北京大学

硕士研究生学位论文

题目：**电信用户网络行为建模及特征分析**

姓　　名：	段婧
学　　号：	1301210610
院　　系：	软件与微电子学院
专　　业：	软件工程
研究方向：	电子商务与物流
导师姓名：	李杰教授

二〇一五年十月

版权声明

摘要

在当前大数据背景下，精准化、个性化的产品服务是必然趋势，各大电信运营商意识到产品服务应由粗放经营转为精细化、个性化，开始建立自己的基于数据挖掘的商业智能系统。但目前各大运营商在流量经营、套餐、网络质量、基站建设等方面正逐步上手研究，而在用户分析和精准营销方面还在起步阶段。所以本次选择此方向进行研究，通过对大量个人上网数据进行深入分析，准确挖掘用户的偏好特性及其他特征，对用户精准化、个性化的服务做数据基础，从而可以对每一位用户提供定制的服务，达到精准服务推送的目的。

本文基于电信用户大数据，对某省移动用户的上网数据进行研究，通过时间序列分析、幂定律、Phantom 联合聚类算法等数据分析方法，对用户访问行为的时间规律、市场竞争结构、用户兴趣的市场细分进行了分析。在时间维度上发现用户对移动互联网日访问量周期为 7 天，周五、周一较多，周末较少；以小时为粒度的时间序列周期为 24，每天早上 7 点左右最低，晚上 20 点左右访问量最高，并建立了 ARIMA 时间序列预测模型。在业务应用市场的竞争结构中发现每类市场都符合幂定律，并对各类市场的竞争激烈程度、垄断情况进行了分析，给出营销建议。在针对用户兴趣的市场细分中，使用 Phantom 双向软聚类算法将用户聚成 16 类，将 16 类用户分成社交型、娱乐型、旅游型、理财型、生活型、新闻型和软件型 7 种类型，针对每种类型每个聚类簇中用户的特点进行了分析，并针对每个用户簇中用户的兴趣偏好不同进行了营销推广建议。

关键词：大数据，数据挖掘，用户分析，市场细分，聚类

The Modeling and Characteristics Analysis of the Internet Behaviors of the Telecom Users

Duan Jing (Software Engineering)
Directed by Professor Li Jie

ABSTRACT

Under the background of the current big data, accurate and personalized products and services are inevitable trends. Operators have realized that they should transform the product services from extensive operation into intensive one and start to build data mining BI system of their own. Even though the operators have gradually carried out the research on traffic data operation, package, network quality and base station construction, the research on the customer analysis and precise promotion is still in its infancy. This paper, through deep analysis of personal network behavior, has precisely captured the customer preference and other characters. It has provided data base for precise and personalized services so as to achieve the goal of precise and personalized promotion.

Based on the big data of the telegram users, this paper has explored the traffic data of mobile users in a certain province. It has analyzed the customer visiting time pattern, market structure and customer preference through the time series analysis, power law and data analysis methods, such as the Phantom combined clustering algorithm. On the time dimension, the paper found that the cycle of mobile internet traffic is seven days as well as Monday and Friday is the most active day with weekend is less active. Taking hours as granularity, the paper found that the cycle is 24 hours with 8 p.m. is the most active and 7 a.m. is the least active. In the analysis of market competition structure, each type of market is in line with the power law, and the degree of market competition and monopoly situation are analyzed. For market segmentation in the user's preference, the use of the Phantom bidirectional soft clustering algorithm put users into 16 classes which can be divided into social, entertainment, tourism, finance, life, news and software types. Each type of cluster are analyzed, and marketing advice has been put forward to meet the user’s preference.

KEY WORDS: Big Data, Data Mining, User analysis, Market segmentation, Clustering

目录

第一章　绪论

近年来互联网发展迅速，2014 年底，联合国国际电信联盟（ITU）公布其年度《衡量信息社会发展报告》，报告中最新研究数据显示，全球网民已突破 30 亿人，2014 年全球互联网使用率继续稳定增长，平均年增长率 6.6%。从 2009 到 2014 年五年时间中，发展中国家的互联网用户数量已经增加了一倍，2/3 网民现在生活在发展中国家。其中移动互联网也迅速发展成为世界第十一位创造万亿美元的新空间。中国移动互联网正处于蓬勃发展的时期，2015 年 4 月工信部副部长在全球移动互联网大会(GMIC)指出我国移动互联网随着 4G 的到来，已经成为全球第二大市场，手机用户超 12 亿，其中有 9 亿是手机用来上网。伴随着移动终端价格下降、智能手机普及、4G 流量费下降以及 WIFI 的广泛铺设，移动互联网如今呈现井喷式的发展，市场潜力巨大。中国电信运营商非常重视互联网市场，密切关注着其可能带来的巨大商机。如何在互联网市场占有一席之地，是中国电信各大运营商所必须考虑且迫在眉睫的问题。

1.1　选题背景及意义

加入世界贸易组织后，中国电信市场逐步开放，随着竞争格局的变化和通信技术的发展，中国电信市场正发生着巨大的变化。现在已经不仅仅是三大运营商的天下，越来越多竞争者的加入使得运营商不能仅仅停留在现有产品上，而不得不考虑转变现在的思路，将标准化的产品、业务、套餐等转向精细化、精准化、个性化。“顾客是上帝”仍是要一直遵守的信条，如何让顾客满意或者说如何使产品被顾客所青睐，在现在看来不仅仅是质量、价格等因素的影响。市场竞争者众多，通信的技术水平又所差无几，在质量价格的基础上消费者从众多的产品中自然选择更加适合自己的。当运营商将消费者转换为自己的用户，再将用户转换为忠实用户、满意用户，收效是加倍的，不仅用户离网率大大降低，而这个用户更会积极的向其朋友推荐该运营商，从而达成良性循环。那如何才能达到这种良性循环，了解消费者的兴趣与特征，针对消费者进行精细化、个各运营商也想接着性化、差异化的服务是各运营商目前的关键工作所在。

信息技术的发展，信息服务水平的提高，五彩缤纷、日新月异的众多应用背后带来的是数据爆炸式的增长，我们已经进入了一个大数据的时代。大数据迅猛来临，而我们对其能带来的价值还是不明觉厉。我们目前面对的是难以想象的数据量级的、种类繁杂的数据，这些都是数据宝藏，我们应该学会利用这些宝藏，寻找巨大的价值，为我们的工作生活进行服务。大数据背后价值我们是否能探究挖掘出来，能够展现的是冰山的一角还是半壁冰山是对我们的一个巨大挑战。而我们所追求的精准营销，期望

的个性化推荐是以大数据为基础的。在大数据营销中，主要的价值体现在（1）用户行为和特征分析，积累了足够的用户数据，分析出用户的喜好和习惯，甚至与比用户更了解用户自己；（2）精准营销信息推送，精准营销在互联网行业中被提及很久，但真正做到的少之又少，绝大多数都是打折精准营销的旗号，仍旧是广撒网式，垃圾信息泛滥，用户越来越不满意；（3）引导产品或营销投用户所好，在生产产品或营销活动进行之前，了解潜在用户的主要特征、需求，针对研究结果寻找营销活动的关键，或产品方案的制定及改进等等。针对这些大数据营销中的价值体现，着重对用户行为和特征、精准营销等进行研究和应用，是运营商目前的目标。

目前，电信运营商应用和研究大数据还处在初级阶段，在流量经营、套餐、网络质量、基站建设等方面正在逐步上手研究，在用户分析和精准营销方面还在起步阶段，都在争相建立数据挖掘智能系统，目的在于了解不同消费群体的不同需求，从而快速、准确的找到他们的需求进行个性化推荐，发展新用户并使老用户的满意度提升，增加其对产品的忠诚度，从而争取从各个方面扩大市场占有量。这也是各运营商从标准化、粗放经营走向个性化、精细管理的一条关键道路。

国内各运营商把目标瞄准了大数据，正基于此紧锣密鼓的进行相关研究和应用，但目前也主要体现在移动设备各个相关方面，如基站建设、通话质量、流量推荐等，虽然数据挖掘智能系统已经在着手研究之列，但还处于初级阶段，在实现和应用上还未达到研究的目标，所以以此为方向，做为本次研究内容，通过分析电信用户数据，发现用户的兴趣和网络行为规律，对用户服务的精准化、个性化提供帮助。

1.2 国内外研究现状

大数据的研究在全世界掀起了一股热潮，从互联网行业到商业智能零售业、咨询服务、物流、交通、医疗卫生、生物科技等等，几乎各行各业都加入到大数据的浪潮中，数据服务意识从认识到了解再被研究应用，大数据正影响着我们的工作和生活。

美国政府 2012 年 3 月宣布了“大数据研究和发展计划”，投资 2 亿美元投入到大数据研究中，为提高从海量数据中提取、发现信息的技术水平，更新相应的技术工具，并增强搜集获取海量信息、解析信息的能力，他们认为这有助于提高未来的国际竞争力，并使得国家安全方面更有保障。在商业应用方面，美国一公司把天气预报的数据放在亚马逊平台进行数据处理，通过大数据的技术工具，将天气信息更加准确的预测，帮助农业种植者提高收益。我们也搭建出了一个汇集几千万中小企业的数据平台，对这些企业的日常数据进行挖掘和解析，有助于对我国经济运行状况作出准确的预警，协助相关的国家机关做出决策。金融行业随着大数据的应用也发生了巨大的改变，技术创新、商业模式创新，互联网的领军企业阿里巴巴、腾讯等凭借其雄厚的数据支持

进军金融行业，开发金融产品新的销售模式，开拓出新的盈利点。

随着大数据的发展，越来越多的人意识到大数据对用户行为分析带来的影响。世界上多个研究机构和大学纷纷开展了用户行为分析的研究项目，探讨新的理论，研究新的技术，用户行为分析的研究也从科研方面转向了商业的应用研究。用户行为分析在以下方面取得了一定的研究进展。信息检索方面，从早期的联机公共检索目录，Anick 研究出专业的交互性反馈机制，基于用户的 Web 日志搜索细化检索目录[1]；Zhang Z 和 Nasraoui O 提出了一种数据挖掘的聚类算法用于信息检索和搜索引擎的查询优化[2,3]；Agosti M、Crivellari F 等人对数字图书馆用户的 Web 日志进行分析处理，对信息管理系统进行了研究[4]。用户群体行为特征方面，Murata T 根据网页和超链接，通过网络间关系对用户进行社团划分[5]；Yu X、Li M 等人通过关联规则对 Web 日志进行数据分析，对用户的行为进行预测[6]；付关友、朱征宇从心理学角度对 Web 用户的浏览行为进行分析，研究用户兴趣与网页浏览行为之间的关系，并用线性回归模型进行两者之间关系的证实[7]。在网站结构的优化上，Li D H、Laurent A 和 Poncelet P 通过序列关联规则对用户的 Web 日志数据进行关联分析，完善网站的网络结构，提升用户体验[8]； R ós S A, Vel ásquez J D 等人通过改善网站的文本内容，对用户浏览状况进行分析，然后对网站进行重构[9]；国内也有相关研究，对海量的用户访问 Web 数据进行关联规则挖掘，建立页面与用户访问的模型，对网站结构进行调整优化。在商业应用上，M üller H、 Pun T 和 Squire D 从内容的图像检索法入手，对市场购物篮问题进行分析研究[10]；高琳琦通过模糊相似度分析用户的网络行为，以关键字偏好表示用户兴趣，并进行动态研究，以进行个性化的新闻推荐服务[11]。

电信行业，大数据也被越来越多的运营商进行应用研究，尤其想依靠大数据进行客户维系、精准营销，以提高市场占有率增加收益。2013 年，电信与媒体市场调研公司（Informa Telecoms&Media）发布了一份电信业的调查报告，报告中显示，全球 120 多家运营商中有大约 48%家正在通过大数据进行业务研究。大数据的实施成本在运营商的总 IT 预算中在未来五年中将由现在的 10%升至 23%，全球运营商都意识到大数据将是一个巨大的潜力股，大数据将成为运营商的一大战略优势。

英国 O2 在英国推出了免费 WiFi 的服务，目的是积累客户以获得更多的用户数据，通过大数据的处理、数据挖掘，进行更精确的广告营销等服务。NTT Docomo 将用户表格进行改进优化，把表格精细制作，获取用户多方面的信息，加强 CRM 系统和知识库的数据积累，对目标用户进行精准营销服务，业务办理的成功率大大提高。西班牙电信在 2012 年 10 成立了动态洞察部门，针对大数据业务对用户行为进行分析。他们推出的一款“智慧足迹”的产品是将用户数据匿名化，对用户的行为足迹与地点人流量等因素进行结合分析，产品面向政企客户。美国的 Verizon 运营商建立了专门的精准营销部门，对用户群进行大数据分析，提供精准营销等服务，再将有价值的信息提供

给政府或企业获得盈利。

中国运营商中四川移动分公司与亚信公司合作，完成了“个人服务顾问”的精准营销项目，项目实施后，数据大幅度提高，增长近 50%，业务推荐成功率也高达 60%，人均办理成功量比项目之前也翻了一番[12]。大数据的精准营销在中国电信业也进入了慢慢进入研究应用之列，某些运营商分公司在进行精准营销服务后取得了良好的效果和收益。

但目前中国运营商在营销方面大多数还是进行的传统营销，采用大众营销的模式，通过传统媒介进行轰炸式、地毯式的广告推送和宣传[13]。这种模式虽然范围广，但成本高并且无法精确定位目标用户，不仅使用户反感，投资回报还不成正比。传统的营销模式已经渐渐不再适用于现在的需求，需要将粗放化转向精细化、个性化。

总体来说，中国的电信运营商在大数据应用方面还处于初级阶段，尤其是面对大数量级的用户数据不知从何处入手，又没有相关的经验，在用户特征分析方面刚刚涉及，不仅要学习国外运营商的成功经验，还要对互联网中其他行业在此方面的经验进行学习。在保障用户个人数据保密及安全的情况下，对用户特征进行分析后的应用主要还是要用于精准营销。虽然中国电信业精准营销项目成功应用的还比较少，但给电信业带来的效果和收益却是毋庸置疑的，各大运营商也坚信这一点，纷纷加入到大数据应用的行列，借助海量的用户数据对用户进行分析，争取获得更多的市场占有率。

1.3 研究目标与内容

大数据给电信业带来全新的发展和变革，各运营商争相投入研究，但都还处于初级阶段。大数据用户分析在电信业应用广泛，精准营销作为其主要的应用，已经被运营商放在重视之列。由于没有足够的经验和技术，取得成功的还在少数，大多还在研究及开发阶段。因为获得了电信某省用户的移动 DPI 数据的研究权限，所以通过用户的访问数据对用户进行分析。

本文的研究目标是分析电信移动用户的网页访问行为，从时间维度、业务应用竞争结构、基于用户兴趣的市场细分三个维度研究用户的网络行为特征规律，建立用户行为模型，结合用户的行为规律进行分析，总结各维度的规律并说明在现实中的特征应用。

本文的研究对象是电信某省的移动用户，为了符合目前各运营商的大数据发展方向，对大数据重要应用精准营销提供依据，通过分析大数量级的用户移动 DPI 数据，选择用户属性对用户进行分析处理。针对数据属性及数据实际情况和对用户行为分析文献的研究，选择用户的访问时间进行统计，研究用户的访问规律；选择业务应用的访问量进行统计，分析每类业务市场的竞争结构；通过网页分类进行用户兴趣的匹配，

对用户添加标签，通过 Phantom 联合聚类算法对用户聚类，进行市场细分，总结每类用户的特征并介绍在精准营销中的应用。本文采用的技术路线如图 1.1 所示。

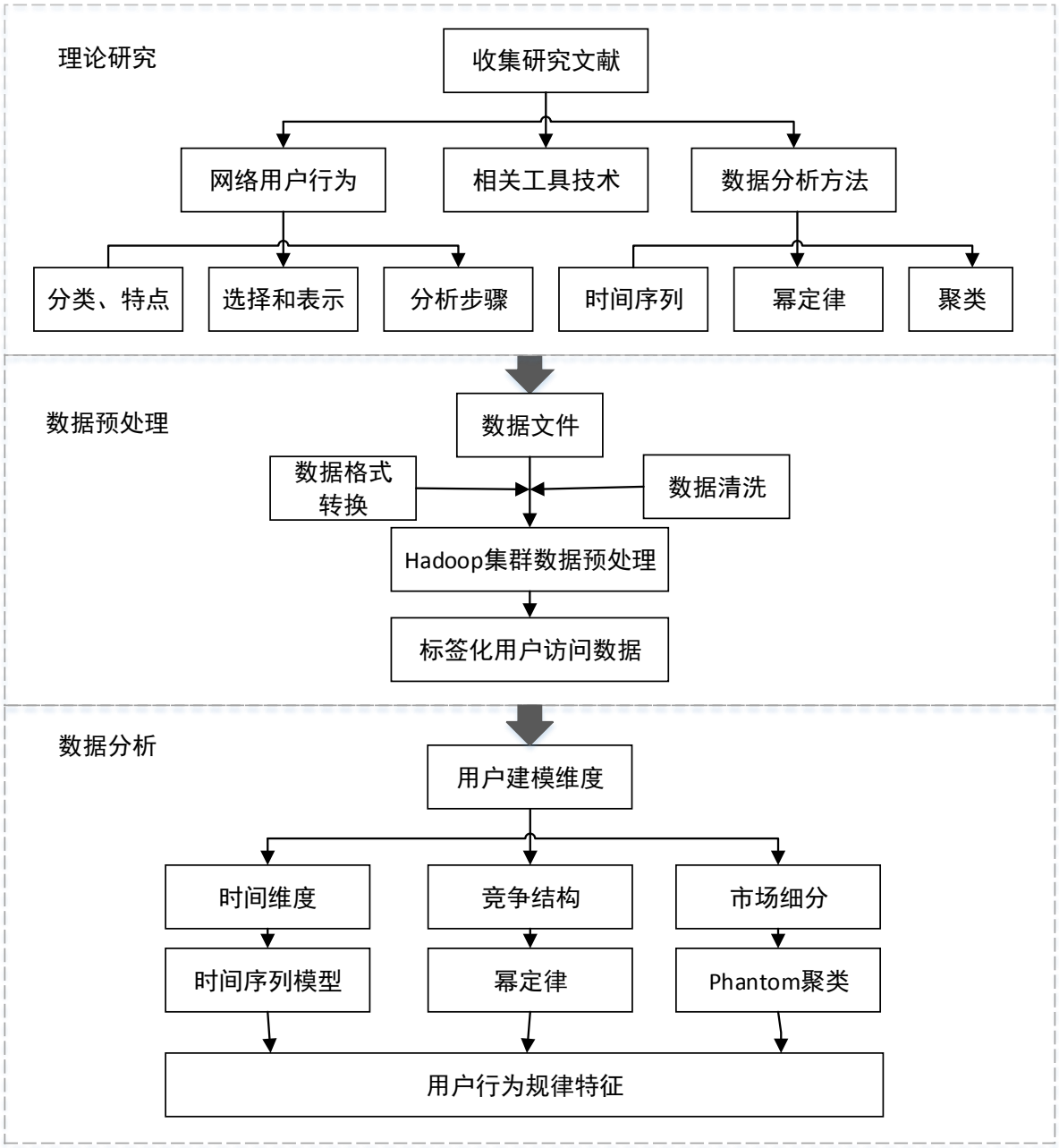

图 1.1　技术路线图

1.4　论文组织结构

全文总共分为七章，具体结构安排如下：

第一章 阐述了研究本文的背景及意义，在大数据发展情况下国内外电信运营商的

发展现状，介绍了本文的研究目标、内容和组织结构。

第二章 相关理论的研究与介绍，从网络用户行为分类、特点、特征选取和表示对用户行为进行了研究，然后对时间序列分析、幂定律、聚类算法进行了研究，接着对Hadoop 技术与 hive 工具进行了研究与介绍。

第三章 研究设计，首先从建模的目的、步骤、维度对用户建模方案的设计进行介绍，然后对数据来源、内容进行了介绍，接着对数据预处理工作的步骤和内容进行了说明，最后是对本文用到的数据分析方法进行了研究。

第四章 时间规律分析 先对用户访问时间序列进行了介绍，接着从日访问规律和小时访问规律对时间序列进行统计分析，然后拟合出时间序列模型，最后对本章得到的规律和特征进行了总结。

第五章 竞争结构分析 首先对业务应用、竞争结构进行了介绍，然后对各业务在其市场的占有率进行了统计分析，接着对各类市场中的竞争结构进行统计，分析出各类业务应用市场的具体情况、竞争激励程度和垄断程度，最后对竞争结构的规律和市场特征进行了总结。

第六章 市场细分 通过用户兴趣偏好对市场进行细分，先对问题进行了描述，然后对聚类算法的步骤及实现进行了说明，接着介绍了市场细分结果，并对结果进行了分析，最后对细分结果进行了规律和特征总结。

第七章 结论，总结了本次论文所做的工作及对后期工作的描述与展望。

第二章　相关理论与技术研究

2.1　网络用户行为研究

网络用户，即为网络的使用者，是指其终端设备（计算机、移动终端等）连接上互联网进行网络活动的人。如今是互联网时代，网络用户大量增长，他们随着互联网的发展，其借助互联网进行的相关行为和活动的行为方式和特征也逐渐呈现，这就是网络用户行为。因为网络用户的行为方式多种多样，所以想对其做一个明确的划分和归纳不太容易，随着大数据的迅猛发展，网络用户行为的研究成为一个热门的课题。

网络用户行为研究与心理学、社会学、哲学、人类学、经济学以及与网络行为相关的一切学科都密切相关。通过对网络用户的行为进行研究，找出其规律特征，用来预测或控制用户的网络行为，并实现其研究目标。具体来说，网络用户行为研究就是分析网络用户的构成、用户特点及其行为规律。

2.1.1　网络用户行为分类

网络用户行为根据对象数目和研究目的的不同分为个体用户行为和群体用户行为。个体用户行为是只单个用户的网络行为，网络用户的行为千差万别，与其个性、心理、性格、价值、兴趣、环境等等多个方面相关联，这些都造成了用户个体的差异和需求的不同，但同时也具有一定的稳定性。短期的个体行为可能总结不出什么规律，但长期的用户个体网络行为会具有稳定性，能够发现一些规律。多个用户个体就组成了群体，根据特定的研究目标，可以将多个用户划分成不同的群体。这些群体在网络中的行为，成为群体用户行为[14]。将群体用户以学科领域进行划分，可以分为社科、文艺、科技、综合等用户群体；以职业进行划分，就有老师、医生、律师等多种职业用户群体；以信息利用状态进行划分，有现时用户群体和潜在用户群体之分；以年龄、性别等进行划分又可划分出不同的用户群体。群体用户行为能够比较明显的表现出规律性，对群体用户网络行为进行分析，能够较好的发现用户网络行为的共性。

根据网络应用的不同，将网络用户行为分为基础网络行为和拓展网络行为，再细分将网络行为分为五类。基础网络行为分为信息获取类和交流沟通类，拓展网络行为分为休闲娱乐类、电子商务类和电子服务类。由于互联网的共享性和开阔性，网络资源得到快速共享，信息获取成为网络用户的主要行为之一，搜索引擎作为基础应用，是网络用户获取信息的重要工具之一。互联网使用户实现实时沟通，交流方式的多样性、时间距离无约束性，使得交流沟通成为网络用户的重要行为。随着互联网的快速

发展，生活水平的提高，网络用户的行为不仅仅停留在基础网络行为上，用户利用网络的方便、快捷、搞笑，越来越多的网络用户在网上进行休闲娱乐、电子商务和电子服务，网络用户的生活也因互联网变得更加丰富多彩。

根据网络访问形式不同，从源 IP 和目的 IP 为切入点，将网络用户行为分为一对一方式、一对多方式、多对一方式和多对多方式。一对一方式表示为某单一用户对同一目的 IP 多次访问，即高频次的访问统一网页，可以表现出此用户对特定网页的喜爱程度。一对多方式就是同一 IP 访问多个目的 IP，即浏览多个网页，根据浏览不同网页的次数不同，也可以分析出源 IP 用户的兴趣偏好。多对一方式是通过用户 IP 与目的 IP 的访问关系，分析网络用户对此目的 IP 对应网站的关系。多对多方式就是对整个网络的内容进行研究。

根据网络用户行为是否有益进行分类，网络用户行为分为有益行为和无益行为。根据网络用户行为是否符合常理，又分为正常行为和异常行为。

2.1.2 网络用户行为特点

由于互联网的开放、自由、免费等特性，是一个虚拟空间并具有特殊性，不同于一般的物理空间，所以网络用户行为不同于一般的用户行为，具有其自身的特点。通过对网络用户行为的研究，将其大致分为以下几点。

（1）知识含量高。网络用户一般都具备一定的知识积累，掌握一定的计算机知识和网络知识，具备使用常规软件的能力，进行网络活动。

（2）隐蔽性强。这种隐蔽性一方面是指用户的隐蔽性，任何人都可以使用网络查询信息、传播信息，无需登记，匿名操作。另一方面网络本身也具有隐蔽性，网络上的信息存储传递都通过计算机语言进行，操作者在数据传输过程中就可以更改信息且不留痕迹。

（3）主动性强。网络行为受用户个体的个性和主观意识，主动的去完成网络行为，较少受到他人影响。

（4）内容丰富。由于互联网的丰富多彩，网络应用的纷繁不一，用户的网络行为也变得很丰富。

（5）特点鲜明。同一网络行为，因为用户的不同，时间地点的不同，最后行为的结果也是各不相同的，由于各种因素，网络行为呈现出不同的鲜明特点。

（6）判断标准不一。互联网是无国界的，它将全球变成了一个地球村，但无国界空间中网络行为，也会涉及到不同国家地区的利益和标准。

（7）随机性和规律性。对于个体网络用户来说，其行为特点、行为属性都存在着较大的随机性，而对于群体网络用户的行为又可总结归纳出一些规律性。

2.1.3　网络用户行为特征选择和表示

网络用户行为的千差万别，决定了网络用户行为特征的表示多元化，网络用户行为特征的表示其实就是特征的选择，选择哪种或者哪些特征，与研究目标与研究场景相关联。

国内对特征选择的研究主要集中在理论和 Web 服务方面，比如马力、焦李成等人系统的对 Internet 用户的行为特征进行了研究，从检测分析的角度，根据用户兴趣度、事务等概念定义出一种网络用户行为的分类和表示方式[15]。

国外对这方面的研究更注重应用，也比较系统和深入。他们注重研究用户行为的角度，定性的选择出用户的行为特征。Paolo G 进行的研究就是对宽带进入家庭后，网络用户的行为发生的变化情况[16]。Xu K、Bhattacharyya S 等人对由于宽带的兴起，带起的 P2P、宽带视频等应用对用户行为特征的影响进行了研究[17]。真正系统的研究网络用户行为的还是比较少，Marques Nt H T 等人从运营商的角度把网络用户分为两类，家庭和办公，分别分析了两类用户的网络会话访问行为[18]。Fukuda K、Cho K 等人分析出日渐增长的流量是由家庭宽带用户为主带来的，70%的用户上网流量是稳定的，并随着用户的生活习惯发生规律性的变化[19]。Maia M、Almeida J 等人对特定网络环境和社交网络下的用户进行了分析，文章提取了 9 个行为特征对网络用户行为进行分析，通过聚类得到了 4 中主流的用户分类，并对结果进行了相应的分析[20]。

由于网络用户行为分析的多样性和复杂性，研究分析会借助一些工具和方法，表示起来一般会采用属性向量的方式。具有 n 个属性的网络行为，可表示为（属性 1，属性 2，…，属性 n），其中 n 个属性都是网络用户行为的特征，根据研究目标，选出特定的几个属性进行分析研究。比如研究网络用户在电子商务中的消费行为，可以选取特征属性为（用户 ID，时间，购买物品名称，物品种类，购买数量，金额等）；比如用户浏览网页的行为，可以选取特征属性为（用户 ID，网页 url，时间，次数等）。选取的特征不是越多越好，根据应用场景和研究目标，取其精华，适当的选取网络行为特征，才能为后续的研究分析奠定基础。

单个用户的网络行为用属性向量进行表示，群体用户的网络行为通过属性矩阵进行表示。k 个用户的 n 个相同网络行为属性，可表示为：

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{pmatrix}$$

其中 n 个属性表示网络用户行为的 n 个特征，x_{kj}表示第 k 个用户的第 j 个属性特征。例如包含 k 个用户的用户群的浏览行为可以表示为：

$$\begin{pmatrix} \text{用户 ID}_1 & URL_{11} & \cdots & \text{时间} & \text{次数} \\ \text{用户 ID}_2 & URL_{21} & \cdots & \text{时间} & \text{次数} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \text{用户 ID}_k & URL_{k1} & \cdots & \text{时间} & \text{次数} \end{pmatrix}$$

2.1.4 网络用户行为分析主要步骤

网络用户行为分析是从确定研究目标到评价模型、分析结果的一系列过程，主要包括以下几个步骤，步骤如图 2.1 所示：

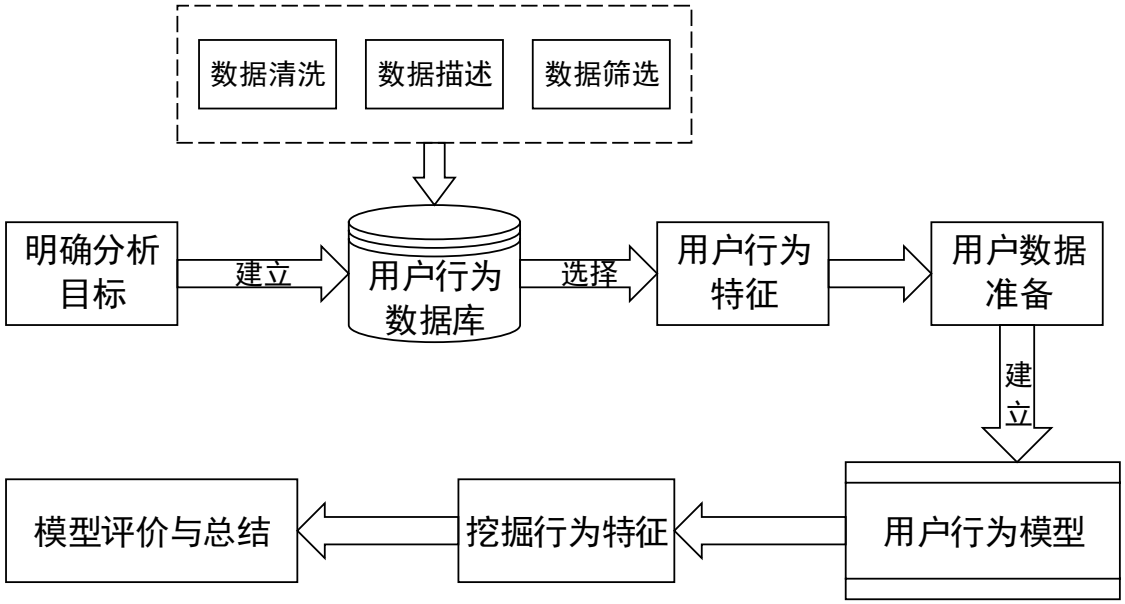

图 2.1 用户行为特征分析步骤

（1）明确分析目标

确定网络用户行为分析的目标是对用户行为进行精准定位和正确分析的第一步，也是至关重要的一步。针对不同场景解决不同问题所建立的用户模型也是大不相同的，所以明确目标是行为分析的基础，只有目标定位明确，模型的建立和分析的正确才会得以实现。

（2）建立用户行为数据库

将所要研究的用户数据导入到服务器或计算机中，形成课题研究单独的数据库，为保证数据的稳定性、一致性、安全性等，只将所用数据单独存放。数据收集后，进行数据清洗、数据描述、数据筛选等工作。

（3）选择用户行为特征

针对分析目标，将数据库中用户行为的特征进行分析，选出最具代表性的数据字段，并导出相关字段的数据，另存到数据库中。

（4）用户数据准备

在建立模型之前最后一步工作，包括提出无用数据记录、选择用户记录、创建变量、转换变量等，将数据更精细化的进行清洗及转换。

（5）建立用户行为模型

基于网络用户行为分析的目标及前期的数据准备，确定模型建立的角度和维度，选择合适的方法或工具，建立符合目标的用户模型。模型的选择和建立直接影响着后期的分析和研究。

（6）挖掘行为特征

通过建立好的模型进行用户行为数据实施，针对不同用户行为的角度的模型，采用合适的方法或工具，分别进行探索和分析，挖掘出用户特征所在。

（7）评价与总结

对模型的结果和分析的特征进行规律和知识总结，解释其价值所在，并对此次分析进行评估。

2.2　时间序列分析研究

时间序列是指将某一统计指标按照时间先后的顺序进行排列，其数据形成的数列即为时间序列。例如国内生产总值作为统计指标，按照年份进行先后排序，形成的数列；统计某类商品的销售数据，按照天或月份进行排序，形成的数列等都称为时间序列。针对本文来说，将电信用户在移动端进行的网络访问量作为统计指标，将统计数据按照天进行排序，形成的序列为时间序列。

时间序列有四种构成因素：长期趋势、季节变动、循环变动和不规则变动。长期趋势是指受某种根本因素的影响，长期内形成的某种固定的总体趋势变化。季节变动是指随着季节因子的变化而形成的规律性的周期变化。循环变动是指若干个变动周期内，相同或相似的重复性规律变化。不规则变动是指没有规律可循的变动，包活严格的随机变化和突发的不规则变化。

时间序列分析是指对某一时间序列进行分析，寻找出四种构成因素的存在形式，建立对应的数学模型。目的在于探索统计数据相对于时间的变化规律，并对将来的数据进行预测。

研究时间序列的规律，需要对时间序列进行分析，寻找出四种因素的存在状态与形式，并建立时间序列模型。时间序列模型一般包括以下四种类型：

（1）自回归（AR）模型

如果一个线性随机过程能用公式

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t \tag{2-1}$$

表达出来，其中 ε_t是白噪声，$\emptyset_i(i=1,\ldots,p)$是回归参数，则这个线性随机过程$y_t$可称为$p$阶自回归模型，表示为 AR($p$)。$y_t$的值为它的$p$个滞后变量加权和与$u_t$的和。

若使用滞后算子表达成

$$\left(1-\emptyset_1 L-\emptyset_2 L^2-\cdots-\emptyset_p L^p\right)y_t=\Phi(\mathrm{L})y_t=\varepsilon_t \tag{2-2}$$

其中$\Phi(\mathrm{L})=1-\emptyset_1 L-\emptyset_2 L^2-\cdots-\emptyset_p L^p$可称作自回归算子或特征多项式。

平稳性问题是常与自回归模型联系到一起，对于 AR(p)，如果该自回归过程的特征方程

$$\Phi(z)=1-\emptyset_1 z-\emptyset_2 z^2-\cdots-\emptyset_p z^p=(1-G_1 L)(1-G_2 L)\ldots(1-G_p L) \tag{2-3}$$

中的所有根的都小于-1 或者大于 1，则自回归过程 AR(p)是平稳的。其中${G_1}^{-1}$，${G_2}^{-1}$，$\cdots{G_p}^{-1}$是特征方程$\Phi(\mathrm{L})=0$的根。

自回归过程 AR(p)中最常用的两个过程是自回归过程 AR(1)和自回归过程 AR(2)。AR(1) 自回归过程是一阶自回归过程，即

$$y_t=\emptyset_1 y_{t-1}+\varepsilon_t \tag{2-4}$$

AR(2) 为二阶自回归过程，即

$$y_t=\emptyset_1 y_{t-1}+\emptyset_2 y_{t-2}+\varepsilon_t \tag{2-5}$$

（2）移动平均（MA）模型

如果我们将时间序列看作白噪声的线性组合，那么即产生出移动平均模型。通常表达成

$$y_t=\mu+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q} \tag{2-6}$$

式中y_t为在t期时间序列观测值，μ是不变常量，q是滑动平均模型阶数，一般应用情况下q的范围在 l~2 之间；表达式中表示过去的q个周期的随机扰动项值通过的加权平均计算就是其时间序列，所以称作移动平均（MA）模型。

（3）自回归移动平均（ARMA）模型

当建立一个实际时间序列的模型时，经常可能产生一个模型，同时包含自回归和移动平均参数。这时该模型为(p,q)阶自回归移动平均模型。该自动回归移动平均模型通常表示为：

$$y_t-\emptyset_1 y_{t-1}-\emptyset_2 y_{t-2}-\cdots-\emptyset_p y_{t-p}=\mu+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q} \tag{2-7}$$

y_t为t时刻时间序列的相应，从中可以看出y_t既受到因变量自己在之前的某个时刻的值的影响，也受到系统对其在某个时刻进入时的扰动影响。记为 ARMA(p,q) 自回归移动平均模型。等式左边部分为 ARMA 模型的自回归，p是一个大于等于 0 的参数，

表示模型自回归的阶数，实参数$(\emptyset_1，\emptyset_2，…，\emptyset_p)$为自回归的系数。等式右边部分是ARMA 模型的移动平均，非负参数q是移动平均的阶数，$(\theta_1，\theta_2，…，\theta_q)$是移动平均的系数。当$p=0$时，此模型为 MA 模型；当$q=0$时，此模型为为 AR 模型。

（4）自回归积分移动平均（ARIMA）模型

前面三种模型都是基于时间序列平稳的这一假设而建立的， 即该随机过程的期望为常数，协方差只和时间间隔有关，这样的随机时间序列就称为平稳的时间序列。但许多现实中的时间序列都不具备平稳性， 它们往往带有趋势性，这时就要用到 ARIMA 模型。

ARIMA 自回归积分移动平均模型，能把非平稳时间序列转化为平稳，能将因变量只对因变量自身滞后值和随机误差项现值与滞后值进行回归。

对于模型 ARIMA $(p，d，q)(P，D，Q)$参数估计，它是指利用有关的样本数据，对已选出的模型参数进行估计，也就是要估计出p个自回归参数$(\emptyset_1，\emptyset_2，…，\emptyset_p)$，$q$个移动平均参数$(\theta_1，\theta_2，…，\theta_q)$，$P$个季节自回归参数 $(\Phi_1，\Phi_2，…，\Phi_P)$以及Q个季节移动平均参数$(\Theta_1，\Theta_2，…，\Theta_Q)$的数值。当时间序列的模型结构和阶次初步确定后，下一步工作就是估计求解模型参数$\emptyset_1，\emptyset_2，…，\emptyset_p，\theta_1，\theta_2，…，\theta_q$。因为模型的结构不同，要求的预测精度不同，需要的特性统计也不同，所以用来估计参数所使用的方法也不一样。常用的三种参数估计方法是非线性最小二乘估计法，广义最小二乘方法，还有矩估计法。非线性最小二乘法的实际计算比较复杂，参数估计时经常需要多次尝试不同的初值，而广义最小二乘法相比较而言更为简单。矩估计需要根据相关的函数来计算估计的各个参数值，但是其不要求通过函数计算而来的参数估计值必须满足特定的最优化的条件，也被称作粗估计。

2.3　幂律分布研究

幂律分布是指以y为因变量，x为自变量，y随着x变化而变化形成的函数关系可以用$y=cx^{-a}$形式表示，成此函数关系符合幂律分布。其中x表示排名，y表示该排名下的统计量的数值，c和a为常数且都大于 0。将函数的等号两边同时取对数，得到$\ln y=\ln c-a\ln x$，此时幂律分布函数变为一条直线，直线的斜率为负，值为$-a$。在双对数坐标下看函数是否为斜率为负的直线，是判断函数分布是否为幂律分布的一个重要方法。

在幂律分布研究中最为著名的是 Zipf 定律和 Pareto 定律，Zipf 和 Pareto 是幂定律研究中的突出代表[21]。其中 Zipf 为哈佛大学的语言学家，1932 年他在研究英文单词在文章中出现的次数时发现，如果把英文单词按照出现的次数从大到小进行排列，单词出现的次数与其所在排名的幂次（幂次为常数）关系呈反比，这个发现后来就被称为

Zipf 定律。Zipf 定律表明在英文中，极少数的单词会被频繁使用，而绝大多数的单词使用次数很少，这种规律在汉语等其他语言中同样存在。Pareto 为意大利经济学家，1877 年他在研究个人收入的统计分布时发现大部分财富流向了少数人的手中，而剩下的小部分财富分到了大部分人的手里。然后他继续研究，发现了著名的二八定律，即社会上 20%的人掌握了 80%的财富，社会中的财富分配是极不均衡的，这种不均衡的现象也称为 Pareto 定律。

幂律分布广泛存在于物理学、计算机科学、生物学、金融学、社会学等众多领域。在统计物理学中，物理统计学家通常把符合幂律分布的现象成为无标度现象，系统中个体的尺度范围很大，个体与个体间尺度相差很多，有竞争、有生命、有进化的地方都会有不同程度的无标度现象。

随着计算机技术迅猛发展，数据工具越来越多种多样，金融数据的研究变得更为容易，更多的学者投入到经济物理领域，研究的数据范围越来越广泛，模型的变化和改进也越来越丰富[22-24]。 Ander 对多个国家的股票市场进行研究分析，发现新兴市场中的投机现象能够用幂定律分布进行解释[25]。司马则茜等人基于银行操作缺失数据进行了 POT 幂律模型的拟合，验证其符合幂定律分布并对操作风险进行了思路说明[26]。吉翔等人把我国股市中泡沫与反泡沫基于对数周期性幂律模型进行了研究，得出我国股市符合对数周期性幂律分布[27]。

除了经济物理学领域，许多其他领域的研究都证明有很多符合幂律分布的规律存在。尤其是在互联网领域中，面对庞大复杂的互联网，有许多方面的规律竟然也与幂律分布十分契合。山石等人在研究中指出互联网中出现新的网页，会有一定几率链接到其他已存在的网页中，如果某个网页与其他网页的连接越多，他与新网页链接的可能性也越大，这样网页链接次数就形成了一个幂律分布[28]。杨波等人对 CNN 模型网络和爵士音乐家网络进行了幂律分布研究，并用最大似然估计的方法对幂函数的指数进行了估计[29]。刘臣等人对近三十年中文期刊的引用关系进行了研究，证实了中国学科知识网络具有幂律分布特征[30]。叶作亮等人对 C2C 商家的交易记录进行统计分析，发现顾客重复购买次数符合幂律分布，并建立了用户购买概率模型[31]。

电信业中对幂律分布的研究还是比较少的，已有的研究也是集中在移动通信方面，对移动信道、移动通信场或对移动通话次数进行的研究，而针对电信用户的网络访问数据的研究几乎没有。本文通过电信用户的上网记录对进行每类业务的总体访问量进行幂律分布研究，从而发现各类业务市场的竞争结构。

2.4 聚类算法研究

用户是任何企业发展的源泉，将潜在消费者转化为自己的用户，并使用户成为企

业或某种产品的忠实用户是企业一直努力的方向，运营商也不例外，尤其是在越来越同质化的产品面前，怎样使用户满意度最高是关键所在。想要挖掘新用户并留住老用户，就要对市场细分，对用户分群。用户分群就是通过用户的行为特征和数据信息对用户的特点进行准确描述，然后对用户进行区分，识别不同类用户的不同需求，以便对用户进行更为准确的服务和营销。

对用户分群通常采用数据挖掘中的聚类算法，通过用户的属性特征将用户分成多个群组，相同群组间的用户属性特征较为相似，而不同群组间的用户特征差别比较大，这样就可以对不同群组的用户分别进行服务和营销，也就是用户的精准营销。

聚类算法是一种常用的数据分析工具，其目的在于将大量的数据点组成的集合分为若干个类，最大程度的使同一类中的数据点相似，而不同类中的数据点不同。聚类算法一般分为层次聚类、分割聚类、基于约束的聚类、机器学习中的聚类和高维数据的聚类[32]。

层次聚类算法通过将数据集合组织成若干组并通过树状图来进行聚类，一般分为自顶向下的分解层次聚类和自底向上的聚合层次聚类，代表算法有 CURE 算法、ROCK 算法、CHAMELEON 算法等。层次聚类算法的优点为适用于任意形状和任意属性的聚类，可以灵活控制聚类层次的粒度，有着强聚类能力；但缺点为时间复杂度高，不能进行回溯处理。

分割聚类算法是先将聚类集合随机分成 k 类，通过某个聚类控制准则从初始的 k 类进行重复迭代，当达到最优为止。分割聚类算法又分为基于密度、基于网格、基于图论、基于平方误差的迭代重分配四种聚类算法，四种算法中的典型算法分别为 DBSCAN 算法、STING 算法、谱聚类算法、k-means 算法等。其中 k-means 算法是目前应用最为广泛的聚类算法，算法的优点为时间复杂度低，运算速度快，对于处理大的数据集时该算法是高效的，当数据为密集的，且不同类之间的差异较大时该算法的效果好；缺点为在聚类之前先要确定聚类的数目 k，异常值对算法的影响大，只能处理数值属性的数据。

基于约束的聚类包括对数据对象的约束、对聚类参数的约束等，因为现实中的问题往往是存在多种约束条件的，如果在处理过程中不能加入这些约束，聚类的结果对现实问题就意义不大。基于约束的聚类算法中比较有代表性的为 COD 算法，其用用两点之间的障碍距离取代了一般的欧式距离。基于约束的聚类算法通常只能处理特定领域中的特定方向的问题，并且需要具备一定的经验知识来对算法进行约束。

机器学习中的聚类算法是指采用了某种机器学习理论的聚类算法，主要包括基于人工神经网络和基于进化论的聚类算法。这些算法的优点为利用相应的启发式算法得到高质量的聚类结果，缺点在于有较高的算法复杂度，并依赖与经验参数的选择。

高维数据聚类是目前数据挖掘领域中重要的研究方向，大数据带来的不仅是巨大

的数据量级，数据属性也更加的多样化，数据的维度也变得更高。高维数据聚类的困难来源于高维属性中无关属性的存在使得数据失去了聚类趋势，且高维数据不同组间的界线变得越来越模糊。对高维数据的处理除了降维，还有子空间聚类和联合聚类等。子空间聚类算法是对同一数据集的不同子空间进行聚类，对数据特征选择的任务进行拓展[33]。联合聚类算法是对数据点和他们的属性同时聚类，如 Spectral Graph Partitioning 算法在解决文档和单词的聚类问题时提出了基于双向划分图和最小分割的方法，并揭示了联合聚类与图论划分的关系。

本文研究的数据为电信业移动用户的上网数据，对用户分群时针对其上网的兴趣进行划分，通过数据预处理后，数据仍具有数据量大、维度高、数据矩阵较为稀疏等特点，对上述五类聚类算法进行比较，选择了高维数据聚类中的联合聚类算法对数据进行聚类。

2.5 相关技术介绍

2.5.1 Hadoop 技术

本文所研究数据的数据存放和数据处理过程都在 Hadoop 平台上，所以先对 Hadoop 进行简单的介绍。

Hadoop 是源于 Apache 下的 Nutch 形成的一个开源分布式计算平台，在 Nutch 项目受到 2004 年 Google 在会议“操作系统设计与实现”中发表的文章的深刻影响后，通过尝试整合了 Nutch Distributed FileSystem 和 MapReduce 这两大核心技术而形成了之后 Nutch 的引擎核心，并且因为两者在其中起到显著的作用，从而把他们分离出来整合成为当今的 Hadoop 平台。随着现在互联网的迅猛发展，每个人在互联网当中的数据信息也越来越多，各种形式的网络信息也在非常迅速的增长，这么多的数据堆放在一起处理起来是非常困难的。而 Hadoop 系统最为善长的就是处理海量的数据。并且已经在现在出现的越来越多的新兴的业务中担任了重要的角色。在处理海量数据时，Hadoop 可以通过分布式特殊的处理方式以及存储方法让数据存储更加快速，更容易扩展；可通过 MapReduce 在计算机集群中高速的并行处理并且能保证数据处理的正确及数据处理的安全。

Hadoop 能够让用户非常轻松的搭建起一个具备处理海量数据处理能力的应用系统，优势有以下几个方面：稳定性，数据存储和数据处理的性能是已经经过检验的；可扩展，可以轻松完成计算机节点高数量级的扩展，并把把 Job 分配到这些计算机集群中快速完成；效率高，系统可以快速的移动各个关联节点间存储的数据，从而能够保证各个计算机节点任务高效完成；可容错，能够自动分配执行过程中出错失败的 Job 和自动完成备份工作。

Hadoop 两大核心系统是 HDFS(Hadoop 分布式文件系统)，具有安全性，也具有可扩展性，并且是基于 Java 开发成的一套系统，具有可以方便的部署在配置较低的计算机上的优点；MapReduce，能完成高效并行执行数据处理任务，可以让用户更加方便的完成并行应用的开发。

分布式文件系统 HDFS 结构为 Master、Slave 模型，可以把需要存储的用户数据信息通过文件形式来管理维护。因为 HDFS 能够在单位时间内处理大量的数据信息，所以该系统结构非常适用于需要处理大量数据的情况。

HDFS 分布式系统的主要特性：文件存储的时候需要以 Block 为最小单位把文件分割，之后保存到一个或者多个文件系统的节点上。其备份也会备份到其他一个或者多个数据节点上，为了数据的安全考虑；数据存储节点间磁盘占用率值保持均衡的状态，Hadoop 可以使用一些命令完成对系统的配置参数 Threshold 进行设置，从而保证每个数据节点的磁盘的数据存放比例，所以当某个数据节点的磁盘占用率已经很高的时候，会把盘上的块移动到其他节点上；上传数据，当用户需要上传数据到分布式文件系统中时，先把一个数据块完整的存储到要传输到的一个数据节点上，之后通过该数据节点备份到其他多个节点上，然后才会开始写入后续的数据。检查数据，检测数据是通过 CRC32 完成的。

整个结构中仅仅只有一个 NameNode，这不仅使的系统结构更加清晰，而且所有的用户对数据信息的操作请求都可以通过 NameNode 管理关联到其他各个节点的存储数据；DataNode，就是数据信息操作的节点。HDFS 的结构如图 2.2 所示。

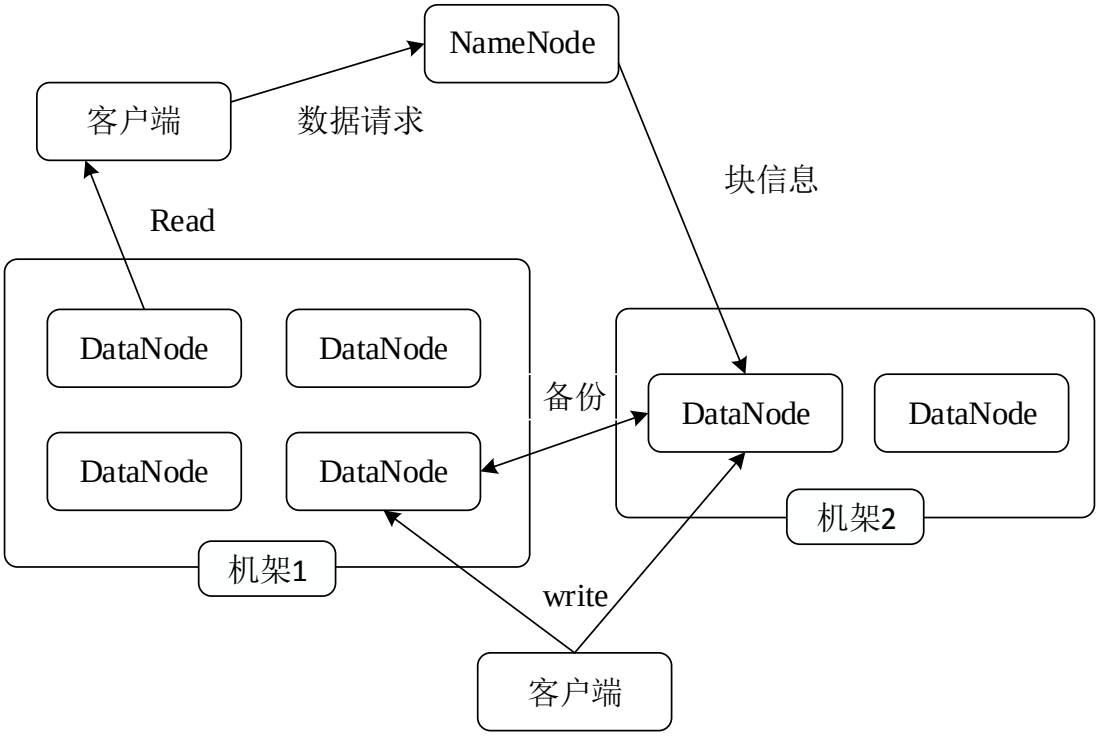

图 2.2　HDFS 体系结构图

MapReduce 的任务处理流程介绍：MapReduce 把用户需要的针对数据的操作分别拆分整合到了 Map 和 Reduce 这两个函数上完成。MapReduce 中管理所有用户提交上来的任务和各个节点的执行情况的为 JobTracker，也就是所说的 Master，负责在各个计算机节点处理执行从 JobTracker 分发下来任务的为 TaskTracker，也就是所说的 Slave。MapReduce 处理数据的一个 Job 的执行过程如下图，把用户数据根据适合的规则切分，由于分布式文件系统可能因为减少文件大小、提高数据节点间传输效率等原因对数据文件进行一些压缩等操作，所以数据处理时如果是压缩文件还需解压。对于普通数据文件可以根据固定大小的数据块，也可以根据文件中的行作为标准，数据切分后，根据对应函数功能完成 Writable 接口类型输入一个键值对到 map 函数中去，map 函数处理完成后按照规则把结果输出到 reduce 函数，完成数据的整理合并，最终输出结果完成任务。

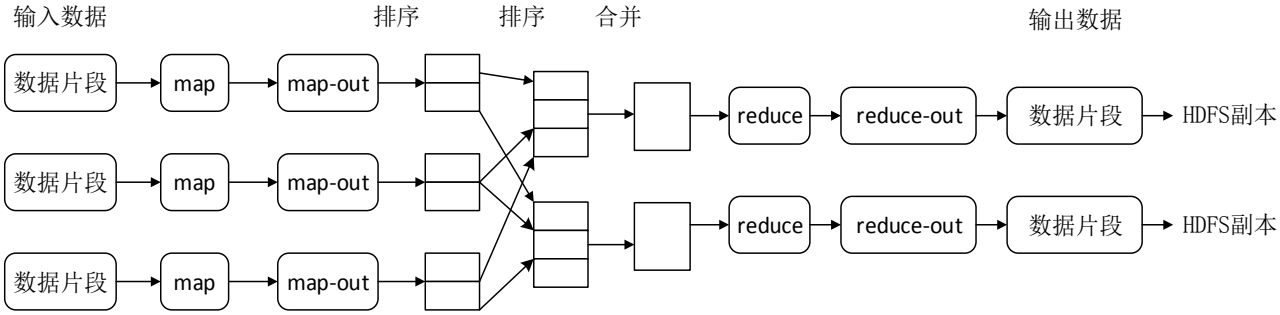

图 2.3　MapReduce 数据流图

一般在利用 JobClient 的 runJob 函数提交并运行 Job 前，必须利用 Hadoop 提供的函数接口确定需要处理的分布式平台上数据的路径，确定数据文件的输入格式，确定 Job 处理结果的输出路径等。调整平台相关配置参数，用 Java 语言实现并完成 Map 以及 Reduce 接口及相关函数功能。完成以上工作后，便可以把程序打包并通过指令在平台上提交运行。

MapReduce 原理如下：

（1）分块保存：使用时我们可以可在每一个计算机节点上进行文件创建、文件内容查看等操作，但事实上这些我们处理中看到的文件都以 Block 为单位保存在磁盘上。当我们需要对数据进行处理时，就要通过名字节点查看文件的 Block 的数量，数据块存储的系统路径，以及先后顺序等。

（2）并行处理：Hadoop 中的运行在主节点的 JobTracker 处理用户提交的 Job，并分发到各个从节点中去，JobTracker 同时维护各个从节点提交上来的处理信息。如果异常停止等情况发生，会重新启动或者将任务交给其他的可用的 TaskTracker 执行。

（3）当前节点处理：当某个节点上的数据需要被用到时，这台机器就负责计算本地的数据，这样可以减少数据传输，减少对网络的压力。虽然计算机节点是可以无限

扩展的，可提供很高的计算能力，但是数据在网络上传输会受到网络带宽的限制。

（4）数据分割整合：数据分割是把 Map 函数运行出来的数据，使用一个自己的规则把 key 分成 Reduce 任务个数份。在分割结果之前，也可以先初步整合结果，即先把相同 key 的值合整合成一个，因为合并的工作和 Reduce 的功能是相似的，所以整合这步可以省略。

（5）Reduce:完成 Map 任务的数据分割和整合操作，之后 Reduce 进一步处理 Map 处理后所导出的数据，完成最终的合并后输出并关闭任务。

2.5.2 hive 工具

Hive 是分布式平台上一个可以用来对大数据进行导入、查询、过滤等操作的数据仓库。Hive 被叫做为 Hive QL，因为 Hive 中为用户提供了类 Sql 语言的各种使用功能，所以掌握 SQL 使用方法的用户可以快速的学会 Hive 的使用。

由于 Hive 本身是可以完成 HDFS 上的数据导入操作的，并且针对 HDFS 上的数据可以使用 MapReduce 完成其大数据处理，所以 Hive 和传统数据库有着很大的区别，最大的区别就是 Hive 依托于 hadoop 平台从而本身具有高效处理海量数据的能力。和传统数据库相比，数据在加载的时候就要确定表的数据信息，包括表中字段数、字段类型、字段长度等等，当要加载数据时会先验证数据合法性才会加载，否则无法加载，这就是写时模式，写时模式在查询上更胜一筹，但是数据加载操作会比较浪费时间。Hive 不会在加载的时候进行检查，这种叫做读时模式，优点是加载数据的时候比较快，但是可能在查询的时候出现未检查的类型错误；由于 hive 的分布式特性，所以操作上会经常对整个表所有记录遍历一遍，不支持表中并发等操作，所以索引、事务都没有提供相应的支持功能；数据类型上，hive 支持 INT、TINYINT、DOUBLE 等基本的原子数据类型，，还支持复杂的数据结构，包括数组 ARRAY，自定义 STRUCT 等。但是原子类型中不支持日期时间等类型。

Hive 的数据可以分从以下三个方面进行管理：

元数据：Hive 的数据库连接方法基本上有三种，可以通过 Single User Mode，是可以连接到一个内存数据库的模式； Multi User Mode，是应用中使用最多的模式，是一个可以通过网络连接进行对数据库的访问来完成各种用户所需要的数据操作的模式；Remote Server Mode，主要使用于客户端并不是 Java 的访问操作，通过一个中间服务来完成。

存储格式：Hive 不像 Mysql 等数据库一样拥有特定的格式存储数据，hive 是在 HDFS 中完成文件的存储工作的，没有特定的格式限制，用户可以根据自定义的数据规则创建表，而且也不用给数据建立索引。Hive 的数据模型有托管表，和一般的数据库中的表类似，每个表都有一个自己数据存储文件的路径；External Table，外部表的数

据不是自身管理的，而且也不会对数据存储路径进行检查；分区，根据日期、组等字段值对表进行一个维度或者多个维度的分区，最为常见的例子是日志的管理，根据记录时间点进行分区，使得同一时间段内的记录都属于同一个分区内，会提高一些表操作的效率；Bucket，桶是为了获得更加高效的查询处理对表或者分区进一步划分，给数据提供额外的结构。

交换数据：交互，主要是 Hive 客户端通过指令输入完成对数据库的查询操作；数据的交换过程如图 2.4 所示。

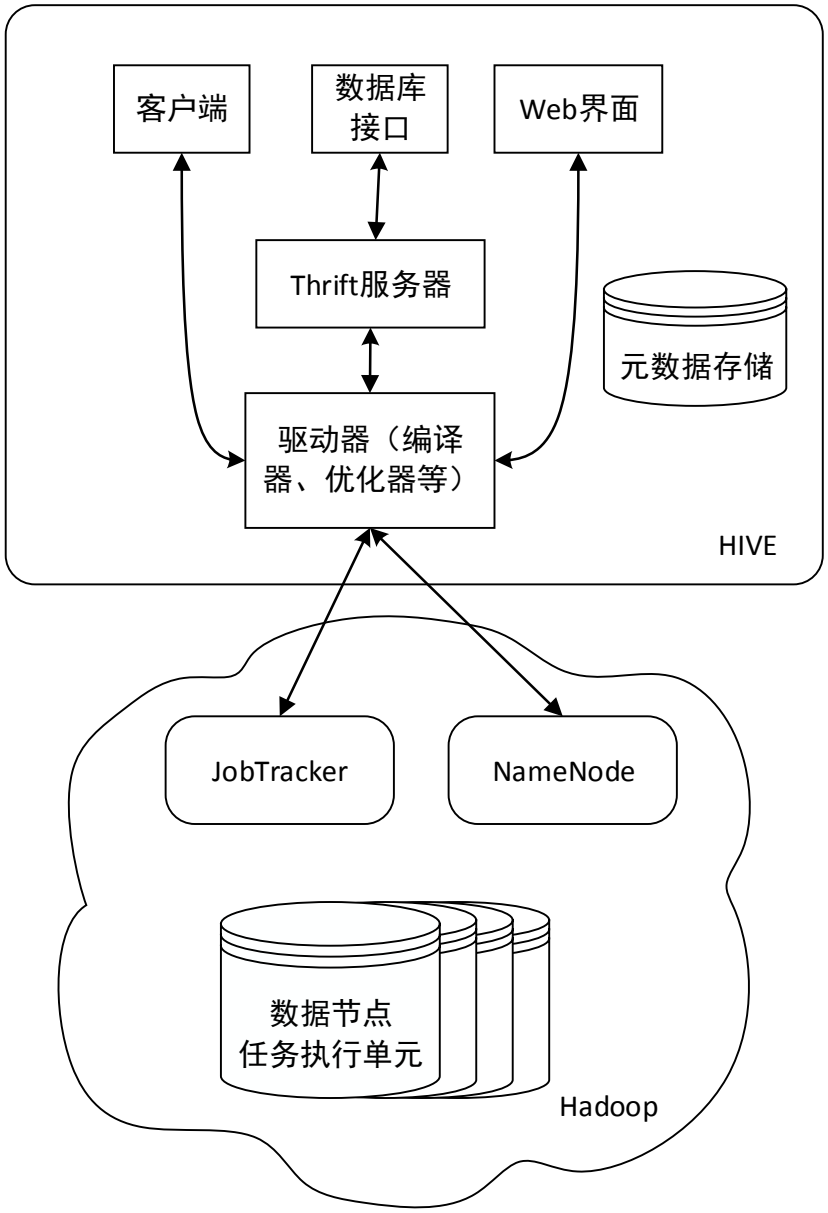

图 2.4　Hive 数据交换图

第三章　研究设计

3.1　用户建模方案设计

3.1.1　建模目的与步骤

本文的研究数据为电信某省用户的移动上网数据，如何对电信用户的网络行为进行有效的认识并对后期的策略制定进行，总结规律是用户行为分析的关键所在。想要对用户网络行为进行规律探索，首先就需要针对收集的现实数据进行用户建模，只有将用户行为数字化、概括化，才能从繁杂的数据中找到研究的切入点，然后从各个切入点入手，寻找符合现实数据的统计分析方法，建立适合电信用户的模型。

用户建模的目的在于用数据分析的方法对电信用户进行规律挖掘，从不同角度分析其网络行为的规律，对得到的规律进行总结，提出针对性的营销方案和精细化的产品方向，发现潜在目标用户的同时，使多种精细化产品更符合不同类别的用户，使用户满意度更高，使得运营商在进行营销时不再是盲目的广撒网，而是在节约成本的前提下提高运营商的收益及市场占有率。

对用户进行建模，首先需要确定分析的角度与用户建模的维度，本文对电信用户的网络行为进行分析，从用户的上网时间、业务应用的选择与访问、网页链接分类三个角度探索用户的时间规律、业务市场的竞争结构和用户的兴趣偏好，其中用户的兴趣偏好作为市场细分的依据；然后从这些维度，结合不同的数据挖掘方法对数据进行分析，其中用户的时间规律选择了时间序列模型统计分析方法，市场的竞争结构用幂定律进行探究，市场细分中采用聚类算法对用户进行规律挖掘，并分别对三个维度得到的结果进行分析，总结为用户的行为特征；最后针对这些用户特征，给出营销建议。具体见图 3.1。

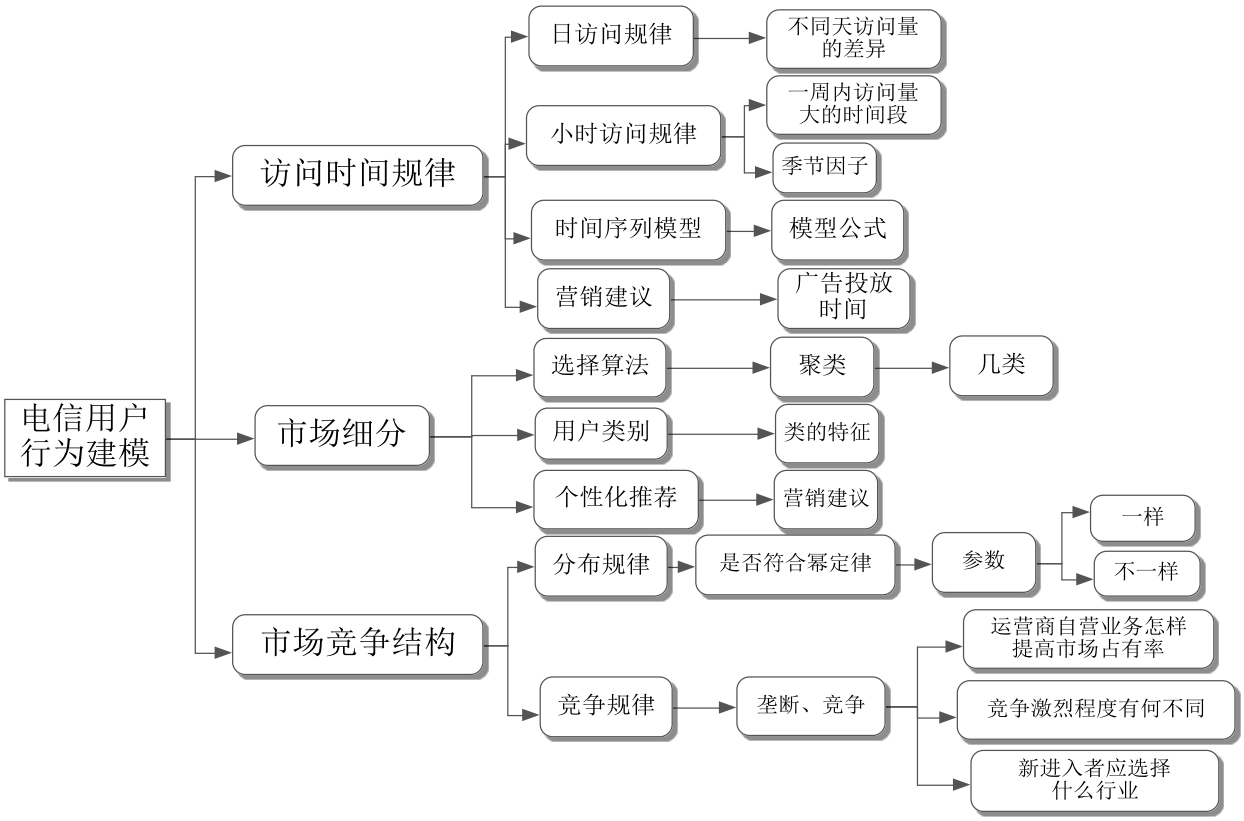

图 3.1　电信用户行为建模

3.1.2　用户建模维度

用户建模的维度取决于可行性及必要性两个方面。

可行性主要基于研究数据的可得性，本次从电信运营商获得了某省移动用户一段时间内上网数据的研究权限，其中从用户数据多个属性中进行挑选，选择出与用户网络行为角度有关联的几个属性：加密手机号、上网时间、业务应用编号、用户访问的目标网站的 URL 等。

必要性主要从需求的角度，哪些方面的规律会对电信业的的营销决策有帮助，哪些维度的分析结果会对目前的电信业有价值，即分析的维度是目前所需要的。进行营销活动时，如果想在广告投放前获取目标用户群体，就需要进行市场细分。首先要了解用户喜欢什么不喜欢什么，即了解用户的兴趣偏好，比较本次研究数据的属性，用户对网站的类型的偏好可以在一定程度上说明用户的兴趣偏好，所以可以从用户访问过的网站中提取出用户的兴趣。广告营销的投放时间也很重要，如果投放的时间不太好，在这个时间段上网的用户不多，看到广告的用户就比较少；如果长时间持续进行广告投放，花费的资源又太多；选择一个大多数用户都在上网的时间，可以达到营销成本不变，但收益增加的效果。从所有用户的业务应用访问上来看，可以得到各类业务应用的市场结构，可以对各类应用市场的竞争结构了如指掌，对于运营商的自营业

务，哪类市场可以进入，哪类市场有提高市场占有率的可能性也就一目了然了。

所以，通过从可行性和必要性两个方面进行分析，本文选择了时间、市场竞争结构、市场细分三个维度对用户建模，进行用户网络的行为分析。

3.2　数据收集

3.2.1　数据来源介绍

本文研究数据是电信某省移动用户的上网访问数据，数据收集工作是由VAS-ODMS 中国电信增值业务运营数据管理系统完成的。此系统在集团层面统一部署，属于增值业务运营体系内自建工程。系统从集团 IT 系统、集团业务平台、SNMS 系统以及集团增值业务管理 NMSC 等多个系统获取基础数据。

文件采集接口协议采用 FTP 方式。采集方式为全量采集，生成文件时存放在临时目录，等待数据文件完全生成后在移动到量化平台的 FTP 服务器指定目录上。

3.2.2　数据文件存储

数据文件格式：数据文件记录方式采用纯文本的格式。

省 DPI 汇聚分析平台支持按照文件生成周期、单个文件大小的数值配置来生成日志文件，且这 3 个条件为或的关系。即生成周期、文件大小同时设置时，如任一指标到达阈值则结束文件。在每个文件生成周期内，如果文件大小超过设定的阀值，则产生多个文件，同一文件生成周期内的多个文件通过文件序号递增进行区分。如下示例：

/home/datamining/flume/yd_DPI/sheng/20150618/FixedDPI.2015061823.1434639601320

/home/datamining/flume/yd_DPI/sheng/20150618/FixedDPI.2015061823.1434639601321

/home/datamining/flume/yd_DPI/sheng/20150618/FixedDPI.2015061823.1434639601322

1、文件生成周期：支持配置，单位为秒，取值 3600（最小 10S）。当前是定 3600s，FixedDPI.2015061823.1434639601320 在文件名第二节最后两位表示 12 点这一小时内的文件。

2、单个文件大小(压缩前)：支持配置，单位为 MB，取值 0-2000（缺省 500M）。当同一小时内一个文件大小达到最大值时，会自动分文件，同一小时内时间，后缀加上 UTC 时间区分。

3.2.3　数据内容

一个文件可包含多项信息，以行区分不同记录，即一行代表一条流的记录；各字段之间以符号“|”分隔各列，代表不同的域;记录之间以 “\r\n”分割。

DPI 设备识别解析的 HTTP GET 报文字段记录，即移动数据包含字段信息如下表：

表 3.1　移动 DPI 规范字段

编号	字段名	字段说明	最大长度	标准协议名称
1	IMSI	用户的国际移动用户识别码	15Byte	Calling-Station-Id，15 位字符串类型
2	MDN	用户手机号码，即 MDN。	15Byte	Served-MDN，用户 MDN 号码
3	ServiceType	业务应用,参见业务应用列表	10Byte	根据端口号和协议特征等分析得出
4	StartTime	业务流开始时间,格式为 yyyymmddhhmmss（24 小时制），如果开启中间记录模式,每条记录都填写相同的开始时间。	14Byte	网络数据流监控统计得出
5	EndTime	业务流结束时间,格式为 yyyymmddhhmmss（24 小时制），如果开启中间记录模式,只在最后一条记录填写结束时间。	14Byte	网络数据流监控统计得出
6	Duration	持续时间，单位毫秒	NUMBER(12)	根据业务流开始和结束统计得出的时间
7	InputOctets	发给用户的业务字节数	4Byte	根据 IP 报文统计得出
8	OutputOctets	用户发出的业务字节数	4Byte	根据 IP 报文统计得出
9	SessionID	session 的 ID 标识号	14Byte	采集设备自动生成，以 session 开始的时间戳为 ID 号，同一个用户中的该 ID 相同表明是同一次 PPP 连接中的不同记录，格式为 yyyymmddhhmmss（24 小时制）
10	UserAgent	User Agent 信息	64Byte	User-Agent，标注浏览器信息，通常指各个手机开发包的名称及版本信息
11	DestinationURL	用户访问的目标网站的 URL。	可变长	根据请求包 GET 字段提取
12	DomainName	外部网站的域名	256Byte	根据请求包 HOST 字段提取

记录中包收集的用户的信息，部分未获取信息为空。在数据处理之前，要先将这几类信息进行数据转换，再进行后续工作，数据转换过程会在本章节的后边提到。

3.3　数据预处理

由于某省用户日访问记录数在 2-5 亿条，收集上来数据文件存储大概在每日 500G 以上，并且包含一些网络原因以及硬件等问题可能导致收集、传输、整理或存储过程中数据损坏，所以对于数据处理上主要需要解决以下三个问题：

1. 数据量大。平均处理每月用户数据文件 10T 左右，选取处理工具、设定处理方案的好坏决定了能否高效、快速的解决问题。
2. 数据规范性验证。数据在收集、传输、存储等过程中由于非可控因素导致的数据破损，数据验证规范上执行不统一等问题，可能导致数据不规范、不合法。会严重影响处理效率和结果正确性。
3. 数据存放平台限制。数据存放的平台不能与外网进行连接、不能下载数据，在进行用户兴趣研究时，需提前建立一个丰富完善的网页 url 库，根据用户访问网页的记录给用户贴上相应标签，完成用户到兴趣标签的映射。

根据电信平台数据的存储规则以及考虑到实际处理过程中可能遇到的难点，设计基于 Hadoop 平台数据处理方案，如图 3.2 所示：

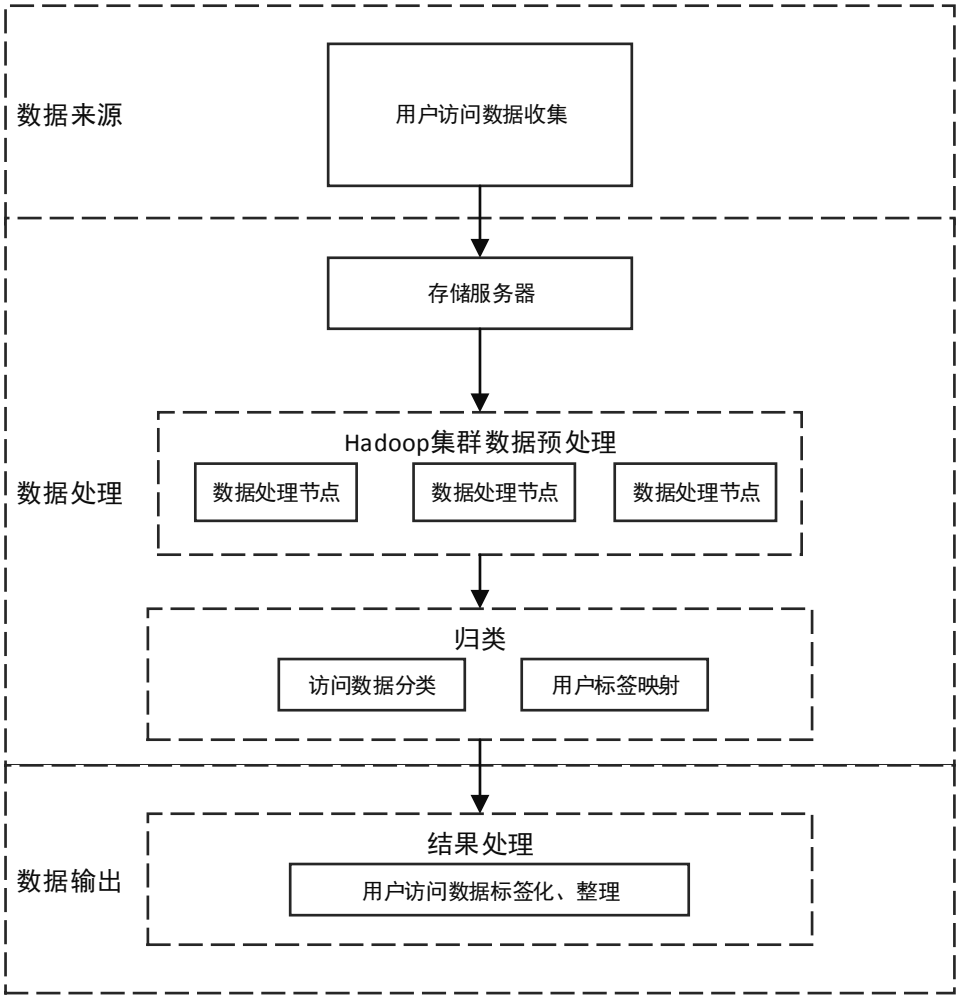

图 3.2　电信用户数据预处理方案

考虑到海量用户数据处理的效率问题，数据处理工作在 Hadoop 平台上完成，Hadoop 平台的在按位存储和处理数据能力上是非常可靠的，可以方便的扩展到多台计算机上，而且对于节点间数据传输处理以及容错性方面都有超强的表现。对于达到 T 量级的文件处理，Hadoop 平台可提供 Hdfs 数据文件存储，可实现对用户数据多处备份、分数据节点存储等功能。利用 MapReduce 可以方便的实现并发处理，在多个数据节点处理数据，最终整合结果，使数据存储、数据处理更加高效、更加安全。

电信用户网络行为特征分析是以访问数据为主的，在确定了 hadoop 环境下处理数据，并且设定了数据收集、hadoop 平台数据预处理、访问数据分析以及标签化的数据处理流程后，根据工作内容分类，设定数据处理的工作步骤。数据处理分为三个步骤：原始数据分析、建立标签库和数据预处理。

1. 原始数据分析中，首先要获取 hadoop 分布式平台 hdfs 上节点数据，并检查数据文件是否完好可读，根据电信公司制定文件格式确定数据文件的导入规则，分析数据文件内数据存储格式和编码格式等特征制定数据序列化与方序列化方法，然后考虑分析工具的选取及处理方法；
2. 建立标签库，针对用户网络访问行为的分析，需要根据已有数据识别该用户的行为特征。完成用户特征识别需要完成访问数据标签库，利用网络爬虫及手工等方式完成网络分类库，包含访问的应用类型分类、URL 分类。
3. 生成用户访问记录标签，首先按照日期方式分组，然后根据用户账号信息为主，确定该用户访问记录，。以用户账号为 Key，完成用户访问记录到标签库的映射关系。然后可从访问时间、用户个人账户、访问域名之间多维关系建立用户特征模型。

具体细化过程如图 3.3 所示：

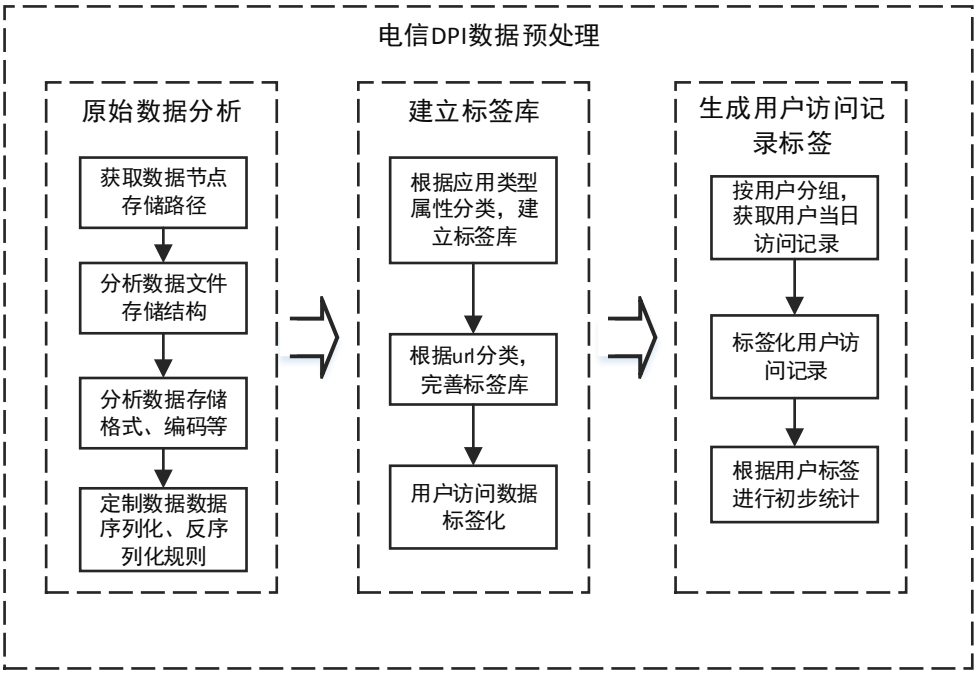

图 3.3　数据预处理具体步骤

3.3.1　原始数据分析

根据数据文件存储数据预处理过程中主要通过 MapReduce 的 java 程序以及 hive 工具等完成数据解码过滤分析等。一个数据文件含多项信息，以行区分不同记录，部分字段采用 Base64 编码，所以实现用于序列化和反序列化工具 serde，Serde 层构建在数据存储和执行引擎之间，实现数据存储和执行引擎的解耦，可作为应用于 MapReduce 程序以及 Hive 等相关工具中通用处理接口实现数据导入和解码功能。

1、使用 hadoop.hive 包中 ObjectInspector 类型参数类型，ObjectInspector 接口可以在输入端和输出端切换不同的输入/输出格式，方便使用不同的数据格式，StructObjectInspector 中又可以（一层或多层的）嵌套任意的 ObjectInspector；

2、CSVWriter 完成数据流到特定格式转换，和编码格式。

对需要深入分析的数据属性进行统计，设计数据合法性检查规则。由于数据是由各区域收集之后传输并存储到平台上的，网络以及数据存储规范等因素可能导致部分数据并不符合使用的条件。所以根据研究所需数据属性，需要对数据的完整性和合法性进行检查过滤，根据制定的筛选规则完成 DataFilter 类实现，筛选出研究所需的数据属性值不为空、值类型正确，值范围正确的访问记录，根据业务流开始时间字段按照实际访问时间分组统计用户。

3.3.2　建立标签库

建立标签库，根据应用类型分类创建标签库，根据 URL 分类完善标签库。细化步

骤如下：

1. 根据应用类型分类，统计已有数据中所有 ServiceType 应用类型字段，及对应访问信息，添加 ServiceType 到其所属类别。
2. 根据 URL 分类。
   a. 热门 URL、热门应用。统计主流门户网站，热门网站、热门应用域名，统计用户访问量较高域名；通过 Jsoup 等工具爬取网页内容，热门应用访问信息整理；分析所获取 Dom 内容树，网页去噪、节点提取；分析整理，添加到分类库。
   b. 二级标签搜索引擎搜索结果。 在谷歌、百度、好搜等搜搜引擎搜索标签内容；获取搜索结果，使用 Jsoup 等工具获取搜索前 20 页内链接；剔除错误及重复链接；整理后添加到 URL 库。
3. 未匹配 URL 处理。
   a. 根据用户访问记录中域名字段访问量排序，并从 hadoop 平台导出。
   b. 获取网页内容，并提取节点“keywords”。
   c. 关键词处理，分析。添加到分类库。

分类库建立第一步应用 MapReduce 程序，通过 JobConf.setInputFormat 方法设置输入格式为 SequenceFileAsBinaryInputFormat.class，设置数据存储路径，在 Map 中数据反序列化后，输出 ServiceType 字段值为 key，当前行数据为 value，recude 中输出并计数，根据该类型对应访问信息进行人工分类。

第二步完成知名热门网站、移动应用访问信息人工收集，完成用户数据中访问量较高域名统计。根据网站多级标题设定自己标签等级，jsoup 爬取收集的网站内容，过滤处理网页信息，获取各级标题相关链接。规则热门的 URL 分类，包括热门网站、主流门户、和热门应用收集整理。例如百度、搜狐、腾讯及新浪等。不同的网站有不同的布局，根据网站分别设计爬取规则，最终获取其网站多级标题以及 URL 的键值对数据；获取对应手机应用访问域名等数据。网站处理部分具体步骤：

1. Jsoup 抓取网页内容。Jsoup 是一款 Java 语言编写的 HTML 分析器。可以用来解析 URL 地址，并提供了较多的 DOM、CSS 以及类似 JQuery 的操作方法来提取操作数据，简单爬虫工具。以收集主流网站为种子，通过 Jsoup 抓取网页内容，

```
Document document = Jsoup.connect(baseurl).get();
```

2. 根据网站标题指定网页处理方法。

根据网页结构以及通过 jsoup 获取的网页节点树，通过对 DOM 信息树分析，获取以及标题以及其连接，获取该子节点，

Elements sublinks = subdocument.select("div.type_list_child");

根据二级标题结构，获取二级标题和连接，String hrefString = subelement.attr("href");

String nameString = subelement.text();

以二级标题下连接为初始种子，重复上述内容，获取最后分类页面内网页信息，门户导航网站各标题下页面结构。将各网站内收集链接加入该URL库，完成所有收集大型网址的分析工作，之后人工对数据进行网址标题到自定义标签的映射工作，完成URL分类库的第一步填充。

分别使用百度、谷歌、好搜等搜索引擎对二级标签搜索，使用 jsoup 工具获取搜索结果网页，示例：百度搜索引擎爬取 URL 为

```
String url = "http://www.baidu.com/s?pn="+ (i*pageNum) + "&wd=" + key;  //百度搜索关键词
```

获取该页面内容，对该网页进行数据清洗，获取页面有效内容。根据百度搜索结果页面结构，提取有效节点，最终获得前 20 页搜索结果链接，由于搜索结果一般是根据相关度排列的，取前 20 页相关度最高，保证正确性。去重，然后把 URL 添加到对应标题下。

第三步根据现有库对用户进行匹配，导出 ServiceType 应用类型属于网页类别，并且未成功匹配的 URL 的 DomainName 域名字段值。此处使用域名处理是因为用户记录量太大，如果对所有 URL 进行文本获取，文本分词处理效率太低，工作量太大，同时电信数据内部网络是不支持连接外网的和没有大量数据导出权限，所以根据域名来做分析，可以只针对域名进行分类导出处理，能解决不能数据大量导出及不能连接外网问题。对于未匹配成功网站，采取二次爬取分类，根据网页内容 keywords 信息，对关键词清理、同义词整合，分析关键词并分类添加到标签库中。

3.3.3　生成用户访问记录标签

上传标签库文件到分布式平台，在 MapReduce 程序中读取标签库内容到内存中，根据标签库数据完成用户访问记录标签匹配步骤：

1、对每条记录提取 MDN(用户手机号码)、ServiceType(业务应用类型)、DestinationURL(用户访问的目标网站 URL)、DomainName(外部网站域名)四个属性值；

2、根据 ServiceType 业务应用类型在标签库中获取特征标签；

3、如果 ServiceType 为 399(网页类型)，根据标签库中网址分类进行全匹配，如果匹配成功，则获取标签；如果匹配不成功，则对其同域名分类库中网址进行模糊匹配，成功则获取特征标签。

4、如以上未获取特征标签，且标签库中域名也对应标签属性，则获取该域名标签，否则忽略该数据。

完成用户数据标签匹配后，可统计某标签类别下用户访问量；不同标签类别访问量

对比和随时间变化；可对用户个人访问兴趣、网站访问的时间规律、某用户访问兴趣的时间规律、某标签类别下用户群体特征等进行进一步的研究，从而挖掘出用户更深层次的需求。

3.4 数据分析方法

3.4.1 时间分析

对时间规律的分析就是从时间序列模型中寻找适合观测数据的模式，从而推导出时间序列的通项公式或递推公式。在分析过程中，除了建立第二章介绍的时间序列模型，还需要一些其他的相关基础知识。

自相关系数一般用来描述时间序列的自身相关结构，用于识别数据的趋势和季节特征。自相关系数r_k的公式为

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2} \tag{3-1}$$

一般，随着k的增加，自相关系数逐渐减少。如果数据序列是随机的，则对于任意k的自相关系数均接近于 0；如果时间序列存在趋势，则当k取 1，2 等较小的值时r_k显著不为 0，且随着k的增加逐渐趋于 0；如果存在季节周期，一般在k取周期的倍数值时r_k显著不为 0。

如果数据序列中存在趋势或季节，则序列不稳定，需要先进行差分去掉趋势和季节。差分方程是变量与它前期值的关系的表达式。间隔为一期的两个序列值之间的减法运算称为一阶差分，即

$$\nabla Y_t = Y_t - Y_{t-1} \tag{3-2}$$

去除趋势一般需要进行一阶差分。当序列存在季节周期s时，需要进行间隔为s期的两个序列值进行减法运算，即

$$\nabla_s Y_t = Y_t - Y_{t-s} \tag{3-3}$$

包含季节的时间序列一般包含趋势 T、季节 S 和不规则因素 I，建立时间序列模型的过程就是寻找趋势、季节因子的过程，就需要对时间序列进行分解。时间序列分解的方法一般有加法模型和乘法模型。加法模型就是三种因素相加，即

$$Y_t = T_t + S_t + I_t \tag{3-4}$$

乘法模型是三种因素相乘，即

$$Y_t = T_t \times S_t \times I_t \tag{3-5}$$

3.4.2　幂定律

市场竞争结构研究建立在一元非线性回归模型幂函数的基础上，分析方法和过程如下：

1、参数估计

当对一个自变量和因变量之间的非线性关系进行研究时通常采用一元非线性回归模型，表达式为

$$Y = f(X,\beta) + \varepsilon \tag{3-6}$$

其中X、Y分别为自变量与因变量，β为参数，ε为随机变量。本文对市场竞争结构进行研究，采用的是一元非线性函数中的幂函数，表达式为

$$Y = \beta_1 X^{\beta_2} \tag{3-7}$$

因为幂函数可以通过简单的变换变为一元线性模型，所以为了更简单的对参数的模型进行估计，首先将幂函数回归模型做一元线性变换处理，处理方式为等式两边同时取对数，式子变为

$$\ln Y = \ln\beta_1 + \beta_2 \ln X \tag{3-8}$$

令 Y=ln$Y$，A= lnβ_1，B=β_2，X= ln$X$，幂函数就转化为一元线性函数

$$\mathrm{Y} = \mathrm{A} + \mathrm{BX} + \delta \tag{3-9}$$

其中δ为误差。

接着用最小二乘法对参数 A 和 B 进行估计，最小二乘法准则$\min \sum(y_i - \hat{y_i})^2$中，$y_i$、$\hat{y_i}$分别为第$i$次观测中因变量的观测值和估计值。回归方程的参数 A、B 的估计公式为：

$$\mathrm{A}=\frac{\sum(x_i-\bar{\mathrm{x}})(y_i-\bar{\mathrm{y}})}{\sum(x_i-\bar{\mathrm{x}})^2} \tag{3-10}$$

$$\mathrm{B}=\bar{\mathrm{y}} - \mathrm{A}\bar{\mathrm{x}} \tag{3-11}$$

其中，x_i为第i次观测中因变量的观测值，$\bar{\mathrm{x}}$、$\bar{\mathrm{y}}$分别表示自变量与因变量的样本平均值。

2、假设检验

得到模型的参数估计值后，需要将参数代入模型进行显著性检验。显著性检验用来判定在统计意义上变量之间的回归关系的是否具有显著性。显著性检验包含两个方面，一个是判断因变量与自变量的显著关系，用 F 检验针对整个方程进行判断；另外一个是在确定方程总体具有显著关系后，判断方程中每个单个自变量是否显著，用 t 检验进行判断。

用于判断整体是否显著的 F 检验为：

$$\mathrm{H}_0\text{：}\ \beta_1 = \beta_2 = \cdots = \beta_\mathrm{p} = 0$$

H_a：至少有一个参数不等于零

原假设H_0如果被拒绝，说明至少有一个参数不为 0，因变量与所有自变量之间在总体上呈显著关系；如果原假设H_0没有被拒绝，说明方程整体上不存在显著关系。

检验的统计量为

$$F = \frac{MSR}{MSE} \tag{3-12}$$

其中，MSR为回归的均方，MSE为误差的均方，

$$MSR = \frac{SSR}{p} \tag{3-13}$$

$$MSE = \frac{SSE}{n-p-1} \tag{3-14}$$

式子中的$n - p - 1$表示自由度，SSR 为总的平方和，SSE 为误差平方和。SSR 和 SSE 的公式为

$$SSR = \sum(\hat{y_i} - \bar{y})^2 \tag{3-15}$$

$$SSE = \sum(y_i - \hat{y_i})^2 \tag{3-16}$$

总平方和

$$SST = SSR + SSE = \sum(y_i - \bar{y})^2 \tag{3-17}$$

如果统计量 F 值足够大，按照临界值法$F \geq F_\alpha$，则拒绝原假设，即回归模型在统计意义上是显著的；否则接受原假设，即回归模型没有意义。在统计软件中，F 检验的结果通常会通过 P 值法表示，一般当 P 值小于 0.05 时拒绝原假设，回归模型是显著的，即模型成立。

F 检验通过后，使用 t 检验用来判定每个变量中单个参数的显著性。对于任一参数β_i，

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

检验的统计量 t 为

$$t = \frac{b_i}{s_{b_i}} \tag{3-18}$$

其中b_i为各个变量参数的最小二乘估计值，s_{b_i}是b_i的估计标准差。如果$t \leq -t_{\alpha/2}$或者$t \geq t_{\alpha/2}$，拒绝原假设H_0，即该变量的参数在统计意义上显著；否则接受原假设H_0，即该变量的参数不具有统计意义。

3、拟合优度判定

做完假设检验后，要判断回归方程的拟合优度。拟合优度是用来判断回归模型与观测值拟合程度，拟合程度越大越好。拟合程度可以用可决系数R^2来度量，公式为

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{3-19}$$

R^2的取值范围为 0 到 1，越接近 1 说明回归方程的拟合程度越好，反之则越差。

为了避免自变量数量对R^2的影响，还需要${R_{adj}}^2$来同时进行衡量，R^2与${R_{adj}}^2$越接近时表明拟合程度越好，

$${R_{adj}}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} \tag{3-20}$$

还有一个指标也用来判断回归模型的拟合优度，这个指标为估计值的标准误差，用 s 表示，

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \tag{3-21}$$

s 用来衡量观察值相对与估计值回归线的变异程度，值越小越好。

3.4.3　Phantom 联合聚类算法

本文在对用户兴趣偏好进行研究时，使用的是用户的网页浏览数据。用户对网页的浏览行为能够反映出用户的兴趣偏好，对实现个性化推荐、精准营销至关重要。Keralapura 等人在研究 3G 用户的网络行为时提到用户和他们浏览的网站是相互影响的，即用户决定了他们浏览的网站，网站反过来也会决定用户，所以在挖掘移动互联网用户浏览习惯的问题建模为联合聚类问题[34]。通过一个矩阵来存储用户和网站信息，以每个用户为一行，每个网站为一列，矩阵中元素表示相应用户访问相应网站的次数。联合聚类最终的目的不仅仅是要分组相似的行和列，而且要将大矩阵分割成许多子矩阵（子簇），每个子矩阵都反映出一定的聚合属性，并提出了 Phantom 联合聚类算法。

在这之前，将矩阵分割的联合聚类算法有 Dhillon 提出的 Spectral Graph Partitioning 算法，其主要解决的是文档和单词的联合聚类问题。文中建立了一个以每个单词为一行，每个文档为一列的矩阵，矩阵元素为对应单词在对应文档中出现的次数，并将聚类问题转化为一个二分图的划分问题，最终通过 Spectral Graph Partitioning 算法将 NP 完全图的划分问题转化为求矩阵的奇异向量问题[35]。

Spectral Graph Partitioning 算法解决文档和单词的联合聚类问题产生了不错的效果，但是在用户网站联合聚类问题却存在一些问题：用户分析时可能会有百万甚至千万的用户和网站，形成矩阵后，计算机由于内存和处理器的限制无法处理；算法要输入聚类产生的簇的数量，但实际操作中很难判断该数量；算法最终结果中，矩阵的一行或

一列只能属于一个簇，但在现实中一个给定的行或列可以属于一个或多个簇。

Phantom 算法通过对 Spectral Graph Partitioning 算法进行了一系列改进，提出并发展了一个扩展的联合聚类算法 Phantom。Phantom 算法通过使用沙漏模型，首先降低了输入数据的规模，在更低规模的数据分别执行一个迭代的改进的自顶而下的层次聚类算法后，再按原始信息将矩阵规模扩展作进一步分析。另外 Phantom 算法提出了 Soft Co-Clustering，使同一行或列可以同时属于两个簇。

本文对用户进行市场细分时选择 Phantom 联合聚类算法对用户进行聚类，并将上述不同算法的聚类结果进行比较，证实了 Phantom 算法的 Soft Co-Clustering 为最优，在细节上对算法进行了小幅改进，应用在电信用户的市场细分中。

第四章　用户访问时间规律分析

4.1　用户访问时间序列介绍

用户访问时间指的是用户访问网络即上网的时间，研究其特性，发现用户访问网络的时间规律对运营商获取特定时间内的目标用户、提高精准营销的命中率等意义非凡。

为观察用户的访问规律，将所有用户每天的数据划分成 24 个小时时段，按照用户上网开始的时间，即移动 DPI 数据中 StartTime 属性字段进行时间标注，按照表 4.1 进行时段划分。每个时段包含起始时间的数据，但不包含时段结束时间的数据，如时段[0:00,1:00)中，0:00 点的数据算在时段 0 中，1:00 点的数据算在下一个时段 1 中。

表 4.1　时段划分

时段	时段名称	时段	时段名称
[0:00, 1:00)	0	[12:00, 13:00)	12
[1:00, 2:00)	1	[13:00, 14:00)	13
[2:00, 3:00)	2	[14:00, 15:00)	14
[3:00, 4:00)	3	[15:00, 16:00)	15
[4:00, 5:00)	4	[16:00, 17:00)	16
[5:00, 6:00)	5	[17:00, 18:00)	17
[6:00, 7:00)	6	[18:00, 19:00)	18
[7:00, 8:00)	7	[19:00, 20:00)	19
[8:00, 9:00)	8	[20:00, 21:00)	20
[9:00, 10:00)	9	[21:00, 22:00)	21
[10:00, 11:00)	10	[22:00, 23:00)	22
[11:00, 12:00)	11	[23:00, 24:00)	23

统计各个时段所有用户的访问数据，共统计了 35 天的数据，部分数据如表 4.1 所示，为避免泄露实际数据，表中数据已乘同一参数，不影响规律探索。

表 4.2　用户访问量统计表

时间	访问量	时间	访问量
2015062412	15561	2015062823	19112
2015062413	18752	2015062900	17086
2015062414	19270	2015062901	16520
2015062415	18698	2015062902	15618
2015062416	19136	2015062903	15065

2015062417	16832	2015062904	13117
2015062418	15585	2015062905	6467
2015062419	14535	2015062906	10474
2015062420	16562	2015062907	20840
2015062421	16637	2015062908	15544
2015062422	16835	2015062909	15936
2015062423	20354	2015062910	18002
2015062500	21113	2015062911	22157
2015062501	21382	2015062912	22522
2015062502	21851	2015062913	20084
2015062503	17386	2015062914	21651
2015062504	7868	2015062915	22293
2015062505	7971	2015062916	18411
2015062506	2343	2015062917	18008
2015062507	7103	2015062918	18844
2015062508	12664	2015062919	21784
2015062509	17678	2015062920	23387
2015062510	22152	2015062921	25203
2015062511	23512	2015062922	21766
2015062512	23246	2015062923	22520

将用户访问量的时间序列画成时间序列图，如图4.1所示

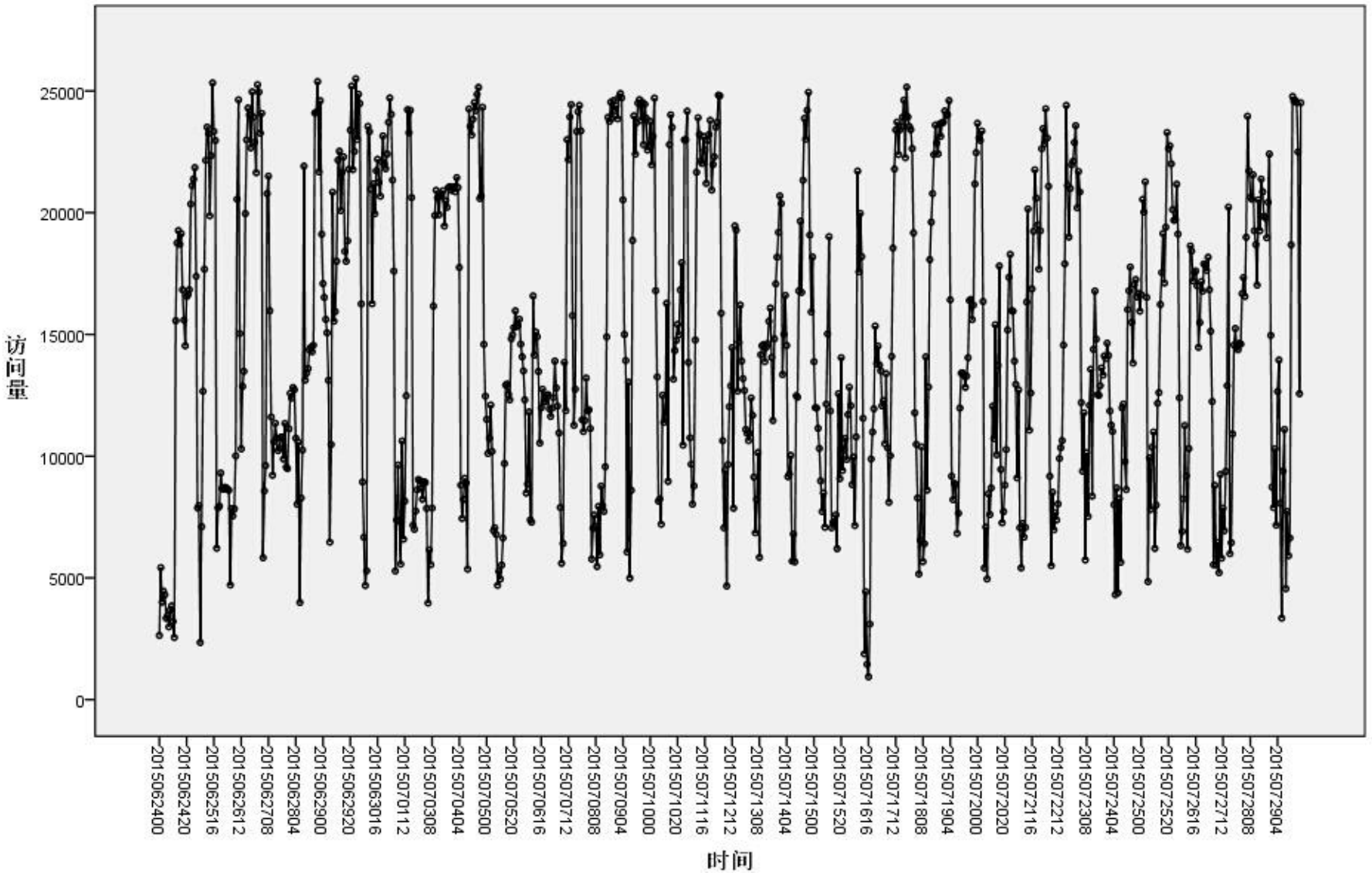

图 4.1　用户访问量的时间序列图

由于移动DPI数据量大，原始数据在接收、传输、上传到平台的过程中存在一些丢包现象，所以统计的数据中包含异常值，图4.1中可看到个别数值小的离群点，在后面建立时间序列模型中也特别注意了异常值的存在。

4.2 日访问规律分析

由于图4.1数据值太多，图中无法肉眼观察到规律，所以把每天的访问量进行加和汇总，组成35个以天为粒度的时间序列，分析日访问量是否有规律，工作日与周末是否有差异。

寻找用户日访问量的规律性，首先寻找其周期性，对时间序列做自相关分析。时间序列的自相关分析是为了了解不同间隔的观察值之间的相关程度，根据自相关函数图进行研究分析访问量的相关性，可观察访问量在不同的时间是否存在周期性及趋势，以天为粒度的时间序列自相关如图4.2所示。

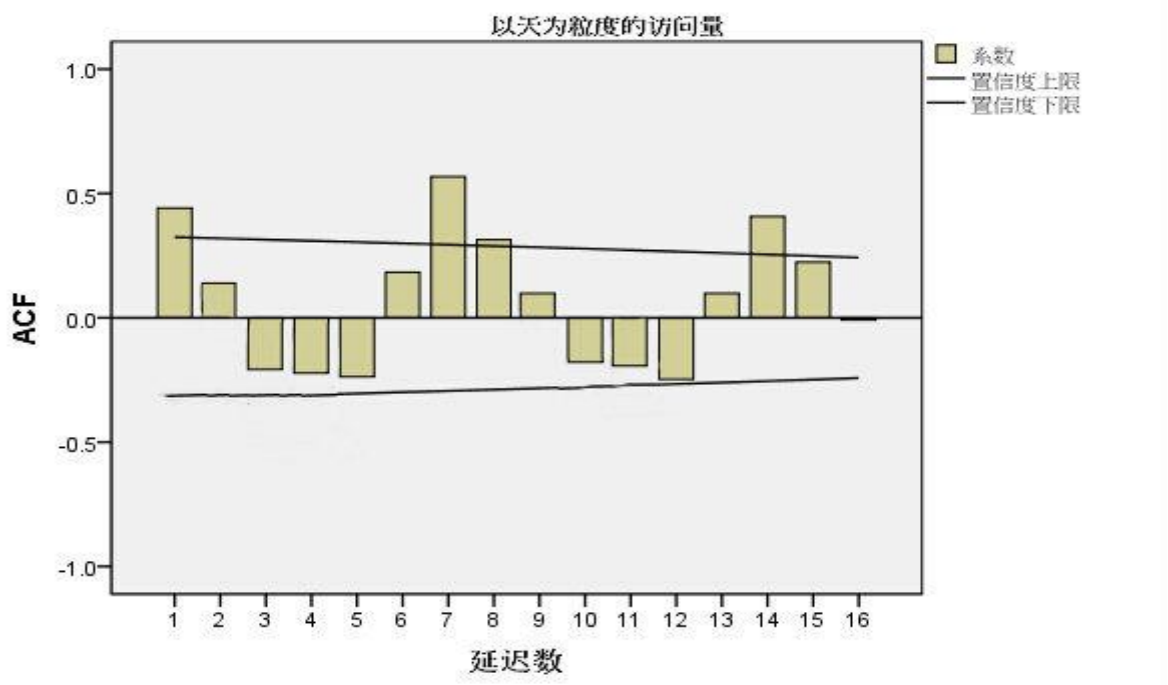

图 4.2 日访问量的自相关函数图

从图4.2中可以明显看出时间序列存在明显的趋势，在第7天和第14天时自相关系数显著不为0，说明周期为7，即7天（一个星期）为周期。下面研究每个周期内用户访问量的变化，将数据按照星期几分为7组，第1组到第7组分别为周日到周六，统计每组数据的最大、最小、四分位、异常值等数据，画成箱图表示，如图4.3所示。

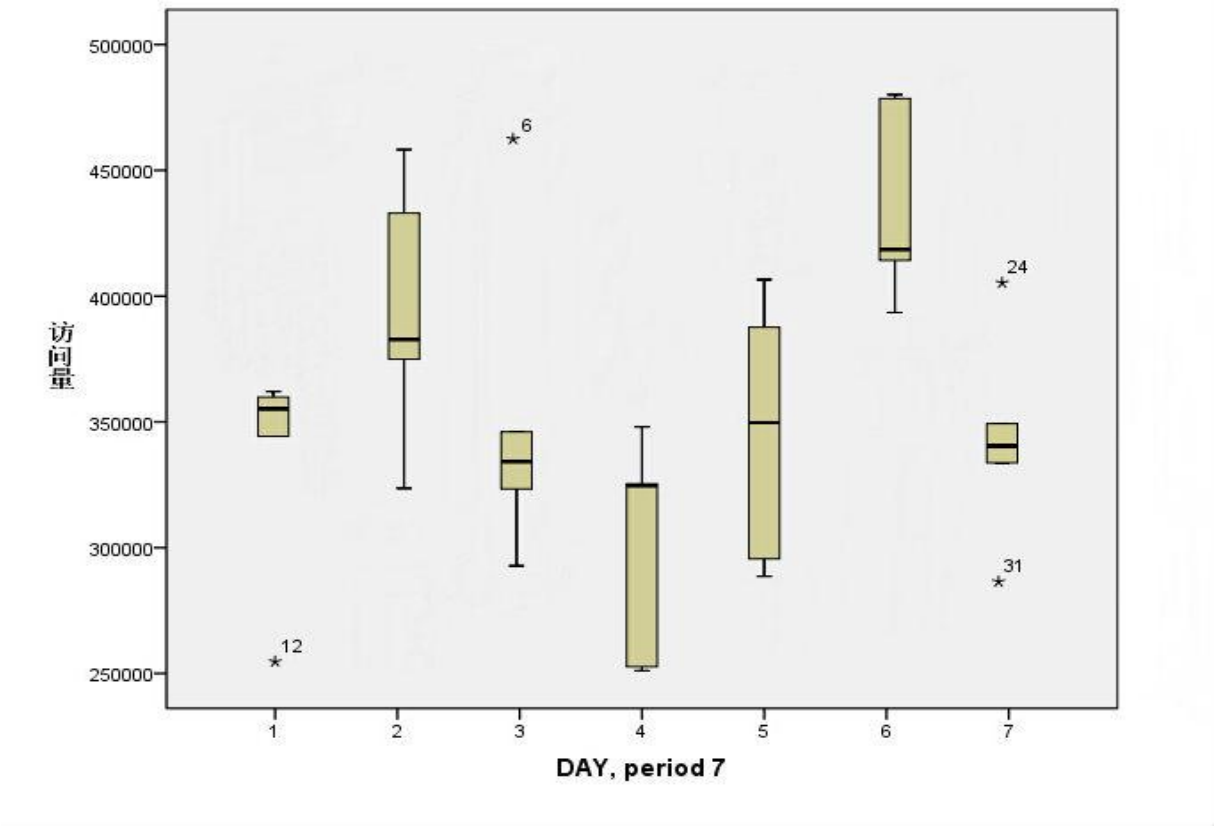

图 4.3　访问量-星期几描述统计图

图4.3中从每天访问量从上到下分别是最大值、上四分位数、中位数、下四分位数和极小值，星型表示极端的异常值。由于原始数据整体存在一些数据丢失，异常值可能多于图中所标出的点，所以主要针对中位数和四分位的数值进行讨论。

首先观察工作日和周末访问量的差异。图中DAY1、DYA7对应着周日和周六，其他几天为工作日，DAY1、DYA7的数据相比工作日的数据处在较低水平，周末的访问量要比工作日的平均访问量要少。按照思维惯式，周末用户的移动互联网访问量应该增加才是，但实际是反而减少了。究其原因，一是数据统计的为用户的流量上网访问量，通过WIFI访问网络的数据不在其中，如果用户周末在家休息，很有可能使用WIFI连接互联网或者使用电脑进行上网活动；二是周末是人们休息放松的时间，空闲时间长，娱乐活动比工作日丰富，不仅仅局限于移动互联网中，这个时间很多人会聚会、出游、看电影等进行各种工作日难以实现的娱乐活动，现实中的社交多了，互联网中的社交媒体使用的就少了。

工作日的几天数据中，DAY2、DAY6比其他组的数据值要大，这两天一个为周一，一个为周五；访问量最少的为DAY4，对应周三；其余两天周二和周四处在中间水平。周五的到来预示着周末马上要来到，周五晚上的第二天为周末，大多数用户不用上班，所以周五的访问量要比其他工作日多一些。周一的访问量较其他几天也属于比较多的，周一属于开始工作的第一天，访问量竟比周末的多一些。说明周一作为工作的第一天，

很多人还没从周末的休闲状态中走出来，工作状态差些，并充分利用碎片化的空闲时间进行上网填充一天的生活。周三的访问量为一周访问量的最低值，这天刚好处在工作日的中间，工作状态达到最佳，期间开小差上网情况就比较少，又因距离周末前不着村后不着店，上班加班后的人们往往身心比较疲惫，选择休息，而进行的上网娱乐休闲就少一些。将每天访问量及特征总结如表4.3。

4.3 用户每周访问量差异表

星期	访问量值	季节因子(%)	特征原因
周一	较大	110.3	还未从休闲状态中走出
周二	较小	88.6	进入工作状态
周三	最小	78.2	工作状态好，距离周末最远，除了工作时间外多选择休息
周四	较小	91.4	工作状态较好
周五	最大	125.8	第二天为周末，工作状态差且晚上熬夜多
周六	一般	100.2	在家用电脑、WIFI 上网多，娱乐活动丰富
周日	一般	103.4	在家用电脑、WIFI 上网多，娱乐活动丰富

4.3 小时访问规律分析

对用户的日访问量时间序列研究后，发现星期几不同用户的访问量是不同的，那每天不同时段用户的访问量是否也有规律，还需对每个小时的访问量进行深入研究。4.1节已经对以小时为粒度的时间序列进行了介绍，本节将针对这个时间序列进行研究，观察不同时间段的用户访问量。

首先进行周期性探索，根据自相关函数图进行研究分析访问量的周期性，可观察访问量在不同的时间是否存在周期性及趋势，自相关函数图如图4.4所示。

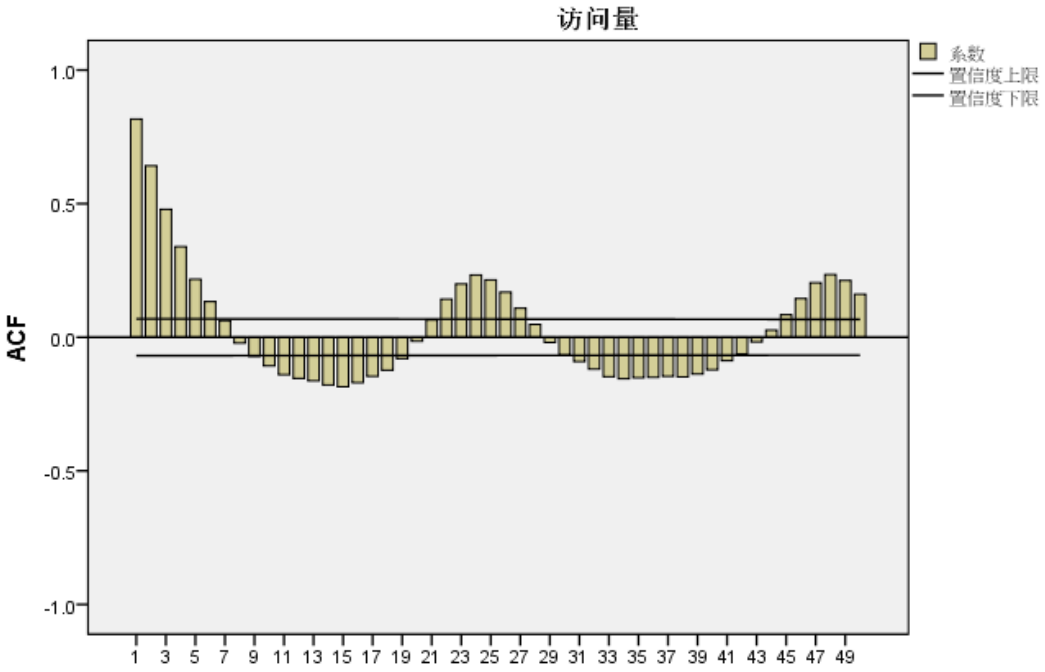

图 4.4　小时访问量自相关函数图

由图4.4看出，访问量时间序列存在一定的趋势，且周期性明显。自相关函数呈现明显的周期性波动，在24和48小时自相关系数显著不为0，说明周期为24，即24个小时（一天）为一个周期。

用户访问量是以24小时为周期的，所以寻找一天内不同时段的访问规律。将所有数据按小时分段，分成24组，做描述性统计，统计每个时段的最大、最小、中位数、四分位、异常值等，结果以箱形图表示，如图4.5。

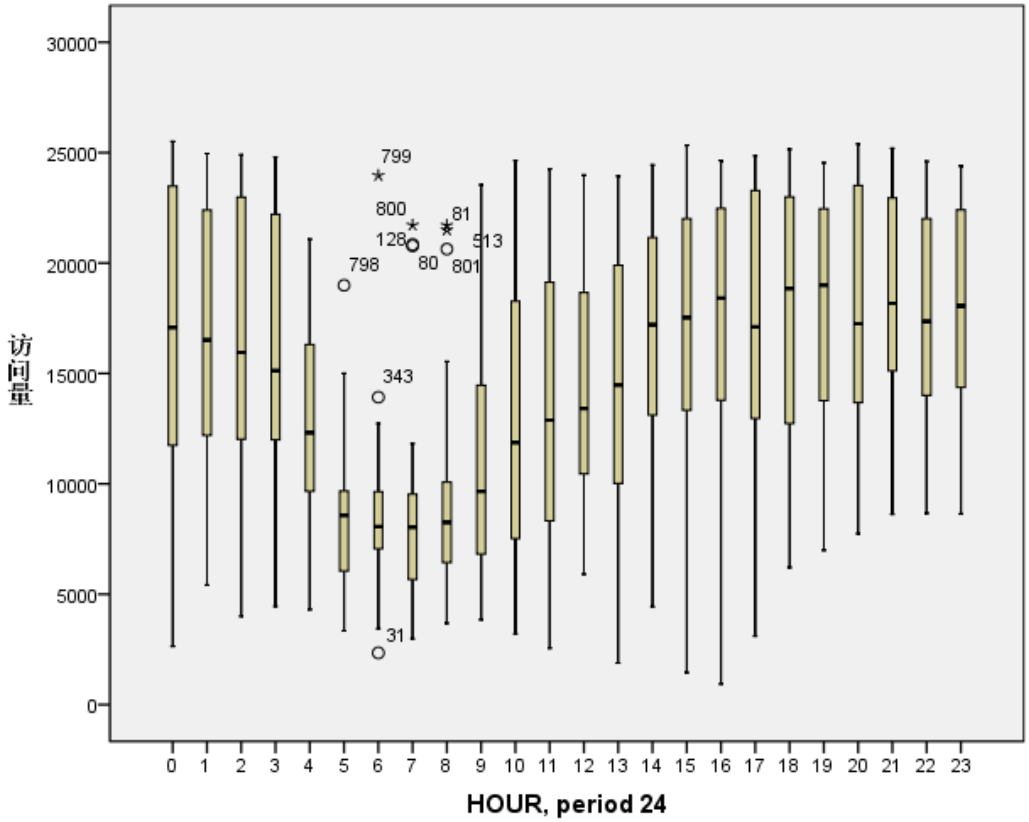

图 4.5　访问量-小时描述性统计图

图中每个小时对应图形中从上到下分别是最大值、上四分位数、中位数、下四分位数和极小值，图中圆圈表示温和的异常值，星型表示极端的异常值。由于原始数据整体存在一些数据丢失，异常值可能多于图中所标出的点，所以规律的寻找主要参考四分位和中位数的数据。

一天24个小时中，从0点开始访问量呈下降趋势，尤其从3点开始访问量急速下降，到6、7点达到最小值，这段时间用户大都处于睡眠状态；8点之后访问量开始回升，用户开始了一天的生活，用户也开始了网络活动；到中午12、13点，访问量又有小幅下降，这段时间有些用户进入午休，暂停网络访问；下午14点开始，访问量开始回升；从20点到24点属于移动网络用户活跃时间，这段时间内用户的访问量属于一天中最多的，说明用户在晚上使用移动网络较多，若选择一个时间段进行营销推广，从整体用户来看，20点-24点这段时间较为合适。

以上分析可以得出，用户访问量的时间序列模型在以天为周期的规律性较强，应该是季节因子与访问量平均值相乘得到的结果。对时间序列进行周期性分解得到季节因子如下：

季节因子

系列名称:　访问量

小时	季节因子（%）	小时	季节因子（%）
1	115.9	13	99.2
2	114.2	14	100.3
3	115	15	111.4
4	111.9	16	114.6
5	85.8	17	119.5
6	58.5	18	113.5
7	59.3	19	114.1
8	57.6	20	114.7
9	60.7	21	121.6
10	74.1	22	120.1
11	87.3	23	117.7
12	94.7	24	118.2

星期几和时段的不同，用户的访问量都存在差异，所以下边将一周内各个时段的访问量进行统计分析。 为了进一步观察一周内不同时间段的访问量差异，将 7 组数据中每个时段的访问量情况做了描述性统计，统计每个时段的最大、最小、中位数、四分位、异常值等，结果以箱形图 4.6 进行展示。

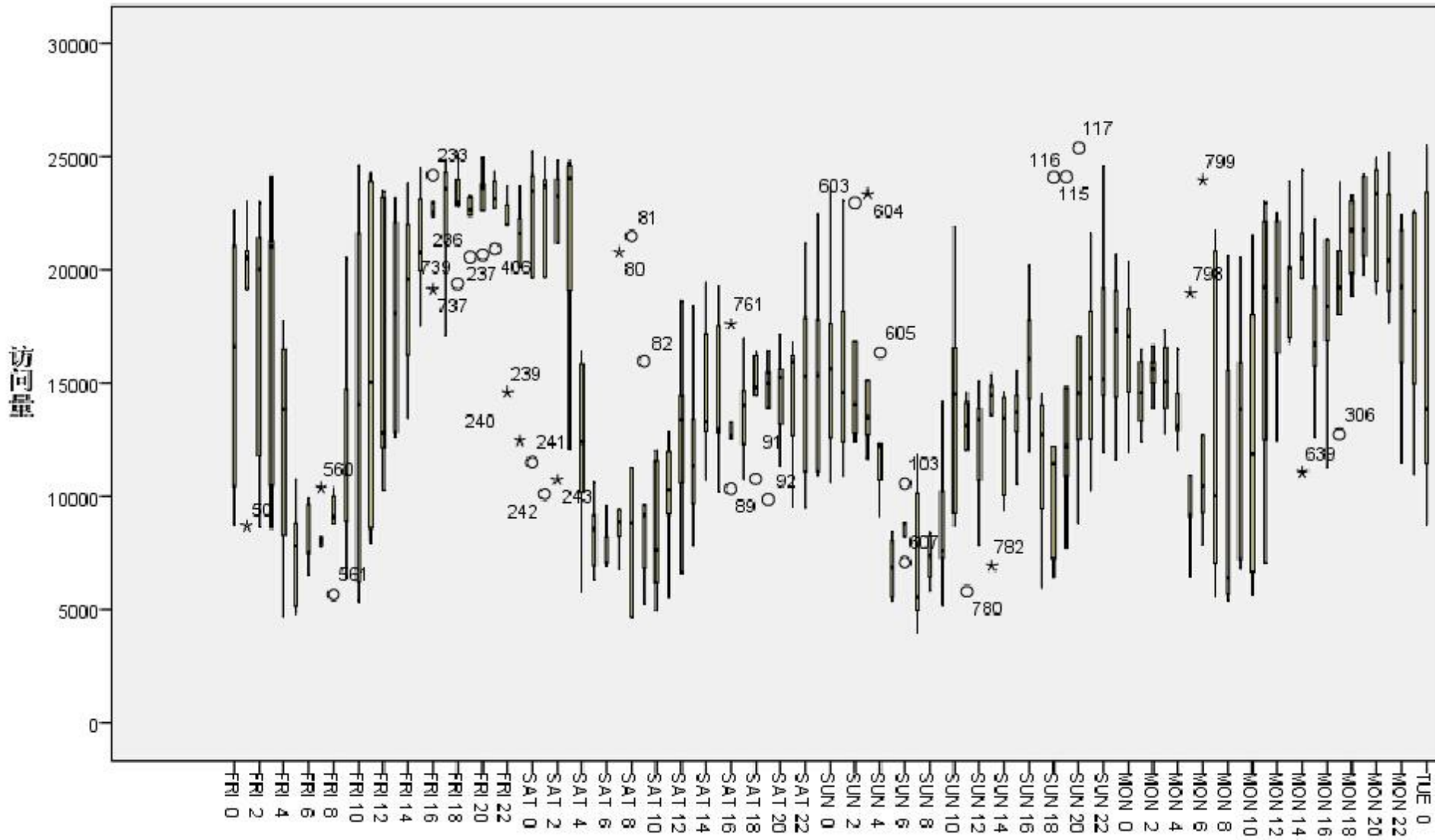

图 4.6　一周内不同时段访问量描述统计图

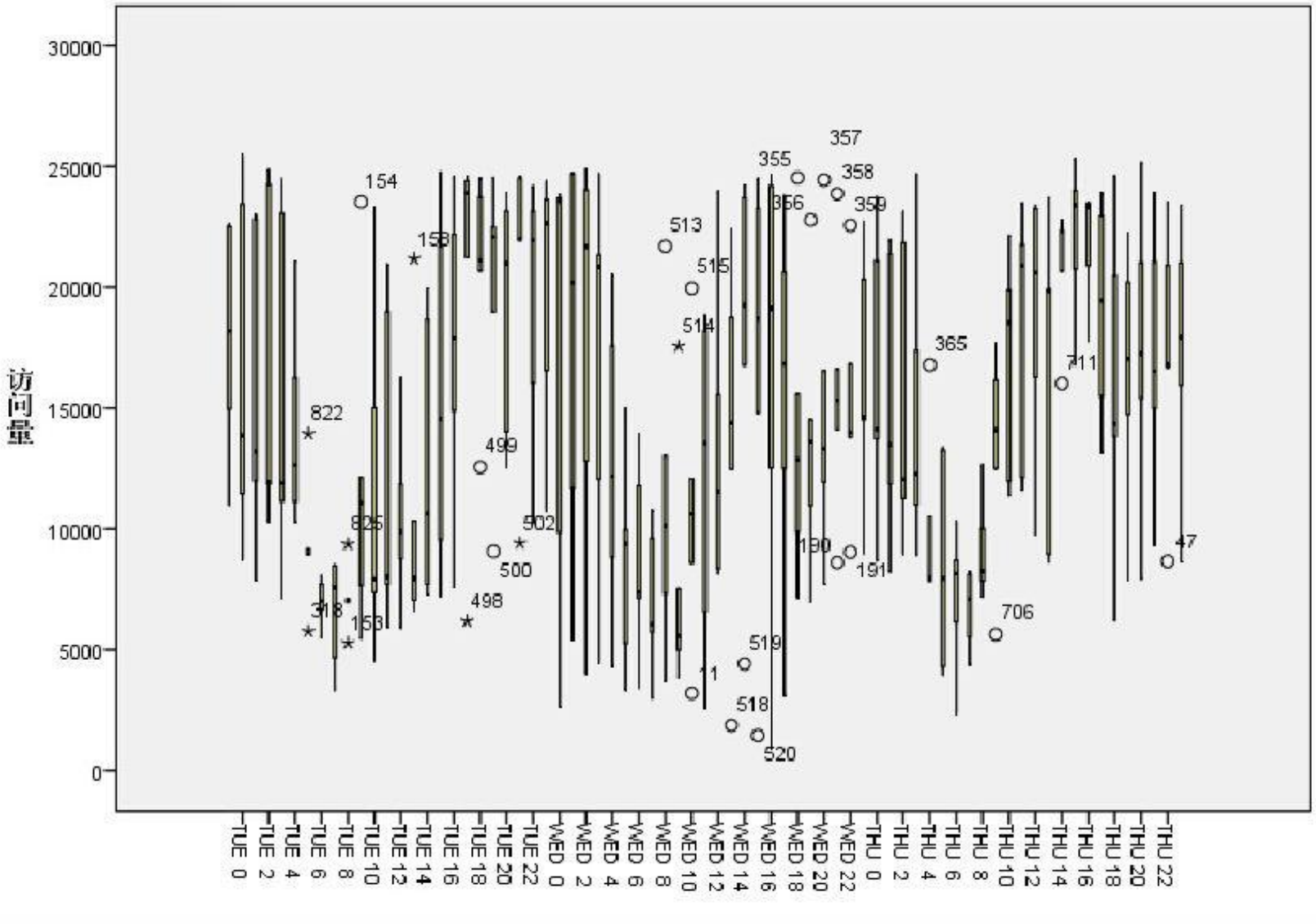

图 4.6　一周内不同时段访问量描述统计图（续）

图4.6中是从周五0点开始一直到周四的23点结束，统计每个时段的访问量。每个小时对应图形中从上到下分别是最大值、上四分位数、中位数、下四分位数和极小值，图中圆圈表示温和的异常值，星型表示极端的异常值。由于原始数据整体存在一些数据丢失，异常值可能多于图中所标出的点，所以规律的寻找主要参考四分位和中位数的数据。

除了前面每天访问量的规律外，一周内的不同时段的访问量也有些差别：

（1）周五与其他几天不同的是，从下午17点开始一直到周六的3点，访问量都比较大，而且这几个时段的访问量中位数、四分位数都很相似，说明这个时段访问量大且波动较小，用户迎来了周末，上网熬夜的人大大增多。

（2）周六周日的访问量从全天来看，各个时段都要比工作日的访问量少，原因在于休息日有大段的空闲时间，部分用户改用WIFI联网或使用电脑上网，部分用户选择其他休闲娱乐方式。一星期中每天凌晨至早上的访问量都处在最低点，工作日中一般在8点左右访问量开始回升，而周六的访问量是在12点左右有所回升，周日的则是在上午10点。说明周末多数人都会多睡一会，周六睡懒觉的时间比周日的要长一些。

（3）周一的访问量较其他几天工作日来说访问量处在较高水平，上午的访问量较其他工作日差别不大，从中午12点开始，一直到晚上24点访问量都处于比较大的时段，且波动比较小，原因在于刚刚进入工作日，工作状态比较差，还处在周末的休闲状态，用户充分利用碎片化时间进行上网娱乐。

（4）周二、周三、周四的访问量曲线除了周三的一段时间有些不同，其他差别不大。周三与其他两天不同的是，从下午17点开始一直到晚上22点，访问量都比较低，形成一个波谷，其他两天在这段时间都处在较大访问量的水平。周三处于一周的中间，是距离周末最远的一天，周三这天工作状态应该比较好，在下班过后身心比较疲惫，选择休息，而进行的上网娱乐休闲就少一些。

通过比较一周内不同天数不同时间段的访问量，将每天较多的时段统计出来，如表4.4。

表 4.4　一周内访问量大的时间段

星期	访问量大的时间段	小时数	访问量相比
周一	12:00-14:00	2	大
	14:00-17:00	3	较大
	17:00-23:00	6	大
周二	17:00-24:00	7	大
周三	1:00-3:00	2	大
	14:00-17:00	3	较大
周四	10:00-13:00	3	较大

	14:00-16:00	2	大
	19:00-23:00	4	较大
周五	0:00-4:00	4	较大
	16:00-24:00	8	大
周六	0:00-3:00	3	大
周日	21:00-23:00	2	较大

4.4　时间分布规律建模

按小时对用户访问量进行排序形成了一个时间序列<y_1，y_2，y_3，…，y_t，…>，其中 t 表示小时，y_t 表示在 t 这个小时内用户的访问量。对这个时间序列建立模型，步骤如下：

（1）判断时间序列的趋势及周期性

画出访问量对应小时的时间序列图、访问量的自相关、偏自相关函数图，在前面小节中已判断出时间序列的趋势及周期性，序列存在趋势及周期性，且周期为 24。

（2）将序列变换位平稳序列

因序列存在趋势及周期性，所以对序列进行差分、季节性差分，然后进行差分后数据的自相关检验，如果不平稳继续差分，直至序列平稳为止。

（3）模型识别

选定一个模型，将不同模型的理论特征作为标准，观测实际序列与标准的接近程度，根据分类比较分析的结果选定模型的类别。

（4）模型建立

对初步选定的模型进行参数估计及假设检验，来判断模型的合理性，如不恰当则返回到上一步，重新选定模型。建立模型时，要从低阶到高阶依次建立，直到增加模型的阶数系数不显著。

第二章对时间序列模型的几种类型做了介绍，结合各模型的特征与实际序列，选择了 ARIMA 模型，通过以上几个步骤，建立了如下模型。下面的表应该有表头

模型描述

			模型类型
模型标识	访问量	模型_1	ARIMA(1,0,0)(1,0,1)

模型拟合度

拟合统计信息	平均值	最小值(M)	最大值(X)	百分位(T)						
				5	10	25	50	75	90	95
平稳的R方	.752	.752	.752	.752	.752	.752	.752	.752	.752	.752

R 方	.752	.752	.752	.752	.752	.752	.752	.752	.752	.752
RMSE	3127.955	3127.955	3127.955	3127.955	3127.955	3127.955	3127.955	3127.955	3127.955	3127.955
MAPE	21.168	21.168	21.168	21.168	21.168	21.168	21.168	21.168	21.168	21.168
MaxAPE	537.707	537.707	537.707	537.707	537.707	537.707	537.707	537.707	537.707	537.707
MAE	2225.004	2225.004	2225.004	2225.004	2225.004	2225.004	2225.004	2225.004	2225.004	2225.004
MaxAE	11304.979	11304.979	11304.979	11304.979	11304.979	11304.979	11304.979	11304.979	11304.979	11304.979
标准化的 BIC(L)	16.242	16.242	16.242	16.242	16.242	16.242	16.242	16.242	16.242	16.242

模型统计

模型	预测变量个数	模型拟合度统计信息				Ljung-Box Q(18)			界外值数
		平稳的 R 方	R 方	MAE	MaxAPE	统计	DF	显著性	
访问量-模型_1	1	.752	.752	2225.004	537.707	18.709	15	.227	7

ARIMA 模型参数

					估算	SE	t	显著性
访问量-模型_1	访问量	不转换	常量		10747.202	797.002	13.485	.000
			AR	延迟 1	.805	.021	37.914	.000
			AR，季节性	延迟 1	-.996	.045	-22.152	.000
			MA，季节性	延迟 1	-.987	.080	-12.347	.000
	HOUR, period 24	不转换	Delay		4			
			分子	延迟 0	185.577	21.914	8.468	.000
			分母	延迟 1	.890	.095	9.352	.000
				延迟 2	-.386	.086	-4.459	.000

界外值

					估算	SE	t	显著性
访问量-模型_1	2	18	瞬时	量级	-15415.824	3121.394	-4.939	.000
				衰变因子	.875	.077	11.366	.000
	4	4	创新		-13196.139	3159.587	-4.177	.000
	7	9	瞬时	量级	15659.432	3131.491	5.001	.000
				衰变因子	.872	.078	11.164	.000
	8	18	瞬时	量级	-12675.101	3115.839	-4.068	.000
				衰变因子	.893	.079	11.376	.000

	10	11	瞬时	量级	17235.443	3128.329	5.509	.000
				衰变因子	.852	.083	10.252	.000
	16	14	创新		12420.115	3124.526	3.975	.000
	33	17	创新		-13579.789	3120.777	-4.351	.000

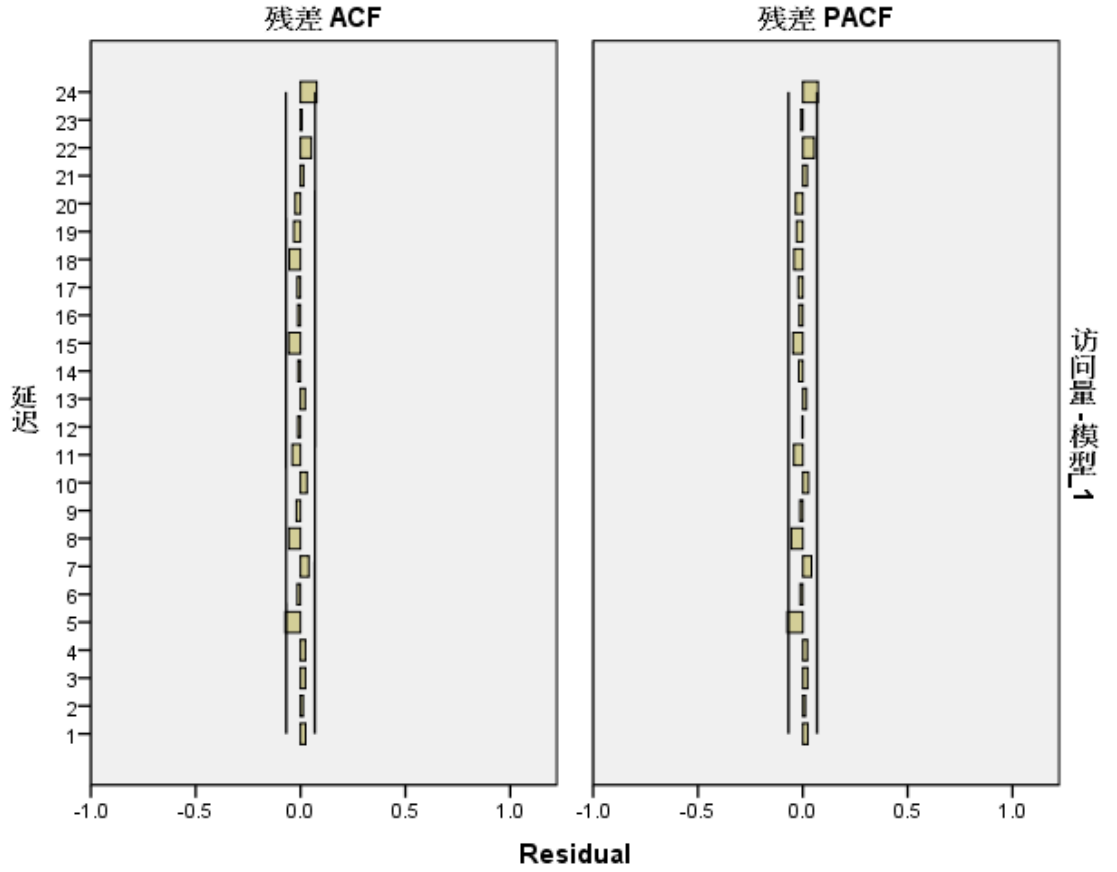

图 4.7　残差自相关图

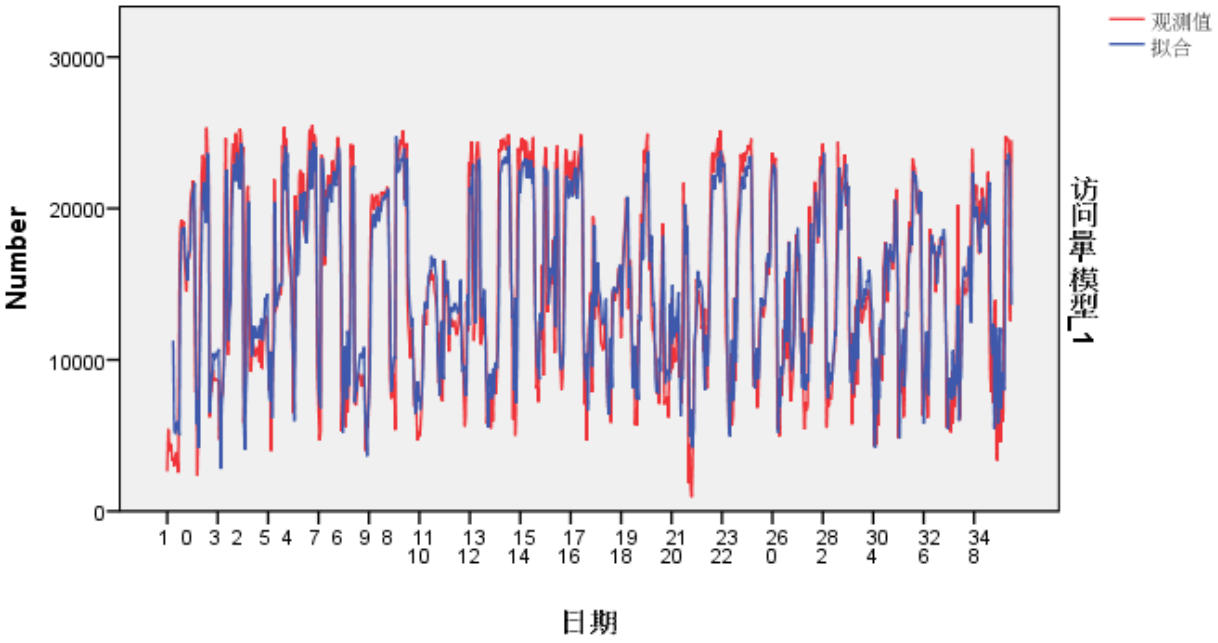

图 4.8　模型拟合比较图

（5）模型评价

通过不断优化建立了ARIMA模型，参数为$(1,0,0)(1,0,1)^{24}$，模型建立的好坏一是通过与建立的其他模型比较，这一步在第四步已做过比较；二是通过相关值的比较：比如残差、R^2、平均绝对误差百分比等。

残差是指因变量的实际值与其估计值之间的离差，如果模型拟合的好，实际值与拟合值之间的差别应该不会太大，对应的的残差绝对值也随之较小，并且围绕着“残差均值=0”这条水平线上下随机分布，残差的自相关就应该没有明显的趋势等特征。从图4.7中可看出，残差ACF和PACF都没有显著的趋势特征，可以初步判断使用的模型是比较恰当的。

可决系数R^2的取值范围为(0，1]，R^2越接近1表示模型拟合的越好，模型中的拟合值越接近实际值。拟合的ARIMA模型的R^2为0.752，再观察图4.8中拟合值和观察值，得出拟合的模型比较成功，下面将模型参数代入到ARIMA模型中，步骤如下：

时间序列对应的模型为$ARIMA(1,0,0)(1,0,1)^{24}$，p=1，d=0，q=0，P=1，D=0，Q=1。p=1时AR(p)为

$$y_t = \emptyset_1 y_{t-1} + \varepsilon_t \tag{4-1}$$

P=1时，AR(P)为

$$y_t = \emptyset_2 y_{t-24} + \varepsilon_t \tag{4-2}$$

Q=1时，MA(Q)为

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-24} \tag{4-3}$$

所以当p=1，d=0，q=0，P=1，D=0，Q=1时，ARIMA模型的公式如下

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-24} + \varepsilon_t - \theta_1 \varepsilon_{t-24} + \mu \tag{4-4}$$

即

$$\text{ARIMA(1,0,0)(1,0,1)24：} y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-24} + \varepsilon_t - \theta_1 \varepsilon_{t-24} + \mu \tag{4-5}$$

代入模型参数，结果为

$$y_t = 10747 + 0.805 y_{t-1} - 0.996 y_{t-24} + 0.987 e_{t-24} + \varepsilon_t \tag{4-6}$$

公式4-6即为用户访问量时间序列模型的公式，通过公式可以预测未来的用户访问量。

4.5　规律总结及营销建议

本章首先对用户访问量的时间序列进行了介绍，在探索时间序列的周期时，先对以天为粒度的时间序列进行了分析，发现周期为7天，每周内不同天内用户的访问量存在差别，访问量由大到小为周五、周一、周六、周日、周二、周四、周三。然后以小时为粒度对时间序列进行分析，周期为24，一个周期也就是一天内的规律为每天7点左右为最小值，20点左右为最大值，从0点到7点，访问量一直下降，从8点到20点大致符合访问量逐渐增加，只有在12点到14点时有稍微下降，20到24点访问量稍有下降但大致相同。对每天的规律进行分析并分解出季节因子后，对一周内不同时间段的访问量差异进行进一步研究，在具体时段上有哪些规律，总结出一周内每天的访问量比较大的时间段，具体规律和特征见图4.9。

图4.9中的时间规律和特征可以应用在广告的时间营销中，时间营销是指改变广告到达传播对象的时间，进而优化营销的效果。营销效果好的时候必定是用户访问量多的时候，看到广告的用户多，转化为目标群体的才会比较多。根据时间规律，如果广告营销的时间是以天为单位的，选择几天进行营销，那么就在周五和周一进行广告营销，因为这两天是一周内访问量最多的时候。如果是选择一天内固定的时间段，那么就选择每天的20点到24点进行广告投放。如果投放广告的时间选择性很灵活，可以任意选择在某天的哪个时段投放，那么首先选择周一的12点到14点、17点到23点；周二的17点到24点；周三的1点到3点；周四的14点到16点，周五的16点到周六的三点几个时间段投放，其次选择周一的14点到17点、周三的14点到17点、周四的10点到13点、周四的19点到周五的4点、周日的21点到23点投放广告。

另外还可根据拟合的公式来预测用户的访问量，结合访问规律，安排网络客服的工作排班情况，在用户访问量大时安排的人数多些，提供优质的客服服务；在访问量比较的少的时候减少客服人员，到达按需分配，尽最大程度提高用户满意度的同时到达资源节约的目的。

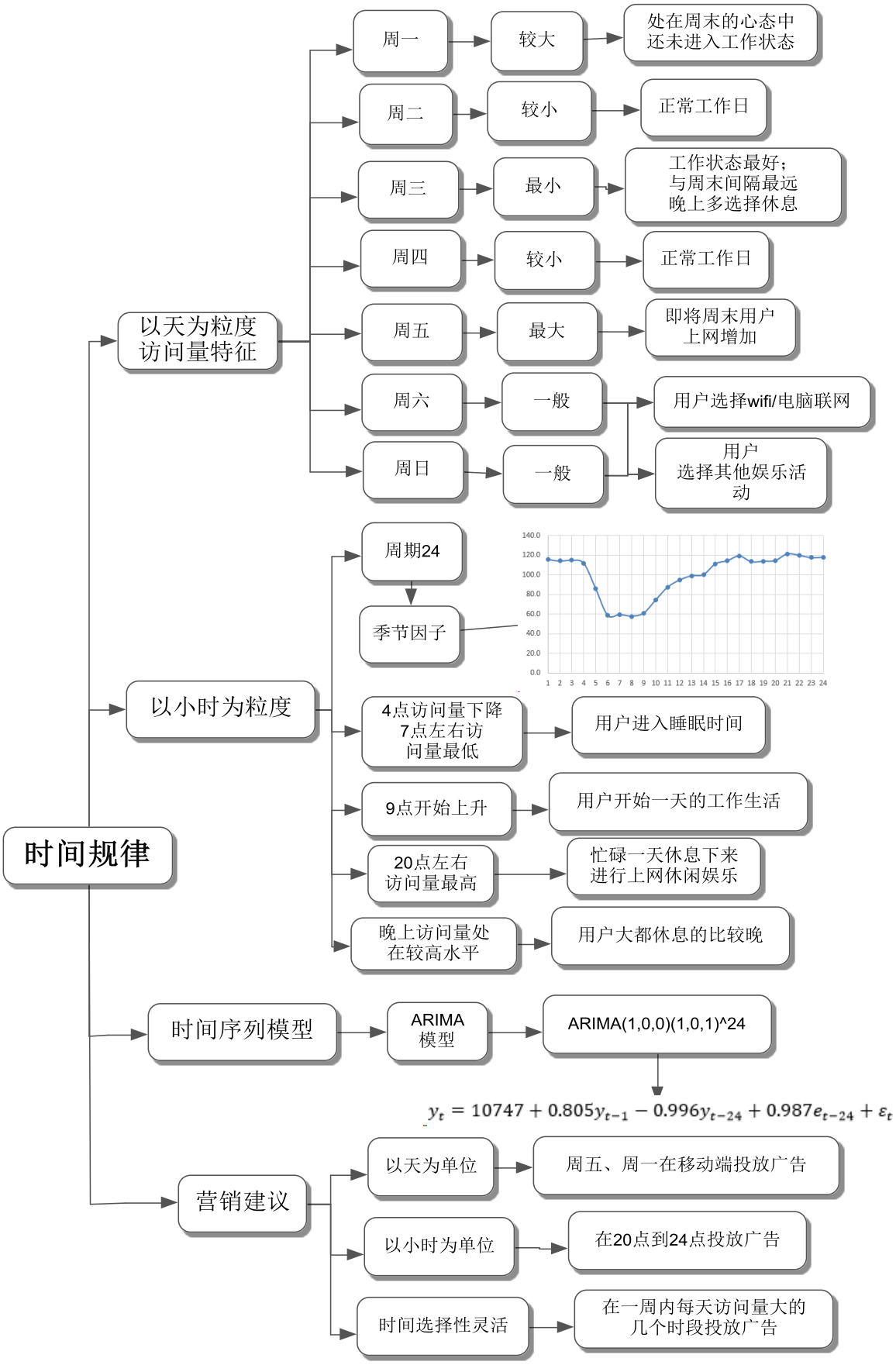

图 4.9　时间规律及特征总结

第五章　竞争结构分析

5.1　业务应用介绍

本章针对各类别的业务应用进行竞争结构分析。竞争结构是指某类产品市场中各企业所占份额以及企业数量分布结构。任何一类市场的竞争结构一般都可以分为完全竞争、垄断竞争、寡头垄断、完全垄断四种。完全竞争是经济学中理想的市场竞争状态，又称纯粹竞争，是一种不受阻碍和干扰的状态，没有企业或消费者可以影响市场。垄断竞争是指许多企业销售相似但不完全相同的产品。寡头垄断是指少数的企业占领市场的大部分份额，新的企业比较难进入；如果是两个企业竞争，占有市场的大部分份额，这种属于寡头垄断的特殊情况，为双寡头垄断。完全垄断是指市场中只有一家企业提供产品和服务，没有替代品，其他企业几乎进入不了这个市场。

业务应用指的是电信用户移动终端里的软件应用，用户在使用APP等软件进行联网时，数据被记录下来，通过对各类业务中各种产品的网络访问量来反映各企业的市场份额。在互联网领域，一般不存在完全竞争和完全垄断，所以业务应用市场竞争结构应该为垄断竞争或寡头垄断。

目前移动互联网主要提供信息即时交互、HTTP应用、电子邮件、VoIP、P2P下载、P2P视频、网络游戏、在线阅读、在线音乐等业务。

信息即时交互服务指的是能够即时发送和接收互联网消息等的业务。如今的信息及时交互服务不仅仅是针对于聊天，而是已经变成综合化的信息交流平台。目前比较流行的软件有QQ、微信、阿里旺旺、LINE、飞信、MSN等。

HTTP应用是当前互联网上最流行的业务，也称作Web业务。目前互联网业务多种多样，但Web业务在互联网结构中的比重依然很大。不受客户端、操作平台的影响，只需一个浏览器就访问Web网站，互联网中的各种资源都可以访问到。如今Web网站不仅内容丰富多彩，信息量巨大，而且图片和动画也异彩纷呈。同时，Web非常易于导航，只需要从一个链接跳到另一个链接，就可以在各个网页或各个站点之间进行浏览了。

电子邮件是一种进行信息交换的电子通讯方式，用户使用网络并以低廉、快捷的方式与世界存在网络的任何一个地方的用户进行联系。不同于普通的纸质邮件，电子邮件除了网络费用，不需要其他费用，并且没有地理位置带来的时间成本，只需几秒就可以到达对方的手中。不仅仅是文字性的东西，图片、音频、视频都可以通过电子邮件进行传递，极大的方便了人们的交流沟通。

VoIP技术是传统通讯行业的巨大变革，它将模拟声音的信号进行数字化转换，然后数据封装在网络上进行传送。如今广泛的应用与Internet中，低成本的传送语音、视频等，VoIP业务有虚拟电话、虚拟语音、网络呼叫中心等。

P2P下载业务，即P2P文件共享。P2P(Peer to Peer，对等网络)技术，是一种直接通过终端进行信息交换和资源共享的应用模式。P2P网络中的任何节点地位都是相同的。P2P不仅仅是一种技术，更体现了互联网自由平等、资源开发共享的特点。P2P下载业务的代表为BT、eMule、迅雷等。

P2P视频的用户不仅是下载，还把媒体流上载给其他的P2P用户，扩大规模，带来更多的资源。用户可以选择直播或者可以控制节目播放顺序的轮播形式，按照自己的喜好进行播放。

5.2 业务应用访问量分布

从访问量来看，移动互联网的业务主要来源于HTTP应用，其次是信息即时交互服务，占总访问量的11.6%，然后是P2P视频和在线音乐，其他业务所占比例都比较小，具体见图5.1。

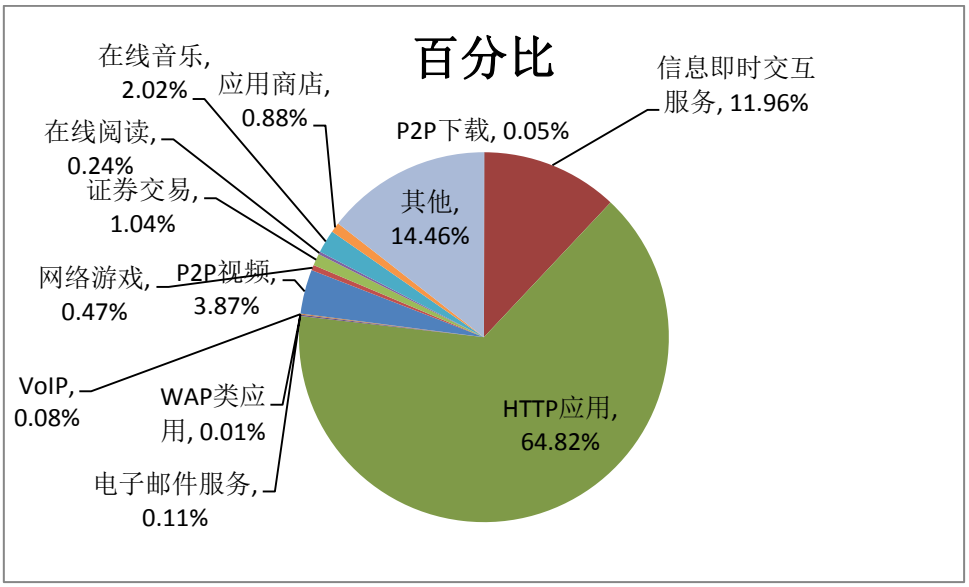

图 5.1 业务访问量百分比

为了进一步研究用户对业务的选择，看每类业务中具体业务的访问量占比情况，统计每个业务的用户的访问量，并按业务类别进行分组。从图5.1可看出，其他业务占比14.46%，比位居第二的信息即时交互服务还要高，所以将其他业务继续细分，添加证券交易、应用商店、社区门户、旅游、动漫等业务类别，所有业务分类见表5.1。

5.1　业务分类表

P2P 下载	信息即时交互服务	P2P 视频	HTTP 应用	电子邮件服务
网络游戏	证券交易	在线阅读	在线音乐	地图导航
应用商店	社区门户	旅游	动漫	云盘
VoIP	支付	生活	购物	其他

每个业务类别为一组，将每组中所有业务的访问量统计出来，换算成占此类市场总访问量的百分比，按所占比例进行降序排名。这时每类业务的用户访问量的百分比就对应为此类业务应用市场的市场占有率，部分业务类别中市场占有率排名前10的业务应用如表5.2所示。

表 5.2　业务应用市场占有率表

P2P 下载			信息即时交互服务			在线阅读		
排名	业务编号	市场占有率	排名	业务编号	市场占有率	排名	业务编号	市场占有率
1	110	51.414%	1	205	70.272%	1	C11	20.475%
2	117	25.987%	2	203	16.659%	2	C17	17.703%
3	111	13.525%	3	201	6.348%	3	C80	15.808%
4	101	8.071%	4	216	3.883%	4	C06	10.278%
5	102	0.794%	5	206	2.352%	5	C02	8.453%
6	104	0.112%	6	243	0.239%	6	C12	6.991%
7	103	0.079%	7	220	0.097%	7	C10	4.700%
8	105	0.014%	8	219	0.073%	8	C46	3.224%
9	112	0.003%	9	227	0.055%	9	C29	1.889%
10	115	0.001%	10	204	0.013%	10	C27	1.682%
P2P 视频			**网络游戏**			**证券交易**		
排名	业务编号	市场占有率	排名	业务编号	市场占有率	排名	业务编号	市场占有率
1	905	35.559%	1	A22	62.599%	1	B01	43.713%
2	902	26.905%	2	A73	9.323%	2	B13	42.971%
3	906	12.579%	3	A01	7.047%	3	B02	7.136%
4	925	6.230%	4	A38	2.248%	4	B10	3.442%
5	918	5.941%	5	A68	2.137%	5	B08	0.847%
6	901	2.384%	6	A25	1.814%	6	B11	0.772%
7	913	2.046%	7	A97	1.781%	7	B04	0.680%
8	904	1.689%	8	A92	1.424%	8	B05	0.337%
9	907	1.097%	9	A03	1.278%	9	B14	0.079%
10	908	0.869%	10	A61	1.103%	10	B07	0.010%
在线音乐			**应用商店**			**社区门户**		
排名	业务编号	市场占有率	排名	业务编号	市场占有率	排名	业务编号	市场占有率
1	D03	43.608%	1	E02	25.773%	1	G51	49.146%
2	D01	20.237%	2	E04	19.246%	2	G82	31.460%
3	D09	17.797%	3	E31	12.769%	3	G34	3.453%
4	D02	8.168%	4	E07	12.203%	4	G83	3.276%
5	D37	2.799%	5	E09	10.645%	5	G55	2.991%

6	D16	2.273%	6	E14	6.872%	6	G38	2.480%
7	D07	1.250%	7	E03	3.787%	7	G87	1.936%
8	D12	1.006%	8	E16	2.310%	8	G72	1.440%
9	D25	0.795%	9	E17	1.825%	9	G84	0.677%
10	D05	0.725%	10	E36	1.513%	10	G10	0.607%

将每组业务应用的市场占有率的排名拟合成曲线，如图5.2。

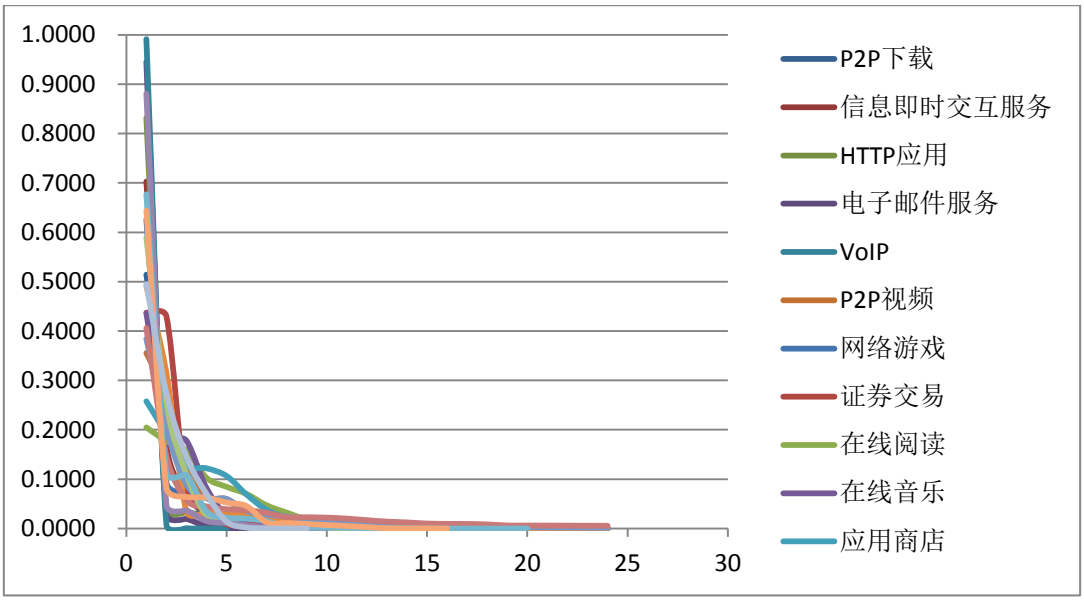

图 5.2　业务排名-访问量占比图

图中每条曲线对应此类别下每个业务的排名与所占份额情况，图中的曲线基本都符合幂定律函数，与幂函数$y = cx^{-a}$曲线相似，初步判定每个业务类别中排名与份额符合幂定律分布，下面通过统计分析来验证其中的幂定律分布规律。

5.3　业务应用占有率分布规律

通过 5.2 节可知，每个类别下的业务排名与所占份额的曲线与幂函数相似，所以下边对各类进行幂函数的拟合，拟合公式为

$$y = cx^{-a} \tag{5-1}$$

经过统计分析步骤的处理，基于幂函数进行拟合，得到了每个业务分类对应曲线的参数，对基于幂定律拟和的分布规律进行进一步分析，得到了 F 检验和 T 检验的 P 值。将业务访问数据基于幂定律拟合的分析结果整理如表 5.3。

表 5.3　业务访问数据幂定律拟合表

业务类别	F 检验	a		c		R-squared
		拟合值	t 检验	拟合值	t 检验	
P2P 下载	0.000	1.431	0.000	0.531	0.000	0.9564
电子邮件服务	0.000	4.861	0.000	0.9448	0.000	0.9997

VoIP	0.000	6.945	0.000	0.9911	0.000	1
P2P 视频	0.000	1.207	0.000	0.3893	0.000	0.9241
网络游戏	0.000	2.39	0.000	0.6236	0.003	0.9942
证券交易	0.000	1.297	0.000	0.4905	0.01	0.7799
在线阅读	0.000	0.9217	0.000	0.2485	0.049	0.8491
在线音乐	0.000	1.339	0.000	0.4516	0.000	0.9578
应用商店	0.003	0.9906	0.003	0.2917	0.002	0.8846
社区门户	0.000	1.49	0.000	0.5154	0.000	0.9236
旅游	0.000	1.302	0.000	0.3961	0.000	0.9851
生活	0.000	1.474	0.000	0.4039	0.000	0.994
地图导航	0.000	1.665	0.000	0.6019	0.000	0.9768
支付	0.000	3.937	0.000	0.8805	0.000	0.9985
购物	0.000	2.202	0.000	0.674	0.000	0.9827
动漫	0.000	2.335	0.000	0.6392	0.000	0.9827
云盘	0.000	1.366	0.000	0.5153	0.000	0.9437
HTTP 服务	0.000	4.012	0.000	0.8315	0.000	0.9971
信息及时交互服务	0.000	2.153	0.000	0.7039	0.000	0.999

通过表 5.3 可以看出，幂定律拟合的可决系数均大于 0.77，且 F 检验和 T 检验的 P 值均小于 0.05，回归拟合结果令人满意，表明拟合模型在统计意义上十分显著。分析结果表明，业务访问量占比对应排名即每个业务类别中业务应用竞争结构符合幂律分布特征，反映出每个领域软件的市场份额与其对应的排名成反比关系，整个市场中大部分的份额集中在少数软件中。

比较各类别中拟合公式的幂指数 a 和系数 c，发现幂指数 a 越大，曲线越陡峭，排名第一的业务占有率越大，此行业的垄断性越大，部分公式的拟合公式曲线如图 5.3。图 5.3 中各曲线从上到下依次是当 a=0.9217、1.431、2.202、3.937，当 a=0.9217 时是图中四个曲线最为平缓的，而当 a=3.937 时曲线就变得很陡峭了。而且 a 越大，第一名的市场占有率就越大，说明行业的垄断性就越大。

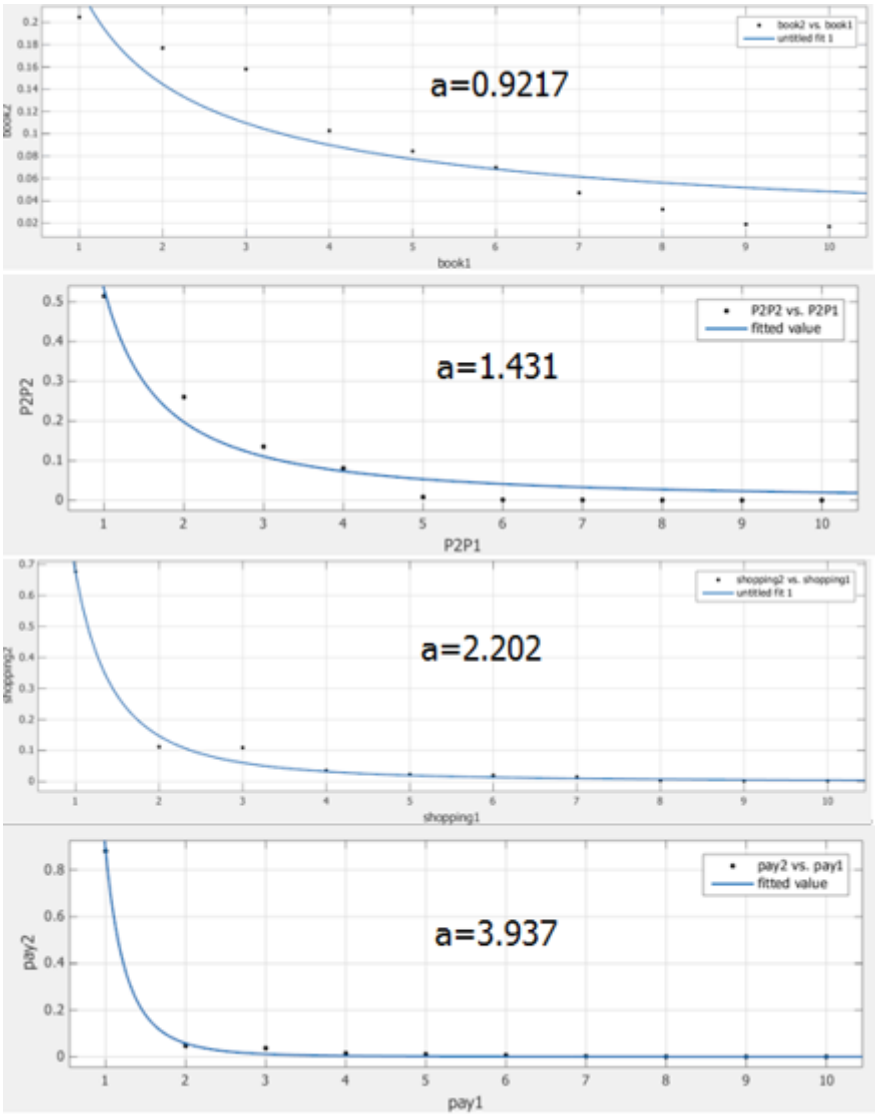

图 5.3　拟合公式曲线图

5.4　各类业务市场竞争结构

计算出每类业务应用的市场占有率公式，分析每类市场的竞争结构，各类业务应用的市场分布情况如表 5.4 所示。

表 5.4　各类业务应用的市场分布

序号	业务类别	公式	占有率高的业务	占有率	业务名称	竞争结构
1	在线阅读	$y=0.2485x^{-0.9217}$	C11 C17 C80 C06	20.475% 17.703% 15.808% 10.278%	掌阅 QQ 书城 豆瓣 起点	竞争
2	应用商店	$y=0.2917x^{-0.9906}$	E02 E04	25.773% 19.246%	安卓市场 AppStore	

			E31 E07 E09	12.769% 12.203% 10.645%	联想 智慧云 豌豆荚	
3	P2P 视频	$y=0.3893x^{-1.207}$	905 902 906 925	35.559% 26.905% 12.579% 6.230%	腾讯视频 爱奇艺 优酷 乐视网	
4	证券交易	$y=0.4905x^{-1.297}$	B01 B13	43.713% 42.971%	同花顺 东方财富网	
5	旅游	$y=0.3961x^{-1.302}$	JB0 J98 JA6	38.406% 20.196% 8.878%	携程旅行网 去哪网 同程旅游	
6	在线音乐	$y=0.4516x^{-1.339}$	D03 D01 D09	43.608% 20.237% 17.797%	QQ 音乐 酷狗 百度音乐	
7	云盘	$y=0.5153x^{-1.366}$	T06 T10 T03	49.521% 27.335% 14.471%	百度云盘 酷云 快盘	
8	P2P 下载	$y=0.531x^{-1.431}$	110 117 111	51.414% 25.987% 13.525%	酷狗音乐盒 Poco 115 优蛋	垄断
9	生活	$y=0.4039x^{-1.474}$	L18 L23	40.593% 14.709%	58 同城 赶集网	
10	社区门户	$y=0.5154x^{-1.49}$	G51 G82	49.146% 31.460%	新浪 百度贴吧	
11	地图导航	$y=0.6019x^{-1.665}$	N09 N05	58.935% 25.159%	百度地图 高德地图	
12	信息及时交互服务	$y=0.7039x^{-2.153}$	205 203	70.272% 16.659%	腾讯(微信、QQ) 旺旺	
13	购物	$y=0.674x^{-2.202}$	R32 R06 R11	67.686% 11.171% 10.905%	淘宝 天猫 京东	
14	网络游戏	$y=0.6236x^{-2.39}$	A22 A73	62.599% 9.323%	新浪游戏 多玩	
15	支付	$y=0.8805x^{-3.937}$	P07	88.075%	支付宝	
16	电子邮件服务	$y=0.9448x^{-4.861}$	507	94.488%	126 邮箱	
17	VoIP	$y=0.9911x^{-6.945}$	706	99.108%	歪歪语音	

表中各类市场按照表 5.4 的排序，占有率从市场中几个企业平分市场到几乎一个企业独占鳌头，竞争结构相对比来说是逐渐从竞争到垄断的一个顺序，市场占有率曲线中幂指数 a 值越小竞争越激烈，幂指数 a 越大可竞争者越少，市场越垄断。

在线阅读、应用商店和 P2P 视频三类业务应用的竞争最为激烈，前四五名业务应用几乎瓜分了整个市场，而且每个的占有率大致相同。在这三类市场中的产品服务差

异性比较小，要努力做出个性化的产品才能更吸引用户，提升市场占有率。新企业或新产品可以进入这样的市场，在竞争激烈的市场中拿出对用户有诚意的产品，可以迅速崛起，争夺市场。

在线阅读中有运营商的自营阅读业务，市场占有率排名排在第五位，占比 8.45%；其前一名市场占有率为 10.28%，如果自营业务想要提升排名，需向第 4 名的市场占有率 10.28%努力。这个行业因为产品差异性小且没有垄断情况存在，争夺市场占有率的机会还是很大的，可以结合用户的兴趣偏好，向用户推送一些感兴趣的书籍，提供个性化的推荐，提升用户满意度，增加用户忠诚度。

应用商店类别下同样有运营商的自营业务，在此类别下排名为 13，市场占有率为 0.64%。应用商店的同质化比较严重，运营商的自营业务也没有特色，所以在市场的竞争力很小，很难在市场占有率上获得提高。如果想要挤入前列，通过应用商店的幂定律公式计算出相应排名的占有率，明确目标后改进产品，突出产品差异性，获取潜在用户，向潜在用户群体推广营销，提升产品的排名与占有率。

P2P 视频中运营商的自营业务排名第 10，市场占有率为 0.8694%，同上面两类业务应用，在竞争激励且产品较为相似的市场中，拿出高质量的产品服务，再冠以个性化的推荐服务，知用户之所爱，会在市场中获得更多的用户，得到更多的收益。

证券交易市场中出现了接近双寡头垄断的竞争结构，同花顺和东方财富网两个业务应用几乎平分秋色的占领了整个市场，在这个市场中的其他企业提升市场占有率的要困难一些，且新的企业和产品不易进入。

旅游、在线音乐、云盘、购物和 P2P 下载几个业务应用市场中，基本都是前三名几乎占据整个市场，接近寡头垄断。其中每类市场的第一名市场占有率都比较大，如果没有一些突发情况的出现，这些市场中的第一名市场占有率会越来越大，强者越强，弱者越弱。第二名追赶第一名比较困难，但也不是完全没有可能改变现状。新的企业或产品可以选择这些市场，但只有一开始展开比较大的规模，获得较多的新用户，否则过高的产品成本、营销成本无法与原有的优秀企业抗衡。

生活、社区门户、地图导航、信息即时交互服务、网络游戏几类市场中前两名占据了大部分的市场，且第一名几乎都超过第二名的两倍还要多。这些市场逐渐走向第一名的垄断，如果排名第一的企业借助大数据的浪潮，想用户推出更精准更个性化的服务，那他们只有可能越来越霸占市场。第二名如果在第一名之前抢占先机，推出与以往产品差异化、个性化的服务，还是有提高市场占有率的机会的。如果是新的企业和产品，建议不要进入这些市场。

支付、电子邮件服务和 VoIP 三类市场中，第一名几乎占据了整个市场，接近与完全垄断的竞争结构。因为统计的都是手机端 APP 业务的网络访问量，所以像 QQ 邮箱这种属于 QQ 等其他软件附加服务的业务访问量就比较少，包括微信支付直接在微信中，

所以支付、电子邮件市场看似接近垄断，其实是指的 APP 的市场占有率。当然同样反映出除了向腾讯旗下的这些附加业务无法统计，支付宝、126 邮箱、歪歪语音在市场中独占鳌头，无法被超越和取代。新的企业和产品一定不要进入这类市场。

5.5　规律总结及建议

本章首先对业务应用、竞争结构进行了介绍，然后对各业务在其市场的占有率进行了统计分析，发现各市场业务的占有率符合幂定律，并对回归公式的参数进行了计算。随后按照幂指数参数将各类市场的进行升序排列，两个指数之间拟合成一条立方模型曲线。最后对各类市场中的竞争结构进行统计，分析出各类业务应用市场的具体情况、竞争激烈程度和垄断程度，幂指数 a 值越小竞争越激烈，幂指数 a 越大可竞争者越少，市场越垄断；并针对运营商的自营业务提升市场占有率的方法进行了说明。本章节统计分析得到的竞争结构的规律和市场特征总结如图 5.3 所示。

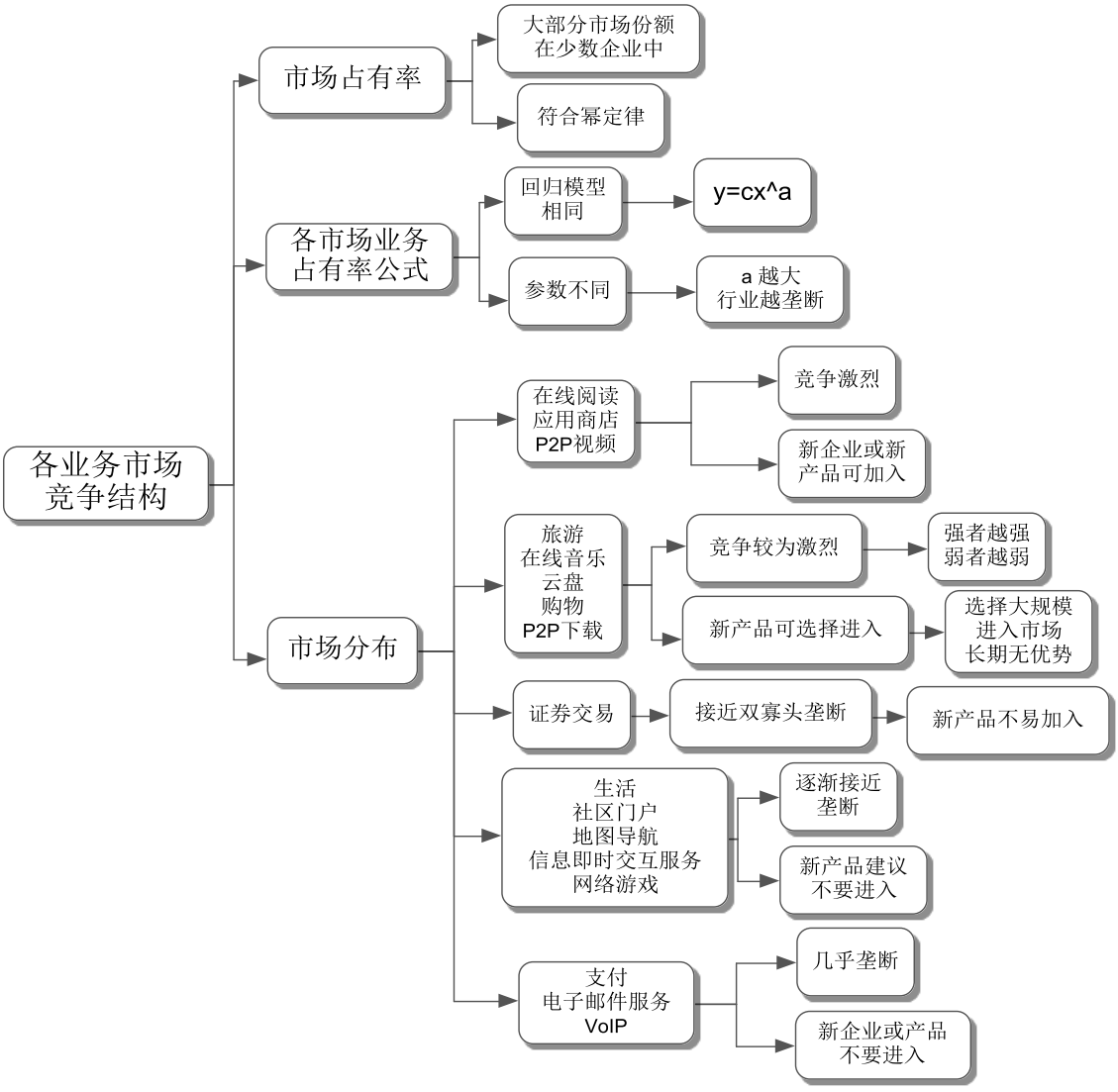

图 5.3　竞争结构规律和市场特征图

在线阅读、应用商店和 P2P 视频三个市场几乎没有垄断，但竞争最为激烈，产品服务差异性比较小，要做出差异化、个性化的产品才能更吸引用户，提升市场占有率。运营商的自营业务可以结合用户的兴趣偏好，将产品针对不同用户群体差异化或提供个性化的服务，提升用户满意度，提高市场占有率。同样如果是新企业或新产品如果能够做出使用户满意的产品，也是能够争夺市场的，可以选择垄断性小、竞争激烈的市场进入。旅游、在线音乐、云盘、购物和 P2P 下载的等市场中，几乎是前三名占据整个市场，排名靠后的企业基本没有竞争力，接近寡头垄断，市场中只会强者越强，弱者越弱。新的企业或产品可以选择这些市场，但需在开始展开大规模，获取很多新用户，否则过高的产品成本、营销成本无法与原有的优秀企业抗衡。其他市场都逐渐接近垄断，第一名很难被后边的企业超越，新进入者也不建议进入这些行业。

第六章　市场细分

6.1　问题描述

本章基于用户兴趣偏好对用户进行市场细分，用户兴趣可通过用户访问的网页反应出来，所以市场细分问题的输入为用户和他们访问各个网站的次数，可以从两个角度来理解这个问题。

（1） 矩阵角度

可以产生一个用户——网站矩阵。矩阵的行和列分别表示用户和网站，矩阵中的元素表示用户访问该网站的次数。

（2） 图论角度

同样，也可以用一个二分图来反映用户和网站的关系。二分图有两组点集，分别代表用户和网站。若两组点集中有两个点之间有边连接，表示用户访问过该类别的网站，边的权重表示用户访问该网站的总次数。

我们使用以上两种等价的方式来描述问题，为了表达更清楚，有时会从某一角度来描述问题。实际上，无论哪个角度都可以等价的按另外一种方式表达。

双边图模型中，图$\mathrm{G}=(\mathrm{V},\mathrm{E})$，顶点集$\mathrm{V}=\{1,2,\ldots\ldots,|\mathrm{V}|\}$，边集$\{\mathrm{i},\mathrm{j}\}$，边的权重为$\mathrm{E_{ij}}$，邻接矩阵 M 定义为

$$\mathrm{M_{ij}}=\begin{cases}E_{ij}, & \text{if edge } \{\mathrm{i},\mathrm{j}\} \text{ exists}\\ 0, & \text{otherwise}\end{cases} \tag{6-1}$$

若给定一个要分割的顶点集V，要分割为 k 个子顶点集$\mathrm{V_1},\mathrm{V_2},\ldots\ldots,\mathrm{V_k}$，我们定义 cut 表示分割后所有属于不同顶点集的边的权重之和，用数学语言可表示为

$$\mathrm{cut}(\mathrm{V_1},\mathrm{V_2},\ldots\ldots,\mathrm{V_k})=\sum_{i<j}cut(V_i,V_j) \tag{6-2}$$

其中

$$cut(V_1,V_2)=\sum_{i\in V_1,j\in V_2}M_{ij} \tag{6-3}$$

结合所分析的用户和网站，将图定义为$G=(U,W,E)$,其中$U=\{u_1,u_2,\ldots,u_m\}$表示用户顶点集，$W=\{w_1,w_2,\ldots,w_n\}$表示网站顶点集，E 表示边集$\{\{u_{\mathrm{i}},w_{\mathrm{j}}\}:\mathrm{u_i}\in U,w_j\in \mathrm{W}\}$。可以注意到用户和用户之间，网站和网站之间是没有边的。连接用户和网站的边代表了其连接的强度，在这里使用用户访问该类别的网络流量所占比例作为其权值。

$$\mathrm{E_{ij}}=\frac{t_{ij}}{\sum_{j=1,2,\ldots\ldots,m}t_{ij}} \tag{6-4}$$

其中t_{ij}表示用户u_{i}访问网站w_{j}产生的总流量，易知$\sum_{j=1,2,\ldots\ldots,m}\mathrm{E_{ij}}=1$。

考虑用户网站矩阵 A 有 m 个用户和 n 个类别，矩阵的每个元素$\mathrm{D_{ij}}$表示权重$\mathrm{E_{ij}}$.则二分图的毗邻矩阵 M 表示为：

$$\mathrm{M}=\begin{pmatrix}0 & A\\ A^{T} & 0\end{pmatrix}$$

注意到 M 的顶点中前 m 个顶点表示用户，后 n 个顶点表示网站。本方法的一个基本假设是用户簇和网站簇之间相互影响，即用户簇决定了网站簇，网站簇同时决定用户簇。比如，给定一组不相交的网站簇$\mathrm{c}_1^{(w)},\mathrm{c}_2^{(w)},\ldots\ldots,\mathrm{c}_{\mathrm{k}}^{(w)}$，则与其相关的用户簇$\mathrm{c}_1^{(u)},\mathrm{c}_2^{(u)},\ldots\ldots,\mathrm{c}_{\mathrm{k}}^{(u)}$可以通过以下方式获得：如果一个给定的用户$\mathrm{u}_i$与网站簇$\mathrm{c}_{\mathrm{l}}^{(w)}$的关联度大于任意其他网站簇，则该用户属于用户簇$\mathrm{c}_{\mathrm{l}}^{(u)}$。

$$\mathrm{c}_{\mathrm{l}}^{(u)}=\left\{u_i:\sum_{j\in c_l^{(w)}}A_{ij}\geq\sum_{j\in c_h^{(w)}}A_{ij},\forall h=1,\ldots\ldots,k\right\}\tag{6-5}$$

对于网站w_j

$$\mathrm{c}_{\mathrm{l}}^{(w)}=\left\{\mathrm{w}_j:\sum_{i\in c_l^{(u)}}A_{ij}\geq\sum_{i\in c_h^{(u)}}A_{ij},\forall h=1,\ldots\ldots,k\right\}\tag{6-6}$$

可以看到，最好的用户网站分类结果是分割后的各个簇之间的交叉权值之和有最小值，即上文提到的 cut 值最小，即：

$$cut\left(c_1^{(u)}\cup c_1^{(w)},\ldots,c_k^{(u)}\cup c_k^{(w)}\right)=min_{V_1,\ldots,V_k}\,cut(V_1,\ldots,V_k)\tag{6-7}$$

其中，$\mathrm{V}_1,\ldots,V_k$是双边图分割后的 k 个部分。

将网站顶点集中的所有网站进行分类，共分为 14 类，具体分类如表 6.1 所示。

表 6.1　网站分类表

编号	类别	内容
1	新闻	新闻资讯，娱乐、体育、军事新闻等
2	阅读	阅读相关网站，包含小说、文学、教育等
3	购物	购物、团购相关网站，各大电商、支付平台等
4	社交	即时通讯软件、交友平台、社区等
5	财经	证券交易、财经理财网站、财经资讯等
6	视频	视频网站，优酷、土豆、爱奇艺、腾讯视频等
7	游戏	手机端游、网页游戏等
8	音乐	音乐盒、在线音乐网站，酷我、酷狗、QQ 音乐等
9	软件应用	手机应用商店、软件网站、手机相关等
10	搜索	搜索引擎、查询工具，百度、好搜、谷歌等
11	旅游	旅游网站、论坛，航空网站、火车票务购买、酒店等网站
12	生活	生活分类信息网站、美食、家居、医疗、宠物等相关网站
13	邮箱	电子邮件相关软件、在线网站等
14	动漫	动漫相关软件、网站等

6.2　算法步骤及实现

本章对用户聚类使用 Phantom 联合聚类算法，前面第三章已经对 Phantom 算法进行了简单的介绍，研究了算法的由来及改进过程。该算法是对 Spectral Graph Partitioning 算法的改进，所以先对 Spectral Graph Partitioning 算法进行详细介绍。

6.2.1　Spectral Graph Partitioning 算法

Spectral Graph Partitioning 算法将用户网站矩阵 A 看做一个二分图，并把聚类问题转化成求最小 cut 的过程，并引入权重 W。W 可以通过不同方式定义，最普遍使用的方法是

$$\mathrm{W_{ii}} = \mathrm{weight(V_i)} = \mathrm{cut(V_1, V_2, .., V_k)} + \mathrm{within(V_i)} \tag{6-8}$$

其中$\mathrm{within(V_i)}$是$\mathrm{V_i}$中顶点所形成的边的权重之和。引入权重后，原 cut 最小化问题转化为以下问题的最小化：

$$\mathrm{Q(V_1, V_2, .., V_k)} = \sum_i \frac{cut(\mathrm{V_1, V_2, .., V_k})}{weight(V_i)} \tag{6-9}$$

上式最小化问题可以通过求分割向量（分割向量将所有顶点划分到不同的簇）解决，而可以证明，归一化特征值问题$\mathrm{Lz} = \lambda \mathrm{Wz}$（L为拉普拉斯矩阵($\mathrm{L} = \mathrm{D} - \mathrm{M}$,D为对角矩阵，$\mathrm{D_{ii}} = \sum_k E_{ik}$），W 为顶点权重对角矩阵）的所有特征值的特征向量代表了分割向量。

以上归一化特征值问题将原问题大大简化。

Dhillon 对该归一化特征值问题经过一系列转化，又将该特征值问题转化为求归一化的原始矩阵A的左右奇异向量问题。

总结后的算法步骤如下：

Step1：给定原始矩阵 A，计算

$$\mathrm{A_n} = D_1^{-\frac{1}{2}} \mathrm{A} D_2^{-\frac{1}{2}} \tag{6-10}$$

其中$\mathrm{D_1}, \mathrm{D_2}$分别是对角矩阵

$$\mathrm{D_1}(i, i) = \sum_j A_{ij} \tag{6-11}$$

$$\mathrm{D_2}(i, i) = \sum_i A_{ij} \tag{6-12}$$

Step2：计算$\mathrm{A_n}$的$\mathrm{h} = \lceil \log_2 k \rceil$个奇异向量$\mathrm{u_2}, \ldots, \mathrm{u_{h+1}}$和$\mathrm{v_2}, \ldots, \mathrm{v_{h+1}}$。生成向量

$$\mathrm{X^T} = [\mathrm{D_1}^{-\frac{1}{2}} \mathrm{U} D_2^{-\frac{1}{2}} \mathrm{V}] \tag{6-13}$$

其中$\mathrm{U} = [\mathrm{u_2}, \ldots, \mathrm{u_{h+1}}]$，$\mathrm{V} = [\mathrm{v_2}, \ldots, \mathrm{v_{h+1}}]$。

Step3：对h维数据X用 k-means 算法得到想要的 k 个簇。

下文称该算法为 Spectral Graph-k-Part 算法。

Spectral Graph-k-Part 算法解决了将原始矩阵分割成许多小矩阵的联合聚类问题，但

该算法存在一些不足，主要体现为以下几个方面：

（1） 矩阵太大时无法处理：用户数量太大，网站数量过多时计算机内存无法承受。

（2） 不知道簇的数量时无法处理，最后只能通过指定簇的数量 k 完成矩阵的分割。

（3） 只能用于"hard" co-clustering，即一行或一列只隶属于一个簇。

6.2.2 Phantom 算法

针对 Spectral Graph-k-Part 存在的一些不足，Phantom 算法提出了一些解决方案。

1、Hourglass Model：沙漏模型，通过降低矩阵规模，使原始矩阵缩小，完成聚类后再将矩阵扩展，再做进一步分析。

2、Divisive hierarchical clustering：自顶而下的层次聚类，通过迭代分层次分组自动决定聚类的数量。

3、Soft Co-Clustering：双向软聚类，使同一行或列可以属于 2 个簇。

下面分别介绍几个解决方案的具体实施步骤。

1、Hourglass Model

将真实数据放入输入矩阵后，矩阵规模会非常大，以致计算机无法处理。所以须将原始矩阵降低规模后输入，完成聚类后再将存储的原始网站信息恢复。

对原始矩阵A（$m \times n$）降低规模，将网站进行分类合并，共分为新闻、阅读、购物、社交等 14 类，具体分类见表 6.1。

假设所有网站分为l个类别，可将原始矩阵A（$m \times n$）形成一个新矩阵A_r，新的矩阵大小为$m \times l$，对A_r使用 divisive hierarchical clustering 方法可以形成若干簇，即若干更小的矩阵。假设形成了 k 个簇，则$C^{(u,l)} = \{c_1^{(u,l)}, c_2^{(u,l)}, \dots, c_k^{(u,l)}\}$，每个簇$c_i^{(u,l)}$都是有相似习惯的用户和有相似用户的网站。$c_i^{(u,l)} = c_i^{(u)} \cup c_i^{(l)}, i = 1, \dots, k.$其中$c_i^{(u)}$和$c_i^{(l)}$分别代表用户和网站类别相应的簇。

对每个$c_i^{(u,l)}$的类别$c_i^{(l)}$扩大，将每个$c_i^{(u,l)}$恢复成原始网站，各自形成新矩阵$A_e(i)$，矩阵各自的大小为$|c_i^u| \times |c_i^{\widehat{w}}|$，其中$\widehat{w}$表示由$c_i^{(l)}$扩展为原始网站后得到的新空间。可以看到新矩阵$A_e(i)$比原始矩阵 A 要小很多，$|c_i^u| \ll |U|$。

改进后，将每个$A_e(i)$再当作 Phantom 的输入矩阵，可以极大地节省了计算机的内存空间和处理时间。

2、Divisive hierarchical clustering

在分类前决定簇的数量并不容易，使用 Divisive hierarchical clustering 算法，通过建立一个二叉树，最终可以自动决定簇的数量。具体步骤如下：

开始把输入矩阵作为一个单一的簇。每一次迭代中，将上一次迭代中获得的簇用 Spectral Graph-2-Part 算法将其分为两个簇。

每次迭代后，对子簇进行检验，如果两个子簇的聚合度都大于指定阈值 T，则划分是好的，可以继续迭代。其中一个簇$\mathrm{c_i^{(a,b)}}$的聚合度的定义为：集合$c_i^{(a)}$和集合$\mathrm{c_i^{(b)}}$的关联权重之和与集合$c_i^{(a)}$和所有顶点的关联度权重之和的比值。数学表达为：

$$\gamma_{\mathrm{i}}^{(a,b)} = \frac{\sum_{h \in c_i^{(a)}} \sum_{k \in c_i^{(b)}} A_{hk}}{\sum_{h \in c_i^{(a)}} \sum_{k \in C^{(b)}} A_{hk}} \tag{6-14}$$

其中$C^{(b)} = \{c_i^{(b)}\}$，$\mathrm{i} \in 1, \ldots, \mathrm{q}$是$\mathrm{c_i^{(a,b)}}$在 b 空间的所有兄弟簇（包括自己）。可以看出，T 值越大，则分类的结果越理想。如果两个子簇之一的聚合度小于指定阈值 T，则分区是不好的，原簇就是不用继续分区的叶子簇。重复上述过程至不能继续分区。叶子簇的数量就是最终簇的数量。

以上改进后，将原算法的输入参数簇的数量 k 替换成阈值 T，避免了原算法中不知道簇的数量时无法分类的缺点。

3、Soft Co-Clustering

以上算法中依然还都是 hard co-clustering，即一行或一列只隶属于一个簇，有时这样生成的分类结果会很不理想，使用 Soft Co-Clustering 的可通过如下方法实现。

在上述 Divisive Hierarchical Co-Clustering 算法的第 3 步，如果两个子簇的聚合度一个小于 T，一个大于 T，则触发 greedy borrowing，使用贪婪算法进行借用，即具有较低聚合度的簇向具有较高聚合度的簇每次借一个使自己的聚合度提升最大的列，直至两个簇的聚合度都大于 T。另外，还需将迭代的停止条件变为两个子簇的聚合度都小于 T。通过使用 Soft Co-Clustering，可以使同一行或列的元素同时属于 2 个簇。

6.2.3　算法实现

1. 实现 Spectral Graph-k-Part 算法，可与后续改进算法比较。
2. 实现改进的 Divisive Hierarchical Co-Clustering 算法，将大矩阵聚类成小矩阵，可以自动决定簇的数量。
3. 增加 Soft Co-Clustering 方法，使一列（一个网站类别）可同时属于两个簇。
4. 将 3 和 4 两方法合并，可以每次同时得到 Hard Co-Clustering 和 Soft Co-Clustering 结果。两方法合并后，算法流程图如图 6.1 所示。

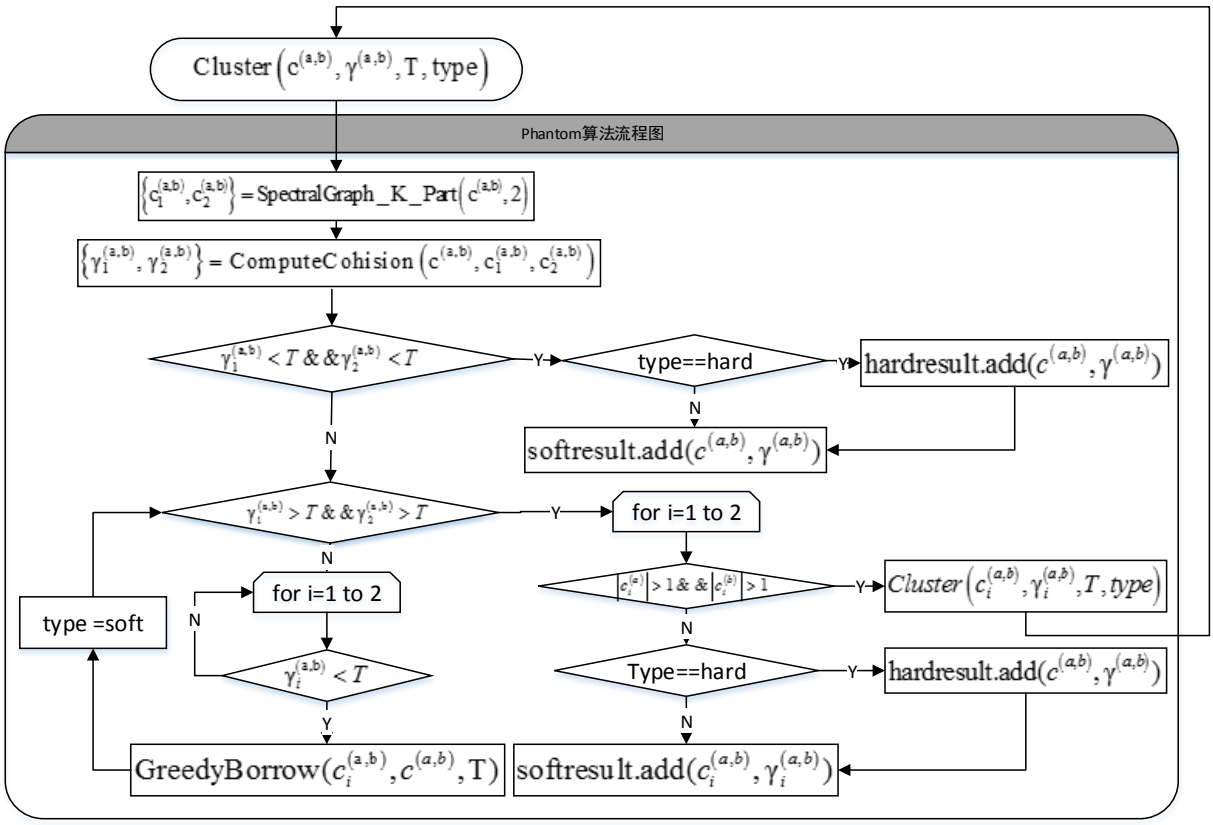

图 6.1　算法流程图

6.3　市场细分结果

前面第三章已对用户数据进行了预处理，将用户已标签化的数据进行聚类，为说明 Phantom（soft）算法的优势，将 Spectral Graph-k-Part、Phantom（hard）、Phantom（soft）三种算法的结果进行比较，各算法结果整理后分别与表 6.2、6.3、6.4 相对应，具体结果如下。

表 6.2　Spectral Graph-k-Part 聚类结果（k=5）

Spectral Graph-k-Part	聚合度	用户数	网站类别
cluster-1	0.891	132741	4
cluster-2	0.780	12875	9
cluster-3	0.873	36152	12，11,8,2
cluster-4	0.789	19474	5
cluster-5	0.562	11392	14,13,10,7,6,3,1
Average Leaf Cluster Cohesiveness = 0.779			
Std Deviation of Leaf Cluster Cohesiveness = 0.131			

表 6.3　Phantom（hard）聚类结果（T=0.86）

Phantom（hard）	聚合度	用户数	网站类别
cluster-1	0.861	207051	14,13,9,7,6,4,3
cluster-2	0.946	28802	12,11,2,1
cluster-3	0.875	20947	8
cluster-4	0.900	24013	10,5
Average Leaf Cluster Cohesiveness = 0.896			
Std Deviation of Leaf Cluster Cohesiveness = 0.038			

表 6.4　Phantom（soft）聚类结果（T=0.86）

Phantom（soft）	聚合度	用户数	网站类别
cluster-1	0.972037	15592	9
cluster-2	0.919645	5296	9,4
cluster-3	0.952775	151364	13,4,3
cluster-4	0.907615	9312	14,7,6
cluster-5	0.937498	31064	4
cluster-6	0.951095	15432	14,4
cluster-7	0.868972	23952	8
cluster-8	0.928763	1260	11,2
cluster-9	0.974090	1300	2
cluster-10	0.906428	10544	12
cluster-11	0.960995	18112	11
cluster-12	0.949206	2004	5
cluster-13	0.993521	7648	10,5
cluster-14	0.973632	1672	1
cluster-15	0.999342	3004	5,1
cluster-16	0.993187	23840	5
Average Leaf Cluster Cohesiveness = 0.949			
Std Deviation of Leaf Cluster Cohesiveness = 0.036			

对比三个算法的结果可知：

从聚类数和用户数来看，Spectral Graph-k-Part 算法聚成了 5 类，第一类用户占总体用户的大部分，且第 5 类用户兴趣显然是没有被分开，聚类效果不是很好。Phantom（hard）算法将用户分成了 4 类，其中第一类用户占很大比重，且用户兴趣没有被细分开，不能分割的原因和阈值 T 的选取有关，若减小 T 至 0.85，则可以将其分开，但会导致类别聚合度较低，聚类效果也不是很好。Phantom（soft）算法将用户分成 16 类，每个类中用户都占一定数量，且不存在兴趣类别没有被分割开的情况，想比较前两种算法，该算法的聚类结果好。

从聚类结果的平均聚合度来看， Spectral Graph-k-Part<Phantom（hard）<Phantom（soft），Spectral Graph-k-Part 的聚合结果只有 0.779，Phantom（hard）算法的提升到 0.896，Phantom（soft）的平均聚合度达到了 0.949，平均聚合度越大表示聚类结果越好。且聚合度标准差也是 Phantom（soft）算法最小，该算法对用户聚类后每个类的聚合度不但高且差异小，说明其中每个类的聚合结果很好。

由此可知 Phantom（soft）算法的聚类效果最好，使用 Phantom（soft）算法的聚类结果，即用户按照兴趣偏好被聚成 16 类，下面 6.4 节对聚类结果进行分析。

6.4 结果分析

通过 Phantom（soft）算法将用户按兴趣偏好聚成了 16 类，将每个类中用户的访问量情况进行详细说明，每类用户在该类别下的访问量占比见表 6.5。

表 6.5 每类用户访问量详细表

簇	网站类型 / 兴趣偏好	新闻	阅读	购物	社交	财经	视频	游戏
1	软件应用	0.14%	0.27%	2.21%	2.19%	0.56%	3.87%	0.54%
2	社交、软件应用	0.35%	0.75%	0.99%	26.07%	0.29%	3.85%	1.23%
3	邮箱、社交、购物	0.38%	0.28%	3.67%	76.05%	0.59%	2.42%	1.12%
4	动漫、游戏、视频	0.32%	0.31%	0.74%	11.30%	0.52%	7.49%	41.27%
5	视频	0.69%	0.37%	0.99%	5.15%	0.54%	77.32%	3.15%
6	视频、社交	1.51%	0.43%	1.15%	37.21%	0.98%	38.23%	4.32%
7	音乐	0.91%	1.13%	0.27%	2.38%	3.56%	1.90%	0.97%
8	旅游、阅读	0.57%	47.32%	0.11%	3.53%	5.19%	2.18%	3.21%
9	阅读	0.38%	75.56%	0.53%	2.38%	4.41%	2.96%	1.02%
10	生活	2.84%	0.47%	0.36%	2.98%	3.40%	2.22%	0.74%
11	旅游	1.79%	0.36%	0.27%	3.33%	3.54%	3.60%	0.58%
12	搜索、社交	0.00%	0.00%	0.45%	20.63%	3.59%	2.69%	0.00%
13	搜索、财经	1.48%	0.49%	1.48%	3.45%	27.59%	10.84%	0.49%
14	新闻	88.59%	0.00%	0.07%	2.98%	2.40%	3.93%	0.51%
15	财经、新闻	40.44%	0.63%	0.08%	13.22%	27.73%	7.12%	0.87%
16	财经	0.51%	0.14%	0.34%	9.44%	73.75%	5.26%	1.61%
簇	网站类型 / 兴趣偏好	音乐	软件应用	搜索	旅游	生活	邮箱	动漫
1	软件应用	5.79%	76.21%	0.00%	5.81%	2.13%	0.08%	0.19%
2	社交、软件应用	6.29%	48.69%	0.00%	8.16%	3.04%	0.07%	0.23%
3	邮箱、社交、购物	5.83%	1.06%	0.02%	4.16%	2.49%	1.45%	0.48%
4	动漫、游戏、视频	9.35%	1.55%	0.01%	3.45%	2.61%	0.00%	21.06%
5	视频	4.88%	1.10%	0.03%	2.99%	2.15%	0.01%	0.65%
6	视频、社交	6.71%	1.35%	0.06%	4.88%	2.67%	0.01%	0.50%
7	音乐	77.54%	0.51%	0.09%	9.87%	0.69%	0.05%	0.12%
8	旅游、阅读	12.35%	0.92%	0.00%	21.95%	2.64%	0.00%	0.04%

9	阅读	2. 95%	0. 75%	0. 04%	2. 01%	6. 98%	0. 00%	0. 03%
10	生活	4. 81%	0. 56%	0. 09%	7. 60%	73. 65%	0. 11%	0. 17%
11	旅游	3. 97%	0. 87%	0. 09%	78. 23%	3. 18%	0. 12%	0. 06%
12	搜索、社交	2. 02%	1. 12%	67. 04%	0. 67%	0. 00%	1. 79%	0. 00%
13	搜索、财经	0. 49%	3. 94%	43. 84%	1. 48%	2. 46%	0. 00%	1. 97%
14	新闻	0. 22%	0. 69%	0. 00%	0. 40%	0. 22%	0. 00%	0. 00%
15	财经、新闻	2. 36%	2. 28%	0. 00%	1. 30%	3. 86%	0. 00%	0. 12%
16	财经	3. 02%	0. 97%	0. 05%	2. 99%	1. 53%	0. 29%	0. 12%

为了更方便观察统计，将表 6.5 画成三维柱形图，如图 6.2 所示。

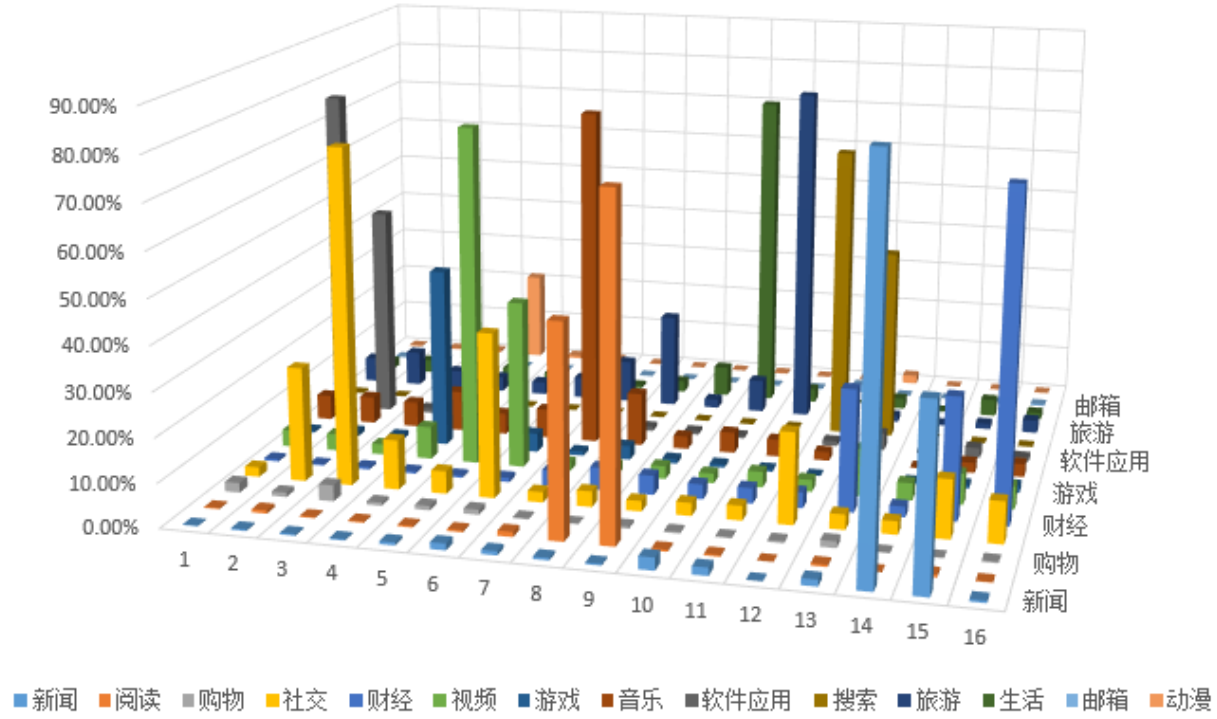

图 6.2　各类用户兴趣偏好

观察 16 个类的用户兴趣特征，比较聚类结果，将 16 个类的用户分为社交型、娱乐型、旅游型、理财型、生活型、新闻型和软件型 7 种，每种类型用户的聚类类别和特征如下：

社交型：3、6、12 类用户属于社交型，他们对社交类应用、网站的访问量比较多。但三类相比不同的是，簇 3 中的用户还比较喜欢在网上购物，使用邮件也比较多，如果针对这类用户进行精准营销，可以在社交网页或软件中进行购物活动的推广，或者邮件推广送营销广告。簇 6 中的用户除了网络社交活动多，看视频的次数也是很多的，对这类用户进行营销时可以进行产社交产品和视频产品的绑定销售或交叉销售。簇 12 用户经常使用搜索引擎，可以将社交产品的营销广告放在搜索引擎中进行推广。

娱乐型：4、5、7、9 类用户属于娱乐型用户，这些用户基本是在网上进行娱乐休

闲活动。簇 4 中的用户是游戏的狂热爱好者，其次是很喜欢动漫，还看一些视频，进行一下网络社交，这部分用户是游戏类和动漫类产品的目标人群，可以在动漫、视频、社交等网站进行游戏推广营销，在游戏、视频、社交网络进行动漫产品的广告推广。簇 5 中的用户绝大部分时间都在看视频，偶尔进行一下网络社交活动，他们是视频类产品的忠实用户，在视频类、社交类网站进行视频类产品的营销推广效果会很好。簇 7 中用户大部分上网都是在听音乐，偶尔会访问旅游类网站，可以对这部分用户进行音乐相关的产品推广。簇 9 中的用户大多在网上进行阅读类休闲活动，这部分人比较偏爱小说、文学等阅读类产品，阅读类市场竞争激烈且产品差异较小，如果能将这部分用户转化成某个产品的忠实用户，市场占有率会得到很大提升，为对这类用户提供个性化的推荐，将阅读类的兴趣偏好进行细分然后进行推荐会取得很好的效果。

旅游型：8、11 类用户属于旅游型用户，这部分用户在这段时间经常进行旅游类相关软件或网站的访问。两者不太相同的是，簇 11 的用户对旅游相关网站情有独钟，偶尔会对新闻、财经、音乐等相关网站进行访问，这部分用户应该近期有旅游出行的准备，应该对他们进行旅游相关产品的推广。簇 8 中用户不但对旅游类网站访问较多，对阅读类软件或网站的访问也很多，对这类用户可以在阅读类网站或软件中进行旅游类产品的广告推广，或者在旅游类产品上进行小说等阅读产品的广告营销。

理财型：13、16 类用户属于理财型用户，对财经、证券交易等网站进行访问比较多。簇 13 中的用户在财经、搜索两类网站的访问比较多，财经类产品的推广营销放在搜索引擎上会比较好。簇 16 的用户大部分都在关注财经类的软件或网站，说明他们对证券、股票、基金等理财类产品很感兴趣，应该对他们进行理财产品的推荐。

生活型：第 10 类用户属于生活型用户，这部分用户大部分的上网时间都在进行生活类网站的访问，他们很可能是居家人士，或者在 58 同城等生活分类网站进行房屋租赁、招聘求职等活动。对这部分用户进行生活类相关产品的推广营销效果会比较好。

新闻型：14、15 类用户属于新闻型用户，对新闻类软件或网站的访问较多。簇 14 中的用户上网的绝大部分时间都在进行新闻资讯的浏览，说明他们对实时的新闻事件、热点话题很关注，他们是新闻类产品的忠实用户。簇 15 中的用户不仅对新闻类网站的访问多，对财经类网站的访问也不少，说明这部分用户不但关注新闻，还对理财方面感兴趣，对这部分用户来说，理财产品的推广营销可以放在新闻等网站或软件中。

软件型：1、2 类用户属于软件型用户，对软件应用商店、手机软件相关网站等访问比较多。簇 1 中的用户访问大都集中在软件应用中，而簇 2 的用户对软件应用和社交访问都比较多。针对推广营销，簇 1 中的用户选择软件应用等产品对其进行推广，簇 2 中用户在社交网站中进行软件产品的推广，或在软件应用商店推荐社交软件，会收到比较好的推广效果。

根据聚类结果还可以进行用户的个性化推荐。每个用户类型中用户的兴趣偏好都相似，都算作相似用户。求得用户的相似度，可以进行基于用户的推荐。把用户对每类网站的兴趣度的集合作为一个用户向量，那用户相似度就转换成在向量空间中用户向量间的夹角余弦，余弦值越大，用户相似度越高。用户余弦相似度$sim(a,b)$的公式如下：

$$sim(a,b)=\frac{\sum_{T_i\in T_{ab}} r_{ai}r_{bi}}{\sqrt{\sum_{T_i\in T_{ab}} r_{ai}{}^2}\sqrt{\sum_{T_i\in T_{ab}} r_{bi}{}^2}} \tag{6-15}$$

其中T_{ab}为用户 a 和用户 b 共同标签的集合，即共同兴趣的集合，r_{ij}表示第i个用户对第j类网站的喜好程度,r_{ai}和r_{bi}分别表示用户 a 和用户 b 对 i 类网站的兴趣度的值。

$$r_{ij}=\frac{C(W_{ij})}{\sum_0^j C(W_{ij})} \tag{6-16}$$

其中，$C(W_{ij})$为用户i对第j类类网站的访问次数，$0\le R_{ij}\le 1$。

基于用户的推荐是目前个性化推荐中实际应用较为成功的，其基本思想是将具有相同兴趣爱好的用户标记为相邻邻居用户，将邻居用户喜欢的产品或服务推荐给用户。基于物品的推荐与基于用户的推荐类似，只是在计算邻居时采用物品本身，即用户对某类物品感兴趣，推荐给用户此类下的其他相似物品。

6.5　各类用户特征总结及建议

本章首先对市场细分进行问题描述，然后对使用的 Phantom 算法进行了详细的说明和介绍，并比较了 Spectral Graph-k-Part、Phantom（hard）、Phantom（soft）三种算法的差异和改进，用三种算法对用户进行聚类，有结果分析得出 Phantom（soft）算法效果最好。对聚类结果进行分析，将 16 类用户分成社交型、娱乐型、旅游型、理财型、生活型、新闻型和软件型 7 种类型，针对每种类型每个聚类簇中用户的特点进行了说明，并针对每个用户簇中用户的兴趣偏好不同进行了营销推广建议，各类型中每个用户簇的特征总结见图 6.3。根据每个簇中用户兴趣求得用户相似度，可以对相似度高的用户进行基于用户的个性化推荐。

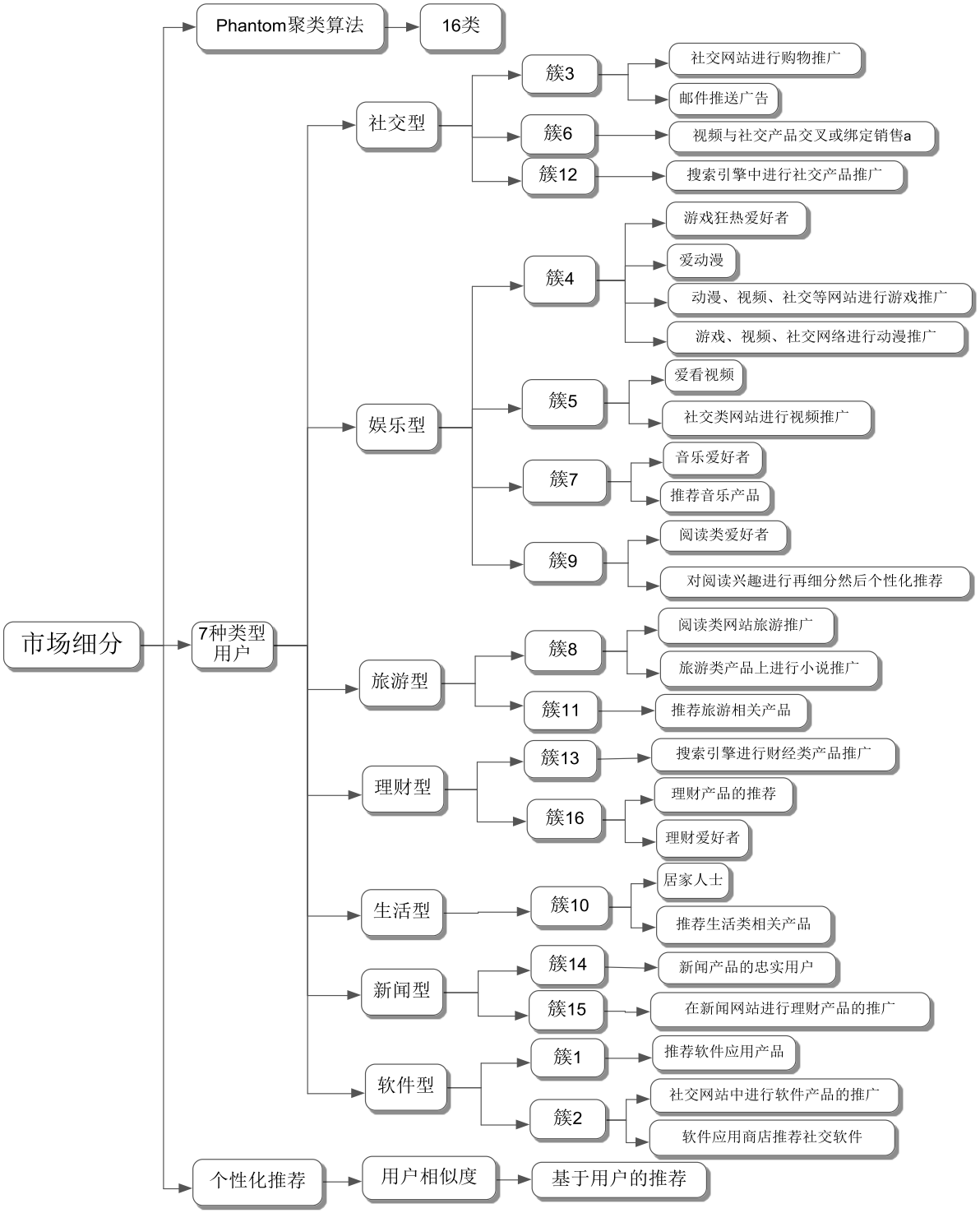

图 6.3　市场细分特征总结

第七章　结论

在科技信息迅猛发展的今天，大数据潮流向我们袭来，它带给各行各业的是潜在的无限价值，各行各业都争相涌进这浪潮中，电信业也不例外。比互联网其它行业起步要晚一些，各运营商也想借着大数据之机稳定自己的市场占有率并争取提高。目前运营商已经意识到要从粗放经营转为精准化、个性化的经营，建立自己的数据挖掘智能系统，不过还在起步阶段。所以本文针对电信行业，对电信某省的移动用户上网数据进行分析研究，了解用户访问网络的特征规律和兴趣偏好，为精准化、个性化服务提供基础。

本文通过对电信用户移动端上网大量数据统计挖掘，研究发现，在时间分布规律上， 以天为粒度时周期为 7 天，一周内用户访问量由大到小为周五、周一、周六、周日、周二、周四、周三；以小时为粒度时周期为 24，一天内的规律为每天 7 点左右为最小值，20 点左右为最大值，从 0 点到 7 点，访问量一直下降，从 8 点到 20 点大致符合访问量逐渐增加，只有在 12 点到 14 点时有稍微下降，20 到 24 点访问量稍有下降但大致相同；对一周内不同时间段的访问量差异进行进一步研究，总结出一周内每天的访问量比较大的时间段；并建立了 ARIMA 时间序列模型；根据这些时间上的分布规律可以进行访问量预测，在广告营销中进行精准的时间营销。如果广告按天进行营销，选择周五和周一；如果是按固定的时间段，选择每天的 20 点到 24 点进行广告投放；如果投放广告的时间选择性很灵活，那么选择周一的 12 点到 14 点、17 点到 23 点；周二的 17 点到 24 点；周三的 1 点到 3 点；周四的 14 点到 16 点，周五的 16 点到周六的三点这几个访问量大的时间段进行投放。在竞争结构分布规律上，发现各类业务应用的市场占有率符合幂定律，并对回归公式的参数进行了计算，市场竞争激烈程度和垄断程度当幂指数 a 值越小竞争越激烈，幂指数 a 越大可竞争者越少，市场越垄断；随后按照幂指数参数将各类市场的进行升序排列，两个指数之间拟合成一条立方模型曲线；接着对各个业务市场竞争情况进行了详细分析，并针对运营商的自营业务提升市场占有率的方法进行了说明；新产品或企业可以加入在线阅读、应用商店和 P2P 视频三类市场，这几类市场竞争激烈但无垄断，谨慎加入旅游、在线音乐、云盘、购物和 P2P 下载这个三足鼎立的市场，其他双寡头或接近垄断的市场不建议加入。针对用户兴趣进行的市场细分中，使用 Phantom（soft）算法对已预处理的用户数据进行聚类，并将聚成的 16 类用户分成社交型、娱乐型、旅游型、理财型、生活型、新闻型和软件型 7 种类型，针对每种类型中不同聚类簇中用户的特点进行了说明，并针对每个用户簇中用户的兴趣偏好进行了特征总结和精准营销建议。总的来说，完成了本文的研究

目标。把三大方面的研究分成三段吧

通过这次研究，达到了预期的目标与效果，同时也对下一步的研究有了启发。因为移动用户 DPI 数据属性的一些局限性，对用户基本信息的了解很少，用户的 ODS 数据中包含大量的用户基础信息，但目前 ODS 数据还在收集过程中。等 ODS 数据收集完毕整合到平台中，与移动 DPI 数据相结合，对用户从多个维度进行数据分析，建立更细致的用户模型，相信对用户精准化、个性化的服务会更有帮助。

参考文献

[1] Anick P. Using terminological feedback for web search refinement: a log-based study[C]. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 88-95.

[2] Zhang Z, Nasraoui O. Mining search engine query logs for query recommendation[C]. Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 1039-1040.

[3] Zhang Z, Nasraoui O. Mining search engine query logs for social filtering-based query recommendation [J]. Applied Soft Computing, 2008, 8(4): 1326-1334.

[4] Agosti M, Crivellari F, Di Nunzio G M. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction [J]. Data Mining and Knowledge Discovery, 2012, 24(3): 663-696.

[5] Murata T. Discovery of user communities based on terms of web log data[J]. New Generation Computing, 2007, 25(3): 293-303.

[6] Yu X, Li M, Paik I, et al. Prediction of web user behavior by discovering temporal relational rules from web log data[C]. Database and Expert Systems Applications. Springer Berlin Heidelberg, 2012: 31-38.

[7] 付关友,朱征宇.个性化服务中基于行为分析的用户兴趣建模[J]. 计算机工程与科学, 2006, 27(12): 76-78.

[8] Li D H, Laurent A, Poncelet P. Mining unexpected Web usage behaviors[M]. Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects. Springer Berlin Heidelberg, 2008: 283-297.

[9] Ríos S A, Velásquez J D, Yasuda H, et al. Web site off-line structure reconfiguration: A web user browsing analysis[C]. Knowledge-Based Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2006: 371-378.

[10] Müller H, Pun T, Squire D. Learning from user behavior in image retrieval: Application of market basket analysis[J]. International Journal of Computer Vision, 2004, 56(1-2): 65-77.

[11] 高琳琦. 基于用户行为分析的自适应新闻推荐模型[J]. 图书情报工作, 2007, 51(6): 77-80.

[12] 佚名. 变革传统 BI 让效益最大化 “个人服务顾问”助四川移动实现精准营销[J]. 中国电信业, 2008, (1):76-77.

[13] 林桂珠, 范鹏飞. 电信企业基于 3G 时代的精准营销[J]. 重庆邮电大学学报: 社会科学版, 2009, 21(4): 22-26.

[14] 何明升. 网络行为的哲学意义[J]. 自然辩证法研究, 2000, 16(11): 56-58.

[15] 马力, 焦李成, 董富强. 一种 Internet 的网络用户行为分析方法的研究[J]. 微电子学与计算机, 2005, 22(7): 124-126.

[16] Paolo G. Applied data mining: Statistical methods for business and industry[J]. 2003.

[17] Xu K, Zhang Z L, Bhattacharyya S. Profiling internet backbone traffic: behavior models and applications[C]. ACM SIGCOMM Computer Communication Review. ACM, 2005, 35(4): 169-180.

[18] Marques Nt H T, Rocha L C D, Guerra P H C, et al. Characterizing broadband user behavior[C]. Proceedings of the 2004 ACM workshop on Next-generation residential broadband challenges. ACM, 2004: 11-18.

[19] Fukuda K, Cho K, Esaki H. The impact of residential broadband traffic on Japanese ISP backbones [J]. ACM SIGCOMM Computer Communication Review, 2005, 35(1): 15-22.

[20] Maia M, Almeida J, Almeida V. Identifying user behavior in online social networks[C]. Proceedings of the 1st workshop on Social network systems. ACM, 2008: 1-6.

[21] 胡海波, 王林. 幂律分布研究简史[J]. 物理, 2005, (12): 889-896.

[22] LILLO F. Limit order placement as an utility maximization problem and the origin of power law distribution of limit order prices [J]. European Physical Journal B -- Condensed Matter, 2007, 55(4): 453-459.

[23] Dominique C R, Rivera-Solis L, Rosiers F D. Determining The Value-at-risk In The Shadow Of The Power Law: The Case Of The SP-500 Index[J]. Journal of Global Business &Technology, 2010, 7.

[24] KOSTANJČAR Z, JEREN B. Emergence of power-law and two-phase behavior in financial market fluctuations. Advances in Complex Systems [J]. Advances in Complex Systems, 2013, 16(1): 1-12.

[25] [31]Anders Johansen, Didier Sornette. Log-periodic power law bubbles in Latin-American and Asian markets and correlated anti-bubbles in Western stock markets: An empirical study[J]. International Journal of Theoretical and Applied Finance, 2001, 4 (6): 853-920.

[26] 司马则茜, 蔡晨, 李建平. 度量银行操作风险的 POT 幂律模型及其应用[J]. 中国管理科学, 2009, (01): 36-41.

[27] 吉翔, 高英. 中国股市的泡沫与反泡沫——基于对数周期性幂律模型的实证研究[J]. 山西财经大学学报, 2012, (12): 27-38.

[28] 山石,邱红. 长尾分布、幂律的产生机制和西蒙模型[J]. 第六届中国管理科学与工程论坛论文集. 上海: 第六届中国管理科学与工程论坛, 2008:886-890

[29] 杨波, 陈忠, 段文奇. 复杂网络幂律函数标度指数的估计与检验[J]. 上海交通大学学报, 2007, (07): 1066-1073.

[30] 刘臣, 单伟, 于晶. 中国学科知识网络的演化研究——基于 1981-2010 年引文数据[J]. 系统工程理论与实践, 2013,(02): 430-436.

[31] 叶作亮, 王雪乔, 宝智红, 陈滨桐. C2C 环境中顾客重复购买行为的实证与建模[J]. 管理科学学报, 2011, (12): 71-78.

[32] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007, 24:10-13.

[33] 郝媛, 高学东, 孟海东. 高维数据对象聚类算法效果分析[J]. 中国管理信息化, 2012, 第 8 期:51-53. DOI:doi:10.3969/j.issn.1673-0194.2012.08.035.

[34] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao. Profiling users in a 3G network using hourglass co-clustering," inProc. ACM MobiCom, 2010.

[35] I. S. Dhillon. Co-clustering documents and words using Bipartite Spectral Graph Partitioning. KDD, 2001.

致谢

首先要感谢我的导师李杰老师对我此次论文的帮助，不仅是对我研究方向的支持、论文思路的指导，还有在我遇到困难或者思路偏差时都是耐心的指正与引导，李老师在研究方面思路清晰、洞察力敏锐，帮我拨开迷雾，让我能够一直走向正确的研究方向，并且积极鼓励我有自己的想法，对我正确的想法予以肯定，鼓励我一路前行。在生活中，李老师也是对我们各种关心和关怀，我们经常与老师谈心，在学习和生活中对我们帮助很大，李老师是我遇到最亲最尊敬的老师，在此对李杰老师表达我最诚挚的感激与祝福。

我也非常感谢我的诸位师兄师姐与同学，在学习和实习中，师兄师姐把他们所了解，对我们有帮助的各种信息都毫无保留的告诉我们，在遇到一些小波折的时候尽力的帮助我们。还有各位实验室小伙伴、班级的同学，在学习上互相探讨，解决课程或科研问题，生活中互相帮助，我们共同成长，争相进步。

校园生活总是最美好的，研究生生涯一晃而过，开学那天还犹如昨天，今天的我们要即将步入社会。在学校的这两年中，不仅在学习上学到了很多，在眼界、胸怀上都感觉变得宽阔了。因为遇到了各位优秀的老师，和一群有理想有志向的同学，我的研究生生活变得丰富充实，我从他们身上也学习到了很多东西，使我受益匪浅，我相信，我会记得各位老师对我们的谆谆教导，记得我们许诺共同努力的目标，希望能够在未来的社会工作中继续认真努力的脚步，踏踏实实的为社会贡献我的一份力量，争取为母校增添一份光彩，为社会做出一份贡献。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日