用于修正物品的相似度，其计算公式如下 [6]：

$$w_{AB}=\frac{\sum_{i\in N(UserA)\cap N(UserB)}\frac{1}{\log 1+|N(Item)|}}{\sqrt{|N(UserA)|\times|N(UserB)|}} \tag{4}$$

式中，分子的倒数惩罚了 *UserA* 和 *UserB* 共同兴趣列表中热门物品对他们相似度的影响。*N*(*Item*) 是对书籍 *Item* 有过行为的用户集合，*N*(*Item*) 值越大表明书籍越热门。

2.2 皮尔逊相似度

在实际应用中若使用余弦相似度计算两者相似性时偏差较大，为了得到更优的结果需要对数据进行标准化处理，通常每个向量减去一个均值，经过处理后就得到了皮尔逊相似度的计算公式：

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\times\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5}$$

表 1 用户 – 图书矩阵

用户 / 图书	*T-Book*1	*T-Book2*	*T-Book3*	*P-Book*	*S-Book*1	*S-Book2*	*S-Book3*
UserA	6	—	—	7	1	—	—
UserD	7	7	5	—	—	—	—
UserC	—	—	—	4	6	6	—
UserB	—	4	—	—	—	—	4

皮尔逊相似度与余弦相似度非常类似，它是余弦相似度在维度值缺失情况下的一种改进。很显然 r_{xy} 为 1 时两者严格正相关，r_{xy} 为 -1 时两者严格负相关，r_{xy} 为 0 时两者完全不相关。

2.3 Jaccard 相似度

Jaccard 相似度也叫杰卡德系数，用于比较有限样本集之间的相似性与差异性。其计算非常简单是将两者的交集数量和并集数量的比值作为二者的相似系数。

$$\mathrm{sim}(A,B)=\frac{P_A\cap P_B}{P_A\cup P_B} \quad (6)$$

UserA 和 *UserB* 的杰卡德系数为 $\mathrm{sim}(A,B)=\frac{P_A\cap P_B}{P_A\cup P_B}=\frac{1}{5}$，也就是说 *UserA* 和 *UserB* 的杰卡德系数是 0.2。由式（6）可以得出，杰卡德系数只统计用户是否评价了某本书籍而忽略了对书籍的评价分数，这种计算方法同样太粗糙，在实际应用中可以根据实际情况加入权重信息修正杰卡德系数，使计算结果更准确。

3 冷启动问题

冷启动问题是协同过滤算法中面临的首要问题，也是推荐系统面临的难题，所谓冷启动就是系统建立之初，并没有足够的数据用于计算，这时利用相关公式进行计算是无法达到期望值，更无法为新用户推荐合适的书籍，这时可考虑以下解决方案 [7]：提供非个性化的方案，比如热门推荐、新书推荐等，这适用于系统冷启动的情况；利用用户注册时填写的信息（性别、年龄等）做粗粒度个性化，这适用于用户冷启动情况；对新加入的书籍，可以利用内容简介、书籍名称、所属类别等信息推荐给相似的用户，这适用于物品冷启动的情况。

冷启动问题是智能推荐面临的难题，多数情况下根据实际情况制订非个性化的综合方案，也有一些观点认为冷启动问题不是核心问题，因为随着数据的不断增长，达到智能推荐系统的基本要求，那么就不需要再担心冷启动的问题 [8]。

4 相似度计算的局限性

相似度是一种特定场景的运算结果，其不具备普适性，在实际的应用系统中单一的相似度计算方法无法满足实际需要，因此相似度度量具备相当大的局限性。余弦相似度对数值敏感度较低，按照表 1 所示 *UserA* 和 *UserB* 两个用户对两个内容的评分分别为（1，2）和（4，5），若采用余弦相似度公式计算的结果是 0.98，这表明 *UserA* 和 *UserB* 极为相似。但从评分上看 *UserB* 喜欢第二个内容的程度要大于 *UserA*，这是余弦相似度对数值的不敏感导致的误差。虽然出现了许多修正余弦相似度的方法，但无法掩盖其在数值敏感度方面的天然缺陷 [9]。

杰卡德相似度常用于反馈用户隐式数据，利用其计算表 1 中的 *UserA* 与 *UserB*、*UserC* 的相似度时，$\mathrm{sim}(A,B)=\frac{P_A\cap P_B}{P_A\cup P_B}=\frac{1}{5}$，$\mathrm{sim}(A,C)=\frac{P_A\cup P_C}{P_A\cup P_C}=\frac{1}{2}$，根据 Jaccard 相似度 $\mathrm{sin}(A,B) < \mathrm{sin}(A,C)$，很显然这是存在问题的 [10]。这也说明单一的相似度计算很难满足复杂的业务需求，在实际应用时通常结合多种算法进行计算，根据业务需求进行相应的取舍或合成。

5 结 语

本文从相似度的度量角度出发，探讨了协同过滤算法在图书推荐中的应用，信息爆炸使人们获取知识的门槛越来越低，用户的个性化的需求越来越突出，在图书阅览中采用智能推荐不但能提高借阅效率、根据用户喜好按需推荐，而且还能提高用户体验。数据挖掘技术和人工智能技术的不断演进，以协同过滤算法为基础的智能算法将会使图书推荐系统更加智能，用户借阅图书的效率越来越高。

参考文献

[1] 刘文佳，张骏 . 改进的协同过滤算法在电影推荐系统中的应用 [J]. 现代商贸工业 ,2018,39(17):59-62.
[2] 黄川林，鲁艳霞 . 基于协同过滤和标签的混合音乐推荐算法研究 [J]. 软件工程 ,2021,24(4):10-14.
[3] 唐高芳 . 基于改进的协同过滤图书推荐算法 [J]. 信息技术 ,2021(3):16-20.
[4] 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.
[5] 李清 . 基于 MovieLens 数据集的协同过滤推荐系统研究 [D]. 西安：西安电子科技大学 ,2014:61-62.
[6] 高阳团 . 推荐系统开发实战 [M]. 北京：电子工业出版社 ,2019:23-25.
[7] 朱明 . 基于协同过滤的教学资源推荐研究 [D]. 北京：北京交通大学 ,2017:36-37.
[8] 邵煜，谢颖华 . 协同过滤算法中冷启动问题研究 [J]. 计算机系统应用 ,2019(28):246-252.
[9] 彭石 . 基于用户兴趣和项目特性的协同过滤推荐算法研究 [D]. 杭州：浙江大学 ,2012:44-45.
[10] 徐啸 . 一种基于神经网络和双重聚类的协同过滤算法研究 [D] 重庆：西南大学 ,2020:33-35.