Computer Science and Application 计算机科学与应用, 2022, 12(6), 1538-1543
Published Online June 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.126153

Hans汉斯

改进的协同过滤混合算法在电影系统中的应用

鞠达旺，孙　杰

天津工业大学，天津

收稿日期：2022年5月8日；录用日期：2022年6月8日；发布日期：2022年6月14日

摘　要

随着现代科技的发展，在当前信息过载的互联网时代，根据用户历史行为数据进行物品推荐已经成为当前研究的热点。近年来，推荐系统不断更新迭代，大多围绕准确性问题进行研究和改进。但是，多样性、新颖性以及长尾物品推荐等方面也同等重要。通过对协同过滤推荐系统的研究分析，引入多样性因子，提出一种基于聚类的混合协同过滤推荐算法，通过在Movielens数据集上验证了基于聚类的混合协同过滤算法较传统的协同过滤在提高了多样性的同时，准确性也得到改善，满足用户个性化需求。

关键词

协同过滤，电影推荐，多样性因子，聚类

Application of Improved Collaborative Filtering Hybrid Algorithm in Film System

Dawang Ju, Jie Sun

Tiangong University, Tianjin

Received: May 8th, 2022; accepted: Jun. 8th, 2022; published: Jun. 14th, 2022

Abstract

With the development of modern science and technology, in the current Internet era of information overload, item recommendation based on user historical behavior data has become the focus of current research. In recent years, the recommendation system has been continuously updated and iterated, and most of them focus on the accuracy problem. However, diversity, novelty and long tail recommendation are equally important. Through the research and analysis of collaborative filtering recommendation system, the diversity factor is introduced, and a hybrid collaborative filtering recommendation algorithm based on clustering is proposed. It is verified on Movielens data set that the hybrid collaborative filtering algorithm based on clustering improves the

文章引用：鞠达旺，孙杰．改进的协同过滤混合算法在电影系统中的应用[J]．计算机科学与应用，2022，12(6): 1538-1543. DOI: 10.12677/csa.2022.126153

diversity and accuracy compared with the traditional collaborative filtering, so as to meet the personalized needs of users.

Keywords

Collaborative Filtering, Film Recommendation, Diversity Factor, Clustering

1. 引言

随着互联网数据爆炸式增长，互联网为人们提供了海量的信息资源，丰富了我们的日常生活。而与此同时，信息过载的问题也日益突出，在海量的信息之中，提取出人们所需要的关键信息变得尤为重要。在这个背景之下，推荐技术逐渐成为当前研究的热点，现已广泛应用于电商，电影，音乐以及图书等平台[1] [2] [3] [4]。

目前，应用较多的有基于统计学的推荐，基于内容的推荐和基于协同过滤的推荐。其中基于协同过滤的推荐是通过用户对商品的行为数据，来预测用户对商品的偏好。但是，由于物品的丰富性逐渐提升，同时又难以显示的对物品进行分类，通过属性标签也难以代表用户对物品的真实偏好，传统的单一推荐算法不能很好地发现目标用户的喜好，同时存在可扩展性差的问题[5]。本文根据电影推荐系统的特点，运用聚类算法以及基于物品和用户的混合推荐算法对用户进行个性化推荐，同时根据用户个人偏好，调整电影推荐类型的多样性。

2. 协同过滤混合算法的设计

2.1. 协同过滤算法的分类

推荐算法发展至今，可以分为基于用户的协同过滤算法(user-based collaboratIve filtering，简称 User CF)和基于物品的协同过滤推荐算法(item-based collaborative filtering，简称 Item CF)，1992 年 Goldberg 与 Nicols 提出了协同过滤算法基本概念，起初是用来过滤用户的电子邮件[6]。随着互联网的不断发展，协同过滤算法现已成为当今主流的推荐算法之一。具体来讲，协同过滤算法(Collaborative Filtering，简称 CF)是根据用户对物品的历史行为数据，通过计算用户或物品之间的相似度，然后查找与目标用户相似度较高的邻近集，并通过近邻用户对物品的评分来达到对目标用户进行有效推荐的目的。

2.2. 对比分析 UserCF 和 Item CF 的优缺点，为融合两种算法提供依据

基于用户的协同过滤算法是推荐和当前用户相似度高的近邻用户产生过行为的物品给当前用户，基于物品的协同过滤算法是推荐和当前用户历史上行为过的相似度高的近邻物品给当前用户。

1) 适用场景上，ItemCF 是利用物品间的相似性来推荐，适用于用户数量大于物品数量的场景，比如电影系统，由于电影数据相对稳定，不仅计算物品相似度时计算量较小，而且不会频繁更新，维护相似度矩阵的代价也同步减小。而 UserCF 更适合做新闻与博客推荐，因为其内容更新频率快，而且具有时效性，则使用 UserCF 在推荐上有更好的可解释性[7]。

2) 推荐多样性上，对于单一用户推荐上的多样性，ItemCF 显然是不如 UserCF 的好，因为 ItemCF

的推荐是基于历史行为物品相似的推荐，相比之下 UserCF 更具有多样性，其根据近邻用户行为过的物品进行推荐，具有发现意外兴趣的能力[8]。在系统多样性上(也被称为覆盖率，指一个推荐系统能否给用户提供多种选择)，在这个指标下，ItemCF 则具有发现长尾物品的能力，UserCF 更多的是推荐不同类和热门的物品。

3) 用户特点对推荐算法影响，对于 UserCF 推荐的原则是依据在那些在已经行为过的物品有着相似喜好，推荐的准确性来源于可信的邻居用户，如果邻居用户不足，会出现准确性较差的情况。ItemCF 是根据行为过的物品进行推荐，只要用户有了行为，就能快速进行相似物品的推荐，但是在物品过多或者意外发掘上效果不太理想。

综上所述，两者在推荐指标上各有优势，在本文的电影推荐系统中，我们针对这两种推荐系统各自的特点进行混合推荐，以提高推荐效果。

3. 个性化的电影推荐系统

3.1. 电影聚类

电影的种类有很多种，一个电影可能同时拥有多个不同属性标签，而聚类是为了将相似的电影尽可能划分在一起，不同类的电影尽可能的分离，在聚类过程中我们不需要考虑电影本身拥有的标签，最终的目的是将输入的电影集划分到不同的类别。聚类算法有多种，本文使用的 K-means 算法是被广泛使用且简洁的一种算法。具体的算法流程如下：

1) 随机选取电影中 K 个点作为聚类的初始质心

2) 对于每个电影数据点，使用 Euclidean 距离计算其与各个质心点之间的距离，并将该点分配到最近的集群，其中 Euclidean 距离相似性度量函数如下：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

其中：x、y 为不同电影项目，和为不同用户对电影的评分分值，n 为用户个数。

3) 对 K 个集群重新计算均值并将均值作为新的集群质心

4) 如果计算得出的新中心点超出了预设的阈值，则转移到步骤 2)，直到算法发生了收敛，则结束迭代。

K-means 聚类具有思想简单，收敛速度快，在对大量数据进行聚类分析时，算法具有聚类效果好且高效的特点。

3.2. 基于项目的协同过滤

通过用户电影评分矩阵建立电影相似度矩阵，根据相似电影预测电影评分，从而进行 topN 电影推荐。两种最常用的相似度度量方式为基于余弦相似度和基于 pearson 相关系数。本次算法我们使用修正的余弦相似度，公式如下：

$$sim(x,y)=\frac{\sum_{i\in I_{xy}}\left(R_{x,i}-\bar{R}_X\right)\left(R_{y,i}-\bar{R}_Y\right)}{\sqrt{\sum_{i\in I_{xy}}\left(R_{x,i}-\bar{R}_X\right)^2}\sqrt{\sum_{i\in I_{xy}}\left(R_{y,i}-\bar{R}_Y\right)^2}} \tag{2}$$

其中，I_{xy} 表示用户 xy 共同评分过的电影集合，R_{xi} 和 R_{yi} 分别表示用户 xy 对电影 i 的评分，$\bar{R}_X$ 和 $\bar{R}_Y$ 分别表示用户 x 与用户 y 对全部电影的评分平均值，通过公式(2)可以算出电影之间的相似度，然后根据相

似度计算得到电影评分预测值。

3.3. 基于用户的协同过滤

通过用户电影评分矩阵建立用户相似度矩阵，然后预测电影评分，本文的实验数据来源于 Movielens 在网上公布的数据，上面丰富的用户信息，可以考虑选择对推荐有可靠依据的用户信息(年龄、职业、性别)进行相似度计算的加权融合[9]，当前我们使用简单传统的实现方式，计算方式和上述方法类似，如公式(3)所示：

$$sim(i,j)=\frac{\sum_{k=1}^{k}\left(R_{i,k}-\bar{R}_i\right)\left(R_{j,k}-\bar{R}_j\right)}{\sqrt{\sum_{k=1}^{k}\left(R_{i,k}-\bar{R}_i\right)^2}\sqrt{\sum_{k=1}^{k}\left(R_{j,k}-\bar{R}_j\right)^2}} \tag{3}$$

其中， R_{ik} 和 R_{jk} 分别表示用户 ij 对电影 k 的评分分值， $\bar{R}_i$ 和 $\bar{R}_j$ 分别表示用户 i 与用户 j 评分平均值。

3.4. 基于聚类的混合电影推荐系统

在电影系统中，电影数目相比用户数量较少，且更新稳定，在准确性上 Item 优于 UserCF，所以对用户原始推荐列表采用基于 ItemCF 的 topN 推荐，同时系统根据用户偏好阈值(threshod)对推荐列表进行调整，目的是构建用户权值矩阵 CW，其中 CW_{ui} 为混合推荐系统向用户推荐的权值，$CW_{ui}[0]$为 ItemCF 的推荐权值，$CW_{ui}[1]$为 UserCF 的推荐权值，算法描述如下：

输入：基物品的协同过滤推荐列表 topN，电影聚类 C，用户列表 U；

输出：用户 u 的聚类权重 CW_u。

步骤 1 利用推荐列表 topN 和电影聚类 C 通过公式 $C_i \cap$topN 计算出用户的初始聚类权重 $CW_{ui}[0]$。

步骤 2 对于电影聚类 C_i 和用户 u，如果 $CW_{ui}[0]$大于多样性阈值 threshold，则选择用户最小项目权值聚类 C_j 使该聚类 $CW_{uj}[1]+1$，而当前聚类 C_i 权值 $CW_{ui}[0]-1$，直到用户的项目权值均不大于 threshod。

步骤 3 对每个用户重复步骤 2，调整所有用户的聚类权值。

上述算法可以得到每个用户在不同聚类中的权值，输入参数 topN 为基于 ItemCF 的前 N 个推荐列表，C 为由 K-meams 算法得到的电影聚类，输出参数 CWu 为不同聚类向用户 u 推荐列表的贡献权值，CWui 是聚类 i 向用户 u 所提供的推荐权重，它是一个二维数组结构，其中 CWui[0]为聚类 i 中的电影基于项目的协同过滤向用户 u 提供的推荐数目，CWui[1]为聚类 i 中的电影基于用户的协同过滤向用户 u 提供的推荐数目，CWui[1]初始化均为 0，最终的推荐列表由 CWu 决定。

步骤 1 计算了用户的初始权重，可以看出用户的初始推荐完全由 ItemCF 决定，其中当 ItemCF 贡献权值在某个聚类中大于阈值 threshold，我们通过降低它的权值，然后随机选择 CWui[0]最小的聚类，使 CWui[1]增加来调整用户推荐列表的多样性，因为超过阈值的聚类权重总会被分配到那些其他类型的聚类之中，并且通过 UserCF 来进行推荐，可以有良好的解释性，使得在提高多样性的同时，准确性也得到保障。

4. 实验

4.1. 数据集

本文采用标准的 Movielens 数据集，该数据集包含了 6040 名用户对 3950 个电影的 100,000 万次评分，评分的范围为 1~5，每一位用户至少评价 10 部以上的电影，同时对每个用户的数据集做随机划分，取 70%作为训练集，30%作为测试集。数据的属性包含用户 ID，电影类别，电影 ID，以及用户对电影的相应评分。

4.2. 评价标准

采取召回率(Recall)以及多样性(Diversity)指标作为检测算法优劣的评价指标，召回率表示预测出的物品有多少在用户真实喜爱列表之中，多样性体现了覆盖用户兴趣点和发觉用户潜在兴趣的能力，公式如下：

$$\text{Recall} = \frac{\sum_{u \in U} \left| R(u) \cap T(u) \right|}{\sum_{u \in U} \left| T(u) \right|} \tag{4}$$

$$\text{Diversity}\left(R(u)\right) = \frac{\sum_{i,j \in R(u), i \neq j} \left(1 - sim(i,j)\right)}{\frac{1}{2}\left|R(u)\right|\left(\left|R(u)\right| - 1\right)} \tag{5}$$

4.3. 实验结果与分析

本文采用一种将 ItemCF 算法为默认推荐，根据用户喜好多样性不同的特点融入 UserCF 算法动态调整推荐列表的一种基于聚类的混合推荐算法(简称 KhCF)，并与基于单一聚类的 UserCF 和 ItemCF 算法进行比较分析，如图 1 和图 2 所示。

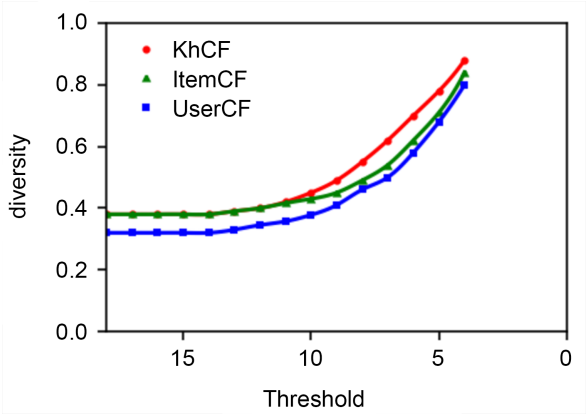

Figure 1. Diversity comparison of different algorithms
图 1. 不同算法的多样性对比

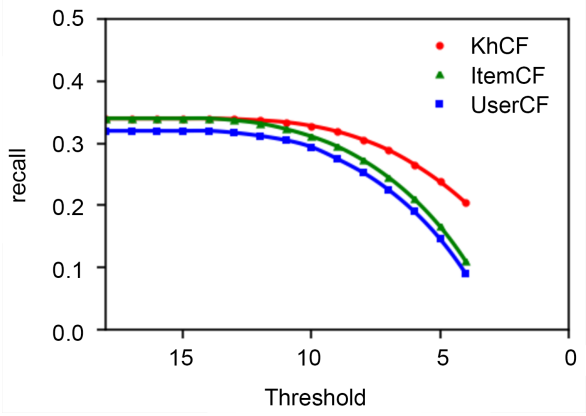

Figure 2. Recall comparison of different algorithms
图 2. 不同算法的召回率对比

由图 1 可以看出，用户整体在阈值变化的情况下，多样性的上升幅度有所提高，由图 2 可以看出召回率的下降程度也得到缓解。同时，由于多样性的调整选取是全体用户，一部分用户可能更加地偏向于个性化推荐，而非多样性推荐，这也正是我们选取 ItemCF 为主推荐算法的原因之一，这部分用户能够在初始阈值的情况下，得到较好的个性化推荐，这是 ItemCF 算法的优势。另一方面，那些拥有多样性喜好的用户，可以根据阈值的调整得到各种项目类型的推荐，而这些项目将由 UserCF 推荐列表所贡献，进而最大程度地满足不同用户的需求。

5. 结束语

本电影推荐系统在设计过程中，使用了 K-means 算法对电影进行聚类，然后根据电影分类结合用户喜好，将基于用户的协同过滤和基于物品的协同过滤相结合进行混合推荐，充分结合电影的相关性与用户行为，并考虑到用户个人兴趣差异进行长尾推荐或多样化推荐，提高推荐的多样性和准确性。

参考文献

[1] 张玉叶. 基于协同过滤的电影推荐系统的设计与实现[J]. 电脑知识与技术, 2019, 15(6): 70-73.

[2] 张鹏飞, 熊娇娇, 罗绳烨, 吴方君. 面向电商的基于协同过滤的个性化推荐[J]. 科技广场, 2016(6): 15-19.

[3] 隋占丽, 李影, 于娟, 王波. 基于协同过滤技术的音乐推荐系统的研究[J]. 福建电脑, 2015, 31(2): 12-13+112.

[4] 赵宇凤. 基于协同过滤的图书推荐系统[J]. 微型电脑应用, 2022, 38(1): 181-184.

[5] 陈彦萍, 王赛. 基于用户-项目的混合协同过滤算法[J]. 计算机技术与发展, 2014, 24(12): 88-91+95.

[6] Goldberg, D., Nicols, D., Oki, B.M. and Terry, D. (1992) Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM*, **35**, 61-70. https://doi.org/10.1145/138859.138867

[7] 韩瑞. 基于协同过滤个性化推荐算法的综述[J]. 商, 2015(51): 216.

[8] 罗文. 协同过滤推荐算法综述[J]. 科技传播, 2015, 7(7): 115+196.

[9] 曹俊豪, 李泽河, 江龙, 张德刚. 一种融合协同过滤和用户属性过滤的混合推荐算法[J]. 电子设计工程, 2018, 26(9): 60-63.