**Statistics and Application** 统计学与应用, 2019, 8(4), 622-630
Published Online August 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.84071

Hans汉斯

# Personalized Recommendation of TV Users Based on Collaborative Filtering Algorithm

**Xingxing Chen, Ruitao Li, Junhua Liao, Yanke Wu***

School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang Guangdong
Email: *yanke.wu@163.com

Received: Jul. 19th, 2019; accepted: Aug. 1st, 2019; published: Aug. 12th, 2019

## Abstract

**In order to integrate and utilize existing data better, and improve the marketing effectiveness of TV program products, this paper processes the data of the watching information of TV users and establishes the user preference model to obtain three preference types of each user, and then uses the collaborative filtering algorithm to carry out personalized recommendation of individual users. In addition, we use K-means algorithm as well as the KNN algorithm to divide the users into groups and obtain the recommendation of each user group, thereby effectively solving the problem of personalized recommendation of the users.**

## Keywords

**K-means, KNN, Collaborative Filtering, Personalized Recommendation**

# 基于协同过滤算法的电视用户个性化推荐

**陈星星，李瑞涛，廖军华，吴延科***

广东海洋大学数学与计算机学院，广东 湛江
Email: *yanke.wu@163.com

收稿日期：2019年7月19日；录用日期：2019年8月1日；发布日期：2019年8月12日

## 摘　要

**为了更好地整合和利用现有的数据，提高电视节目产品营销效益，本文通过对广电网络公司电视用户的收看信息数据进行数据处理，建立用户偏好模型，得到各个用户的三个偏好类型，然后使用协同过滤算**

*通讯作者。

文章引用: 陈星星, 李瑞涛, 廖军华, 吴延科. 基于协同过滤算法的电视用户个性化推荐[J]. 统计学与应用, 2019, 8(4): 622-630. DOI: 10.12677/sa.2019.84071

**法进行单个用户的个性化推荐，以及使用K-means算法和KNN算法将用户进行分群，得到用户群的推荐，进而有效地解决了用户个性化推荐的问题。**

## 关键词

**K-means聚类分析，KNN算法，协同过滤，个性化推荐**



## 1. 引言

在“数字融合”和“三网融合”引领下，电视产业进入了大数据时代，数据挖掘和数据资料的发现与利用开始为电视行业的多项工作提供更为科学的决策依据，为电视业的题材选择和媒介营销带来了巨大的变革。通过对大量观看数据的整合和利用，可以分析用户的特征和偏好，根据用户的收视偏好为用户贴上合适的“标签”，尽可能精确地勾勒出用户的“画像”[1]。继而对产品进行分类，将合适的产品在合适的时间推送给合适的用户，实现市场营销的精准化。

在现代许多平台，例如淘宝、京东、抖音和快手等等，在进行用户的兴趣分析以及类型推荐的时候，都使用了协同过滤算法。迄今协同过滤算法已经非常成熟，主要包括了基于物品的协同过滤算法和基于用户的协同过滤算法，它们都可以很好地描绘产品的属性以及用户的兴趣爱好，所以其在推荐系统的算法中具有十分重要的研究价值。

很多学者对于这个问题都做出了一定的研究。例如，林霜梅[2]提出了一种改进的向量空间模型(VSM)用户单兴趣表示法及其动态学习算法，用以捕捉用户最新的兴趣需求。沈建军[3]采用协同过滤算法对电视用户个性化推荐进行研究，对用户产品推荐算法提供了不少创新思想。在高肖俊的《基于用户收视兴趣模型的广电客户分群及精确化营销系统建设》中，作者根据用户以往收看信息建立收视兴趣度和忠诚度，以及电视用户对节目的评分，使用协同过滤算法产生推荐结果。在周虹君的《Spark 框架下的受众分群及矩阵分解的推荐算法研究》中通过找到目标用户的最近邻居集，以及通过最近邻居集去计算预测评分，从而解决了评分矩阵过于稀疏和冷启动的问题，提高了推荐结果的准确度。《个性化推荐系统综述》主要介绍了各种推荐算法，包括协同过滤算法、关联规则和深度学习等等的发展历程，更好地了解推荐算法的优缺点以及适用范围。在夏欢的《基于组合策略的 IPTV 节目推荐》中，其提出了两种组合策略的用户模型更新的方法相结合，并且这两种方法实现了协同过滤部分和基于内容的推荐部分，也很好地解决了评分矩阵过于稀疏和冷启动的问题。在《推荐系统实践》中，作者通过实际的推荐例子，突出了协同过滤算法对比于其他推荐系统的优势，也为协同过滤算法在各大平台上的使用奠定了基础。

虽然推荐系统的研究都有较为完善的理论，但是给用户推荐的节目被该用户所喜欢的成功率较低，夏欢[4]认为，结合相应节目的评分方法，可以有效地提高用户接受的成功率。

基于以上研究现状，本文的剩余内容将按以下结构展开：第 2 部分为对广电网络公司的电视用户收看数据进行数据预处理，第 3 部分建立了用户偏好模型提取各个用户的偏好类型，第 4 部分使用了协同过滤算法对各个电视用户进行了节目的推荐，第 5 部分是使用 K-means 算法和 KNN 算法对用户进行分群，在考虑到节目包的总时长和经济效益的基础上，有效地推荐节目。

## 2. 协同过滤算法

协同过滤算法是最早诞生、较为著名的推荐算法。其主要的功能为预测和推荐，该算法依据对用户的历史行为数据进行挖掘，发现用户的喜爱偏好，对于偏好不同的用户进行有效地个性化推荐。协同过滤算法主要分为两种，分别是基于用户的协同过滤算法和基于物品的协同过滤算法，可以简单地描述为：人以类聚，物以群分。协同过滤算法的实现过程如图 1 所示，假设用户 1 的偏好产品为产品 1、产品 2、产品 3 和产品 4，用户 3 的偏好产品为产品 2 和产品 3，若把用户 3 作为目标用户，则用户 1 是他的邻居用户，所以把产品 1 和产品 2 推荐给用户 3，从而实现了协同过滤。

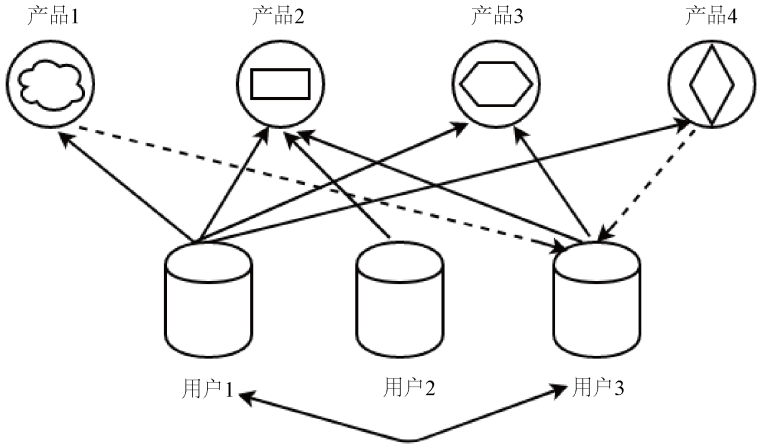


**Figure 1.** Classification of television programs
**图 1.** 电视节目的分类

在本文中，用到的是基于用户的协同过滤算法，下面主要介绍基于用户的协同过滤算法的步骤：

1) 建立用户模型

通过查阅资料和数据，得到各个用户对于各个物品的评分，若某个用户对某个物品无评分，则为 0，便得到了用户评分矩阵，构造出用户模型。一般来说，对于评分不为 0 的产品，评分越高，则代表用户对该产品的偏好度越高，否则越低。

2) 寻找目标用户的邻居

根据上面的用户评分矩阵，并使用皮尔逊相关系数、余弦相似度或者修正余弦相似度，可以计算各个目标用户与用户群中其他用户的相似度，取相似度最大的 $k$ 个用户作为该目标用户的最近邻居集。在本文中我们采用的是修正的余弦相似度公式(1)进行计算。

$$Sim(i,j)=\frac{\sum_{q=1}^{15}\left(R_{i,q}-\overline{R_i}\right)\left(R_{j,q}-\overline{R_j}\right)}{\sqrt{\sum_{q=1}^{15}\left(R_{i,q}-\overline{R_i}\right)\overline{R_j}}\sqrt{\sum_{q=1}^{15}\left(R_{j,q}-\overline{R_j}\right)^2}} \tag{1}$$

其中：$Sim(i,j)$ 表示用户 $i$ 和 $j$ 的余弦相似性；$R_{j,q}$ 表示用户 $j$ 对 $q$ 类型的评分。

3) 产生目标用户的推荐产品

根据目标用户的最近邻居集，并使用式子(2)计算出目标用户未观看过的产品类型的预测评分，然后把预测评分较高的产品类型，优先推荐给目标用户。

$$P_{i,q}=\overline{R_i}+\frac{\sum_{j\in U}sim(i,j)\times\left(R_{j,q}-\overline{R_j}\right)}{\sum_{j\in U}\left(\left|sim(i,j)\right|\right)} \tag{2}$$

其中：$P_{i,q}$ 表示用户 $i$ 对类型 $q$ 的预测评分；$R_{j,q}$ 表示用户 $j$ 对类型 $q$ 的评分。

## 3. 数据来源及预处理

本文采用了第六届泰迪杯数据挖掘挑战赛 C 题的原始数据，进而对电视用户收看信息的原始数据进行数据清洗、属性规约和数据变换三个步骤。数据清理主要针对以下两个方面进行：

1) 去除回看信息重复和用户点播信息重复的数据、观看时间为空的记录和观看时长小于一分钟的记录。

2) 将其中冗余的属性以及与挖掘过程不相关的属性剔除得到处理后的数据，如表 1 所示。

**Table 1.** Form of data processing
**表 1.** 数据预处理的形式

| 正题名 | 分类名称 | 时长 |
|---|---|---|
| 职场是个技术活(39) | 电视剧场\大陆剧场 | 38:53. |
| 06 月 28 日 自然：猫——隐匿的野性 | 科学教育 | 52:36. |
| ... | ... | ... |
| 小马宝莉特别篇 失控彩虹 | 电视剧场\欧美剧场 | 44:06. |

在处理后的数据中，用户收视信息和用户回看信息的统计日期均采用 YMD 格式。我们发现，每个频道一周的节目表当中，周一至周五的节目表相似，周六与周日的节目表相似。因此将用户收视信息和用户回看信息的统计日期转换为星期的格式，然后将数据按周一至周五、周末分为两类，分别写入两个表格当中。

## 4. 建立用户偏好模型

### 4.1. 构建用户偏好类型

首先，我们通过 python 爬虫技术，爬取各个频道分别在工作日及周末的两个节目表，经过分析和统计得出每个频道节目表中各个节目类型所占时间比例。在此基础上，可以初步确立节目分为综艺、财经、体育、电影、电视剧、科教、新闻、少儿动画、音乐、生活、戏剧、军事、纪录片、法律和广告购物，共 15 种节目类型，如图 2 所示。


**Figure 2.** The type of TV program
**图 2.** 电视节目的分类

然后把各个电视用户的观看各种类型的时间计算出来，得到“用户－类型－观看时长”矩阵，如表2所示。

**Table 2.** Matrix: “user-type-duration”
**表 2.** “用户－类型－时长”矩阵

| | 综艺 | 财经 | 体育 | 电影 | | 广告购物 |
|---|---|---|---|---|---|---|
| 10001 | 15.5 | 0 | 0 | 0 | ... | 0 |
| 10002 | 1396.4 | 4.2 | 263.0 | 300.4 | ... | 72.2 |
| 10003 | 183.8 | 4.1 | 0 | 21.9 | ... | 26.0 |
| ... | ... | ... | ... | ... | ... | ... |
| 11329 | 249.9 | 0.3 | 0 | 753.1 | ... | 24.9 |

## 4.2. 用户类型的偏好度

引入兴趣度和忠诚度，并且用它们之间的加权和(即用户偏好度)来刻画用户对各种节目类型的喜欢程度和偏好程度[1]。

兴趣度：用户观看某种节目类型的总时长与该用户观看总时长的关系。代表该用户对某种节目类型的感兴趣程度[1]。

$$interest_{(i,q)} = \frac{\sum_{k=1}^{n} t_{label(i,q)}[k]}{\sum_{q=1}^{m}\sum_{k=1}^{n} t_{label(i,q)}[k]} \tag{3}$$

其中：$interest_{(i,q)}$ 表示用户 $i$ 对节目类型 $q$ 的兴趣度；$\sum_{k=1}^{n} t_{label(i,q)}[k]$ 表示用户 $i$ 观看节目类型 $q$ 的时长；$\sum_{q=1}^{m}\sum_{k=1}^{n} t_{label(i,q)}[k]$ 表示用户 $i$ 观看所有节目类型的总时长。

忠诚度：用户收看某节目类型的总时长与该节目类型的播放时长的关系。代表该用户对该节目类型的忠诚程度[1]。

$$Loyalty_{(i,q)} = \frac{\sum_{k=1}^{n} t_{label(i,q)}[k]}{\sum_{k=1}^{n} t_{program(i,q)}[k]} \tag{4}$$

其中：$Loyalty_{(i,q)}$ 表示用户 $i$ 对节目类型 $q$ 的忠诚度；$\sum_{k=1}^{n} t_{label(i,q)}[k]$ 表示用户 $i$ 收看节目类型 $q$ 的总时长；$\sum_{k=1}^{n} t_{program(i,q)}[k]$ 表示节目类型 $q$ 的播出总时长。

用户偏好度：描述的是用户对某一类型节目的喜好和偏好程度，一般为兴趣度和忠诚度的加权和[5]。

$$Label_{(i,q)} = \omega_1 Loyalty_{(i,q)} + \omega_2 interest_{(i,q)} \tag{5}$$

为“用户－标签”偏好集合中不同特征赋予权值，且权值之和等于 1，以便区分不同集合中其模型的贡献度。本文假定模型的权值为 $\omega_1 = 0.7$，$\omega_2 = 0.3$。

然后通过计算每个用户对于各个类型的偏好度，得到“用户－类型－偏好度”矩阵，选取了每个用户的偏好度最高的三个标签作为该用户的偏好标签，如表3所示。

**Table 3.** These preference types for each user
**表 3.** 各个用户的三个偏好类型

| 设备号 | 编好类型 | | |
|---|---|---|---|
| 10001 | 电视剧 | 新闻 | 综艺 |
| 10002 | 电视剧 | 少儿动画 | 综艺 |
| ... | | ... | |
| 11329 | 电视剧 | 新闻 | 电影 |

## 5. 协同过滤算法对单个用户的推荐

### 5.1. 建立用户模型

由于用户偏好度代表的是用户对某一种类型的喜欢和热爱程度，一般来说用户对某一类型的偏好度越高，相应的该用户对这种类型的评分(0~1 之间)也会越高，因此我们近似地把“用户－类型－偏好度”矩阵看做用户对节目类型的评分矩阵。

### 5.2. 寻找目标用户的邻居

在这一阶段，主要完成目标用户的邻居的寻找。将全部用户视为一个用户群 $U$，每一次拿出一个用户作为目标用户，并且在其余的用户中寻找到该目标用户的邻居 $u$ [6]。然后使用式子(1)计算每两个用户之间的相似度，得出相似度矩阵，在矩阵中找出与目标用户相似度最高的用户作为目标用户的邻居。

### 5.3. 产生目标用户的推荐类型

对于目标用户，找到了他的邻居之后，使用式子(2)计算该目标用户未观看过的类型的预测评分[7]，然后选取其中目标用户中的预测评分中最高的那一个节目类型作为该目标用户的推荐类型。最后，把目标用户的三个偏好类型加上这一个推荐类型，共为四个电视节目类型，作为目标用户的全部推荐类型。

### 5.4. 实现用户的节目推荐

推荐指数的定义：某一节目产品的推荐指数等于该用户对这种节目类型的偏好度与这一节目产品的评分的加权和[8]。

我们通过 python 爬虫技术，在豆瓣网爬取了各个节目的标签和评分，以便获得更精确的产品营销推荐。每一个用户都有四个推荐类型，对应这四个推荐的节目类型，再结合节目的评分的高低和用户的偏好度的大小，进而将节目精准地推送给各个用户，得到每个用户的节目推荐表，如表 4 所示。

**Table 4.** Recommended program for individual user
**表 4.** 各个用户的推荐节目

| 设备号 | 节目名 | 推荐指数 |
|---|---|---|
| 10001 | 琅琊榜 | 1.00 |
| 10001 | 西游记续大闹披香殿加防抖 | 1.00 |
| 10001 | 白夜追凶 | 0.99 |
| 10002 | 西游记续大闹披香殿加防抖 | 1.00 |
| 10002 | 秦时明月 | 0.97 |
| ... | ... | ... |
| 11329 | 百鸟朝凤 | 0.93 |
| 11329 | 碟中谍神秘国度调整字幕 | 0.92 |

## 6. 用户的打包推荐

### 6.1. K-means 聚类分析与 KNN 算法进行用户分群

首先，通过所得到的“用户－类型－观看时长”矩阵，针对全部用户，利用 K-means 聚类算法，分类得到 100 个聚类中心，然后运用 KNN 最近邻分类算法将全部的待测用户找到所隶属的聚簇，将同一个

聚簇内的用户归为一个用户群。

通过 K-means 聚类分析算法得到 100 个聚类中心如表 5 所示。

**Table 5.** Clustering center of K-means algorithm
**表 5.** K-means 算法的聚类中心

| | 聚类中心 | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1156.3 | 29.0 | 304.1 | … | 210.0 | 31.6 |
| 2 | 2900.2 | 123.8 | 3091.5 | … | 974.5 | 68.1 |
| … | … | … | … | … | … | … |
| 99 | 4427.7 | 225.2 | 5302.3 | … | 618.5 | 45.7 |
| 100 | 9520.2 | 310.5 | 2950.1 | … | 1475.9 | 22.0 |

通过 KNN 最邻近分类算法，将各个用户都分类到所隶属的聚簇用户群中，所得到的结果，如表 6 所示。

**Table 6.** User groups of individual user
**表 6.** 各个用户所在的用户群

| 机顶盒号 | 分类用户群 |
|---|---|
| 10001 | 68 |
| 10002 | 25 |
| … | … |
| 11328 | 24 |
| 11329 | 87 |

找出各聚类中心最近的 3 个测试变量，分别将这 3 个测试变量作为该聚簇用户群的三个偏好类型。结果如表 7 所示。

**Table 7.** Preference types of each user group
**表 7.** 各个用户群的偏好类型

| 用户群 | 偏好类型 | | |
|---|---|---|---|
| 1 | 电视剧 | 综艺 | 新闻 |
| 2 | 少儿动画 | 电视剧 | 体育 |
| … | … | … | … |
| 99 | 体育 | 电视剧 | 综艺 |
| 100 | 电视剧 | 电影 | 新闻 |

## 6.2. 节目产品的打包和推送

为了便于产商在产品包内植入广告从而获取最大的经济效益，将每一个产品包的节目总时长限制在一个合理的范围内，并且优先考虑到高评分的产品[9]，建立聚类推荐模型，实现了节目的打包推送问题。

由上述得到的各个用户群的三个偏好类型，匹配到相应类型的电视节目产品，从中选择评分较高的

电视节目产品，由于考虑到每个节目产品包的互补性与互斥性[10]，规定每一种类型的节目产品不超过三个。为了便于产商在产品包内植入广告从而获取最大的经济效益[11]，将每一个产品包的节目总时长限制在 $200\,\text{min} \le T \le 720\,\text{min}$ 。

对电视节目产品打包并且将产品包推送给每一个用户群，并且计算出每个节目包的总时长，推送结果如表 8 所示。

**Table 8.** Recommended program package for each user group
**表 8.** 各个用户群的推荐节目包

| 用户群 | 打包的推送产品 | 产品包总时长/分钟 |
|---|---|---|
| 1 | 琅琊榜 | |
| … | … | |
| 1 | 广视新闻 | 524.4 |
| 2 | 艾可魔法少女 | |
| … | … | |
| 2 | 国际田联砖石联赛奥斯陆站 | 482.3 |
| … | … | |
| 100 | 琅琊榜 | |
| … | … | |
| 100 | 广视新闻 | 679.7 |

## 7. 结论及展望

本文通过对用户观看数据进行预处理、挖掘出各个用户的观看偏好，并且通过爬虫技术，爬取得到节目的类型，进而把节目推荐给感兴趣的用户，使得广电网络公司在电视节目的营销上更加有针对性，但是模型主要存在以下两个不足之处：

1) 各个电视台的节目类型时间比例是根据各个电视台(分周一至周五和周末)的常规播放规律来进行计算的，与准确的节目类型播放规律会有一定的误差。

2) 评分较高的节目推荐给目标用户，往往会导致推送的不精准和有一大部分评分较低的节目没有用户去收看。

对于上述的不足，可以通过更精准地利用爬虫技术对电视台的播放节目类型时间比例和节目类型、评分进行爬取，以减低推荐的误差，希望以后的学习可以在这方面做出一定的研究。

## 参考文献

[1] 周虹君, 殷复莲, 陈怡婷, 周嘉琪, 伊成昱. Spark 框架下的受众分群及矩阵分解的推荐算法研究[J]. 互联网 + 健康, 2016, 20(2): 139-141.

[2] 张晓阳. 基于受众收视行为分析的电视节目编排策略刍探[J]. 企业家天地, 2011, 3(1): 187-188.

[3] 高肖俊, 丁云强. 基于用户收视兴趣模型的广电客户分群及精确化营销系统建设[J]. 视听界, 2016, 3(2): 1-3.

[4] 夏欢. 基于组合策略的 IPTV 节目推荐[D]: [硕士学位论文]. 昆明: 云南大学, 2016.

[5] 沈建军. 面向互动电视的影视节目推进系统研究与实现[D]: [硕士学位论文]. 杭州: 复旦大学, 2012.

[6] 钟智, 朱曼龙, 张晨, 黄樑昌. 最近邻分类方法的研究[J]. 广西师范学院, 2011, 5(1): 2-3.

[7] 赵营. 基于协同过滤算法的习题推荐[J]. 科技经济导刊, 2016, 6(5): 136.

[8] 李楚桐, 莫赞. 基于协同过滤算法的推荐系统研究[J]. 信息通信, 2018, 1(2): 38-39.
[9] 王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7): 66-67.
[10] 于玉龙, 王秀芳. 改进的协同过滤推荐算法[J]. 互联网 + 通信, 2016, 1(1): 12.
[11] 张岩. 基于协同过滤的个性推荐算法研究及系统实现[D]: [硕士学位论文]. 北京: 北京化工大学, 2014.