电视大数据的用户收视行为分析系统设计与实现

张冠巍

院 （系）：计算机科学与技术学院　　专　　业：计算机科学与技术

学　　号：1130310616　　指导教师：高宏

2017 年 **6** 月

哈爾濱工業大學

毕业设计（论文）

题　　目 电视大数据的用户收视行为分析系统设计与实现

专　　业 计算机科学与技术

学　　号 1130310616

学　　生 张冠巍

指 导 教 师 高　宏

答 辩 日 期 2017 年 6 月 25 日

摘　要

随着数字电视的普及，多样化的网络应用服务极大地丰富了用户的视听体验，用户在收看电视时有了更多节目选择，并可以享受更为丰富的个性化服务，用户的收视行为也变得更加复杂。作为传统的数据集中行业，电视媒体拥有着基数庞大的用户群体，几乎每时每刻都有庞大的数据产生，如同一座产量巨大的矿藏等待着挖掘。如何利用这些用户的收视行为数据，实现创新与发展，是电视媒体提高自身竞争力、寻求新突破的关键。

本文研究了对实际用户收视行为的分析方法，以江苏广电提供的 2016 年 5 月的用户原始数据为研究对象，分别以频道/节目为中心和以用户为中心这两个角度入手，构建了类别知识库，提出了二次分类方法对频道和节目划分为不同类型；统计不同条件下的收视情况，观察各个频道、节目的收视规律，并做出一些猜想和结论；进行了热点分析，研究各时段热点频道、热点节目、热点词的变化情况；绘制了用户行为模式图，分析并总结了几种常见的用户行为模式，据此提出了识别用户行为和分析用户偏好的算法，并设计实验验证了算法的效率。

关键词：收视行为；节目分类；行为模式生成；行为识别；偏好分析；

Abstract

With the popularity of digital TV, a variety of network application services greatly enriched the user experience. There are more programs to choose and richer personalized services for users, and the behaviors of users are becoming more complicated. As a traditional data-centered industry, television media has a large user base. There is a huge amount of data genenrated nearly every moment, which is just like a great treasure waiting to be unearthed. How to utilize the data of users' behavior, to achieve innovation and development, is the key for TV media and corporations to improve their competitiveness and seek for new breakthoughs.

This paper studies the methods to analyze actual TV viewers' behavior. Basing on the original viewing data of May 2016, provided by Jiangsu Broadcasting, it starts with two aspect: conserning on channel/program and conserning on users. And then the paper puts forward to establish a knowledge base for classification, and proposes a method of two-lyer classification to classify channels and programs. The paper analyzes the ratings in different condition, to observe regular patterns of viewing each channel and program and draw some conclusions and conjectures, and do the hot topic analysis, to study hotspots of channels, programs, and keywords in different period. It also generates the user's behavior model, and conclude several most common behaviors. And based on these work, it proposes the method to recognize user's behavior and analyze user's preference. There are also some experiment designed and conducted to test the efficiency.

Keywords: Viewing Behavior, Program Classification, Behavior Model Generation, Behavior Recognization, Preference Analysis

目　录

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

近年来，数字电视得到了极大得普及。根据 Digital TV Research 于 2016 年针对于全球电视用户进行的一项调查报告显示[1]，全球数字电视的普及率由 2010 年的 40.4%猛增至 2015 年末的 74.6%，报告还指出，至 2021 年，全球数字电视的普及率将会达到 98.3%。与此同时，随着技术的发展，网络电视、智能电视也逐渐走近千家万户，多样化的网络应用服务极大地丰富了用户的视听体验。然而相比于互联网、电信网的迅猛发展，越来越多的用户转而倾向于使用电脑或移动设备收看节目，电视对于用户的吸引力逐渐降低，导致大量的用户流失，在三网融合的趋势下，广电行业面临着前所未有的机遇与挑战，如何更好地满足用户需求、提高自身的竞争力，成为其不得不面对的一个重要问题。

1.1.1 课题来源

本课题使用由江苏广电提供的数据，通过数据挖掘技术，筛选有价值的信息，分析隐藏在数据背后的用户收视行为，从而推测用户喜好，以便于更加精准的节目内容投递以及广告投放。

1.1.2 电视媒体数据分析的意义

适逢大数据时代的到来，数据处理技术和能力上的质的飞跃给人类社会带来了革命性的变化。作为传统的数据集中行业，电视媒体拥有着基数庞大的用户群体，几乎每时每刻都有庞大的数据产生，如同一座产量巨大的矿藏等待着挖掘。面临着更为严峻的挑战和前所未有的发展机遇，如何妥善利用自身数据集中优势，借助数据分析技术，抓住历史性的机遇实现创新发展，成为电视媒体提高自身竞争力、寻求新突破的关键。

1.1.3 用户行为的研究意义

一般来说，用户行为可以看作由几个元素构成：时间、地点、任务、交互以及交互内容等。对用户行为进行分析，要将其定义为各种事件：如哪一个用户、在什么时间、发生了什么事件、事件内容是什么等等。有了这样的时间以后，就可以把

用户行为连起来观察，构成用户行为的轮廓模型，方便预测用户未来可能发出的事件动作。[2]

以购物网站为例，当网站运营者掌握了足够的用户行为数据，根据不同的需求，将数据按时间、用户、事件等不同的级别进行分类、整理、建模和分析，由此将其应用于网站运营策略的制定中，应用场景包括但不限于[2]：

（1） **拉新** 策划制造足够有吸引力的因素以获取新用户；

（2） **转化** 提高订单转化率，访客转化为常驻用户进而转化为消费用户；

（3） **促活** 刺激用户活跃程度，使用户频繁使用自家产品；

（4） **留存** 提前发现可能流失的用户，采取补救措施降低流失率。

（5） **变现** 发现高价值用户，提高销售效率

总之，有效地利用用户行为分析结果，可以帮助运营商制定合适的运营策略，吸引、留存并转化用户，增加用户粘性和用户活跃度，从而提高自身产品的竞争力。

而对于电视媒体而言，用户收视行为的分析不但可以帮助运营商把握整体形势，对用户流动以及用户的整体需求等有一个全局上的认识，同时也有助于运营商和广告商了解每个个体用户的收视习惯，投其所好地提供个性化服务。例如，在奥运会期间，同一时段可能有不同项目的比赛同时进行，需要不同的频道进行转播，运营商便可以根据以往用户群体的对不同体育项目的关注程度进行安排，热门项目安排在主要频道重点转播，而冷门项目安排在次要频道转播甚至不转播；而对于某一个体用户而言，运营商可以根据其以往的喜好，向其优先推送他最有可能想要收看的项目转播，并根据用户最终选择收看的项目对原有的行为模型进行调整。

1.2 国内外研究现状及分析

本课题以江苏广电提供的电视数据作为研究对象，虽然目前对于广电数据挖掘的系统的、有针对性的研究并不是很多，但仍可以其他领域在诸如偏好挖掘、推荐系统等方面的研究状况进行参考。

1.2.1 大数据时代下电视媒体发展方向的探索

刘飞等人重点阐述了数据挖掘技术在广电行业可应用的领域，并根据在实际应用过程中遇到的困难，总结了数据挖掘在广电行业应用所遇到的障碍[3]。他们认为，结合广电行业数据特点，可将数据挖掘技术应用于客户细分、客户流失分析和动态预警分析、客户维系、客户欠费分析和动态防欺诈分析、市场发展分析等，并指出当前广电行业中存在的不同地区数据不共享、数据质量和完备性、相应人员素

质、应用周期与成本投入等阻碍。而 Spangler 等人也指出，面向观众的营销中要求广播公司理解用户需要的、想要的以及其行为，以使观众满意、使自身获利，并由此开发了一个广告投递系统（ADS）[4]。李忠哗等人指出了引入数据挖掘技术之后的有线电视管理系统的体系结构的三个部分，即原始数据、数据挖掘应用服务器和用户端，并以 FP-Growth 算法为例，进行关联规则分析为，以寻找所有的频繁项集，并在此基础上生成强关联规则[5]。刘峰则选定了网络整合营销的 4I 原则作为研究视角，分别从 Interesting（趣味原则）、Interests（利益原则）、Interaction（互动原则）、Individuality（个性原则）四个方面对大数据时代的电视媒体营销创新做出而来分析，并在此基础上，从电视媒体的传统营销方式、新媒体整合营销体系以及“大数据营销”三个层面探讨了如何构建大数据时代电视媒体的整合营销体系[6]。

通过对这些文献的总结，可以看出，将数据挖掘技术应用于电视媒体数据，主要可以从两方面入手：

一方面是从宏观上来看，运营商可以通过研究用户群体的行为变化进行系统的改进，如根据用户收视率、用户实时在线趋势等调整节目档期、改变广告结构，也可以以此作为客户流失分析、动态预警分析的基础；而另一方面，从微观的用户个体来说，通过研究个体用户的收视习惯、兴趣爱好，可以便于运营商提供个性化、私人定制化服务，以此极大地提高用户体验，增强自身的竞争力。

虽然上述研究都阐述了大数据时代下电视媒体的发展方向，但仍有些许不足。刘飞等人用业界的眼光展开讨论，但并没有提出具体的可执行方法；Spangler 主要关注了对于用户个体的研究，而没有把用户看成一个宏观整体；李忠哗等人虽提出了系统体系结构，但并未在技术层面上予以实现；刘峰的则更多的倾向于从营销学的角度进行讨论，而未涉及更多技术问题。

1.2.2 用户行为、偏好相关算法研究

何速研究了社会电视中的用户行为[7]。对网络电视用户行为分析则主要包括了用户在线时长分布、基于在线时长的用户分类、分类用户在线时长演化、用户地理分布可视化和用户到达率等。并以用户收看湖南卫视的数据为基础，以用户在线时长描述用户观看湖南卫视的持续时间，将用户按忠诚度分为轻度、中度、重度收看者三种类型，并分别对其进行了用户在线时长的演化分析，反映了 24 小时内观看湖南卫视用户在各个时间点的数量分布。

Spangler 等针对 PVR(个人视频录像，即时移功能)的出现导致视频流中插播广告收看率降低的背景，开发了一套基于数据挖掘的观众分析系统——广告投递系

统 ADS，根据用户的收视特点来判断其人口学特征和心理学特征，向用户投放与其兴趣相关的广告以增加广告的收看率。该系统以用户收看节目类型、每个节目的收看频数、收看时间点以及收看时长等因素作为用户收视特点的判断依据，将用户所对应的子集反馈给广告商[4][8-10]。

Chanza 辩证地回顾了用于在大型数据库中发现有意义的信息的数据挖掘技术，想要制定一种适合分析二元电视收视数据的聚类分析方法。研究使用了南非广电公司的收视数据，比较了分裂聚类与层次聚类两种方法，还检验了二元数据聚类中应用的距离度量法，着重考量了确定最合适的聚类划分数目的方法，并最终基于聚类分析结果，确定了四种不同的用户轮廓，使得南非广电公司能够提供靶向观众的节目编排[9-12]。

Holland 等人提出了一种较为新颖的、基于严格偏序偏好的用户日志数据偏好挖掘技术。设计了集中算法用于探测分类的、数值化的、复杂的偏好，其原型实现在数据库服务器上执行所有的数据敏感操作，并展现了不错的效率。其主要优点在于偏好挖掘结果的语义表达。系统的实现使用了高级 SQL 语句在数据库层面执行所有数据敏感操作，在大规模的日志数据集仍能运行良好[13-16]。

Levene 等人则提出了一个数据挖掘模型，用来捕获用户导航行为的特征，该导航会话被建模为一个超文本概率文法，文法中更高概率的字符串对应于用户更加倾向的踪迹。根据 Ngram 模型假定最后仅 N 个被浏览的网页对下一个可能访问的页面的概率产生影响。此外还提出了使用熵作为评价语法的统计学特征的标准。实验结果表明算法运行时间是线性的，语法的熵值可以很好地评估被挖掘的踪迹的数量，且现实数据规则证明了模型的有效性[17-20]。

Smyth 等人描述讨论了了一个一完全部署的内容个性化系统，ClixSmart，采用学习过的用户轮廓将合适的内容靶向发送给个体端用户。其内容个性化引擎主要负责两方面工作：轮廓管理器负责监控在线活动并自动构建用户轮廓以捕捉他们的行为偏好；而个性化管理器则采用了两种不同的内容筛选策略向用户推荐条目：一个基于内容的筛选方法以向用户推荐其链接过的条目的相似条目，以及一个与之相对的协作推荐方法以向某给定的用户推荐与其相似的用户也喜欢的条目[21-23]。

上述研究中所提出的系统、算法与模型，何速、Spangler 分别以收视时间、PVR 功能的使用作为用户分类的主要维度，而 Chanza 使用聚类的方法来挖掘有意义的信息，但对于数据的处理仍是二元的，三者都缺少了对于复杂情况下不同因素的综合影响的考虑；Holland、Borges、Smyth 等人分别在挖掘用户行为的方法上做出了不同尝试，虽然其研究内容并不针对电视媒体数据，但依然能给本课题提供不错的

借鉴。

1.3 本文的主要研究内容

本课题拟设计一个基于电视大数据的用户收视行为分析系统，根据江苏广电提供的数据，对原数据进行适当整理后选取合适的维度进行分类、分析与总结，并最终构建用户行为模型、确定用户偏好以便电视台制定推荐策略。

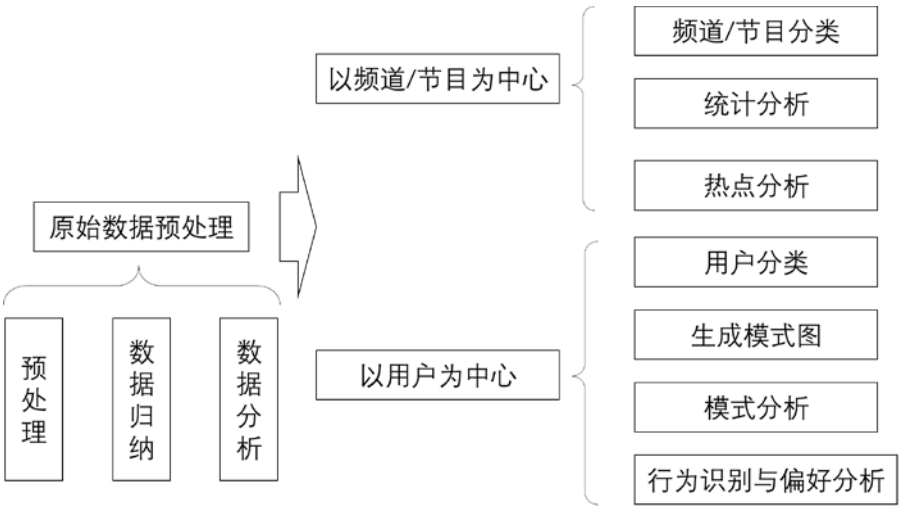


图 1-1 主要研究内容

要制定分析策略，首先要选择合适的分析维度，如时间、用户、节目类型、事件类型等等，这些分析维度主要可以归纳为两大类：

一种是以频道或节目类型为中心，对于指定的频道或节目，在宏观上分析其收视率如何，不同时间的热度如何，不同频道与节目间有无相互影响，不同的节目如何归类等等。

另一种是以用户为中心，围绕个体用户触发的不同事件构建用户行为轮廓，如用户倾向于收看电视的时间、行为习惯、经常观看的节目类型等等。

因此，本课题的研究内容主要包括如图 1-1 所示的以下几个方面：

（1） 原始数据的分析与预处理；

（2） 以频道/节目为中心的收视行为分析；

（3） 以用户为中心的收视行为分析

1.3.1 原始数据的分析与预处理

数据处理可分为三个预处理、数据归纳和初步结果分析三个阶段。

（1） 预处理阶段 主要进行数据格式整理、数据除杂、异常值去除，并进

行一些简单的统计、分类和排序工作。

（2）数据归纳阶段 主要进行一些简单的挖掘工作，分析单个用户收视类型、归纳不同的节目类型、筛选可以体现用户偏好的条目等

（3）初步分析阶段 对前两阶段的处理结果进行总结分析，并最终生成便于下一步制定分类策略的数据格式。

1.3.2 以频道/节目为中心的收视行为分析

以频道和节目为研究对象，对数据进行分类、分析和总结，主要内容包括：

（1） 频道/节目分类 对已知频道和节目进行不同层次的分类，以构建知识库，用于对给定的目标频道和名称进行分类处理；

（2） 统计分析 一系列的统计分析工作，如各频道的收视规律、不同节目间收视的相互影响等，从中挖掘有价值的规律以供研究；

（3） 热点分析 分析不同时段的热点频道、热点节目和热点词，从以从宏观上反映用户群体的收视偏好。

1.3.3 以用户为中心的收视行为分析

将用户作为主要研究对象，对单个或一组抽样用户进行研究，主要研究其收视行为及偏好，以便电视台可以制定个性化推送服务。

（1） 用户分类 将数据条目按照 CA 卡号进行归类，得到用户在某一时段的时序事件序列，方便抽样，行为模式分析正式基于此特征展开；

（2） 生成模式图 将用户的行为模式抽象为一个有向图，对于用户的连续动作，以事件 ID 为节点，相邻事件的先后顺序作为有向边，算法构建状态转移图，删去出现次数过少的边，以贴合用户的主要行为模式图

（3） 模式分析 结合模式图与具体数据流，通过观察与分析，得出如下几种最常见的行为模式，如浏览行为、时移行为、VOD 点播行为、中间件行为等；

（4） 行为识别与偏好分析 根据得到行为模式，算法构建自动机模型进行行为识别，并分析用户偏好的节目类别。

第 2 章 原始数据分析与预处理

2.1 引言

本章主要介绍了对原始数据的初步处理与分析过程。

本文使用了由江苏广电提供的数据作为研究对象，原始数据共 440GB，文本文件形式存储，数据说明补全、数据质量较差，因此需要对数据进行初步处理与分析，以了解数据中各个字段的意义，并筛选、清理出潜在的有意义的条目为之后的研究工作做准备。

本章结构如下：2.2 节分析了单条目中个字段的意义并筛选出研究中可能需要的关键字段；2.3 节分析了各个事件的具体意义并刷选出有潜在研究意义的候选事件；2.4 节详细介绍了数据预处理的过程和内容；2.5 节给出了事件与对应 Event ID 对照表；2.6 节为本章小结。

2.2 字段分析

每个数据条目均由若干不同字段组成，其中包括固定字段和可选字段，可选字段的组成由固定字段中的指定字段确定。

2.2.1 固定字段

对于某一条目，其固定字段包括：

（1） Message ID 相当于事件的序号，用于标识同一用户在同一时间段内连续若干条目的先后顺序；

（2） Event ID 事件 ID，用于区分不同事件，条目中的可选字段的组成由该字段唯一确定；

（3） 随机序列 对于同一个用户，只有同一时间段内的一系列连续事件才拥有相同随机序列；

（4） CA 卡号 区分用户身份的唯一标识，同一时间段内，某一用户所产生的条目可能会出现在 10 个服务器中的任意一个，但其 CA 卡号相同；

（5） 序列号 通常情况下有 CA 卡号确定

（6） 服务器时间 服务器接收条目的时间，由于延迟等缘故，并不能代表时间顺序。

2.2.2 可选字段

可选字段由固定字段中的 Event ID 唯一确定，视事件不同而包含了不同的属性字段，以下仅列出一些关键的、值得重点关注的字段：

（1） 结束时间 单个事件发生（或者说结束）的时间；

（2） 开始时间 在有两个或以上事件所组成的连续动作中，动作发生时的开始时间，如对于频道退出事件，其开始时间字段即为频道进入事件的发生时间；

（3） 频道名称 有关频道的事件中标识频道名称

（4） 节目名称 有关节目的事件中标识节目名称

2.3 事件分析

共有 35 种可能出现的事件，用于描述包括电视机状态设置、节目播放、用户个性化等不同功能操作的发生和结束，其中，与研究相关的几个主要事件如下：

（1） 21|进入频道事件、5|退出频道事件 进入/退出频道，包括频道名称、节目名称、时间等关键信息，通常情况下两事件是对应关系；

（2） 96|VOD 点播事件 Video On Demand，视频点播，跟时移不同的是并不是录制，而是直接点播，比如电影，播放过程中将不断重复该事件直到结束；

（3） 97|时移节目播放（云 2） 播放录制的时移节目，播放过程中将不断重复该事件直到结束；

（4） 6|EPG 显示事件 在显示的 EGP（电子节目指南）上查看节目信息；

2.4 事件 ID 对照表

表 2-1 给出了数据集中出现的所有事件与其在数据条目中的 Event ID 的对应关系。

2.5 数据预处理

根据后续研究对数据的要求，需要以不同的方法将数据处理为不同的格式，此外，由于原始数据的数据质量较差，存在大量丢失、错误数据，在预处理过程中应当予以清除。

2.5.1 简单的数据清洗

针对原始数据集中出现的质量问题，在预处理过程中，采取相对简单有效的方式进行处理，以如下几种常见数据质量问题为例：

表 2-1. 事件 ID 对照表

ID	事件	ID	事件
1	开机事件	21	频道进入事件
2	待机事件	23	喜爱键
3	搜台结束事件	24	功能键按键事件
4	单频道信号质量事件	25	零频道进入事件
5	频道退出事件	26	链接地址跳转事件
6	EPG 显示事件	27	VOD 点播故障
7	音量调节	28	双向页面抛送页面跳转事件
8	菜单事件	29	网络连接故障
9	Portal 进出事件	33	移动外设内容播放
10	二级链接地址事件	34	截屏事件
13	数据广播事件	35	录制事件
14	如加广告事件	37	语音事件
15	机顶盒单向故障	38	XMPP 消息接受事件
16	VOD 点播事件	39	移动外设内容播放（云 2）
17	时移事件	41	移动设备或机顶盒接入 wifi 事件
18	USB 事件	96	VOD 点播事件（云 2）
19	中间件事件	97	时移节目播放（云 2）
20	心跳事件		

（1） 编码错误　　数据中的一些条目，由于丢包、损坏等原因无法解码，在处理过程中可直接跳过；

（2） 服务器数据丢失　　由于服务器问题导致服务器中的数据全部丢失，在抽样过程中，如 5 月 1 日丢失了两组服务器数据、5 月 29 日 10 组服务器数据全部丢失。对于此类问题，在抽样过程中，应尽量避开对此类数据的抽样；

（3） 字段的文字错误　　字段中存在可修改的文字错误，则对其进行修改，例如，在某些条目中，频道名称中出现的 CCTV=1，即可直接修改为 CCTV-1；

（4） 字段的约束错误　　字段中存在约束错误，例如，在某些条目中，结束时间早于开始时间、时间超出数据集范围（2016 年 5 月）等，对此类数据，同样采取删除或跳过的措施；

（5） 字段部分内容丢失 字段中部分内容丢失，在不影响实验分析的情况下，可忽略丢失情况，正常分析。例如，在一些条目中，节目名称由于过长，存在丢包和被截断的情况，但是，在分类中，仍然可以使用剩下的部分进行分词分析，所以可直接当作正常数据进行处理。

2.5.2 以频道/节目为中心的数据预处理

（1） 频道/节目提取与分类 提取数据条目中出现的频道或节目名称，为一次分类和二次分类提供源数据；

（2） 归类数据集 按每日、每小时为单位对数据进行归类，并按照频道名称分类存储，以供不同时段、不同频道的流量统计与分析使用；

（3） 频道/节目/关键词频次统计 各时段的频道、节目、关键词的频次统计，为热点分析和词云展示做准备。

2.5.3 以用户为中心的数据预处理

（1） 用户提取 提取用户列表，并统计各个用户的事件触发频次，以便用户抽样；

（2） 抽样用户分类 抽取若干用户，以每日、每小时为单位对数据条目进行归类，按照用户进行分类存储，并以 Message ID 排序，得到用户各个时段的时序事件序列；

2.6 本章小结

本部分主要工作内容为对原始数据的分析与预处理。理清数据条目所代表的实际意义，以确定主要研究对象和研究方法；筛选出可供研究的事件和字段，进行重点研究；对丢失、损坏的数据进行处理，使其不至于影响后续研究；归类整理数据条目，以满足算法设计，提高程序效率。总之，本章的主要工作就是为后文的研究作好准备。

第 3 章 以频道/节目为中心的收视行为分析

3.1 引言

本章以频道/节目为中心，进行了一系列的统计与分析，并由此得出一些结论。

本章的结构如下：3.2 节描述了对频道/节目类别的划分工作；3.3 节分别介绍和描述了围绕频道/节目所进行的统计分析并给出相应结论和猜想，以及对猜想的验证；3.4 节描述了对指定日期的频道与节目的热点分析；3.5 节为本章小结。

3.2 频道/节目分类

本节主要对数据中出现的频道与节目进行了分类，主要用于构建一个频道-节目-类型的知识库，以方便之后对频道和节目流量及受欢迎程度的分析、用户行为偏好的分析等。分类工作初步将频道分为综合、财经、综艺等 20 多种类别，对于某些频道类别，如电影、音乐等，频道播放的内容比较单一，可直接作为节目类别对待，而对于另一些频道，如卫视频道、综合频道等，后期需要按照节目的不同进行二次分类。由于频道名称、节目名称与其类别并无严格的规律可寻，因此只能主要由手工构建知识库，由于数据庞大，本文的二次分类只选取了一部分样本进行，如 CCTV-1、北京卫视、江苏卫视等。在一些企业，此类工作一般由专门的数据员完成。

3.2.1 一次分类：按频道分类

遍历收集原始数据中所有出现的频道名称，并将其分为 28 类，分类结果如附录 2 所示。

在分类中，类别划分及类别名称的确定主要以中央电视台频道分类为标准，如综合、财经、综艺等，再根据常见的频道类型进行补充，如旅游、汽车、政务等。

待分类频道中，中央台频道、卫视频道等，除普通频道外，还有与普通频道对应的高清频道，在本次分类中，将对应的高清频道与普通频道归为同一个频道，不作多个频道处理。

3.2.2 二次分类：按频道+关键词分类

如上文所提到的，对于一些频道，如综合频道、卫视频道等，其所播放的节目种类很多、综合性较强，包含了多种不同类别，因此无法仅凭频道名称来为其分类，

因此需要更为细化的、以节目为判断依据的二次分类进行确定。

但是，节目名的变化性较大，同一类节目、甚至同一节目可能会出现不同的名称，而同一名称的节目在不同频道又可能指代了完全不同的节目。例如，对于一些节目，可能会出现全名与缩写等不同形式，如“奔跑吧！兄弟”与“跑男”是同一节目的不同名称；而如“欢乐颂”这一节目名，在浙江卫视等频道指代了某一电视剧，但在音乐频道里指代的则更可能是世界名曲。又例如，对于一些电视剧，名称中可能会出现第一集、第二季、大结局等变化的字段；对于综艺节目，会根据嘉宾的不同出现不同的简介性质的名称等等。

为了解决这些问题，本文采用了频道名+关键词共同作为二次分类依据来进行分类。在构建知识库时，对某一数据条目，得到其频道名和节目名后，使用 NLP 方法对节目名进行分词切割，得到若干关键词，并与频道名分别组成元组作为键值，而节目原本的类别即作为键值所对应的类别。

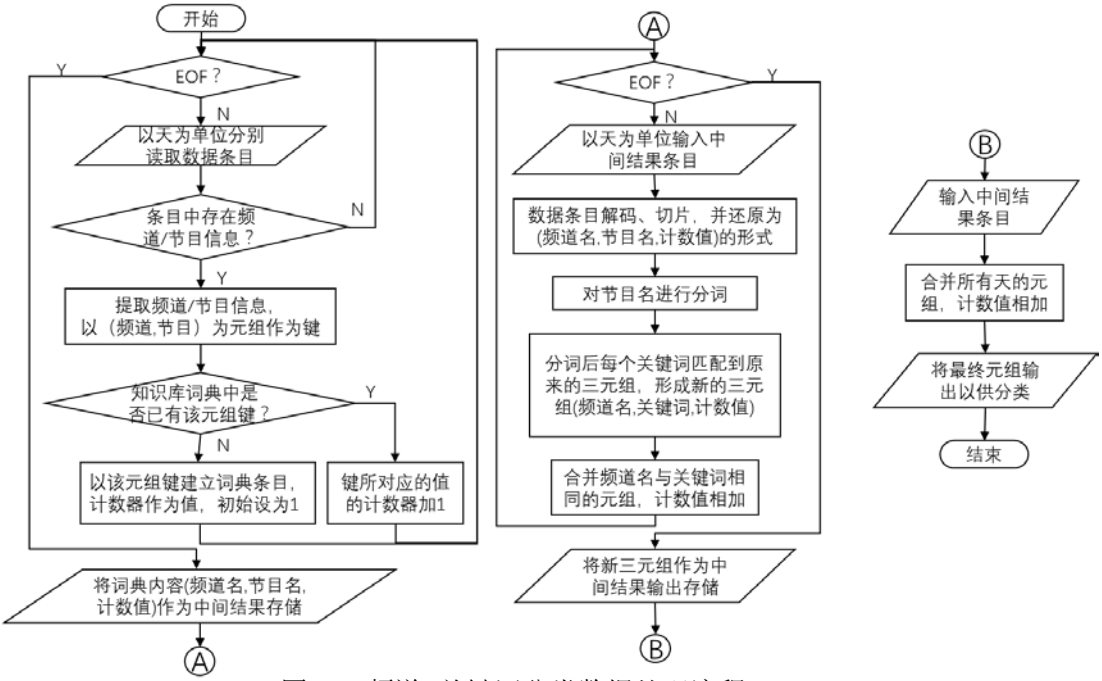


图 3-1 频道+关键词分类数据处理流程

程序处理流程如图 3-1。该流程可分为三部分：

（1） 预处理 扫描数据条目，以（频道，节目）元组为键，建立词典并保存，以便下一步的分词的进行，同时统计各元组的出现频次以供热点分析；

（2） 分词与合并 以每日为单位，枚举字典中的元素，对节目名进行分词处理，得到若干（频道，关键词，频次）三元组，将这些元组中相同的元组进行合并，频次加和，本部分可多线程实现以提高效率；

（3） 结果输出 读取并合并上一步中所有日期的三元组，得到汇总过的三

元组集合，将这些三元组按序输出，再进行分类后，即可完成知识库的建立。

表 3-2 二次分类结果（部分）

频道	关键词	频数	类别	频道	关键词	频数	类别
CCTV-1	爱国	23457	音乐	北京卫视	暗花	6293	电视剧
CCTV-1	百变	4190	少儿	北京卫视	出生入死	10685	电视剧
CCTV-1	报告会	147461	新闻	北京卫视	春妮	104855	电视剧
CCTV-1	贝宁	622015	法治	北京卫视	打狗棍	20298	电视剧
CCTV-1	拨响	19602	音乐	北京卫视	大剧	107178	电视剧
CCTV-1	不是	1130	电视剧	北京卫视	大首	77744	电影
CCTV-1	蔡明	3019	综艺	北京卫视	档案	356870	科教
CCTV-1	参考	12686	生活	北京卫视	电视	11007	电视剧
CCTV-1	茶韵	3887	纪录	北京卫视	电影节	1172	电影
CCTV-1	朝闻	2991594	新闻	北京卫视	独角兽	1792	少儿
CCTV-1	呈现	45032	纪录	北京卫视	独狼	12300	电视剧
CCTV-1	臭味	116662	纪录	北京卫视	法治	57788	法治
CCTV-1	出彩	311211	综艺	北京卫视	服务	95720	天气
CCTV-1	出发	16933	音乐	北京卫视	父亲	412357	电视剧
CCTV-1	揣着	21204	音乐	北京卫视	歌声	2720	综艺
CCTV-1	传家宝	991629	收藏	北京卫视	歌王	5190	综艺
CCTV-1	春日	10336	少儿	北京卫视	关注	554350	新闻
CCTV-1	春晚	9923	综艺	北京卫视	光阴	54251	纪录
CCTV-1	大道	1331010	综艺	北京卫视	广宣	2771	广告
CCTV-1	大国	17967	纪录	北京卫视	红星	49837	电视剧
CCTV-1	大赛	214833	综艺	北京卫视	欢乐	167506	电视剧
CCTV-1	大头	8592	少儿	北京卫视	甲级联赛	128074	体育
CCTV-1	稻之恒	1373	纪录	北京卫视	姐妹	140604	电视剧

二次分类结果如表 3-2 所示，由于篇幅限制，本文只展示了 CCTV-1 和北京卫视的部分分类结果。

3.2.3 频道+节目类别判定

将以上一次分类与二次分类的结果分别入库，即可用于对任意一个(频道，节目)二元组类别的判断。

对于给定的(频道，节目)二元组，首先查询能否由一次分类判断类别，若能，则返回相应类别；若不能，则对节目名进行分词，对于分词后的每一个关键词与频道名组成的二元组(频道，关键词)，查询是否存在该元组对应的类别，将所有关键词对应的类别列出，并计算其所占比例作为判断该元组为该类别的概率。

判断流程如下图 3-2 所示。

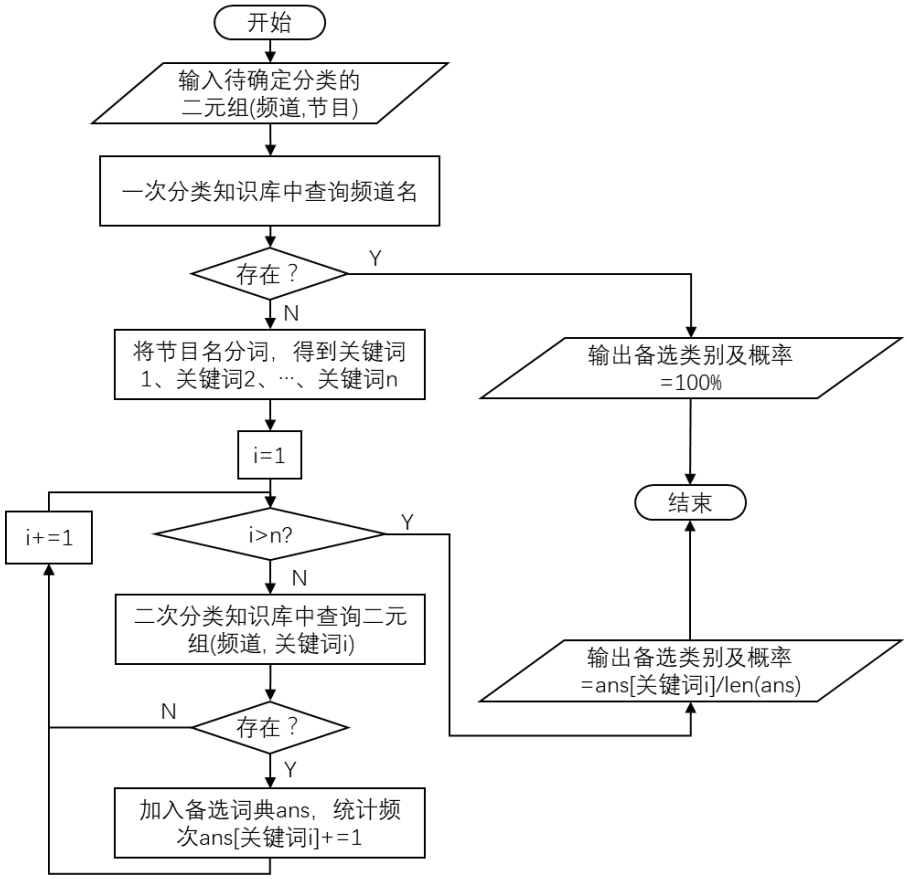


图 3-2 频道+节目类别判定

3.3 以频道/节目为中心的统计分析

本节主要介绍了围绕频道与节目所进行的一系列统计与分析工作，并描述了由此得出的结论和猜想。

3.3.1 普通、高清、杜比频道收视率对比

对于某些频道，如各央视频道、各卫视频道等，都有普通频道与高清频道之分，一些频道如 CCTV-1、CCTV-5 还开设有高清 Dolby 频道。本阶段以 CCTV-1 为样本，统计并对比 5 月 2 日-8 日这一周中普通频道、高清频道及杜比频道的收视率（流量）情况，以决定在后续统计中，应如何对待同一频道的三种不同频道，如合并、拆分亦或是直接删除。

由图 3-3 可以看出，三种频道的流量关系基本稳定，普通频道流量远高于高清频道，高清频道又远高于高清 Dolby 频道。

由图 3-4 可以看出，三种频道的单日频道几乎是数量级的差距，这样的差距放在统计中几乎不会影响统计结果和整体趋势。因此，在后续统计工作中，除非特别提及，否则选择忽略不计高清 Dolby 频道所产生的流量，而对于流量同样稀少但又明显多于 Dolby 频道的高清频道，在后续统计中会将其合并到普通频道中进行统计。

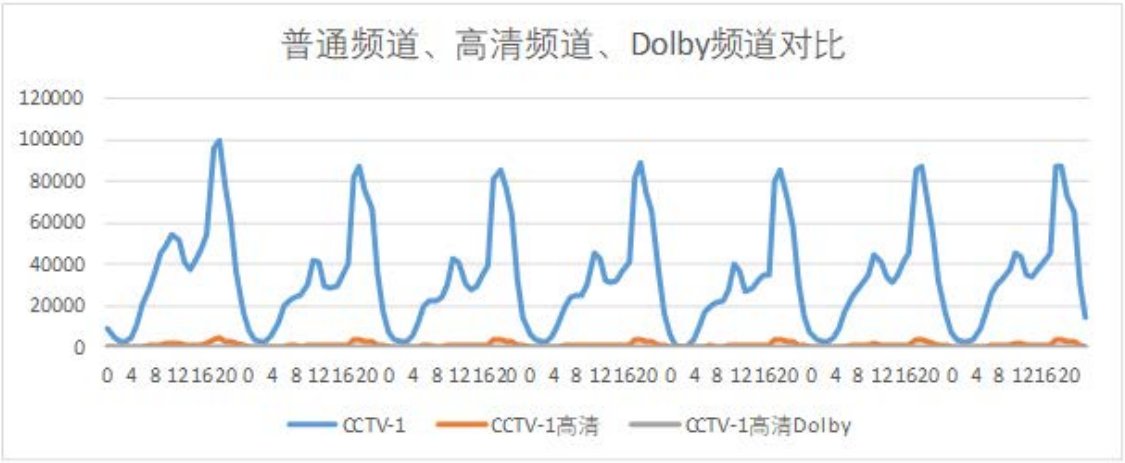


图 3-3 CCTV-1 三种频道流量对比

图 3-4 CCTV-1 三种频道单日流量

3.3.2 央视频道每小时流量分析

以 5 月 1 日-3 日为样本统计各央视频道的访问流量随小时的变化，只要相应的时段内产生了某用户对某一频道的访问，就认为该用户产生对该频道的访问流量。

由图 3-5 可以看出，每日的流量变化趋势都大体相同。且对于所有频道，都会在中午 10-12 点、晚上 17-21 点附近会出现明显的波峰，每天说明在这两个时间段内收看电视的总人数增多，结合实际情况，中午高峰段刚好对应午休时间，而晚上高峰段的上升段始于下班晚高峰、到黄金时段 19 点左右达到最高、而后随着人们陆续入睡而急剧下降。

在这些频道中，CCTV-1 的流量总体领先于其他频道，其次为 CCTV-4，在某些时段甚至超过 CCTV-1，而 CCTV-9 纪录片频道由于其频道的特殊性质，收视始终较为低迷。

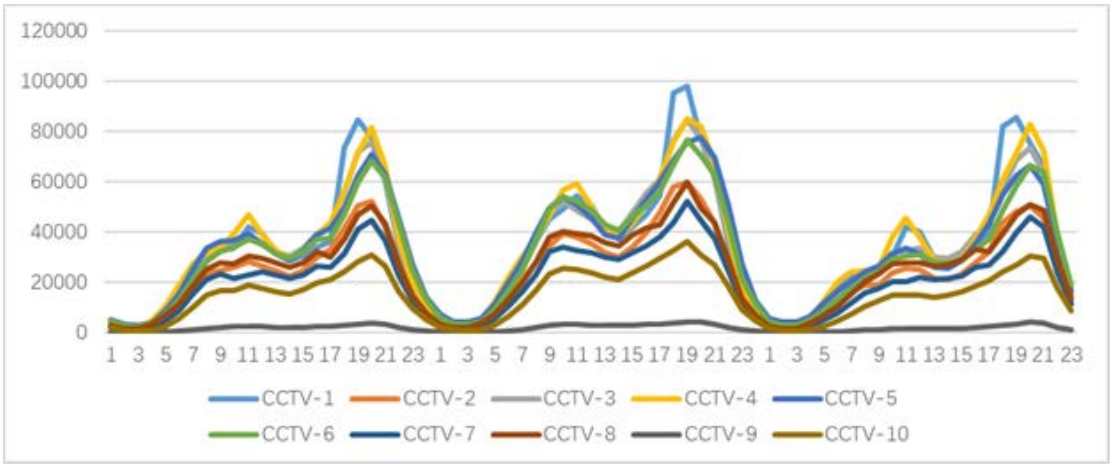


图 3-5 央视频道小时流量统计

3.3.3 央视与卫视热门频道周末至周一流量变化对比

选取 CCTV-1 及江苏卫视等几家热门卫视频道，以周五（5 月 6 日）至周一（5 月 9 日）为例，统计并绘制每日的流量变化曲线，如图 3-6 所示。

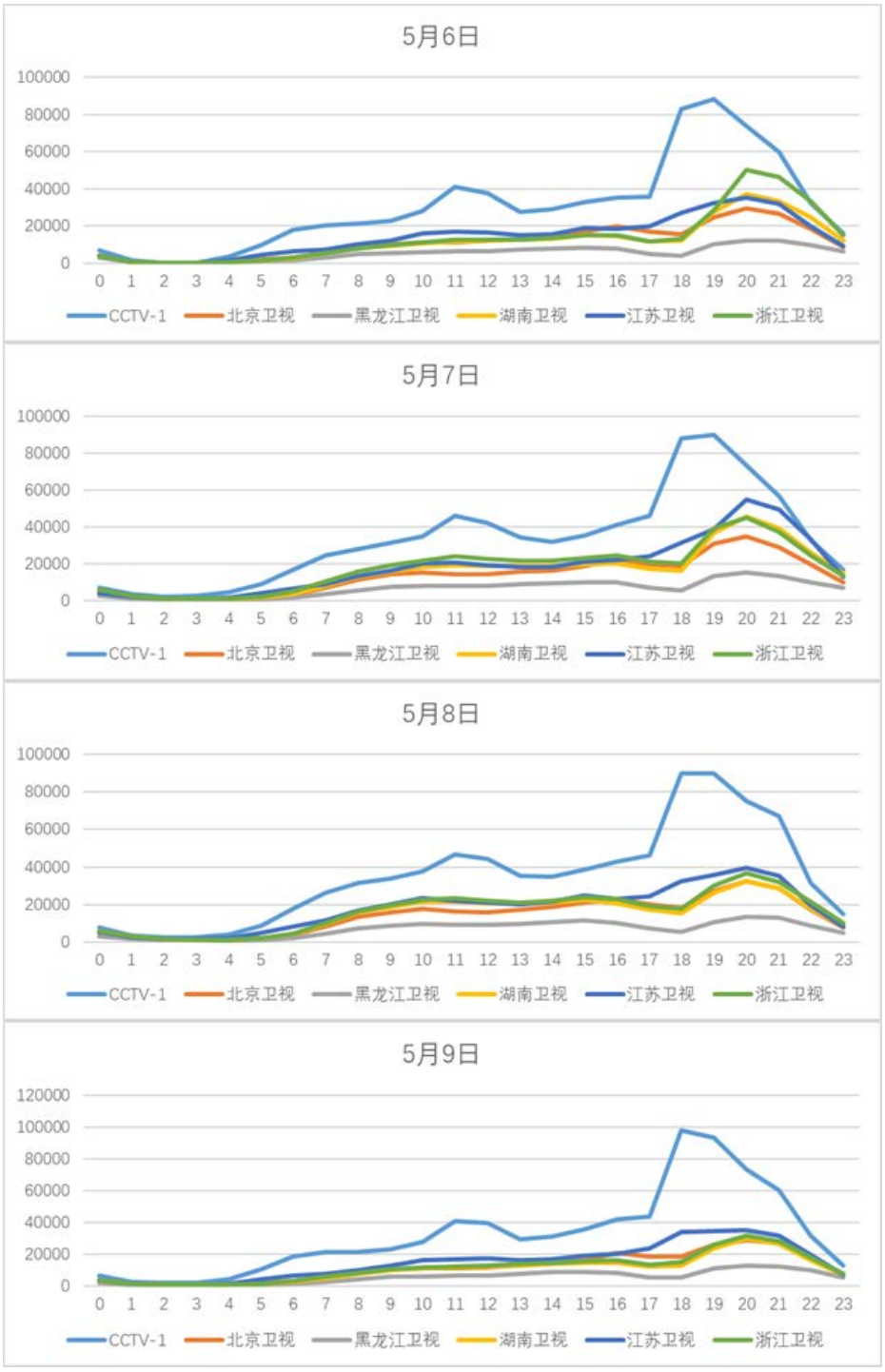


图 3-6 周末至周一每日流量变化

总体来看，CCTV-1 的流量依旧远远领先于其他卫视频道，而所有频道的流量波动随时间的变化趋势都是大体相同的，但由于卫视频道节目之间的相似性和竞争性，在同一时段内，根据各自节目的受欢迎程度，流量的波动情况也会有所变化，并有相互影响的趋势。如 5 月 6 日，浙江卫视在黄金时段 18-20 点内流量陡增，伴

随而来的是其他几家卫视流量的增长变缓；类似的还有 5 月 7 日，江苏卫视的曲线斜率在 19-20 点突然增大，说明此时流量激增，而相对应的，其他几家卫视的曲线斜率都或多或少地突然减小，说明此时江苏卫视的节目更受欢迎。

值得一提的是，从统计数据来看，周一（5 月 9 日）的黄金时段，CCTV-1 的流量最值是样本统计的几天中最大的，但其他频道并不具备这一特征，下文也将验证这一点。

3.3.4 央视与卫视热门频道连续两周流量波动对比

选取 CCTV-1、CCTV-4、湖南卫视、江苏卫视等热门频道，统计自 5 月 2 日至 15 日连续两周的各频道流量变化情况，以观察以日为单位的流量变化情况，如图 3-7 所示。

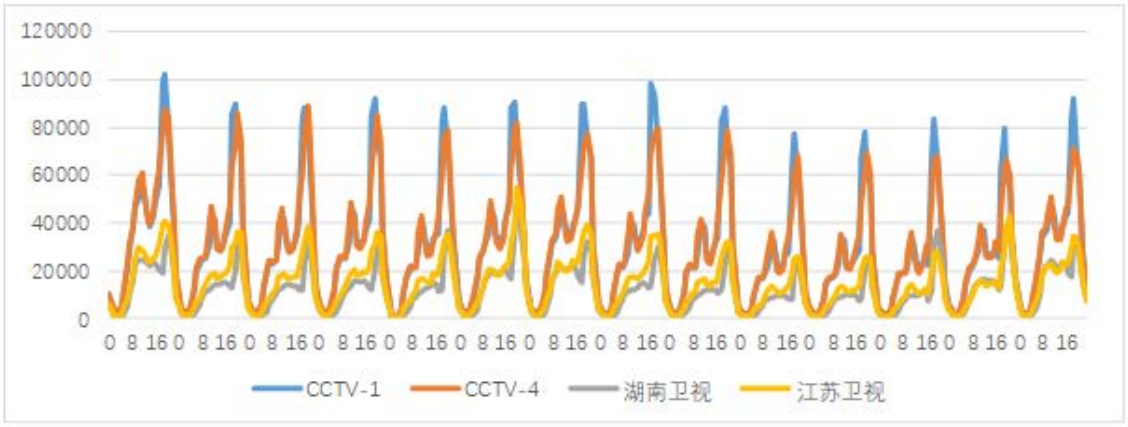


图 3-7 连续两周流量变化

如上文所说，各个频道每日的流量波动情况都是大体相似的，由图 3-7 可以进一步确认这一点。但对于不同频道，虽然每日流量峰值并不相同，但也能观察到一些潜在的规律。例如，上文图 3-6 中发现，CCTV-1 在周一黄金时段的最值会达到最高峰，此处以 CCTV-1 和江苏卫视为例，取其每日的最值点，绘制曲线。如图 3-8 所示，实线为连接各个最值点所得到的实际折线，虚线为拟合的趋势曲线，可以看出，对于 CCTV-1，其流量规律为：周一达到最值，周中减缓下降至平稳，周末反弹上升，直至下个周一达到最值。而对于江苏卫视，一样具有相似的规律，只不过最值点在周六出现。

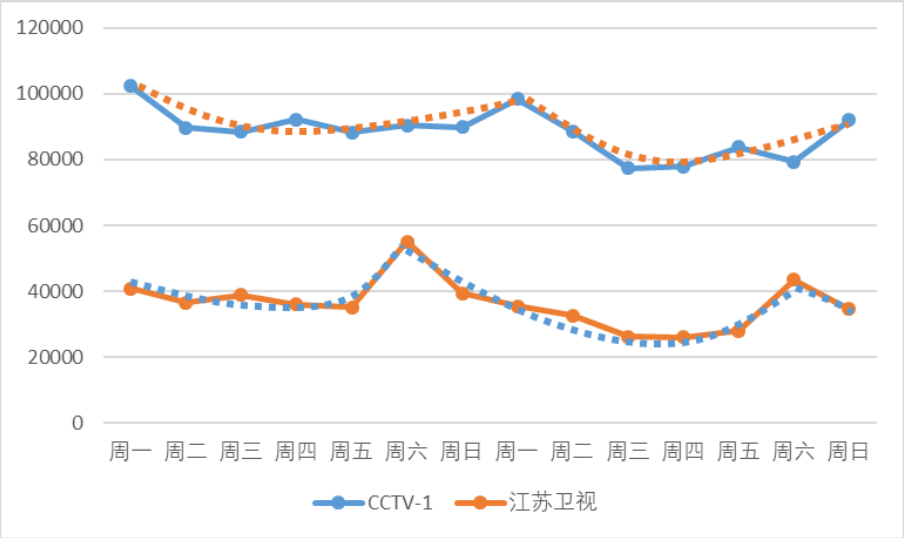


图 3-8 最值趋势

3.4 制定时段的热点分析

本节主要介绍了对指定时段热点频道、热点节目和热点关键词的分析，为了展示方便，结果主要采用词云的形式来直观地呈现。将统计得到的结果映射为词云中词语的字体大小，字体越大，则说明该词出现的频次越高，即热度越高。

3.4.1 热点频道分析

以 5 月 1 日为样本，统计当天的热点收视频道，并以词云的形式直观地展示。

词云与折线图的统计方法略有不同：折线图中按“人”统计，同一用户的重复多次访问同一频道只贡献 1 次流量、以此反映出单个频道的受欢迎程度；而此处的词云则按“次”统计，用户的多次访问都会贡献到流量中、以此反应出所有频道中热点频道的相对繁忙程度。如用户上午 8 点收看了 CCTV-5 的节目，折线图中只统计了一次（1 个用户看了 CCTV-5），而词云中可能会统计多次（翻到别的频道再返回、反复访问等）。

对所有数据条目进行切割，并提取出现的频道名称进行统计，将统计得到的频道出现频次大小映射为词云中词的字体大小。

图 3-9 热点频道词云

如图 3-9 所示，5 月 1 日当天，CCTV-5 位最大热点，其次为 CCTV-4，南京新闻综合频道热度要高于 CCTV-1，这与数据集本身取自江苏广电有关，因为所有观众都来自江苏，节假日会更加关心本省资讯。

3.4.2 热点节目分析

以 5 月 1 日黄金时段为例，统计方法与统计频道热词方法类似。

此处实际直接提取并统计了 5 月 1 日黄金时段（19 点-21 点）的热点节目名（而非关键词）。

由图 3-10 可知，此时段“五一”特别节目有着较高的收视率，但并没有明显的优势，电视剧、体育比赛和其他综艺节目也分享了大部分热点。

图 3-10 黄金时段热点节目词云

3.4.3 工作日、非工作日不同时段热点分析

为比较分析工作日与非工作日的热点节目异同情况，根据先前统计数据，选取周末 5 月 8 日（周日）和周中 5 月 11 日（周三）的四个高峰时段（10:00-13:00, 15:00-18:00, 18:00-21:00, 21:00-00:00），分别统计这几个时段的热点节目，并以词云的形式展现，越高的出现频次反映出越高的热度，词云中词的字体也会越大。

由图 3-11 中的对比可以看出，在午休高峰时段（10:00-12:59），周日的热点节目以综艺节目（奔跑吧兄弟、快乐大本营）、体育竞赛（NBA 季后赛）、电视剧（《四大名捕》）等休闲节目为主，而周三则为新闻（中国新闻、新闻直播间）及生活资讯类（家装 100 问）等为主；

而随着黄金时段（15:00-17:59、18:00-20:59）的到来，周日的热点节目趋向繁杂，没有特别突出的热点节目，说明各个节目的热度相似，而周三则依然以新闻为主（直播南京、新闻联播），但少儿（第一动画乐园）、综艺类（非常 6+1）、影视（机器侠、红高粱）等有了更加突出的表现。

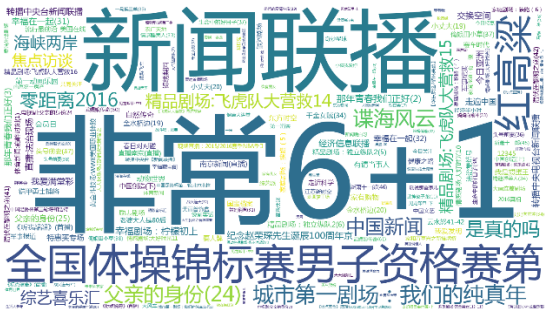
(1)5 月 8 日 10:00-12:59

(2)5 月 11 日 10:00-12:59

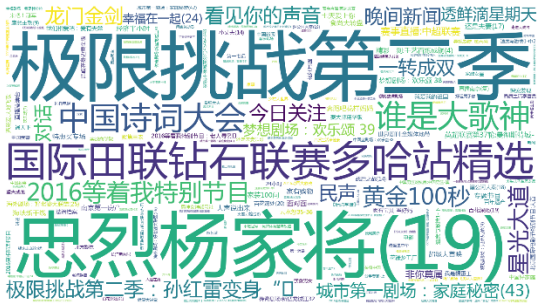
(3)5 月 8 日 15:00-17:59

(4)5 月 11 日 15:00-17:59

(5)5 月 8 日 18:00-20:59



(6)5 月 11 日 18:00-20:59



(7)5 月 8 日 21:00-23:59

(8)5 月 11 日 21:00-23:59

图 3-11 工作日、非工作日的不同时段热点节目对比

随着黄金时段渐渐结束（21:00-23:59），周日的热点节目又渐渐显现，周日与周三的热点节目都以综艺（极限挑战、我爱满堂彩）、影视剧（《忠烈杨家将》、《阳光宝贝》）、体育（钻石联赛、体育世界）等休闲节目为主流，可能对应于观众睡前放松的心理。

3.4.4 热点关键词分析

在 3.2.2 中，本文曾提到，由于节目名称变化性较大，不利于分析，因此采用了关键词的方法作为替代。本小节分析了 5 月份每日全日热点关键词的情况，由于篇幅原因，此处仅列举 5 月 1 日、2 日、11 日、12 日、21 日、22 日的分析情况

如图 3-12 所示，5 月 1 日适逢五一假期以及 NBA 比赛，因此雷霆对马刺这一场球赛成为最大热点；而 5 月 2 日收假第一天，五一特别节目仍然热播，甚至超过了 5 月 1 日当天，这与假期出行和收假返家、节目重播收视率走高有关；

5 月 11 日、12 日为月中周中，因此并无特别瞩目的热点。可以看出，11 日各种关键词出现频次相对平均，12 日周四适逢央视热播节目《回声嘹亮》播出，所以带动了一定的高潮；

(1)5 月 1 日　　(2)5 月 2 日

(3)5 月 11 日　　(4)5 月 12 日

(5)5 月 21 日　　(6)5 月 22 日

图 3-12 热点关键词分析

5 月 21 日、22 日适逢周末，可以看出 21 日的综艺、电视剧等节目，22 日的购物、体育类节目成为热点，这些节目都属娱乐、休闲性质，与周末人们放松身心的心态有关，因此收视需求多为此类。

3.5 本章小结

本章围绕频道与节目这一中心进行了一系列的收视行为统计与分析，其分析结果可以反映出用户群体在宏观上的收视行为特征与规律，根据这些特征与规律，包括频道流量变化、热点变化、频道间相互影响趋势、收视率受时间的影响变化等

等。这些都是宏观角度分析结果，可用于电视台对频道节目安排的宏观调控，但还无法实现针对个体用户的个性化服务。

第 4 章 以用户为中心的收视行为分析

4.1 引言

本章以用户个体为研究对象，进行了针对用户个体的收视行为分析，以帮助电视台进行个性化服务。

本章的结构如下：4.2 节描述了按用户 CA 卡号分类整理存储的数据条目特征；4.3 节详细介绍了由数据自动生成用户行为模式图的方法和实现；4.4 节提取并分析了了三种较为主要的用户行为模式；4.5 节介绍了如何算法识别用户行为并推测用户偏好；4.6 节为本章小结。

4.2 按用户分类分析

将数据条目按照 CA 卡号进行归类，并以 Message ID 为依据，按时间顺序排列这些条目，得到用户在某一时段的时序事件序列，即用户在该时段内所进行的一系列完整动作，方便抽样，下文的行为模式分析正式基于此特征展开。

表 4-1 5 月 7 日用户 825010213880008 的某一连续动作中的第 30-34 条动作

30|5|BlQaSdgECUPZJlYop|825010213880008|99756614330019580|2016.05.07
09:19:28|2016.05.07 09:19:23|443|204|658000|家家购物|以播出为准
|1|99|91|0|5|20160507091827529
31|6|BlQaSdgECUPZJlYop|825010213880008|99756614330019580|2016.05.07
09:19:28|479|215|554000|风尚购物|以播出为准|1|99|96|-1|20160507091829345
32|6|BlQaSdgECUPZJlYop|825010213880008|99756614330019580|2016.05.07
09:19:31|296|114|586000|山西卫视|NULL|1|99|97|-1|20160507091829019
33|6|BlQaSdgECUPZJlYop|825010213880008|99756614330019580|2016.05.07
09:19:34|443|204|658000|家家购物|以播出为准|1|99|98|-1|20160507091834772
34|21|BlQaSdgECUPZJlYop|825010213880008|99756614330019580|2016.05.07
09:19:34|443|204|658000|家家购物|以播出为准|1|99|98|0|20160507091838247

根据统计，仅 5 月 1 日一天，就有约 46 万个 CA 卡号（即用户）出现在数据流中，将其完全分类存储显然是一项极其艰巨而很不划算的任务，因此，在实验过程中，我们仅随机选取 31 位用户的数据进行实验和分析，提取出他们 5 月份每一天的动作序列，从中筛选出适合分析的部分（数据质量好、序列号连续、特征明显、

具有代表性等），为行为模式分析等后续处理的进行做准备。

表 4-1 展示了按用户分类整理数据后生成的结果示意，此处截取了 5 月 7 日 CA 卡号为 825010213880008 的用户在随机序列 BlQaSdgECUPZJlYop 所标识的一系列连续动作中的第 30-34 条动作。

4.3 用户行为模式图的生成

为了得到一个初步的、整体的用户行为模式画像，以便更好地分析用户行为，本节将用户的行为模式抽象为一个有向图，对于用户的连续动作，以事件 ID 为节点，相邻事件的先后顺序作为有向边，设计并实现了一个自动构建用户行为模式有向图的算法，构建出一个完整详尽的状态转移图，并试图寻找一个阈值，以删去出现次数过少的边，保留出现次数更多的边，以贴合用户的主要行为模式图。

算法 4-1 使用伪代码的形式详细描述了对于给定用户行为模式图的生成过程。程序共分为两部分：第一部分根据数据条目生成有向图的点和边，第二部分则根据以后的点和边绘制有向图。

算法 4-1 (a) 用户行为模式图的生成与绘制：生成点集和边集

```
INPUT:   dataset D
OUTPUT: V, E of graph G (V, E)

for each user ∈ TargetUsers do
    cycle_group = ∅
    cycle = []
    for each item ∈ D do
        if item 与上一个 item 在同一个 cycle 中 then
            将 cycle 加入 cycle_group
            cycle_items = []
        将 item 加入 cycle
    将 cycle 加入 cycle_group

nodes = ∅
edges = {}
for each cycle ∈ cycle_group do
    for each item ∈ cycle do
        将 item.eventID 加入 V
        if item 与上一个 item 的 Message ID 相邻 then
            if 两 item 间存在边 then
                E[边] += 1
            else 将该边加入 E
return V, E
```

算法 4-1 (a)用于生成有向图的点集和边集。根据输入的用户数据条目中的随机序列字段，将这些数据划分为不同的 cycle，以保证每个 cycle 都是用户在一段时

间内的连续动作。然后，依次处理每个 cycle，以出现的事件 ID 作为点，相邻两事件间建立一条边，对于重复出现的边，统计其出现次数作为边上的权值。最终得到用于绘制有向图的点集和边集。

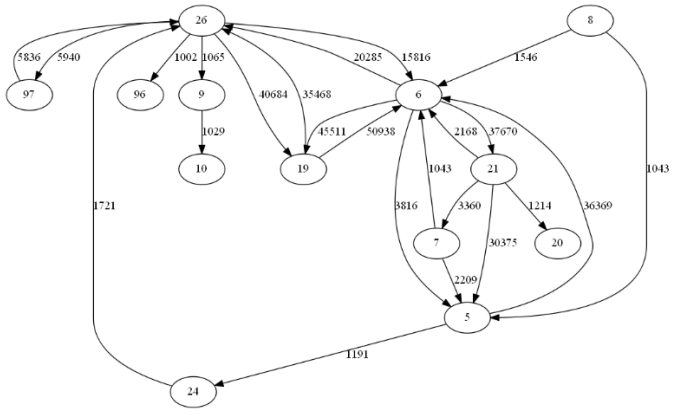
算法 4-1 (b)用于绘制有向图，根据上一步得到的点集和边集，将点和边依次绘制出来。

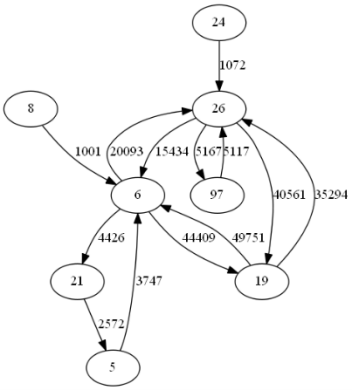
算法 4-1 (b) 用户行为模式图的生成与绘制：绘制有向图

```
INPUT:    V, E of graph G (V, E)
OUTPUT: graph G (V, E)

1.  input nodes, edges
2.  设置阈值 threshold
3.  初始化有向图 G
4.  将所有的点 nodes 加入 G
5.
6.  for each edge ∈ edges.items() do
7.      if edge.counter<threshold then
8.          continue
9.      将 edge 加入 G
10. 绘制 G
```

根据算法 4-1，对随机抽取的若干用户进行了模式图的绘制，如图 4-1 所示，举例了 CA 卡号分别为 825010269689626、825010354067360、825010367831984、825010373957410 四个用户（设为 A，B，C，D）三天内的整体的行为模式图。





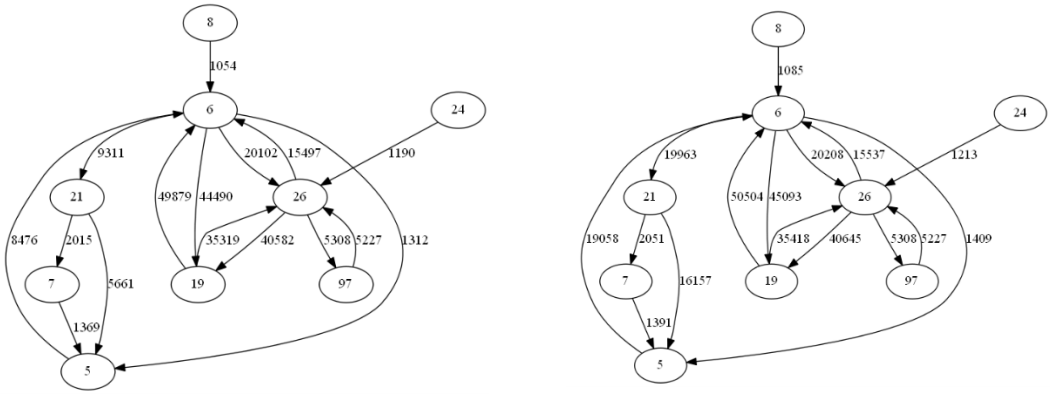


图 4-1 四个用户三天内的行为模式图

左上：用户 A 右上：用户 B

左下：用户 C 右下：用户 D

此外，还可以将所有给定用户的行为累加在一起，得到整体用户模式图，这样的好处是，相比于具有明显偏好和习惯的个体用户模式图，整体用户模式图中的行为方式因为是所有用户的加和，因此更具有代表性和普遍性。图 4-2 给出了对随机抽取的 34 名用户在整个五月份的整体用户行为模式图，阈值分别设定为 500, 1000, 1500, 和 2000。

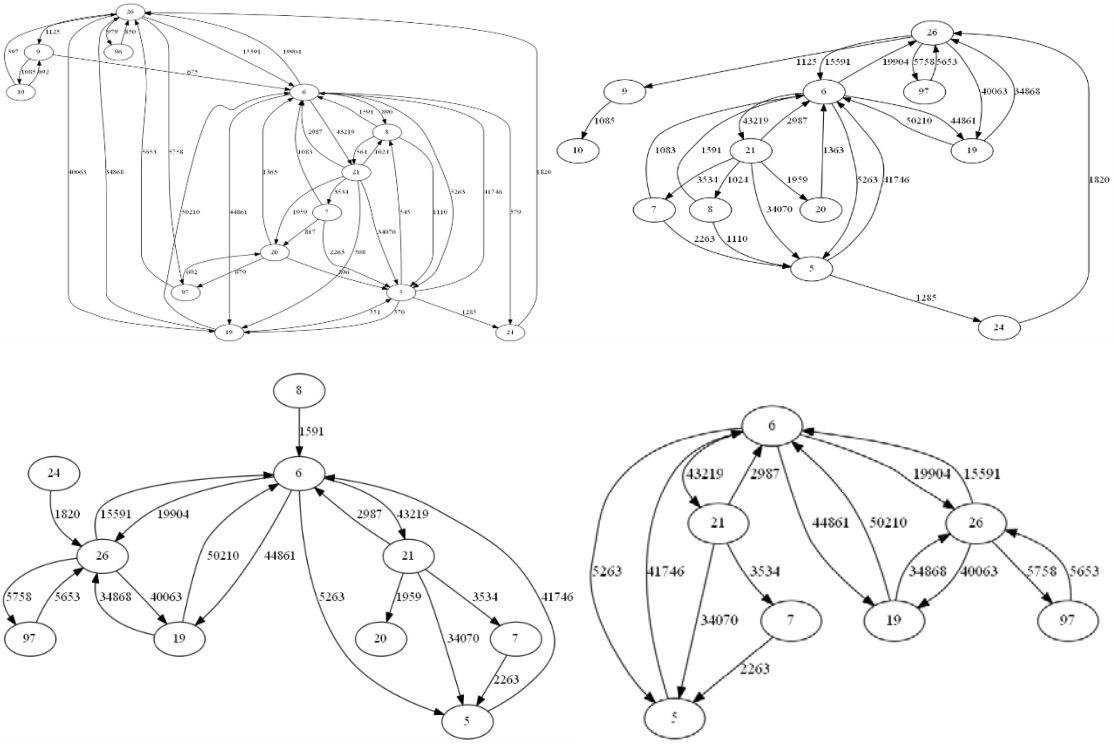


图 4-2 整体用户行为模式图

左上：阈值 500、右上：阈值 1000、左下：阈值 1500、右下：阈值 2000

由图 4-2 可以看出，在众多的事件中，有一些事件是用户经常触发的，如 5|频道退出事件、6|EPG 显示事件、7|音量调节、19|中间件事件、21|频道进入事件、26|链接地址跳转事件、97|时移节目播放等等，当阈值设置的较大时，出现较少的事件被筛去，而出现频次较高的事件得以保留。这些事件即为用户最常出发的事件，所反映的便是用户经常进行行为模式。

但将阈值设置为 2000 时，可以发现，整个有向图可以被分为三条支路：

（1） 6|EPG 显示事件—21|频道进入事件—5|频道退出事件—6|EPG 显示事件 事件间为单向关系，有循环的迹象，其中，21 与 5 之间偶尔会有 7|音量调节等动作。很明显，该支路描述了用户换台、收看节目、继续换台的这一过程，该模式为本研究中最重要、也是出现频次最高的模式之一。

（2） 6|EPG 显示事件—19|中间件事件—26|链接地址跳转事件 事件间为双向关系，该支路描述了用户在收看节目的行为和使用应用程序的行为之间的转换过程，因此出现频次极高。

（3） 6|EPG 显示事件—26|链接跳转事件—97|时移节目播放事件 事件间为双向关系，该支路描述了用户收看时移节目的过程，同样是研究用户偏好的重要内容。

而当阈值减小时，一些出现次数较少的事件逐渐显现。这些事件中，有些是对已有行为模式的补充，例如，当阈值设置为 1500 时，将出现表示用户按下功能键的 24|功能键按键事件、表示用户呼出菜单的 8|菜单事件以及长时间无操作时告知服务器当前用户仍然在线的 20|心跳事件等；此外，还有一些事件是对新的行为模式的描述，例如，当阈值设置为 500 时，将出现支路：26|链接跳转事件—96|VOD 点播事件，该支路描述了用户使用 VOD 点播功能点播并收看节目的过程，虽然在数据中出现的次数较少，但作为研究用户行为与偏好的重要因素之一，同样值得我们关注。

4.4 用户主要行为模式分析

通过前两节的分析，结合对具体数据流的观察与分析，我们可以得出如下四种最常见的行为模式：

（1） 浏览行为 如图 4-3 (a)，以 6|EPG 显示事件→21|频道进入事件→5|频道退出事件为主干，三种事件总是有序地、周期地交替出现，间或穿插如 7|音量调节事件、8|菜单事件、20|心跳事件等。用户浏览至某一频道时，会退出原本所在的频道，然后首先显示 EPG 信息，相当于预览作用，随后进入该频道，此时这一系

列动作是自动完成的，用户可能只需按一下换台键，而不需要进行其他操作。当用户准备换入下一频道、执行其他功能、或跳转到其他页面时，便会退出该频道。

（2） 时移节目播放行为 如图 4-3 (b)，以 26|链接跳转事件→97|时移节目播放事件→26|链接跳转事件为主干，97|时移节目播放事件会连续重复出现若干次，贯穿于整个行为的始终，中间可能穿插 7|音量调节事件、8|菜单事件等，此外，该行为开始时，在 26|链接跳转事件之前，多由 24|功能键按键事件开始，而再行为结束后，一般会先退出到 6|EPG 显示事件，后接浏览行为或其他行为。当用户想要观看时移节目时，按下功能键，选择要播放的时移节目，经过若干次链接跳转，开始播放时移节目，时移节目播放结束后，再经过若干次链接跳转，跳转到下一个时移节目，或跳转回一般电视节目。

（3） VOD 点播行为 如图 4-3 (c)，以 6|链接跳转事件→96|VOD 点播事件→26|链接跳转事件为主干，97|时移节目播放事件会连续重复出现若干次，贯穿于整个行为的始终，中间可能穿插 7|音量调节事件、8|菜单事件等。用户选择要点播的节目，通过若干次链接跳转，开始播放点播节目，点播结束后，经过若干次链接跳转，跳回原来的位置。

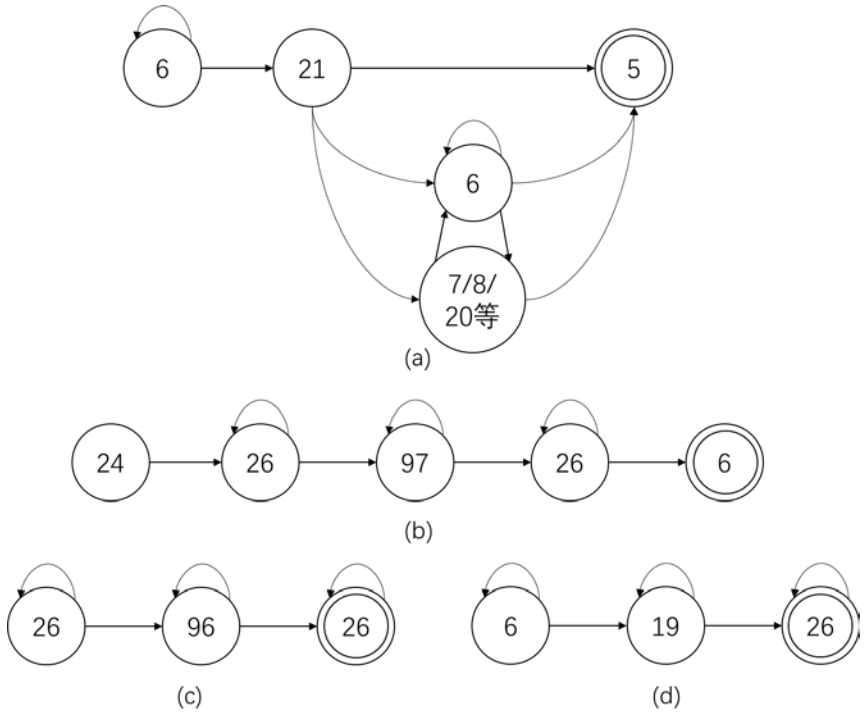


图 4-3 用户主要行为模式

（4）中间件过度行为 如图 4-3 (d)，6|EPG 显示事件→19|中间件事件→26|链接地址跳转事件为主干，19|中间件事件可连续重复出现，若干次，其后可跳转

回 6|EPG 显示事件，也可跳转到 26|链接地址跳转事件。用户做出相应操作时，该行为自动执行，以实现一般电视节目的收看与应用程序的使用的相互转换过度。

4.5 用户行为识别与偏好分析

根据上节得到的行为模式，构建自动机，即可算法实现对用户行为与偏好的识别。在上述四种行为模式种，前三种是要重点研究的，第四种中间件过度行为，由于用户实际参与较少，与用户的行为偏好关联不大，因此暂不提供识别。

算法 4-1 以伪代码的形式描述了具体的识别过程。输入数据可采用先前已经处理过的按用户 CA 卡号分类存储的结果数据，对于给定用户的连续条目，首先根据随机序列字段将这些条目切割为不同 cycle，其中都是一段连续事件序列。对于每一个 cycle，时序扫描 cycle 中各条目的事件 id，并根据构建的自动机进行状态转移，当转移到识别处某一行为的状态时，则做输出处理，输出其行为名称、起止时间、经历时间、频道名、节目名、频道与节目所对应的可能分类及其概率等行为偏好描述信息，同时统计该用户在类别的累计时长。当所有的 cycle 扫描完毕后，行为已经全部识别，此时根据统计的各个分类的累计时长，便可推测处用户最有可能偏好的频道节目分类，即对于某一分类，用户在其上的累计时长越长，则说明用户越有可能喜好这一分类的频道节目。

算法 4-2 用户行为识别与偏好分析算法

```
INPUT:    Dataset D (Classified with CACardNo)
OUTPUT: User_Behaviors, TOP_Favorites

1.   for each user ∈ TargetUsers do
2.        cycle_group = ∅
3.        cycle = [ ]
4.        ClassStat = ∅ /* 各分类的总时间 */
5.        for each item∈ D do /* 划分 cycles */
6.             if item 与上一个 item 在同一 cycle 中 then
7.                  将 cycle 加入 cycle_group
8.                  cycle = [ ]
9.             将 item 加入 cycle
10.       将 cycle 加入 cycle_group
11.
12.       初始化栈 stack
13.       for each cycle ∈ cycle_group do /* 利用自动机识别行为 */
14.            for each item ∈ cylce do
15.                 switch item.eventID:
16.                      case '21':
17.                           item 入栈
18.                      case '5':
19.                           while 栈中频道名与节目名与 item 不符 do
```

算法 4-2 用户行为识别与偏好分析算法（续上）

```
                出栈
            if item.duration>Threshold then
                keywords = cut(item.programName) /* 对节目名分词 */
                知识库中查找(频道名,关键词)对应的类别 porgramClasses
            output 识别到的浏览行为
            将行为的持续时间加到 ClassStat 中的对应类别
            出栈
        case '26':
            if 栈顶的事件 ID 为 96 或 97 then
                寻找栈中第一个 96 或 97 号事件
                if item.duration>Threshold then
                    keywords = cut(item.programName) /* 对节目名分词 */
                    知识库中查找(频道名,关键词)对应的类别 porgramClasses
                output 识别到的时移播放行为或 VOD 点播行为
                将行为的持续时间加到 ClassStat 中的对应类别
                出栈
            else
                item 入栈
        case '97':
        case '96':
            item 入栈
        case '24':
            初始化栈 stack 并将 item 入栈
        case '6':
            if 栈顶的事件 ID 为'26' then
                初始化栈 stack 并将 item 入栈
            else
                item 入栈

    对 ClassStat 排序
    return 输出 ClassStat 中的类别
```

算法 4-2 的结果包括两部分：其一为对用户的行为识别结果，其二为用户对于频道节目类型的偏好分析结果。表 4-2 展示了对指定用户的行为识别的不同结果的举例，图 4-4 则展示了对其中四位用户的偏好分析结果。

如表 4-2 所示，用户行为识别的结果由以下字段组成：

（1）Behavior 用户行为种类判定，结果中主要分为三类：浏览行为（Look Through）、时移播放行为（Time-shifted Play）以及 VOD 点播行为（VOD Play）。根据上文中的自动机状态的转移得出结果。

（2） Start Time、End Time 当前行为的起止时间，对于浏览行为，可从 5|频道退出事件中的字段中直接获取，而对于时移播放行为和 VOD 点播行为，则需要根据第一次和最后一次出现的时移播放/VOD 点播事件得出。如公式（4-1）、（4-2）即：

$$\text{StartTime} = Event_1.Time \quad (4\text{-}1)$$

$$\text{EndTime} = Event_n.Time \quad (4\text{-}2)$$

（3）Duration 当前行为所持续的时间，即起止时间之差，如公式（4-3）：

$$\text{Duration} = \text{EndTime} - \text{StartTime} \quad (4\text{-}3)$$

（4） Channel Name、Program Name 当前行为所涉及到的频道名和节目名们。对于时移行为和 VOD 播放行为，其频道名为非必要字段，有时为空，此时将频道名设为 None。

（5） Keywords 由频道名分词得到的若干关键词，如公式（4-4）。

$$\text{Keywords} = \text{cut}(\text{ProgramName}) = [\text{Keyword}_1, \text{Keyword}_2, \ldots, \text{Keyword}_n] \quad (4\text{-}4)$$

（6） Possible Class(es) and Possibility 列出所有可能的节目分类及其概率。设 f(x)为 Keyword 到 Class 的映射，则有公式（4-5）、（4-6）：

$$\text{Class} = f(Keyword_i) \quad (4\text{-}5)$$

$$\text{P}(\text{Class}_i) = \frac{num(Class_i)}{\sum_{k=1}^{n} num(Class_i)} * 100\% \quad (4\text{-}6)$$

表 4-2 用户行为识别结果举例

(a)	(b)	(c)
----------Find Behavior---------- Behavior: Look Through Start Time:2016-05-02 09:55:21 End Time:2016-05-02 10:06:45 Duration: 0:11:24 Channel Name: CCTV-1 Program Name:2015 出彩中国人 Keywords:2015,出彩,中国 Possible Class(es) and Possibility: 综艺 100.0% ----------End----------	----------Find Behavior---------- Behavior: Time-shifted Play Start Time:2016-05-02 00:15:21 End Time:2016-05-02 00:16:27 Duration: 0:01:06 Channel Name:江西卫视 Program Name:财经新风向 Keywords:财经,新风 Possible Class(es) and Possibility: 新闻 25.0% 财经 75.0% ----------End----------	----------Find Behavior---------- Behavior: VOD Play Start Time:2016-05-20 17:21:58 End Time:2016-05-20 18:02:05 Duration: 0:40:07 Channel Name: None Program Name: 欢乐颂(更新中) 第 42 集 Keywords:欢乐颂,更新,42 Possible Class(es) and Possibility: 电视剧 100.0% ----------End----------

对于表中所给出的结果，表 4-2 (a)中识别到了用户的浏览行为，用户在 2016 年 5 月 2 日 9:55:21 至 10:06:45 历时 11:24 的时间里，收看了 CCTV-1 的 2015 出彩中国人节目，经分类分析，该节目 100%可能性为综艺节目；表 4-2 (b)中识别到

了用户的时移播放行为，用户在 2016 年 5 月 2 日 00:15:21 至 00:16:27 这 1:06 的时间里，通过时移播放功能收看了江西卫视的财经新风向节目，而根据分类分析结果，该节目 25%的可能性为新闻类节目，75%的可能性为财经类节目；表 4-2 (c)中识别到了用户的 VOD 点播行为，用户在 2016 年 5 月 20 日 17:21:58 至 18:02:05 历时 40:07 的时间里，使用 VOD 点播功能收看了欢乐颂第 42 集，根据分类分析，该节目 100%为电视剧类节目。

图 4-4 中，以用户收看某一类节目的总时长作为判断其偏好类别的依据，如公式（4-7）：

$$\eta_i = \frac{\sum_k duration_{k,Class_i}}{\sum_i \sum_k duration_{k,Class_i}} * 100\% \qquad (4\text{-}7)$$

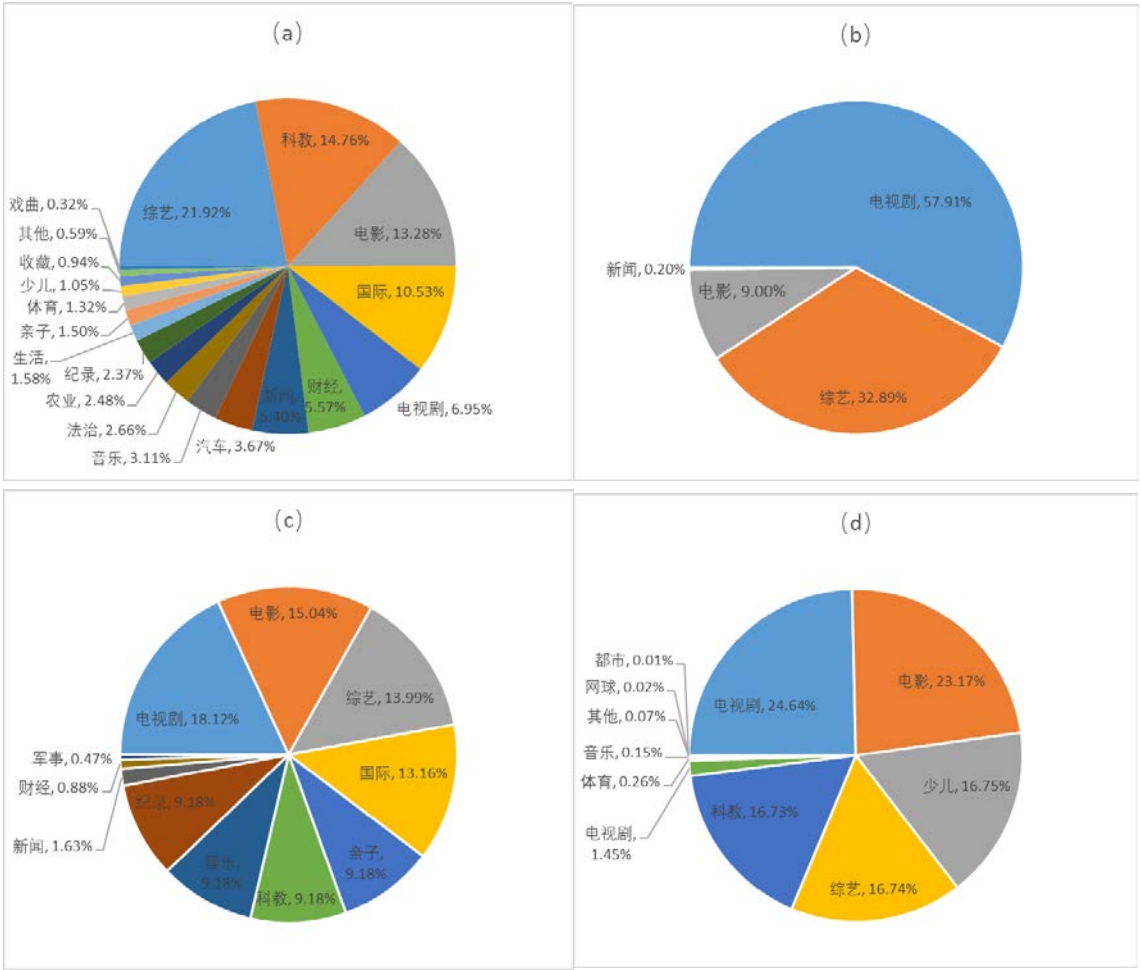


图 4-4 用户偏好分析

图 4-4 (a)中，用户所涉猎的节目内容很多，其中，综艺（21.92%）、科教（14.76%）与电影（13.28%）三种节目类型占据了超过 50%的收看时长，说明该用户更倾向

于收看这三种节目；图 4-4 (b)中，用户整个 5 月份只收看了 4 种类型的节目，且电视剧占据了 57.91%的绝对优势，很显然，电视剧为该用户最爱收看的节目类型；图 4-4 (c)中，用户收看的节目类型丰富且分布较为平均，其中比例较大的，如电视剧（18.12%）、电影（15.04%）、综艺（13.99%）、国际（13.16%），可以推断用户所偏好的节目类别更可能出现在这些节目类型中；图 4-4 (d)中，电视剧（24.64%）、电影（23.17%）占据了更大比重，可以此断定为用户的偏好类型。在实际操作中，我们规定：设用户所收看节目的所有类型的集合为 U，则用户偏好集合 A 应满足如下条件：（1）A 为 U 的子集，元素个数不超过 5；（2）将 U 中所有元素按收视时长百分比从大到小排序，A 中各元组应在前 5 名之内；（3）A 中元素收视时长百分比之和应大于 50%；

即有公式（4-8）：

$$\mathrm{A} = \{\mathrm{c}|\mathrm{c} \in \mathrm{U}, \mathrm{P}(c_i) = p_i, i \in [1, n]\} \quad (4\text{-}8)$$

且满足：

$$c_i \in U,$$

$$\forall \mathrm{i}, \mathrm{j} \in [1, \mathrm{n}], i \geq j, \text{则} p_i \geq p_j, \text{且}$$

$$\forall \mathrm{a} \in \mathrm{P}(\mathrm{U}) - \{\mathrm{p}|p_i, \mathrm{i} \in [1, \mathrm{n}]\}, p_i > a,$$

$$\mathrm{n}\ \text{满足} \sum_{k=1}^{n} p_k \geq 50\% \text{或} n = 5$$

4.6 本章小结

本章从用户个体的角度出发，绘制了用户个体与总体的行为模式画像，研究了对于指定用户个体，其可能存在的行为模式，并总结出四个最为常见用户行为模式，其中三个可用于用户偏好分析，而后又算法实现了对指定用户的个性化的行为识别和偏好分析。

第 5 章 行为识别与偏好分析算法的性能测试

5.1 引言

针对上问题基础的行为识别与偏好分析算法（算法 4-2），设计实验对其性能进行测试。

本章结构如下：5.2 节算法性能分析；5.3 节实验测试设计；5.4 节实验结果分析；5.5 节本章小结。

5.2 算法性能估计

对于算法 4-2，可根据功能的不同将其大致分为三个部分：cycle 的划分、自动机识别和偏好排序。

5.2.1 cycle 的划分

如上文所述，对于一个用户，其数据条目中，随机序列字段相同的所有连续条目被视为一个 cycle，在该时段内发生的这一系列事件均出自用户同一连续的动作，亦即用户一次收看电视所触发的所有事件。

cycle 划分过程的关键在于寻找到随机序列字段发生变化的起始位置。在算法 4-2 中，采用遍历的方式寻找总数为 n 条的条目中不同 cycle 间的分割点，时间复杂度为 O(n)。

5. 2. 2 自动机识别

对于每个 cycle，在使用自动机进行行为识别时，根据 cycle 中条目所表示的事件进行状态转移，并将识别到的所有行为输出。设共有 m 个 cycle，其中第 i 个 cycle 中有 ni 条，则要识别所有 cycle 中的所有行为，所需要的进行的状态转移数为 m 个 cycle 中的事件条数之和，约等于数据条目的总数 n，即有公式（5-1）：

$$n \approx \sum_{k=1}^{m} n_i \qquad (5\text{-}1)$$

因此，自动机识别阶段的时间复杂度亦为 O(n)。

5. 2. 3 偏好统计与排序

在偏好分析中，最为关键的部分即为对用户收视偏好进行时长统计和排序，以得到用户收看时间最长的节目类别作为偏好分析的重要依据。

5.2.3.1 哈希字典实现的时长统计

时长统计的关键问题在于，需要为(类别,时长)二元组进行寻找合适的结构，以减少大量的访问更新带来的时间开销。若使用最为朴素线性表进行存取，对于 m 个二元组，每次访问都需对线性表进行遍历，时间复杂度为 O(m)；则总共 n 次访问所带来的时间开销为：O(m*n)。

作为优化，算法 4-2 在实现时，采用哈希字典来代替线性表完成对二元组的存取更新，则每次访问所需的时间开销降为 O(1)，全部 n 次访问的时间复杂度为 O(n)。

5.2.3.2 排序算法的选择

为了得到用户收看时间最长的前若干个节目类别，可将 5.2.3.1 中的结果进行排序，算法 4-2 的实现中采用了时间复杂度为 O(nlogn)的快速排序。而由于实际要排序的类别数远远小于要处理的条目数，其时间开销在整个算法的时间开销中可看作一个非常小的常数而忽略不计。

5.2.4 总开销估计

由上面的分析可以看出，算法的整体时间开销是线性的，当数据条目数为 n 的时候，算法的时间复杂度可近似表示为 O(n)。

5.3 实验设计

上一节中，我们判断算法 4-2 的时间开销是线性的，且与处理的数据条目数 n 相关。但是，在实际的算法执行中，n 并不是一个给定的值，每个用户在不同时段所涉及到的条目数是完全随机的，因此，为了更好的验证算法的线性关系，本文采取了控制变量的方法来设计实验。

5.3.1 实验环境

实验采用 python 2.7 64 位进行程序编写，程序运行在 Inter Core i5 处理器，内存 16GB，windows 10 64 位操作系统下。

5.3.2 以时间跨度为自变量的性能测试

随机抽取 200 名用户，分别测试算法在处理这 200 名用户在 1~10 天的时间跨度内所涉及到的所有条目所需要的时间，并绘制曲线。测试数据的具体规模见表 5-1。

表 5-1 以时间跨度为自变量实验数据规模

用户人数/人			200		
时间跨度/天	1	2	3	4	5
总条目数/行	3,803,484	3,805,065	3,805,425	3,805,691	3,807,884
用户人数/人			200		
时间跨度/天	6	7	8	9	10
总条目数/行	3,810,136	3,812,182	3,813,233	3,816,698	3,819,008

5.3.3 以用户人数为自变量的性能测试

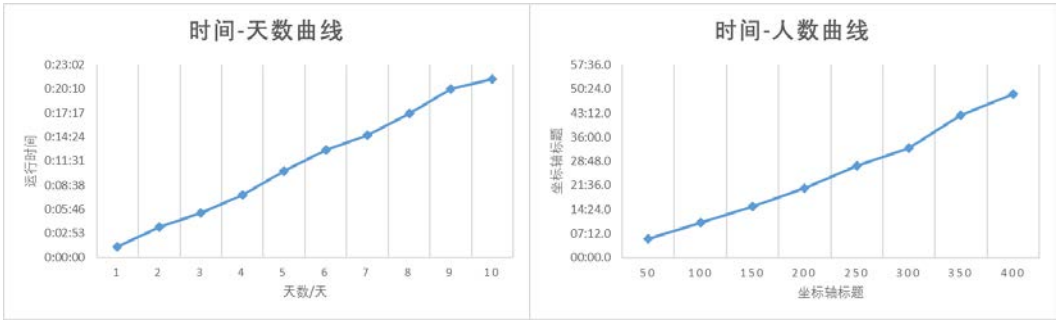
随机抽取 10 天，分别测试算法在处理 50~400 名用户（步长为 50）在 10 天内所涉及到的所有数据条目所需要的时间，并绘制曲线。测试数据的具体规模见表 5-2。

表 5-2 以时间跨度为自变量实验数据规模

时间跨度/天			10	
用户人数/人	50	100	150	200
总条目数/行	730,326	1,623,854	3,088,947	3,819,008
时间跨度/天			10	
用户人数/人	250	300	350	400
总条目数/行	5,253,417	6,134,247	7,103,323	8,235,920

5.4 实验结果分析

如图 5-1 (a)展示了以时间为自变量的测试结果，图 5-1 (b)展示了以人数为自变量的测试结果。



(a) (b)

图 5-4 实验结果曲线

可以看出，无论是以天数或人数为自变量，其运行时间曲线均近似为一条直线，即运行时间随着自变量的增长而呈线性增长趋势。由此可以推断，当总条目数 n 增长时，算法运行时间 t 也将随之线性增长，因此，算法的时间开销是线性的。

5.5 本章小结

本章研究了前文提出的行为识别与偏好分析的性能问题。经过对算法每一部分时间开销的分析与估算，得出算法的时间复杂度为 O(n)，是线性的；随后，我们设计了一组实验对这一结论进行验证，通过对不同的自变量下，运行时间随自变量的变化曲线的绘制，可以直观地观察其变化趋势；最终，通过对实验结果的分析，可以确定，之前的估算正确，该算法的时间开销与算法处理的数据条目数确实呈线性相关。

结　论

本文研究了在现实的电视数据中进行用户收视行为分析的方法。以江苏广电提供的 2016 年 5 月的用户作为研究对象，分别从以频道/节目为中心和以用户为中心这两个角度入手，对不同用户、不同频道、不同节目、不同时间等多个维度上的用户收视情况进行研究分析，得出以下成果：

（1）对原始数据进行了较为细致的分析和处理，为后续交接及其他方向的研究工作打下基础。例如，本文研究重点在用户与其收看的频道、节目间的关系，以对个性化节目推送方面进行探索，而在未来的其它研究可研究不同的方向，如通过研究用户对音量调节的习惯，以实现智能化音量控制等等，本文对原始数据的分析和处理同样可服务于此类后续研究。

（2）提出了通过两次分类来判断节目类别的方法。通过一次分类：按频道分类以及二次分类：按频道+关键词分类，便可以判断用户所收看的节目属于哪一类别，并建立了一个小型的频道-关键词分类知识库以供测试。但是，目前的研究仅将节目划分为诸如音乐、电影、体育等等较为粗粒度的类别，在未来的研究工作中，可以更加深层次地对给定节目进行进一步细致的分类，如将音乐分为现代音乐、古典音乐和流行音乐，将电影分为动作电影、爱情电影、科幻电影等等。这需要更多层次的分类来实现，本文的分类方法将是这些多次分类的基础工作。

（3）得到了关于频道及节目收视率的一些统计性结论和猜想，如：

1. 普通频道流量>>高清频道流量>>杜比频道流量，分析用户偏好时可算作同一频道；

2. 各频道每日流量随时间变化波动趋势相同，午间达到小高峰，黄金时段达到最高峰；

3. 不同频道间同一时段的节目的收视情况存在相互影响，具体的定量研究可在未来研究中确定；

4. 频道的每日收视峰值随日期呈现周期性变化，未来研究中可对此猜想进行验证，并且可作为研究用户流失情况的参考之一：当频道每日峰值的周期性变化被打破、或同期环比程下降趋势时，说明用户正在流失；

5. 通过热点分析，可直观地观察各时段的频道、节目、关键词热点及其变化情况。

（4）算法自动生成了用户行为模式图以供分析，总结出了四种常见的用户行为模式：浏览行为、时移节目播放行为、VOD 点播行为和中间件过度行为，其中，

前三者为用户收视行为分析的重点。

（5） 提出了用户行为识别与偏好分析的算法，并实验验证其效率，证明可在线性时间内完成对用户的行为识别和偏好分析。

在未来的研究工作中，可继续就节目分类、行为识别、偏好分析、用户流失等方面做更加深入、细致的研究工作，以更加迎合广电企业的实际需求，并服务于其中。

参考文献

[1] Digital TV Research. Digital TV World Databook[R]. 2015

[2] 36Kr. 为什么要做用户分析[OL]. http://36kr.com 2017.02.21

[3] 刘飞，马力维．数据挖掘在广电行业的应用[J]. 有线电视技术，2008, 15(10):69-71.

[4] Spangler W E, Gal-Or M, May J H. Using data mining to profile TV viewers[M]. ACM, 2003.

[5] 李忠哗, 冯素勤, 刘志江. 数据挖掘技术在有线电视领域的应用[J]. 中国有线电视, 2006(16):1580-1582.

[6] 刘峰. 大数据时代的电视媒体营销研究[D]. 华东师范大学, 2014.

[7] 何速. 社会电视用户行为分析[D]. 国防科学技术大学, 2011.

[8] Apte C, Liu B, Pednault E P D, et al. Business applications of data mining[C]// Communications of the Acm. 2002:49-53.

[9] Baudisch P, Brueckner L. TV Scout: Lowering the Entry Barrier to Personalized TV Program Recommendation[C]// International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. Springer-Verlag, 2002:58-68.

[10] Chickering D M, Heckerman D. A Decision Theoretic Approach to Targeted Advertising[C]// Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc. 2000:82-88.

[9] Chanza M M. Profiling television viewing using data mining[J]. 2013.

[10] Anderberg M R. Cluster Analysis for Applications,[J]. Probability & Mathematical Statistics New York Academic Press, 1973:347–353.

[11] Andritsos B P. Data Clustering Techniques[M]// Machine Learning for Multimedia Content Analysis. Springer US, 2007:37-70.

[12] Linoff G S, Berry M J A. Data mining techniquesfor marketing, sales, and customerrelationship management[J]. 1997, 17(1):1 - 8.

[13] Holland S, Ester M, Kießling W. Preference Mining: A Novel Approach on Mining User Preferences for Personalized Applications[C]// Knowledge Discovery in Databases: Pkdd 2003, European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings. DBLP, 2003:204-216.

[14] Beeferman D, Berger A. Agglomerative clustering of a search engine query log[C]// ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2000:407-416.

[15] Estivill-Castro V, Houle M E. Robust Distance-Based Clustering with Applications to Spatial Data Mining[J]. Algorithmica, 2001, 30(2):216-242.

[16] Joachims T. Optimizing search engines using clickthrough data[C]// Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002:133-142.

[17] Levene M. Data Mining of User Navigation Patterns[C]// Revised Papers from the International Workshop on Web Usage Analysis and User Profiling. Springer-Verlag, 1999:92-111.

[18] Bchner A G, Baumgarten M, Anand S S, et al. Navigation Pattern Discovery from Internet Data[J]. 2000:74-91.

[19] Levene M. Mining association rules in hypertext databases[C]// International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1998:149-153.

[20] Levene M. Heuristics for Mining High Quality User Web Navigation[J]. 1999.

[21] Smyth B, Cotter P. A personalized television listings service[J]. Communications of the Acm, 2000, 43(8):107-111.

[22] Smyth B, Cotter P. Surfing the Digital Wave - Generating Personalised TV Listings using Collaborative, Case-Based Recommendation[C]// 1999:561--571.

[23] Jennings A, Higuchi H. A user model neural network for a personal news service[J]. User Modeling and User-Adapted Interaction, 1993, 3(1):1-25.

哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《电视大数据的用户收视行为分析系统设计与实现》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：　　日期：　　年　　月　　日

致　谢

转眼间，大学四年的最后一段时间也接近了尾声。回首往昔，如若昨日，四年的本科生活即将结束，在此向所有关心和帮助过我的人致以由衷的谢意。

衷心感谢高宏老师对我的毕业设计的精心指导，感谢王宏志老师指引和帮助，感谢左旺孟老师、孙春奇老师的关心与教导，良师们以身作则践行着“规格严格，功夫到家”的箴训，他们为人师表、为学生甘于奉献，他们的言传身教将使我终生受益。

感谢丁晓欧学姐的耐心帮助与引导，感谢过云燕学姐、罗长春学长、任天萌学长、孙彤学长，在这四年的学习与生活中，学长学姐们的关心、帮助与鼓舞，激励我不断进步，他们用对科研的认真严谨、对技术的执着追求与对困难的积极乐观，为我们树立了榜样。

感谢王新达、段艺、李天宝等挚友们的一路相伴，旅途漫漫，道阻且跻，他们的陪伴和支持让我有了足够的力量一路披荆斩棘，勇往直前。

感谢海量数据计算研究中心、微软俱乐部和 pureweber 的老师、同学们，给了求学在外的我以家般的港湾。

感谢父母的默默付出，为我撑起了最为坚实的后盾，让我能够安心前行。

此去经年，虽好景虚设，亦将不负师恩，不忘同窗，不辜父母，尽我所能，攀向高峰。同时，也祝老师们桃李天下、硕果累累，同窗们前程似锦、成为栋梁，父母身体健康、平安幸福！