计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):182−187 [doi: 10.15888/j.cnki.csa.006726]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于物品的改进协同过滤算法及应用①

邓园园, 吴美香, 潘家辉

(华南师范大学 软件学院, 南海 528225)
通讯作者: 潘家辉, E-mail: panjh82@qq.com

摘　要: 针对电视产品信息资源量过载导致用户选择困难的问题, 本文主要研究了基于物品的协同过滤算法在电视产品推荐系统中的改进及应用, 将个性化推荐技术和电视产品系统有机结合来满足用户和运营商的需求. 在推荐过程中, 首先收集用户的偏好建立数据模型, 以用户观看电视产品的时长作为用户偏好的显式特征, 然后在传统的协同过滤算法中引入点播金额权重进行改进, 并采用欧几里德距离法计算物品相似度, 最后根据邻居集合预测目标用户对电视产品的观看时长, 得到推荐结果. 实验表明, 通过引入点播金额权重这一改进能够提高推荐的准确性.

关键词: 电视产品推荐系统; 推荐算法; 协同过滤算法; 点播金额权重

引用格式: 邓园园,吴美香,潘家辉.基于物品的改进协同过滤算法及应用.计算机系统应用,2019,28(1):182–187. http://www.c-s-a.org.cn/1003-3254/6726.html

Improved Item-Based Collaborative Filtering Algorithm and Its Application

DENG Yuan-Yuan, WU Mei-Xiang, PAN Jia-Hui

(School of Software, South China Normal University, Nanhai 528225, China)

Abstract: Aiming at the problem that the overload of information resources of TV products leads to the difficulty of user selection, this study mainly focuses on the improvement and application of article-based collaborative filtering algorithm in television product recommendation system, and combines the personalized recommendation technology with TV product system to meet the need of users and operations. In the recommendation process, the user's preference data model is first collected, and the duration of the user watching the television product is taken as the explicit characteristics of the user's preference. Then, it is improved by introducing the weight of on-demand amount in the traditional collaborative filtering algorithm, and the Euclidean distance method is used to calculate the similarity of the items. Finally, the viewing time of the target user on the television products is predicted according to the neighbor set, and a recommendation result is obtained. Experiments show that the introduction of on-demand amount weights can improve the accuracy of recommendations.

Key words: TV product recommendation system; recommendation algorithm; collaborative filtering algorithm; on-demand amount weight

1　引言

1.1　研究背景

当今时代, 互联网发展迅猛, 带动了电视产业的发展, 电视信息资源爆炸式增长导致了人们接触的电视产品信息量过载, 使用户多样化需求与电视产品海量资源之间的矛盾日益突出. 此时, 各种基于用户喜好的

① 基金项目: 国家自然科学基金青年科学基金 (61503143); 广州市科技计划项目珠江科技新星科技创新人才专项 (201710010038); 广东省自然科学基金博士科研启动项目 (2014A030310244)
Foundation item: Young Scientists Fund of National Natural Science Foundation of China (61503143); Special Fund for S&T Innovative Talent under Pearl River S&T Star, Science and Technology Plan of Guangzhou Municipality (201710010038); Doctorate Scientific Research Start-up Program of Natural Science Foundation of Guangdong Province (2014A030310244)
收稿时间: 2018-07-17; 修改时间: 2018-08-09; 采用时间: 2018-08-15; csa 在线出版时间: 2018-12-26

电视产品推荐系统应运而生, 推荐精准度成为衡量各大推荐系统的关键手段, 而推荐精准度依赖于推荐算法. 现流行的推荐算法有基于内容推荐、基于协同过滤推荐、基于关联规则推荐以及组合推荐等算法, 其中, 基于协同过滤的推荐算法对推荐对象无结构要求, 广泛应用于电影推荐中.

1.2 研究现状

(1) 国内外研究现状

国外的协同过滤算法首次提出是在二十世纪九十年代, 根据用户评分进行推荐, 这是最原始的推荐依据. 当时的用户数据还是比较稀疏的, 为了解决稀疏性问题, 专家们提出了可以利用用户的历史行为信息来间接获取数据的技术[1]. 至今, 推荐系统已经成为数据挖掘的主要服务对象. 国外著名的推荐系统有 GroupLens、PHOAKS 和 Ringo. 其中 GroupLens 是一个基于群体的共同偏好的新闻推荐系统. PHOAKS 通过记录在线用户所发帖子中的网站数量, 进行排名, 给相关用户推荐数量较高的网站, 达到网站推荐目的. Ringo 是一个音乐推荐系统, 通过对音乐人的评分对用户进行分组, 根据用户组内互推达成推荐目的. 由此可见, 国外的协同过滤技术已经较为成熟和先进.

虽然推荐系统已经应用在我国的电子商务和社交方面, 但是我国的推荐系统相关技术还处于初步阶段, 以推荐系统为关键词的论文发表也大大落后于国外, 并且大部分是参考国外的先进成果[2]. 2009 年, 国内首个个性化推荐系统团队成立, 主要致力于个性化推荐系统的研究. 目前, 我国推荐系统做得较好的有豆瓣网、淘宝网、爱奇艺等各大视频网站等.

(2) 关于协同过滤算法的改进研究

传统的协同过滤算法存在冷启动、稀疏等问题导致推荐精度偏差. 为了提高推荐精度, 相关学者也进行了一些优化处理. 重庆理工大学黄贤英等人[3]提出结合用户兴趣度聚类的协同过滤算法, 将用户-项目矩阵、项目-关键词矩阵结合成用户-关键词矩阵, 使系统发现隐藏的用户间的关系, 平均绝对误差 (MAE) 值降低到 0.643, 提高了推荐准确率. 浙江大学刘晓等人[4]提出电视剧推荐系统计算相似度时引入热门电视剧的热门程度考虑, 避免热门电视剧与其他电视剧相似度偏高, 获得大量推荐的问题, 准确率提高至 39.7%. 上海理工大学实验室[5]提出在皮尔逊相似度原理上添加热门因子来优化皮尔逊相似度计算, 改进后算法的 MAE 值减少到了 0.758, 推荐准确率得到提高.

基于以上的研究成果, 同样考虑到电视产品推荐的冷门项目问题, 本文提出基于物品的协同过滤算法在电视产品推荐系统中的研究应用, 并在此基础上根据用户的历史点播记录引入了点播金额权重这一个隐式特征进行改进, 使得推荐精度更加准确.

2 基于协同过滤的推荐算法

由于协同过滤的推荐算法对推荐对象无结构要求, 广泛应用于电影推荐中, 本文采用基于协同过滤的推荐算法来解决电视产品推荐问题. 协同过滤推荐算法主要分为两类: 基于用户的协同过滤算法和基于物品的协同过滤算法[6].

2.1 基于用户的协同过滤算法

基于用户的协同过滤推荐算法的基本思想是基于用户对物品的偏好找到用户的邻居用户, 然后将邻居用户的偏好推荐给当前用户, 偏好可以通过对用户的历史行为数据 (如商品购买、收藏、分享、评分、观看时长等) 挖掘而来. 在计算上, 根据不同用户对相同物品的偏好程度计算用户之间的关系, 利用有相同偏好的用户来预测当前用户的偏好, 然后选择预测喜爱程度最高的若干个推荐对象反馈给用户. 基于用户的协同过滤算法的推荐流程如图 1 所示.

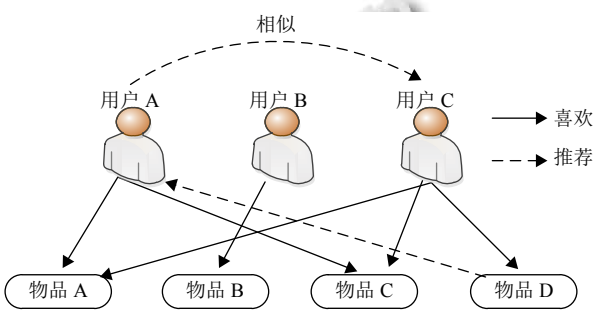

图 1 基于用户的协同过滤算法的推荐流程图

2.2 基于物品的协同过滤算法

基于物品的协同过滤推荐算法的原理和基于用户的协同过滤推荐算法相似, 将物品和用户对换. 在计算时计算的是物品之间的关系, 而不是用户之间的关系, 从物品本身出发, 基于用户对物品的偏好找到相似的物品, 然后利用 K 个最近邻居物品的加权来预测当前用户对这 K 个邻居物品的喜好程度, 从而将喜好程度高的若干个物品推荐给用户. 基于物品的协同过滤算法的推荐流程如图 2 所示.

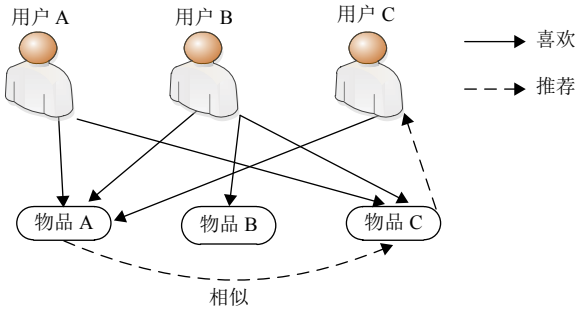

图 2 基于物品的协同过滤算法的推荐流程图

2.3 两种协同过滤推荐算法的比较

基于用户的协同过滤推荐算法和基于物品的协同过滤推荐算法都各有优势, 对于电视产品的推荐来说, 客户量的数据远远少于产品 (物品) 数据, 在这种非社交网络的电视产品推荐引擎中, 内容内在的联系的推荐原则比基于相似用户的推荐原则更加有效, 故在电视产品推荐上选择基于物品的协同过滤算法. 在给用户推荐电视产品时, 给用户推荐与该用户历史观看的电视节目相似度高的电视产品, 和推荐相似用户观看的电视产品给用户相比, 显然前者更加具有说服力.

3 引入点播金额权重后的协同过滤推荐算法

本文采用基于物品的协同过滤推荐算法来解决给用户推荐电视产品的问题, 并在传统的基于物品的协同过滤推荐算法中引入点播金额权重进行改进, 用于提高推荐系统的准确率. 首先根据用户的收视信息来分析用户的收视偏好, 将用户观看某个电视节目的时间长短作为用户喜好的显式特征, 根据电视产品的营销特性, 建立好数据模型后加入电视产品的点播金额权重, 再计算相似度. 然后根据电视产品之间的相似性和用户的历史观看记录来预测目标用户对未观看过的电视产品的观看时长, 产生推荐结果, 寻找 k 最近邻构成推荐矩阵. 算法流程图如图 3 所示.

3.1 收集用户偏好

用户收视信息数据主要包括用户的机顶盒号、电视节目名称, 对应电视节目观看时长、电视产品点播金额, 将用户观看某个电视节目的时长作为衡量用户喜好程度的依据. 根据用户机顶盒号、电视节目名称、观看时长这三个数据生成矩阵, 作为基于物品的协同过滤算法模型的输入数据, 建立数据模型.

在建立完数据模型后, 对构建好的初始矩阵 (用户-观看时长矩阵 **X**) 采取矩阵相乘的方式进行点播金额权重 (电视产品-点播金额权重矩阵 **M**) 的赋权, 得到最终数据矩阵 **W**, 如式 (1). 因为流行物品往往和任意物品的相似度都很高, 现实中基于物品的协同过滤推荐算法应用往往会增加对流行物品的惩罚度[7]. 对于电视产品的推荐来说, 有些电视产品是需要点播金额才可以观看的, 根据用户的消费心理, 点播金额小的电视节目往往会更受欢迎, 也就是所谓的“流行产品”, 在进行推荐时, 这些流行产品会对推荐结果引起干扰, 造成推荐结果不准确. 而且点播金额还能在一定程度上反映用户的偏好, 点播金额昂贵的电视产品用户仍愿意点播说明用户很喜爱这个电视产品. 所以在构建数据模型时引入点播金额权重, 可使推荐结果更加准确.

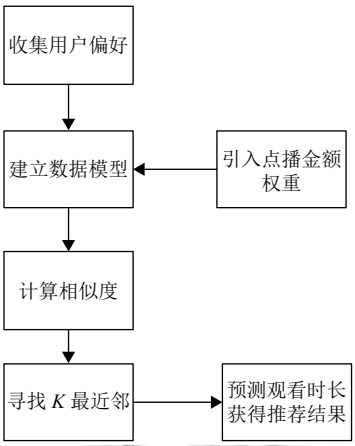

图 3 引入点播金额权重后协同过滤算法的流程图

$$\mathbf{W}=\mathbf{X}\cdot\mathbf{M}=\begin{bmatrix}X_1\\X_2\\\vdots\\X_n\end{bmatrix}\cdot\begin{bmatrix}M_1 & M_2 & \cdots & M_t\end{bmatrix}=\begin{bmatrix}X_1\cdot M_1 & X_1\cdot M_2 & \cdots & X_1\cdot M_t\\X_2\cdot M_1 & X_2\cdot M_2 & \cdots & X_2\cdot M_t\\\vdots & \vdots & \ddots & \vdots\\X_n\cdot M_1 & X_n\cdot M_2 & \cdots & X_n\cdot M_t\end{bmatrix} \tag{1}$$

3.2 计算相似度

计算物品相似度是基于物品的协同过滤算法的核心, 计算电视产品之间的相似度即为计算向量间的距离, 距离越近相似度越大, 主要有欧几里得距离、皮尔逊相关系数、余弦相似度[8]这三种方法.

由于电视产品推荐和其他用评分作为用户偏好的推荐系统不同, 不存在“分数膨胀”的问题, 电视产品的

收视数据比较密集和完整, 距离数据非常重要, 故最终选择欧几里德距离法计算电视产品之间的相似度.

欧几里德距离法: 最初用于计算欧几里德空间中两个点的距离, 假设 x, y 是 n 维空间的两个点, 它们之间的欧几里德距离是:

$$d(x,y)=\sqrt{\left(\sum(x_i-y_i)^2\right)} \tag{2}$$

当用欧几里德距离表示相似度, 一般采用式 (3) 进行转换: 距离越小, 相似度越大.

$$sim(x,y)=\frac{1}{1+d(x,y)} \tag{3}$$

3.3 寻找 K 最近邻

计算出电视产品之间的相似度后, 保留与目标电视产品相似度最大的 K 个电视产品作为其最邻居电视产品集合.

3.4 获得推荐指数

根据上一步获得的最邻居电视产品集合, 结合邻居电视产品对目标用户未观看过的电视产品的观看时长以及相似度来计算目标用户观看未看过的电视产品的预测观看时长, 通常采取中心加权平均值[9]的方法计算目标用户 v 对未看过电视产品的预测观看时长 $p(v,i)$, 记 r_i、r_j 分别表示观看过电视产品 i、电视产品 j 的用户的观看时长平均值, 如式 (4).

$$p(v,i)=\overline{r_i}+\frac{\sum_{j\in N_i}\left(r_{v,j}-\overline{r_j}\right)\times sim(i,j)}{\sum_{j\in N_i}|sim(i,j)|} \tag{4}$$

将预测的观看时长归一化后作为电视产品的推荐指数, 将推荐指数的 Top10 作为推荐结果反馈给用户.

4 实验结果及效果分析

4.1 实验数据集

本文的实验采用 Python 语言, 泰迪杯比赛提供的广电平台收视信息数据作为实验数据, 共 452 455 条数据, 主要是用户观看电视产品的数据, 包含用户机顶盒号、电视节目名称, 对应电视节目观看时长等数据 (如表 1), 还有电视产品对应的点播金额数据 (如表 2). 将这些数据的 80% 作为训练集, 20% 作为测试集.

表 1 广电平台的收视信息数据表

机顶盒号	电视产品名称及其观看时长 (h)		
10242	疯狂婚礼: 0.33	叶问: 2.98	80 天环游世界: 1.84
10246	奔跑吧: 收官之战！邓超遇王力宏变迷弟: 3.6	萨米大冒险 2: 0.47	反恐特战队之猎影: 0.03
10249	绝世高手: 3.58	神偷奶爸: 11.3	碟中谍 5: 神秘国度: 5.36
10255	哈利波特 6 混血王子: 18.27	异种: 5.66	射雕英雄传: 0.74
10257	特别呈现: 传家本事—天地的承诺: 0.98	楚乔传: 0.02	人文地理: 风花雪乐: 0.12

表 2 电视产品点播金额数据表

影片名	点播金额
神秘事件	108
王牌对王牌 第二季: 王源演技爆发成烈士	99
我们的爱	66
我家住在大海边	49
胜利的游戏	99

4.2 评估标准

通过比较推荐结果和实际观看的电视产品集的偏差, 来衡量一个推荐系统的好坏. 本文用平均绝对误差 (MAE)[3]和平均预测覆盖率 (APC)[10]两个评估指标来对比分析传统的基于物品的协同过滤算法和优化后算法的效果.

(1) 平均绝对误差

平均绝对误差是一种常用的用于衡量统计的准确性和比较的度量方法, 通过计算预测的用户观看时长与实际用户观看时长之间的偏差度来度量预测的准确性. MAE 值越小, 推荐准确度就越高. 假设系统预测推荐的电视产品集合为 $(x_1, x_2,\cdots,x_n)$, 用户实际观看电视产品集合为 $(y_1, y_2,\cdots,y_n)$, 计算公式如下:

$$MAE=\frac{\sum_{i=1}^{n}|x_i-y_i|}{n} \tag{5}$$

(2) 覆盖率

覆盖率表示一个推荐系统对长尾商品的挖掘能力, 设推荐给一个用户的电视产品集合为 R, 该用户实际观看电视产品集合为 A, 采用 APC(平均预测覆盖率) 计算平均整体的推荐效果, 如式 (6).

$$APC = \frac{1}{N}\sum_{i=0}^{N}\frac{R_i \cap A_i}{|R_i|} \quad (6)$$

4.3 实验结果及效果分析

本文设置两组实验, 实验 1 用来确定最优的点播金额权重值以达到最佳效果, 实验 2 用来验证优化后算法的有效性.

实验 1 在其他参数相同的条件下, 只改变点播金额权重, 点播金额权重值从 0 开始取值到 1, 步长为 0.1, 使用 *MAE* 指标来衡量不同点播金额权重下的推荐效果. 实验结果如图 4 所示, 由图 4 可看出, 当点播金额权重值为 0.3 时 *MAE* 值最小, 推荐精度最高.

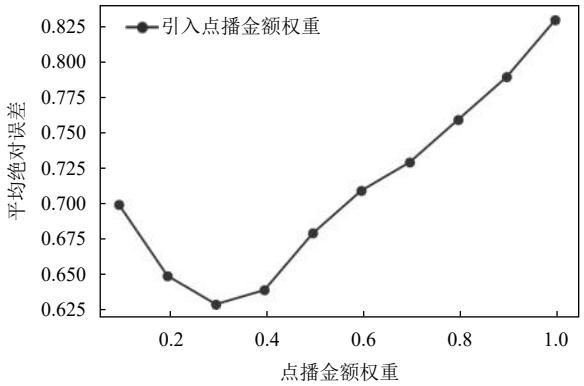

图 4 不同点播金额权重的平均绝对误差对比图

实验 2 在相同参数的环境下, 将本文改进算法 (点播金额权重取实验 1 得到的最优值 0.3) 同时与传统协同过滤算法和文献[5]改进算法进行对比, 产生推荐结果 (如表 3). 将三种算法得到的推荐结果与实际兴趣集合进行比较, 用 *MAE* 和 *APC* 两个评估指标衡量, 实验结果如图 5 和图 6 所示.

由图 5 可知, 相似邻居数目的变化对三种算法均有影响, 随着邻居数目的增加 *MAE* 达到最小值并逐渐趋于稳定. 在邻居数目为 15 时, 三种算法的 *MAE* 均达到最小值, 此时推荐准确度最高. 无论邻居数目为何值, 本文改进算法和文献[5]改进算法都比传统算法的 MAE 值低, 而本文改进算法的 *MAE* 值总体上略优于文献[5]改进算法, 说明准确率较高. 由图 6 可看出, 本文改进算法的平均预测覆盖率均高于传统算法和文献[5]改进算法, 说明引入点播金额权重后能够提高推荐系统的覆盖率. 文献[5]改进算法引入物品热门因子来改进协同过滤算法, 达到优化相似度的计算的目的, 但在本文电视产品推荐应用中, 由于电视产品的特性, 点播金额比物品热门因子衡量物品热门程度的效果更好, 故推荐效果更佳.

表 3 推荐结果部分示例数据表

机顶盒号	电视产品名称	推荐指数
10299	音悦 V 榜内地: TFBOYS 三周年唱响北京, 舞台惊喜连连	0.999 99
10299	美国骗局	0.967 54
10299	神奇大块头	0.907 63
10299	真相: 跨湖桥-未了的迷局	0.887 65
10299	盗剑 72 小时	0.854 27
10299	特别呈现: 搏击玫瑰	0.838 27
10299	正骨	0.802 76
10299	时代: 倔老头还乡	0.756 09
10299	名利场	0.743 29

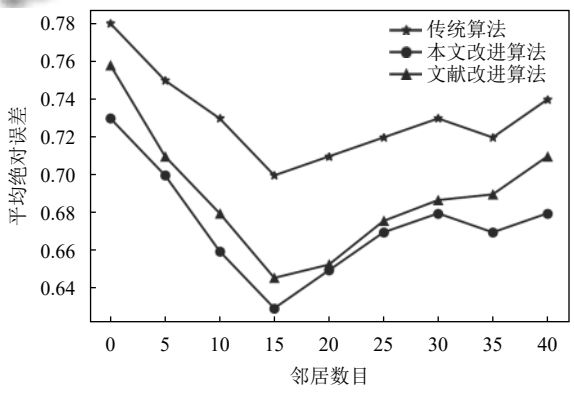

图 5 优化前后平均绝对误差对比图

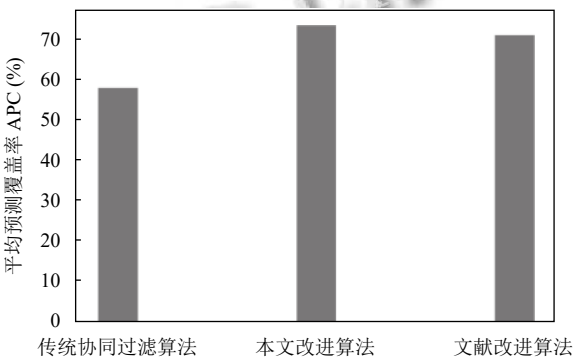

图 6 优化前后平均预测覆盖率对比图

5 结束语

本文主要研究了基于物品的协同过滤算法在电视产品推荐系统中的改进及应用, 在传统的协同过滤算法中引入点播金额权重来避免流行物品对推荐结果产生的干扰, 使推荐结果更加准确. 实验表明, 优化后的推荐算法的平均绝对误差和平均预测覆盖率均比传统

的协同过滤算法高, 说明优化后算法的推荐结果准确率得到提高. 个性化电视产品的推荐不仅能够给用户带来获取信息的便利性, 还能给运营商带来巨大的经济利益, 达到一石二鸟的效果.

参考文献

1 Ying ZY, Zhou ZR, Han FJ, *et al*. Research on personalized Web page recommendation algorithm based on user context and collaborative filtering. Proceedings of the IEEE 4th International Conference on Software Engineering and Service Science. Beijing, China. 2013. 220–224.

2 张光前, 雷彩华, 吕晓敏. 电子商务推荐的研究现状及其发展前景. 情报杂志, 2011, 30(12): 60–65. [doi: 10.3969/j.issn.1671-3982.2011.12.021]

3 黄贤英, 龙姝言, 谢晋. 结合用户兴趣度聚类的协同过滤推荐算法. 计算机应用研究, 2019, 36(9).

4 刘晓. 基于隐式反馈的电视剧推荐系统[硕士学位论文]. 杭州: 浙江大学, 2015.

5 孙红, 韩震. 融合物品热门因子的协同过滤改进算法. 小型微型计算机系统, 2018, 39(4): 638–643. [doi: 10.3969/j.issn.1000-1220.2018.04.004]

6 Zheng S, Shen YJ, Zhang GD, *et al*. A collaborative filtering recommendation algorithm based on dynamic and reliable neighbors. Proceedings of the IEEE 6th International Conference on Software Engineering and Service Science. Beijing, China. 2015. 690–693.

7 曹景振, 贾新磊, 李松丹. 基于物品的协同过滤算法在ACM在线评测推荐系统中的改进及应用. 无线互联科技, 2018, 15(5): 135–136, 139. [doi: 10.3969/j.issn.1672-6944.2018.05.062]

8 张国凯. 基于协同过滤的个性化服务推荐算法研究. 软件导刊, 2015, 14(10): 43–44.

9 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. 现代商贸工业, 2018, 39(17): 59–62.

10 孙权, 贺细平. 协同过滤算法在ACM在线评测推荐系统中的应用研究. 电脑与信息技术, 2015, 23(6): 11–14. [doi: 10.3969/j.issn.1005-1228.2015.06.004]