第六届“泰迪杯”
数据挖掘挑战赛

优
秀
作
品

作品名称：基于收视记录的电视产品营销及用户画像

荣获奖项：一等并获信诺创新奖

作品单位：河海大学

作品成员：方思越 钟雨纯 杨洋

指导老师：施国良

基于收视记录的电视产品营销及用户画像

摘要：三网融合的背景下，一方面越来越多的家庭对机顶盒电视实际上有了更高的需求，希望快速获得目标资源；另一方面，电视服务供应商也希望有效挖掘出用户需要和感兴趣的资源、信息，实现大数据分析，形成个性化的产品营销及有偿服务。

该论文基于用户收视信息、产品信息及用户套餐数据，利用协同过滤及分类的思想，先为用户做产品推荐，再进行分类推荐，并对套餐设置提出建议。

首先对数据进行预处理，对数据进行清洗、转换，利用爬虫捕捉网络数据进行数据补充。

对于第一题，主要使用“用户点播信息”与“用户单片点播信息”表，基于用户的协同过滤与基于电视产品的协同过滤相结合，为用户推荐电视产品。

对于第二题，在第一题的基础上结合三个附件，构建用户标签体系与产品标签体系，按照标签体系对用户和产品进行分类，利用分类后的用户-产品类别矩阵再次进行协同过滤，实现分类推荐。

分析现有套餐设置结构，并对套餐的销售品字段进行拆分。利用套餐拆分后的结果和用户画像的结果进行关联规则，发现不同用户画像与套餐内容之间的规律，为套餐推荐提出建议。

文章最后对此次实验做出总结，浅谈收获并对未来做出展望。

关键词：Excel；协同过滤；爬虫；Python

Television Product Recommendations and User Portraits Based on Viewing History

Abstract: In the integration of three networks age, on the one hand, more and more families actually have higher demand for set-top TVs and hope to quickly obtain target resources; on the other hand, TV service providers also hope to effectively tap user needs and senses. Resources and information of interest, realizing big data analysis, and forming personalized product marketing and paid services.

This paper is based on user viewing history information, product information and user package data. It uses the idea of collaborative filtering and classification to make product recommendations for users first, and then makes recommendations for categorized users. Also, we make suggestions for package settings.

First, the data is preprocessed. The data is cleaned and converted, and we use the crawler captures network data to supplement the data.

For the first question, the last two sheets in the first annex are mainly used, and user-based collaborative filtering is combined with collaborative filtering based on television products to recommend television products for users.

For the second question, our method is based on the results of the first question and the three attachments were combined to build a user tag system and a product labeling system. The users and products were classified according to the tag system, and the classified user-product category matrix was used again. We again use collaborative filtering to achieve classification recommendations.

Analyze the existing package setup structure and split the sales item field of the package. Using the result of the splitting of the package and the result of the user's portrait, the association rule is found, and the rules between the portraits of different users and the contents of the package are found, and suggestions are given for the package recommendation.

At the end of the article, a summary of this experiment is given, which is about harvesting and looking forward to the future.

Key words: Excel; Collaborative filtering; Web crawler; Python

目录

一、引言

1.1 背景

随着技术的发展，宽带网络和电视机顶盒的普及，一方面越来越多的家庭对机顶盒电视实际上有了更高的需求，体现在要求简化对电视节目的检索操作，快速获得感兴趣的资源；另一方面，电视服务供应商也希望有效挖掘出用户需要和感兴趣的资源、信息，实现大数据分析，形成个性化的产品营销及有偿服务。

个性化推荐是数据挖掘与分析的主要应用之一，关于个性化推荐，目前常用的算法有基于内容的推荐、基于协同过滤的推荐以及混合型推荐系统。其中协同过滤又有多个子类别，主要包括基于用户的协同过滤、基于物品的协同过滤及基于模型的协同过滤等。目前，个性化推荐系统已广泛运用于电子商务网站、社交应用及视频、新闻门户网站中。

从本质上而言，电视节目的推荐与其他节目领域具有相似性，然而它有其他推荐领域不具备的特点。一是电视节目通常面向家庭，不同的家庭成员兴趣偏好可能不同，造成推荐结果的不准确；二是电视节目往往具有时效性，除去用户的自身兴趣，用户常被较新的电视节目所吸引；三是电视节目具有延续性，同名的电视节目可能有多集多系列。

1.2 研究内容

本文主要利用提供的收视数据，针对电视节目推荐的特点：

（1）基于用户点播观看行为，利用基于用户的协同过滤与基于物品的协同过滤方法为用户推荐电视产品。

（2）基于用户收视、回看及点播观看行为进行用户画像，给用户打上分类标签。

（3）利用产品信息及网络爬虫获取的数据对产品进行分类打包，结合用户画像，给出营销推荐方案。

（4）对分类后的用户信息做关联规则，为电信公司的电视套餐推荐提供一定的建议。

1.3 研究思路

为了实现研究内容，本小组首先对数据进行分析，从用户和产品两个角度讨

论如何实现标签化。我们认为用户的收视行为能够侧面反映出用户的性别、年龄等特征，能通过用户的收视时段等特征判断用户的家庭构成，电视产品也有针对某类用户的特性。为此我们也在网页、论文库中找到一些资料辅佐我们的看法。

同时，本小组讨论认为在本题中，使用基于内容的协同过滤算法难度较大，首先是补全数据需要大量的时间，其次文本分析的复杂度大，电脑可能难堪重任。最后我们选择基于用户和基于物品的协同过滤算法，利用用户的历史数据，同时使用用户的平均打分补全空缺值消除冷启动的影响。

由于现有的套餐区分度不高，本小组希望通过关联规则发现不同画像与套餐内容之间的关系，给电信公司的套餐设置与推荐以一定的启发。

二、数据预处理

2.1 附件 1：用户收视信息

2.1.1 用户收视信息

（1）对频道名进行去除重复值操作，得到已有收视记录的电视台，将频道号与频道名对应。

（2）利用收看开始时间与收看结束时间计算收视时长，添加到原表中。

（3）除去收看时长大于 16 小时的收视记录。（由于只关电视忘关机顶盒也被记录在收视记录中，但对于用户而言这段时间的收视是无效的；采用 16 小时是假设用户除了基本生理需求外都在看电视）

（4）为收看开始时间与收看结束时间打上星期号的标签。

（5）在“电视猫”网站上获取各电视台节目播放表，统计播放星期、开始时间、节目名称及栏目类型。

2.1.2 用户回看信息

（1）去除表中的重复数据。

（2）利用回看开始时间与回看结束时间计算回看时长。

（3）除去收看时长大于 16 小时的收视记录。

（4）为回看开始时间与回看结束时间打上星期号的标签。

2.1.3 用户点播信息

（1）删除节目名称中含有乱码（&、*、？等符号）的节目。

（2）去除题名相同的节目名称的集数（如“超时空男臣（05）”改为“超时空男臣”）。

2.1.4 用户单片点播信息

（1）删除影片名称中含有乱码的节目。

（2）去除题名相同的影片名称的集数。

（3）计算用户观看时长。

（4）去除重复值。

（去除重复值后数据共有 34187 条、去除名称乱码的异常数据后有 34181 条）

2.2 附件 2：电视产品信息数据

提取分类名称、连续剧分类、字母语种、声道语种、地区参数字段作为用户画像和产品分类参考字段。

（1）将连续剧分类标准化并补充缺失值。

（2）在“爱奇艺”网站中利用爬虫抓取综艺、纪录片、动漫、电影及电视剧标签信息，做已有产品信息的参照。

2.3 附件 3：用户基本信息

（1）以 2017/10/31 作为基准时间（所给收视记录按时间排序最后一天的时间）。利用基准时间与状态改变时间计算用户使用当前套餐时间。

（2）以基准时间与入网时间计算用户入网时长。

（3）对用户使用的套餐、销售品及资费做处理，得到当前业务套餐结构。

以上过程大部分可以使用 excel 中的函数直接得到结果。对数据进行预处理后的结果见**附件：预处理后数据**。

三、研究方案及实施

3.1 问题一

本题要求识别用户偏好并为用户推荐附件 2 中的产品。经思考，小组决定主要使用“用户点播信息”与“用户单片点播信息”表。两表的不同之处在于前者需要用户单独付费，而后者包含在用户套餐中，无须单独付费。本题分别对“用户点播信息”及“用户单片点播信息”构建协同过滤算法，最后为每位用户推荐 10 个电视产品。

3.1.1 基于用户的协同过滤

1.构建用户-节目评分矩阵

分别为“用户点播信息”及“用户单片点播信息”构建评分矩阵，如表 1 所示。

表 1 用户评分矩阵

	Program 1	Program 2	……	Program m
User 1	$R_{1,1}$	$R_{1,2}$	……	$R_{1,m}$
User 2	$R_{2,1}$	$R_{2,2}$	……	$R_{2,m}$
……	……	……	……	……
User n	$R_{n,1}$	$R_{n,2}$	……	$R_{n,m}$

其中 User 即电视用户，Program 即用户收看的电视节目，$R_{i,j}$ 表示用户对观看节目的偏好打分。在“用户点播信息”表中，$R_{i,j}$ 为用户收看对应节目的频次；在“用户单片点播信息”表中，$R_{i,j}$ 为用户收看对应节目的时长。

2.计算相似度

计算用户间的相似度有很多种方式，本文采用皮尔森相关系数计算两个用户之间的相关性，公式如下：

$$sim(u,v)=\frac{\sum_{i\in I_{u,v}}(R_{ui}-\bar{R}_u)\cdot(R_{vi}-\bar{R}_v)}{\sqrt{\sum_{i\in I_{u,v}}(R_{ui}-\bar{R}_u)^2}\cdot\sqrt{\sum_{i\in I_{u,v}}(R_{vi}-\bar{R}_v)^2}}$$

其中 I_{uv} 代表用户 u 和用户 v 共同收看的节目集合，R_{ui} 代表用户 u 对节目 i 的打分，$\bar{R}_u$ 表示用户 u 收看的节目的打分平均值，R_{vi} 代表用户 v 对节目 i 的打分，$\bar{R}_v$ 表示用户 v 收看的节目的打分平均值。

3.计算节目评分

根据相似用户的评分来计算目标用户对所有节目的打分，公式如下：

$$\hat{R}_{ui}=\bar{R}_u+\frac{\sum_{v\in U(u,K)\cap N(i)}S_{uv}(R_{vi}-\bar{R}_v)}{\sum_{v\in U(u,K)\cap N(i)}|S_{uv}|}$$

$\hat{R}_{ui}$ 代表目标用户 u 对项目 i 的预测评分，$\bar{R}_u$ 表示用户 u 收看的节目的打分平均值，$U(u,K)$ 是目标用户 u 最相似的 K 个用户的集合，$N(i)$ 表示对节目 i 有所评分的用户集合，S_{uv} 表示用户 u 和用户 v 的相似度，R_{vi} 代表用户 v 对节目 i 的打分，$\bar{R}_v$ 表示用户 v 收看的节目的打分平均值。

采取 Top-N 推荐，即选取评分最高且目标用户没有产生过收视行为的 N 个节目推荐给用户。

3.1.2 基于电视产品的协同过滤

1.用户-节目评分矩阵

利用基于用户的协同过滤中构建的表 1 所示矩阵。

2. 计算相似度

采用皮尔森相关系数计算两个电视节目之间的相似性，公式如下：

$$sim(i,j)=\frac{\sum_{u\in U_{i,j}}(R_{ui}-\bar{R}_i)\cdot(R_{uj}-\bar{R}_j)}{\sqrt{\sum_{u\in U_{i,j}}(R_{ui}-\bar{R}_i)^2}\cdot\sqrt{\sum_{u\in U_{i,j}}(R_{uj}-\bar{R}_j)^2}}$$

U_{ij} 表示对节目 i 和节目 j 都有评分的用户集合，R_{ui} 代表用户 u 对节目 i 的打分，$\bar{R}_i$ 表示节目 i 的打分平均值，R_{uj} 表示用户 u 对节目 j 的打分，$\bar{R}_j$ 表示节目 j 的打分平均值。

3.计算节目评分

根据步骤 2 中的计算出的最相似集合来计算目标用户对目标节目的预测打分。评分公式如下：

$$\hat{R}_{ui}=\bar{R}_i+\frac{\sum_{j\in U(i,K)\cap N(u)} S_{ij}(R_{ui}-\bar{R}_j)}{\sum_{j\in U(i,K)\cap N(u)}\left|S_{ij}\right|}$$

$\hat{R}_{ui}$ 表示目标用户 u 对节目 i 的预测打分，$\bar{R}_i$ 表示节目 i 的打分平均值，$U(i,K)$ 是与目标节目 i 最相似的 K 个节目的集合，$N(u)$ 表示用户 u 打过分的节目集合，S_{ij} 表示节目 i 和节目 j 的相似度，R_{uj} 表示用户 u 对节目 j 的打分，$\bar{R}_j$ 表示节目 j 的打分平均值。

采取 Top-N 推荐，即选取评分最高且目标用户没有产生过收视行为的 N 个节目推荐给用户。

3.1.3 方案实施及测试

本小组使用 python 语言作为工具，实现以上算法，代码见附件：**产品协同过滤算法.ipynb**，读入文件见附件：**jiemu_shichang1.xlsx**。

经过程序运行，我们选取与用户相似的前 20 位相似用户推荐用户可能感兴趣的 10 个电视节目，利用十字交叉验证法。最后得到**查准率约为 5.1%，查全率约为 9.9%**。

图 1 Jupyter 运行截图

In [18]:
```
#获取测试集中的各用户推荐电视节目,取前最相似的前20个用户
test_users=test_users_items.keys()
#获取测试集中每个用户最相似的20个用户收看的电视节目，生成候选节目
candi_users_items={}
rec_users_items={}
for test_user in test_users:
    sim_users=user_similarity_norm[test_user].argsort()[::-1][:21][1:]
    candi_users_items[test_user]=[]
    rec_users_items[test_user]=[]
    for sim_user in sim_users:
        candi_users_items[test_user].extend(time_matrix[sim_user].nonzero()[0].tolist())
    #对候选项目按照出现次数进行排序，取次数统计为前10个的项目生成推荐项目
    user_item_set=set(candi_users_items[test_user])
    tmp={}
    for item in user_item_set:
        item_count=candi_users_items[test_user].count(item)
        tmp[item]=item_count
    tmp_sorted=sorted(tmp.items(),key=lambda x:x[1],reverse=True)
    for tmp_items in tmp_sorted[:10]:
        rec_users_items[test_user].append(tmp_items[0])
# 计算准确率和召回率
print('测试结果的precison为 %.4f' % precison(rec_users_items,test_users_items))
print('测试结果的recall为 %.4f' % recall(rec_users_items,test_users_items))
```

```
测试结果的precison为 0.0505
测试结果的recall为 0.0987
```

最后得到**附件：问题一结果推荐表**。结果如图 2 所示

图 2 电视节目推荐结果表

id	program	rec_rank
10001	超时空男臣	10
10001	同盟	9
10001	踩过界	8
10001	我的前半生	7
10001	赌城群英会	6
10001	射雕英雄传	5
10001	大军师司马懿之军师联盟	4
10001	蒙面唱将猜猜猜	3
10001	嫁到这世界边端	2
10001	使徒行者2	1
10013	极限挑战	10

rec_rank 代表推荐的优先级，**数值越大代表推荐力度越大**。

3.2 问题二

3.2.1 构建用户标签体系（用户画像）

标签系统是用户动态画像的核心，标签化是对用户特征的符号表示，标签化的用户画像，既方便计算机进行计算分析，又方便人们对用户画像的理解。[4]

本小组通过对附件 1~3 的分析，给出四级用户标签分类体系（详见**附件：标签体系.csv**），其中一级标签为基本特征与收视偏好。基本特征下的二级标签有收视时段、付费意愿、性别偏向、年龄偏向；收视偏好下的二级标签有频道偏好、节目偏好、直播/回播/点播偏好及工作日/周末偏好。

标签含义及计算方法

（1）收视时段

收视时段标签基于附件 1 中的“用户收视数据”，由于收视数据包含用户收看直播的时间，可以基于此计算出用户通常在什么时候收看电视。收视时段分为凌晨[02:00-06:00)、上午[06:00-11:00)、中午[11:00-13:00)、下午[13:00-17:00)、傍晚[17:00-19:00)、晚上[19:00-22:00)及深夜[22:00-02:00)。

（2）付费意愿

付费意愿标签基于附件 1 中的“用户点播数据”，由于点播数据中的节目是套餐外单独付费的节目，通过统计用户是否有此付费行为给用户打上“是”或“否”的标签。

（3）频道偏好

频道偏好基于附件 1 中的“用户收视信息”及“用户回看信息”，统计用户

收看频次最高的 3 个电视台作为其收视频道偏好标签（取 3 是因为据《中国家庭发展报告 2014》，平均每户家庭人口为 3.02，四舍五入为 3，假设每位家庭成员都有一个最常收看的电视台）

（4）节目偏好

节目偏好基于附件 1 中四张记录，并与依据附件 2 产生的产品标签产生关联。对于“用户收视信息”及“用户回看信息”，小组利用网络资源找到 142 个电视台对应的节目表，并对各时段栏目进行标签化处理，如“星光大道——综艺”，采用的标签与产品标签体系中的三级标签对应。然后将节目播放时段与收看时段一一对应，得到用户直播及回播时段收看的是何种类型节目。

对于“用户点播信息”及“用户单片点播信息”，小组利用预处理后数据对单个节目做标签化处理，然后将用户收视记录与标签对应，得到用户点播时收看的是何种类型的节目。

（5）直播/回播/点播偏好

分别对用户的直播、回播、点播时间进行统计，给予用户收视时间最长的操作标签。由于点播信息不包含节目时长，小组先利用附件 2 及网络资源将节目时长付给点播信息中的节目，再进行计算。

（6）工作日/周末标签

分别对用户工作日及周末的收视时长进行统计，给予用户平均收视时间较长的时段作为标签。

（7）性别偏向

性别偏向基于用户的节目偏好进行推测，本小组对此做出一张映射表，详见附件：标签体系。

（8）年龄偏向

年龄偏向基于用户节目偏好、收视时段、直播/回播/点播偏好、工作日/周末进行推测，本小组对此做出一张映射表，详见附件：标签体系。

汇总用户的收视时段偏好、频道偏好、周末/工作日收视偏好、直播/回播/点播收视偏好以及付费意愿偏好，得到**附件：用户收视偏好标签.csv**。如图 3 所示。

图 3 用户收视偏好标签

id	shiduan	pindao1	pindao2	pindao3	shichang	leixing	fufeiyiyuan
10001	晚上	中央1台-高清	翡翠台	东方卫视	周末	回播	否
10002	晚上	浙江卫视-高清	湖南卫视-高清	中央少儿-高清	周末	直播	否
10003	下午	中央1台-高清	安徽卫视-高清	央广购物	工作日	直播	否
10004	深夜	中央1台-高清	东方卫视-高清	广东卫视-高清	工作日	回播	是
10005	深夜	浙江卫视-高清	东方卫视-高清	湖南卫视-高清	工作日	直播	否

3.3.2 构建产品标签体系

本小组根据附件 2 中电视产品信息数据及利用爬虫得到的节目信息标签为电视产品构建了四级标签体系（详见**附件：标签体系.csv**），一级标签为基本特征，二级标签为节目类型、地区及语种。

对电视频道也构建了一个标签体系，电视频道的标签体系与电视产品的三级标签相对应。产品标签体系通过映射关系与用户标签体系之间联系起来（详见**附件：标签体系.csv**）。

图 4 映射关系（部分）

		性别偏好		年龄偏好
电视剧				
偶像		女		少儿、青年
青春		女		少儿、青年
都市		女		青年
犯罪		男		青年、中年

最终我们在电视产品信息表后为节目打上了大类、小类、适用性别及年龄段的标签，见**附件：产品数据标签.csv**。如图 5 所示。

图 5 产品数据标签（部分）

正题名	创建日期	导演	演员	出品年代	内容描述	总集数	分类名称	连续剧分类	声道语种	地区参数	大类	小类	性别	年龄段
职场是个技	########	李牧鸽	王耀庆，潘	2017	讲述了一位	30	电视剧场\	职场是个技	国语	大陆	电视剧	剧情		中年
06月28日	########	无	无	2017	06月28日	1	科学教育	自然	国语	大陆	纪录片	探索		
深海利剑（	########	赵宝刚	高昊睿，文	2017	该剧以海军	34	电视剧场\	深海利剑	国语	大陆	电视剧	偶像	女	少儿、青年
04月26日	########	无	无	2017	04月26日	1	综艺娱乐	金星秀(热	国语	大陆	综艺			
赛小花的近	########	甘露	李晟，魏千	2017	赛小花的近	42	电视剧场\	(热)赛小花	国语	大陆	电视剧	偶像	女	少儿、青年
赛小花的近	########	甘露	李晟，魏千	2017	赛小花的近	42	电视剧场\	(热)赛小花	国语	大陆	电视剧	偶像	女	少儿、青年

3.2.3 电视产品分类推荐

按照用户标签体系，可以为每个用户绘制用户画像，具有相同标签的用户可以被认为是基于标签中的一类，如节目偏好均为“电视剧”的用户可以被看作一类。

按照产品标签体系，我们可以将电视产品进行分类，对于电视剧、电影产品细化到四级类，其他到三级类。

在问题一中，我们已经得到“问题一推荐结果表”，其中我们为每一位用户做了电视产品的推荐及给出了推荐指数。在问题二中，我们利用问题一中的用户-节目评分矩阵，做用户-节目类别评分矩阵，如表 2 所示。

表 2 用户-节目类别评分矩阵

	Class 1	Class 2	……	Class m
User 1	$S_{1,1}$	$S_{1,2}$	……	$S_{1,m}$
User 2	$S_{2,1}$	$S_{2,2}$	……	$S_{2,m}$
……	……	……	……	……
User n	$S_{n,1}$	$S_{n,2}$	……	$S_{n,m}$

其中 User 即电视用户，Class 即用户收看的电视节目类别，$S_{i,j}$ 表示用户对节目类别的偏好打分。节目类别的偏好打分的值为相同类型的节目偏好打分和的平均数，即

$$S_{i,j}=\frac{\sum_{p\in C}R_{i,j}}{|p|}$$

其中 p 代表节目，C 代表类别，$|p|$ 代表节目个数，$R_{i,j}$ 表示用户对节目的打分，沿用问题一中的分数。

这样我们对每一位用户都有画像，并能做出分类产品的推荐。

在做推荐时我们对各个标签采用的是平级处理。

同样适用 python 语言作为工具实现以上算法，代码见**附件：分类协同过滤推荐.ipynb**，读入文件见附件：zhongfenleijiemu_shichang1.xlsx。

经过程序运行，我们选取与用户相似的前 20 位相似用户推荐用户可能感兴趣的 10 个电视节目，利用十字交叉验证法。最后得到**查准率约为 10.1%，查全率约为 39.8%**。

图 6 Jupyter 运行截图

```
In [44]: #获取测试集中的各用户推荐电视节目,取前最相似的前20个用户
test_users=test_users_items.keys()
#获取测试集中每个用户最相似的20个用户收看的电视节目，生成候选节目
candi_users_items={}
rec_users_items={}
for test_user in test_users:
    sim_users=user_similarity_norm[test_user].argsort()[::-1][:21][1:]
    candi_users_items[test_user]=[]
    rec_users_items[test_user]=[]
    for sim_user in sim_users:
        candi_users_items[test_user].extend(time_matrix[sim_user].nonzero()[0].tolist())
    #对候选项目按照出现次数进行排序，取次数统计为前10个的项目生成推荐项目
    user_item_set=set(candi_users_items[test_user])
    tmp={}
    for item in user_item_set:
        item_count=candi_users_items[test_user].count(item)
        tmp[item]=item_count
    tmp_sorted=sorted(tmp.items(),key=lambda x:x[1],reverse=True)
    for tmp_items in tmp_sorted[:10]:
        rec_users_items[test_user].append(tmp_items[0])
# 计算准确率和召回率
print('测试结果的precison为 %.4f' % precison(rec_users_items,test_users_items))
print('测试结果的recall为 %.4f' % recall(rec_users_items,test_users_items))
```

```
测试结果的precison为 0.1054
测试结果的recall为 0.3975
```

最后得到**附件：问题二结果推荐表.csv**，如图 7 所示。

图 7 分类推荐结果表

id	program	rec_rank
10001	电影	10
10001	电视剧	9
10001	喜剧	8
10001	动画	7
10001	英语	6
10001	普通话	5
10001	科幻	4
10001	综艺	3
10001	粤语	2
10001	纪录片	1
10013	电视剧	10
10013	古装	9

rec_rank 代表推荐的优先级，**数值越大代表推荐力度越大**。

四、电视套餐建议

分析现有套餐设置结构，并对套餐的销售品字段进行拆分。利用套餐拆分后的结果和用户画像的结果进行关联规则，发现不同用户画像与套餐内容之间的规律，对没有购买套餐的用户进行套餐推荐。

4.1 分析套餐结构

对附件 3 进行处理后的套餐结构如表 3 所示：

表 3 电视套餐结构表

套餐系列	主产品	宽带	电视包	增值产品

乐惠套餐	互动+联合宽带（月包）	6M	标准包	
		8M	49 元包	
		10M	49 元包	
		12M	49 元包	
		15M	59 元包	
		20M	59 元包	
		25M	59 元包	
		30M	59 元包	
	互动+联合宽带（年）	6M	标准包	优惠购机（200/280/380）若办理 2 年，花费金额 2160 以上送一年套餐
		8M	标准包	
		10M	标准包	
		12M	标准包	
		15M	标准包	
		20M	标准包	
		25M	标准包	
		30M	标准包	
【互动+宽带】	互动+联合宽带(年)	6M	标准包	优惠购机
		12M	标准包	
【互动+宽带】	互动+联合宽带（包月）	6M	标准包	捆绑全频道回看
		12M	标准包	捆绑快乐点
【互动+宽带】	互动+联合宽带+3G（包月）	6M	标准包	捆全频道回看
融合套餐	互动+联合宽带	6M	标准包	捆捆全频道回看；前三个月返还
		10M	标准包	
月享套餐	互动+联合宽带	10M	49 元包	
		30M	59 元包	
		50M	69 元包	
13 年底促销	互动+联合宽带	6M	标准包	捆绑 200 元增值费
		10M	标准包	
联通渠道（员工价）	互动+宽带	10M	标准包	捆绑一年时移回看

4.2 对内容相同的套餐进行合并归类

按照套餐结构对附件 3 的套餐销售品字段进行字段分割，结构如下图。

图 8 套餐销售品字段分割

用户号	业务品	状态改变时间	预存款	套餐	带宽	时间	销售品	入网时间
11183	互动电视	20160718	120	乐惠套餐	15M	包月	互动+联合宽带	20160718
10536	互动电视	20140630	0	乐惠套餐	12M	2年	互动+联合宽带	20140630
10783	互动电视	20170125	0	乐惠套餐	15M	1年	互动+联合宽带	20151230
10381	互动电视	20160123	1	融合套餐	6M	包月	互动+联合宽带	20131213
10546	互动电视	20141004	64.27	乐惠套餐	12M	1年	互动+联合宽带	20141004
11323	互动电视	20160509	60	乐惠套餐	25M	包月	互动+联合宽带	20160509
10435	互动电视	20171029	270	乐惠套餐	8M	包月	互动+联合宽带	20140118
10003	互动电视	20090228	0	乐惠套餐	15M	包月	互动+联合宽带	20070729
10191	互动电视	20130908	0	融合套餐	6M	包月	互动+联合宽带	20130908

4.3 形成用户数据表

根据第一题中得到的用户画像结果，以用户标签为唯一标识，在附件 3 中进行标签匹配；利用计算出的入网时长判断新老用户，将 3 年以上的用户标记为老用户，其余为普通用户。最终形成用户数据表,详见**附件：用户数据表**。

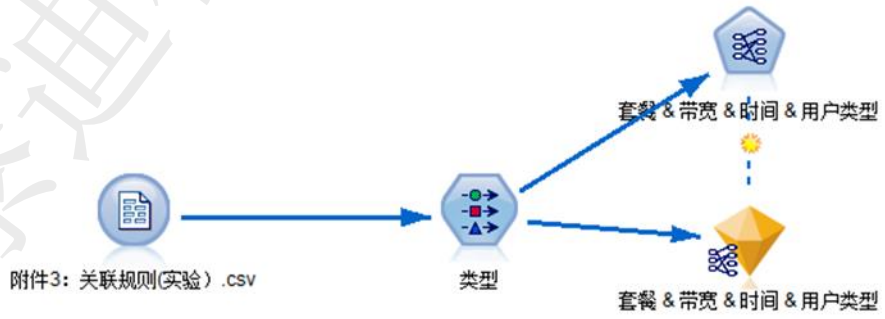
图 9 用户数据表

机顶盒编	用户类型	性别	年龄段
11274	普通用户	男	中年、老年
11273	普通用户	女	中年、老年
11272	老用户	男	中年、老年
11271	老用户		
11270	普通用户	女	少儿、青年、中年
11269	老用户		
11268	老用户		
11267	老用户	男	中年、老年
11266	普通用户	女	少儿、青年、中年
11265	老用户	男	少儿、青年、中年

4.4 实施关联规则

使用关联规则对用户数据表进行处理，计算用户基础标签和现有套餐中某一业务的支持度与置信度。该过程采用 SPSS molder 软件进行实现，软件实施模型如下。

图 10 软件实施模型



N（A），N（B）分别表示用户基础标签 A 和业务 B 在整个数据集中出现的次

数，N（AB）表示共同出现的次数。

规则支持度：共同出现的频繁程度。

$$Support(A \longrightarrow B) = \frac{N(AB)}{N}$$

规则置信度:某产品出现在另一个产品中的概率。

$$Confidence(A \longrightarrow B) = \frac{N(AB)}{N(A)}$$

对用户数据表的销售品字段和套餐进行字段分割，以”-”为分割符号进行分割。并进行标准化处理。根据得出的用户画像结果，以用户标签为唯一标识构建字典如：用户号{电视包业务，宽带，捆绑，用户标签}

假设用户套餐记录为 N

1、初始单遍扫描数据集，确定每个项的支持度。得到所有频繁 1 项集 F1；

2、使用上一次迭代发现的频繁(k-1）项集，产生新的候选 k 项集；

3、再次扫描数据集，确定候选 k 项集的支持度；

4、删除不是频繁的候选 k 项集；

5、循环 2-3 步，当没有新的频繁项集产生，则算法结束。

下表为利用 SPSS molder 软件计算得出的相关关联规则：

表 4 用户特征与购买套餐业务的关联规则

后项	前项	支持度	关联度
时间 = 包月	用户类型 = 普通用户 and 性别 = 男	32.48882265	66.97247706
套餐 = 乐惠套餐	年龄段 = 中年、老年	47.54098361	67.71159875
套餐 = 乐惠套餐	年龄段 = 中年、老年 and 性别 = 男	36.06557377	68.18181818
套餐 = 乐惠套餐	性别 = 男	59.31445604	69.09547739
时间 = 包月	年龄段 = 中年、老年 and 用户类型 = 普通用户	25.78241431	69.36416185
时间 = 包月	性别 = 男	59.31445604	72.86432161
时间 = 包月	年龄段 = 中年、老年	47.54098361	74.92163009

时间 = 包月	年龄段 = 中年、老年 and 性别 = 男	36.06557377	75.20661157
时间 = 包月	用户类型 = 老用户	44.26229508	77.44107744
时间 = 包月	用户类型 = 老用户 and 性别 = 男	26.82563338	80
时间 = 包月	带宽 = 6M and 用户类型 = 老用户	28.76304024	89.11917098
套餐 = 乐惠套餐	用户类型 = 普通用户	55.73770492	97.59358289
套餐 = 乐惠套餐	年龄段 = 中年、老年 and 用户类型 = 普通用户	25.78241431	97.68786127
套餐 = 乐惠套餐	用户类型 = 普通用户 and 性别 = 男	32.48882265	98.16513761
时间 = 包月	套餐 = 融合套餐 and 用户类型 = 老用户	27.57078987	99.45945946

从上表可以归纳出以下几方面的规则：

1.家庭构成以男性为主的家庭在购买套餐时，更偏向与采用包月的方式更喜欢购买乐惠套餐。

2.以中年老年用户为主的家庭在购买套餐时更倾向于采用包月的方式。

3.电视入网时间较短的普通用户更倾向与购买乐惠套餐。

4.电视入网时间较长的老用户更倾向与采用包月的方式购买电视套餐。

五、 总结

5.1 结论

本文首先对数据进行预处理，去除异常值，补充缺失值，利用爬虫网上抓取电视节目分类数据，并查找归纳电视台节目播放规律。

为了给用户推荐电视产品，我们利用用户点播数据，组合进行基于用户和基于产品的协同过滤，先计算用户相似度，利用皮尔逊系数计算出与某用户相邻的20个相似用户，再推荐10个用户可能感兴趣的电视产品，按兴趣度从大到小排序。利用十字交叉验证法计算得出查准率约为5.1%，查全率约为9.9%。得到**附件：问题一推荐结果表.csv**。

为了构建用户标签体系，我们先利用现有数据经过处理后为用户打上收视时段、付费意愿、频道偏好、直播/回播/点播偏好及工作日/周末标签，得到**附件：用户收视偏好标签.csv**，利用频道标签和产品与用户年龄与性别的映射（基于论

文+常识判断）为用户打上性别偏向、年龄偏向的标签，得到**附件：用户画像**； 为构建产品标签体系，我们利用附件 2 提取出产品的名称（“连续剧分类”字段），然后将其与爱奇艺上爬虫得到的电视产品数据进行比对，为产品依次打上“节目类型”标签。再利用产品标签和产品与用户年龄与性别的映射（基于论文+常识判断）为产品打上性别偏向、年龄偏向的标签，得到**附件：电视产品数据标签.csv**。另外利用“电视猫”上的电视台节目播放表得到附件 1 中直播与回播用户的收视规律，为每个时段打上节目类型的三级标签，得到**附件：各频道数据.zip**。

按照标签体系，具有相同标签的用户可以被认为是基于标签中的一类，具有相同标签的产品可以被认为是基于标签的一类。我们构建用户-产品类标签再次运用综合的协同过滤算法实现节目分类推荐，按兴趣度从小到大排序。利用十字交叉法计算得出查准率约为 10.1%，查全率约为 39.8%。得到**附件：问题二推荐结果表.csv**。

通过关联规则的实施，我们归纳得出四条规则：1.家庭构成以男性为主的家庭在购买套餐时，更偏向与采用包月的方式更喜欢购买乐惠套餐。2.以中年老年用户为主的家庭在购买套餐时更倾向于采用包月的方式。3.电视入网时间较短的普通用户更倾向与购买乐惠套餐。4.电视入网时间较长的老用户更倾向与采用包月的方式购买电视套餐。

5.2 回顾与展望

在本次实验的过程中，发现了一些问题，也得到了一些启发，主要可以概括为以下几点：

（1）电视用户为家庭用户，为了实现精准营销，有必要为用户进行用户画像推测家庭结构，还可以利用不同时段的家庭收视记录在不同的时段为用户推荐合适的电视产品。

（2）电视节目具有时效性，一些“过时”的电视节目对于用户的吸引力实际上有所下降，而我们在推荐的过程中没有将这种差别考虑在内。

（3）在数据处理的过程中没有完全考虑用户的有效收视，虽然简单剔除机顶盒未关情形，但对于时长较短的收视记录均默认为有效的。

（4）基于用户的协同过滤与基于电视产品的协同过滤的效果在大量数据支撑的情况下会得到较好的效果，在数据量较少的情况下效果不太理想。可以考虑使用基于内容的协同过滤，但由于时间有限并未实现这一算法。

本小组能力有限，在实践有限的情况下并没有能实现所有的想法。协同过滤的实现和网络信息的查找与爬取并没有想象中那么简单。通过老师、学长指导和看视频、找网页学习这些技术，能力得到了锻炼。

数据挖掘是一样很有意思的工作，思路和方法很重要。小组想到了很多思路，但是没有能力去实现思路，这让我们深刻认识到自身的局限性，感受到需要学习的东西还有很多。值得庆幸的是，小组的氛围一直很好，即使感受到困难做不下去，还是一直坚持着，团队的分工协作与互帮互助使我们完成了这一问题。团结合作与坚持使成功达成目标成为可能。

致谢

在这篇报告完成之际，首先感谢我队导师全程的指导、帮助与监督，导师认真负责的精神感染我队每一位成员。

其次感谢大赛组委会给我们一次锻炼的机会，给予我们创新与展示的平台。

最后感谢小组的每一位成员对此次比赛的付出，以及成员的亲友对成员的支持，我们将继续努力。

Acknowledgement

At the point of finishing this report, firstly we would like to express our sincere thanks to our tutor. Thanks for his guidance, help and supervision. His being responsible affected every one of us.

Secondly we would like to thank the organizing committee of this competition. Thanks for giving us the chance to practice, to show ourselves.

Lastly, thanks for every team member’s effort as well as our backup friends and relatives. We will still try our best.

参考文献

[1]孙光浩,刘丹青,李梦云.个性化推荐算法综述[J].软件,2017,38(07):70-78.

[2]赵培. 面向家庭用户的电视节目动态推荐方法研究[D].合肥工业大学,2017.

[3]喻玲. 面向家庭用户的互联网电视资源推荐模型研究[D].华中师范大学,2015.

[4]王冬羽. 基于移动互联网行为分析的用户画像系统设计[D].成都理工大学,2017.

[5]余远洁.基于大数据技术的广电用户收视行为建模[J].新媒体研究,2017,3(11):65-66.

[6]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].中国新通信,2016,18(11):139-141.

[7]丁伟,王题,刘新海,韩涵.基于大数据技术的手机用户画像与征信研究[J].邮电设计技术,2016(03):64-69.

[8]沈菲,陆晔,王天娇,张志安.新媒介环境下的中国受众分类:基于 2010 全国受众调查的实证研究[J].新闻大学,2014(03):100-107.

[9]顾阳.南京云媒体电视用户的节目精确营销[J].市场周刊(理论研究),2014(01):73-74.

[10]冯哲辉,娄阔峰.当代青少年收视行为分析[J].当代电视,2010(10):21-24.

[11]蒋力.老年心理与收视行为特征简析[J].当代电视,2004(03):46-48.

[12]姜明求,申慧善,吴昶学,杨秀英,白雯英.中国电视观众的电视剧消费口味[J].全球传媒学刊,2015,2(02):14-38.

[13] Naemura, Masahide；Takahashi, Masaki; Clippingdale, Simon; Yamanouchi, Yuko; Fujisawa, Hiroshi.Constructing personalized user profiles through TV viewing.[C]IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB, v 2016-July, July 25, 2016.

[14] Kuan-Chung Chen；Wei-Guang Teng .Adopting user profiles and behavior patterns in a Web-TV recommendation system[C]Digest of Technical Papers - IEEE International Conference on Consumer Electronics, p 320-324, 2009.

[15] Chang, Ray M.; Kauffman, Robert J.; Son, Insoo. Consumer micro-behavior and TV viewership patterns: Data analytics for the two-way set-top box[J]ACM International Conference Proceeding Series, p 272-273, 2012.