文章编号：2095-2163(2023)01-0195-03 中图分类号：TP391 文献标志码：A

基于用户和项目的协同过滤算法的比较研究

罗 洁[1,2]，王 力[1,3]

（1 贵州大学 大数据与信息工程学院，贵阳 550025；2 毕节工业职业技术学院，贵州 毕节 551700；
3 贵州工程应用技术学院 信息工程学院，贵州 毕节 551700）

摘 要：基于用户和基于项目的协同过滤算法作用相似，结果却各有不同。为了明确两种算法的使用情景，按照相似度比较对象的不同，本文通过对协同过滤算法从用户和项目两个方面进行比较，分析两种算法的优缺点；根据客户的历史信息，对客户的喜好进行分析，推荐和用户喜好相似的内容给用户，通过在相同条件下进行电影推荐实验，证明基于项目的协同过滤算法精确率高，而基于用户的协同过滤算法效率高，满足了无明确观看电影用户的选择需求。

关键词：协同过滤；基于用户；基于项目

Comparative analysis of collaborative filtering algorithms based on user and item

LUO Jie [1,2], WANG Li[1,3]

(1 College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China;
2 Bijie Technical College of Industry, Bijie Guizhou 551700, China;
3 Guizhou Institute of Engineering and Applied Technology, Information Engineering, Bijie Guizhou 551700, China)

【Abstract】 User-based and project-based collaborative filtering algorithms have a similar effect, but their results are different. In order to clarify the use of the two algorithms, according to the similarity of the comparison object, through the collaborative filtering algorithm from the user and project two aspects of the comparison, analysis of the advantages and disadvantages of the two algorithms, according to the customer's historical information, analyze customer preferences and recommend content that is similar to user preferences. Through the experiment of movie recommendation under the same conditions, the results show that the project-based collaborative filtering algorithm has high accuracy, while the user-based collaborative filtering algorithm has high efficiency.

【Key words】 collaborative filtering; user-based; project-based

0 引 言

随着时代发展，信息量极大膨胀。用户在面对海量信息时，不能快速从中获取自己有用的信息。针对这种现象，智能算法应运而生。近年来有关个性推荐算法的应用越来越广泛，根据用户的历史行为，对用户的喜好和目标行为，为用户推送信息，极具商业价值和挖掘价值。协同过滤算法最大的优点在于对推荐的对象没有特殊要求；能够有效处理非结构化的复杂的对象，避免了内容的分析不完全性和不精确性，根据用户的历史行为推荐个性化的信息。目前有很多学者对协同过滤算法进行改进并应用，孙传明等[1]针对数据稀疏性和推荐范围问题，提出了一种混合协同过滤推荐算法；荣以平等[2]针对电力大用户选择交易对象的问题，提出了基于用户协同过滤的购电推荐算法；孟晗等[3]针对对恶意用户进行区分的问题，提出了一种改进的新型信任关系度量的推荐算法；夏景明等[4]针对数据稀疏导致的推荐不准确问题，提出了一种基于用户和商品属性挖掘的协同过滤算法。

本文针对协同过滤算法的两种不同对象，基于用户和基于项目，对其进行比较分析，从用户数大于

基金项目：贵州省教育厅创新群体重大研究资助项目（黔财教合[2016]118）；贵州省首批国家级新工科研究与实践资助项目（黔教高函[2018]209 号）。

作者简介：罗 洁（1992-），女，硕士研究生，助理讲师，主要研究方向：计算机视觉、数据挖掘；王 力（1971-），男，学士，教授，主要研究方向：信息系统分析、设计与开发、数据挖掘。

收稿日期：2022-04-09

项目数和用户数小于项目数两方面进行实验,验证了两种不同对象的协同过滤算法的特性。

1 相关知识

协同过滤算法由 3 个部分组成:通过用户评分行为得到用户—项目评分矩阵、计算相似度、根据相似度进行推荐。

1.1 用户评分行为

用户评分行为是通过用户对项目的打分,构成用户—项目评分矩阵 $\boldsymbol{R}$, 式(1)所示,行向量表示用户对项目的评分,列向量表示某个项目得到用户的评分。

$$\boldsymbol{R}=\begin{bmatrix} R_{m1n1}\cdots R_{m1n_v} \\ \vdots \qquad\quad \vdots \\ R_{m_un1}\cdots R_{m_un_v} \end{bmatrix} \tag{1}$$

其中 m 表示用户;n 表示项目;m_u 表示第 u 个用户;n_v 表示第 v 个项目;Rm_un_v 表示第 u 个用户对第 v 个项目的评分,其数值的大小表示用户对项目的兴趣程度。

1.2 相似度计算

采用余弦相似度找到与目标用户兴趣相似的用户集合,利用不同用户对项目评分数的相似度计算出用户的兴趣相似度。余弦相似度是用户向量 $\boldsymbol{i}$ 和用户向量 $\boldsymbol{j}$ 之间的向量夹角大小,夹角越小,余弦相似度越大,两个用户越相似。余弦相似度公式为

$$sim_{ij}=\frac{\sum\limits_{m\in M}R_{m,i}\times R_{m,j}}{\sqrt{\sum\limits_{m\in M}R_{m,i}^{\ 2}}\times\sqrt{\sum\limits_{m\in M}R_{m,j}^{\ 2}}} \tag{2}$$

其中,$R_{m,i}$ 表示用户 $\boldsymbol{i}$ 对项目的评分,$R_{m,j}$ 表示用户 $\boldsymbol{j}$ 对项目的评分。

相似度越高则用户间的喜好相似性越高。公式(2)中的分子代表评价向量,分母代表评分值。基于项目的协同过滤算法同样采用余弦算法计算项目间相似度。

1.3 推荐

利用 k 最近邻算法思想,找到相似度最高的前 k 个用户,通过这些用户的相似度权重以及其对项目的偏好,计算得到一个项目排序列表进行预测推荐。用户 u 对项目 i 的预测评分,为式(3)

$$R_{u,v}=\frac{\sum\limits_{i\in k}sim(m_u,m_i)\times R_{m_in_v}}{\sum\limits_{i\in k}sim(m_u,m_i)} \tag{3}$$

其中,k 是相似度最接近的向量的集合;$\boldsymbol{i}$ 是任意一个用户;$sim(m_u,m_i)$ 表示最近邻 $\boldsymbol{i}$ 和用户 u 的相似度乘上最近邻 $\boldsymbol{i}$ 对项目 v 的评分。

与基于用户的协同过滤算法相似,基于项目的协同过滤算法是通过项目的相似度矩阵乘上评分矩阵得到推荐列表,来为用户推荐其有兴趣但还未涉及的项目。

2 基于用户的协同过滤

采用不同用户对项目的评分作为用户-项目评分矩阵,以此计算用户的相似度,根据相似度给用户推荐和其兴趣一致的用户的其他项目。其过程如图 1 所示。

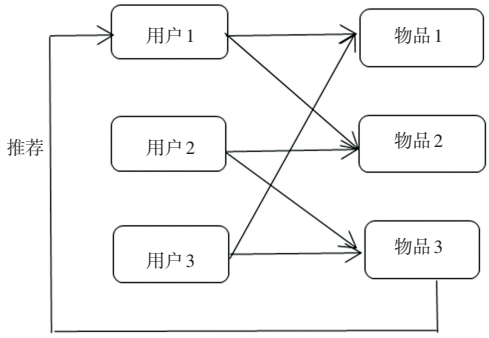

图 1 基于用户的协同过滤过程

Fig. 1 The process of user-based collaborative filtering

该算法利用了用户和用户间的相似性来为用户推荐其感兴趣的信息,通过评分达到筛选信息的目的,但是这个算法存在两个难解决的问题:

(1)稀疏性,即:用户评价信息量少,很难发现用户行为的相似性;

(2)随着项目和用户数量的增多,可扩展性变差。针对这两个问题,一方面可以通过改进相似度计算方法来改善数据稀疏性;另一方面,可以采用分布式编程来提高算法的可扩展性。

3 基于项目的协同过滤

将用户对不同项目的评分行为用矩阵来表示,以此计算项目之间的相似度,根据相似度排序为用户推荐与用户偏好相似度高的项目。每个用户操作独立,有独立的特征向量,不受相邻用户的偏好影响,可以为目标用户推荐其感兴趣的、新的、冷门的项目,使算法不受冷启动和稀疏性问题的影响,过程如图 2 所示。

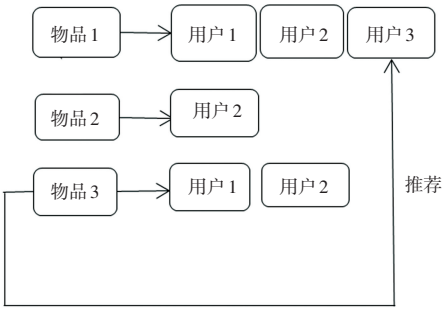

图 2　基于项目的协同过滤过程

Fig. 2　Project-based collaborative filtering process

4　实　验

4.1　实验环境

实验环境为：Inter(R) Core(TM) i5-2410M CPU @ 2.30 GHz；8 GB 内存；操作系统是 Winsdow10 64 位，利用 Jupyter Notebook 进行编程。实验数据集从 MovieLens (https://grouplens.org/datasets/movielens/) 中抽取。电影的评分范围为[1,5]区间所有整数值，用户对电影的喜好程度由 1 到 5 逐渐递增，数值越大，喜欢程度越深。实验数据集包含了用户信息，评分信息，电影信息。

4.2　评价指标

4.2.1　召回率(Recall)

召回率(Recall)又称为查全率，表示样本中正例被预测正确的比例，召回率为

$$Recall = TP / (TP + FN) \tag{4}$$

其中，TP 表示预测结果为正，实际结果为正；FN 表示预测结果为负，实际结果为正；$TP + FN$ 表示实际结果为正的样例。

4.2.2　精确率(Precision)

精确率(Precision)又称为查准率。表示预测为正的样本中正样本的比例，精确率为

$$Precision = TP / (TP + FP) \tag{5}$$

其中，TP 表示预测结果为正，实际结果为正；FP 表示预测结果为正，实际结果为负；$TP + FP$ 表示预测结果为正的样例。

4.2.3　覆盖率(coverage)

覆盖率(coverage)是度量测试完整性的手段，覆盖率为

$$Coverage = 覆盖数 / 总数 \tag{6}$$

4.3　实验结果

实验 1　用户数大于项目数

将两种算法对同一数据集，6 040 个用户对 3 925部电影共1 000 209条评论信息进行实验，实验结果见表 1。实验证明基于项目的协同过滤算法准确率更高，而基于用户的算法召回率、覆盖率更高，从时间上看基于用户的算法效率更高。

表 1　用户数大于项目数

Tab. 1　The number of users is greater than the number of items

算法	时间/s	精确率	召回率	覆盖率
基于用户	497.75	0.35	0.08	0.33
基于项目	1 087.01	0.36	0.07	0.17

实验 2　用户数小于项目数

将两种算法对同一数据集，610 个用户对 9 742 部电影的评论信息进行实验，结果见表 2。实验证明基于项目的协同过滤算法精准率、覆盖率更高，而基于用户的算法召回率更高，从时间上看基于用户的算法效率更高。

表 2　用户数小于项目数

Tab. 2　The number of users is less than the number of items

算法	时间/s	精确率	召回率	覆盖率
基于用户	9.63	0.29	0.07	0.04
基于项目	144.61	0.30	0.06	0.08

结论：

(1)从精确率来说，基于项目的协同过滤算法质量更高。

(2)从时间成本来说，基于用户的协同过滤算法效率更高。

5　结束语

信息大爆炸时代，面对如此庞大数量的信息，如何有效筛选有用信息是个性化推荐算法的主要目的，也极具商业价值。本文就协同过滤算法的选择对象不同，对基于项目和基于用户的协同过滤算法进行了比较分析研究，实验表明两种算法各具其特色，从精确率角度，基于项目的协同过滤算法质量更高；从时间成本角度，基于用户的协同过滤算法效率更高，应该在适宜的情况下，用相应的算法。当考虑精确率时，就使用基于项目的协同过滤算法，当考虑时间成本时，就使用基于用户的协同过滤算法。

参考文献

[1] 孙传明，周炎，涂燕. 基于混合协同过滤的个性化推荐方法研究[J]. 华中师范大学学报(自然科学版)，2020，54(6)：956-962.

[2] 荣以平，张鹏，朱伟义，等. 基于用户协同过滤的购电推荐算法[J]. 电力需求侧管理，2020，22(5)：58-62.

[3] 孟晗，高岑，王嵩，等. 结合信任关系的用户聚类协同过滤推荐算法[J]. 计算机系统应用，2020，29(8)：224-229.

[4] 夏景明，刘聪慧. 一种基于用户和商品属性挖掘的协同过滤算法[J]. 现代电子技术，2020，43(23)：120-123.