分类号 TP393 学号 09050014

UDC 000 密级 公 开

工学硕士学位论文

社会电视用户行为分析

硕士生姓名 何 速

学 科 专 业 控制科学与工程

研 究 方 向 信息系统工程

指 导 教 师 王晖 教授

国防科学技术大学研究生院

二〇一一年十一月

User behavior Analysis of Social TV

Candidate：Su He

Advisor：Prof. Hui Wang

A thesis

Submitted in partial fulfillment of the requirements

for the degree of Master of Engineering

in Control Science and Engineering

Graduate School of National University of Defense Technology

Changsha，Hunan，P.R.China

November，2011

目 录

表 目 录

图 目 录

摘 要

随着信息技术的发展，特别是 WEB 2.0 的推出及网络电视的普及，越来越多的用户通过互联网来获取信息，并在互联网上发表有关社会、经济、政治等方面的舆论。互联网给人们带来便利的同时也伴随着安全问题，在互联网上，用户能更容易的组成团体，歪曲事实，对一些敏感话题能够起到推波助澜的效果，所以对互联网上的用户行为进行鉴别的分析尤为重要。

本文主要研究社会电视中的用户行为，以新浪微博用户及 P2P 网络电视用户为基础，进行了数据挖掘技术在用户行为分析中的应用研究。主要内容包括：

第一，微博用户行为分析。该内容主要包括焦点人物分析和活跃用户行为分析两部分，其中焦点人物分析，主要对网络中粉丝数量众多，在信息转播中起到关键作用的节点用户，进行专门的分析，主要包括焦点人物粉丝地理位置分布分析；关注度分析和基于活跃度的用户分类。另外活跃用户行为分析，主要针对关注某一事件，发帖量和转发量较多的前若干用户，就其发帖量分布、单用户发帖行为和用户网络拓扑结构进行分析。

第二，网络电视用户行为分析。该内容主要针对用户在线时长分布、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用户行为进行了深入研究。这四个方面的内容都是以 PPlive、PPStream 和 UUSee 三个平台的湖南卫视频道的数据为基础展开的，其中用户在线时长分布主要描述了用户观看湖南卫视的持续时间，反映了用户对湖南卫视频道的忠诚度； 根据用户在线时长的长短，本文将用户分为三类：轻度收看者、中度收看者和重度收看者；针对三类不同的用户，本文分别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频道用户在各个时间点的数量分布；网络电视用户地理位置可视化利用 GoogleMap API，将用户节点描绘在地图上；最后本文对用户到达率进行了研究。

第三，将微博平台用户行为与 P2P TV 平台用户行为进行对比分析。主要就用户在线数量、用户在线时长分布、用户在线时间演化和用户地理分布四个方面展开，分析就某一事件，两个网络平台用户行为表现上的异同。

主题词：社会电视、微博用户行为、网络电视用户行为、数据挖掘

ABSTRACT

With the development of information technology, especially the launch of WEB 2.0 and the popularity of network TV, more and more users exploit the Internet to get information, and published public opinion online on social, economic, political, and other aspects. However, the Internet brings not convenience but also safety problem. On the Internet, the user can more easily grouped, distort the facts, and become propellent of some sensitive topics, so is especially important to analysis and identify the behavior oInternet user. In the paper we analyze the user behavior of social TV based on Microblog and P2P TV dataset. The main content is:

First, the Microblog user behavior analysis.The content mainly includes two parts: key character behavior analysis and active user behavior analysis. The former mainly focused on the node user who have lots of network fans and played a key role in the information broadcast. Including the fans' geographical location distribution analysis and public attention and active degrees analysis. The later focused on some certain event and the users whose post and transmit number were on the top.

Second, the network TV user behavior analysis. The main content included the user's online time distribution analysis, online hour based user classification analysis, the online hour evolution of classificated user, the network TV user geographical location evolution analysis, user arrival rate and other user behavior analysis. Those four analysis all based on the date of Hunan Satellite TV Channel. The distribution of user online hour described the continue watching duration, reflected user' loyalty to Hunan Satellite TV Channel. According to the length of online hour this paper divided network TV user into three kinds: the mild viewer, moderate viewer and severe viewer. It also reflected the distribution of Hunan Satellite TV Channel user in different time point.The network TV user geographical location evolution is using GoogleMaps API to node the user location in the map. At last, the article studied the user arrival rate.

Third, the comparison and analysis of Microblog platform user behavior and P2P TV platform user behavior.Mainly included the number of online users, the distribution of user online hour, the online hour evolution and the geographic distribution.And analysis certain event to shoe the difference and similarity of those two network platform.

Key Words：Social TV,User behavior of Microblog, User behavior of P2P TV, Data mining

第一章　绪论

1.1　论文背景及目的

社会电视（Social TV）是指在电视收看过程中，同步支持与电视内容相关的社会交互的一种新型电视服务。本文主要就社会电视用户在选秀或选举类节目中的行为展开研究，主要涉及网络电视用户和网络电视相关的微博用户。随着网络电视逐渐融入到网民的生活，越来越多的用户通过网络电视观看自己喜欢的节目，并发表评论。而微博作为新型的消息传送平台，以用户数量众多，消息传递速度快为基本特点，已经成为了人们关注时事消息的一个主要平台，并已深入渗透到网民的日常生活中。在微博中用户没有任何信息接入的门槛，用户不需要经过其他微博主的同意就可以对其添加关注，而且微博主发布的信息会自动在关注者的页面上显示，关注者看到信息后也可以同时成为发布者。在微博的传播环境下，人人都可以发声，人人都可以选择自己想要听取的和不想听取的意见，受众接收到的意见有来自传统媒体设定好的，而更多地是来自其他微博用户的，受众不是单纯地接收媒体或者其他用户给予的意见，而是可以直接参与到意见的讨论中，发表自己的意见，根据自己的思想倾向形成自己的舆论圈。在社会电视日益兴起的时代，处于舆论报道尾部的普通大众越来越有力地影响着人们的意见选择和舆论走向。近年来由于微博而吵的沸沸扬扬的事情比比皆是，如“药家鑫杀人事件”、“郭美美事件”、“日记门”等，所以针对社会电视用户行为分析十分必要。

本文以新浪微博和网络电视的用户为基础，运用数据挖掘技术，对用户的行为进行建模分析和挖掘，找到用户参与热点事件的规律。

1.2　研究意义

典型的社会电视系统集成语音通信、文本交流、语境感知、节目推荐和评价等功能,如Alcatel公司推出的Amigo TV[1]，和Philo Media公司的Philo[2]社会化电视系统。2010 年，美国麻省理工大学《Technology Review》杂志将社会电视列入当年 10 项最重要的新兴技术，预测它将成为下一代的主流电视媒体形式。无独有偶，美国著名网络电子类杂志《Wired》在对 2011 年六大技术趋势的预测中，将社会电视排在其中第三位。而著名媒体公司Endemol集团总裁Ynon Kreiz 2011 年出席慕尼黑的国际数字生活设计(Digital Life Design, DLD)大会时称：在线社会媒体和电视的融合，为电视观众提供了交互，联络，推荐和共享收看体验的能力，将大大提振整个电视产业的前景。

目前社会电视的实现方式主要包括两种：一种是沿袭早期交互式电视的设计思路，将传统电视屏幕与在线论坛相结合，如Alcatel的Amigo TV, Microsoft Labs的Media Center Buddies[3],荷兰TNO ICT研究中心的ConnecTV[4],以及Motorola的STV[5]等。从已有文献和资料看，这些研究大都处于实验室验证阶段，尚没有面向公众的大规模的运营服务。另一种实现方式得益于网络电视和在线社会媒体的迅速发展，人们在互联网上收看电视节目的同时，通过关联的在线社会媒体与异地朋友进行实时地讨论和交流，并且互相评价和推荐节目内容，不仅让人们可与亲人朋友或兴趣相近的其他人分享视觉体验，而且在日益增长的电视频道和内容中，能更容易地找到适合观看的节目。如直接面向消费者的社会化电视媒体平台Philo[3]，Tunerfish和FanTalkTV[6]等，采用了与地理位置“签到”（checking in）类似的媒体签到概念，通过关联微博，让朋友知道他在看什么电视节目并进行互相交流，从而构建了一幅视频领域的社交图谱。此外，还有一些面向独立电视网运营商提供品牌社会电视服务的社会电视平台，如LiveHive Systems的tvClickr[7]和Ex Machina的PlayToTV[8]。在国内，著名网络电视运营商PPTV（PPLive）于 2010 年推出PPTV财经直播频道，也将新浪微博的互动性融入电视频道中，构成了面向金融财经类受众的社会电视平台。由于在线社会媒体和网络电视在网络用户中的高覆盖率，在可以预见的将来，这种结合网络电视和在线社会媒体的社会电视形式，极有可能成为事实上的社会电视主流实现方式，并成为一种新型的基于互联网的大众媒体。从受众的可测性和媒体平台的大众性出发，本文所关注和研究的是社会电视的第二种实现方式，即网络电视与微博的结合，之所以选择微博作为在线社会媒体主要由其自身特点决定。根据新浪微博 2010 年的统计数据微博用户数量 10 个月内增加了 3 倍，呈现短时间内增长迅速的特点，且用户分布相对集中。根据艾瑞网民连续用户行为研究系统iUserTracker[9]数据统计显示到 2011 年 2 月中国微博用户规模已达 2 亿，且仍呈增长趋势，其用户群体之大是其他在线社会网络所不能比的，另外微博通过加关注来确定两个用户之间的关系，而且不同以往的在线社会网络，即不需博主确认即可建立联系。图 1.1 为 2009 年至 2011 年中国微博用户发展情况。

图 1.1 微博用户发展趋势

与传统的社交网站用户相比，微博呈现出其独有的特点，具体表现如表 1.1 所示：

表 1.1 微博与传统博客、社交网络的特点对比

类 别	微 博	博 客	社交网络
建立沟通	单方面行为（即使用关注无需博主确认就能建立联系）	开放式行为（通过搜索引擎或链接可以直接查看博主文章）	互动式行为（一方提出申请后经另一方确认方可建立联系）
内容发布	简单随意发布	深思熟虑发布	随手转帖发布
传播速度	极快	一般	较快
传播方式	垂直辐射	点对多	点对点
传播范围	较广	极广	一般

由于微博自身的特点使得消息在微博中传播十分迅速，受众接受消息快捷简便。这里所说的受众是指媒体传播内容的接受者，受众研究对于媒体具有重要的意义，主要体现在两个方面：首先，受众是市场经济环境下媒体能够获取广告经营收入的关键，受众的收看行为对媒体内容的安排、时间调配、广告价格等，有着极大的影响。其次，受众是通过媒体进行信息和舆论传播的主体。在社会危机事件信息沟通、重大公共安全信息扩散等过程中，大众媒体具有“议程影响”功能，受到某种议程影响的受众成员会按照该大众媒体对这些问题的重视程度调整自己对问题重要性的看法。长期以来，大众媒体的受众行为研究，及其对受众群

体的影响力研究一直是传播学关注的热点。

作为一种新型的大众媒体，结合网络电视和在线社会媒体二者于一体的社会电视的出现，不仅改变了受众收看电视的行为和习惯，同时在对受众的传播影响能力和方式上，也存在着与传统电视平台的显著不同，因而社会电视受众行为及其影响力的研究面临新的挑战，主要体现在以下几个方面：

1、社会电视受众复合行为数据的多源性与异类性

大众媒体受众行为包括媒体接触性行为和对媒体内容的心理认知行为。社会电视平台包含网络电视和在线社会媒体两个功能模块，对网络电视模块的测量可以得到社会电视受众的媒体接触行为，对于在线社会媒体的测量可得到社会电视受众心理认知行为。但是目前社会电视中网络电视和在线社会媒体两个功能模块在架构上存在较强的独立性，使得分别测量得到的受众行为数据，存在多源异类的特点，而且在空间上和时间上都不存在显式的关联，难于得到完整的、一致的社会电视受众复合行为信息。如何实现多源受众行为数据的同步测量，并且在时空一致的基础上实现多源异类的受众行为数据的关联和融合，从而获取完整可用的受众行为数据，是社会电视受众行为测量所面临的关键问题。

2、社会电视受众行为的多态性与复杂性

与传统电视或其他传统媒体不同，社会电视具有高度的交互性。受众不仅仅是被动地封闭地接受信息，而且同时通过关联的在线社会媒体共享收看体验，包括推荐和评价节目内容，互相交流态度和情感倾向等。从信息传播和舆论形成角度看，强调受众交互和参与的社会电视也具有一定程度的“自媒体”特性。因而在受众行为上表现出许多不同于传统电视受众的地方，其行为特征和机理都具有复杂、多态的特点，从而大大增加社会电视受众行为分析和理解的难度。例如，社会电视受众行为是否包含自发行为和外力驱动的推手行为等典型模式？这些行为模式分别具有什么样的动力学特征和机理？这些问题都是传统电视受众分析中不存在，而在社会电视受众行为研究中需要解决的新问题。

3、社会电视信息传播过程的开放性

社会电视与传统电视媒体的最大不同是，它更加强调受众在信息传播过程中的参与和交互。社会电视的信息传播过程不再是一个以电视台为主体的封闭系统，受众不仅仅是被动的信息接受者，而是积极地参与到信息传播过程中，受众行为在一定程度上甚至可能影响传播效果。通过测量方法研究和应用，社会电视平台已经能够提供用户网络接触行为和收看的心理态度两方面的数据，基于这些测量数据，我们在研究社会电视（或其中的具体节目内容和频道）影响力的精确评估方法时，面临的问题是，作为一种强调用户交互和参与的大众媒体，社会电视的信息传播和舆论传播过程，在多大程度上会受到推手行为的影响？社会电视中具

有外力驱动型行为模式的受众（推手）的存在，是否会干扰我们对其认知域影响力的评估？因而，需要在考虑典型受众行为干扰的情况，研究和提出新的社会电视认知域影响力评估方法，以便对社会电视的认知与影响力进行全面、及时、准确地判断。

本文提出的社会电视平台（包括具体的电视内容，如节目和频道）的受众行为研究，以及其认知域影响力评估的研究，是精确全面地了解社会电视这一新型媒体形式的传播特点和传播规律的必要手段，同时也是对人类行为动力学的一个实例研究，有助于定量的了解人类行为，为众多由人类行为发起的系统的复杂性提供基础但却重要的理解。本文的研究成果不仅能为社会电视的技术改进、运行管理以及广告营销方面提供更多的理论性参考，同时还能服务于政府舆论导向，以及国家、军队实施认知域影响力作战等方面的需求。本文研究的开展，对于构建和谐有序的网络舆论空间，为国民经济发展和社会公共安全创造良好的网络信息环境，具有重要的意义。本文主要结合大陆主流的网络电视平台 PPLive、PPStream 和 UUSee 以及在线社会媒体新浪微博，分析这两种社会媒体各自的用户行为，并将其进行对比分析，找出其内在联系和规律。图 1.1 给出了网络电视用户与新浪微博用户之间的关系。本文先各自研究网络电视用户和新浪微博用户在选秀类或选举类节目中用户的行为，然后重点分析对比公共区域 B 中用户的行为，对于某一事件来说，区域 B 中的用户是连接两大社会媒体的桥梁，这类用户既观看网络电视，同时又在微博中发表自己观看网络电视内容的观点和看法。他们的行为更易影响到其他用户对事件的看法。

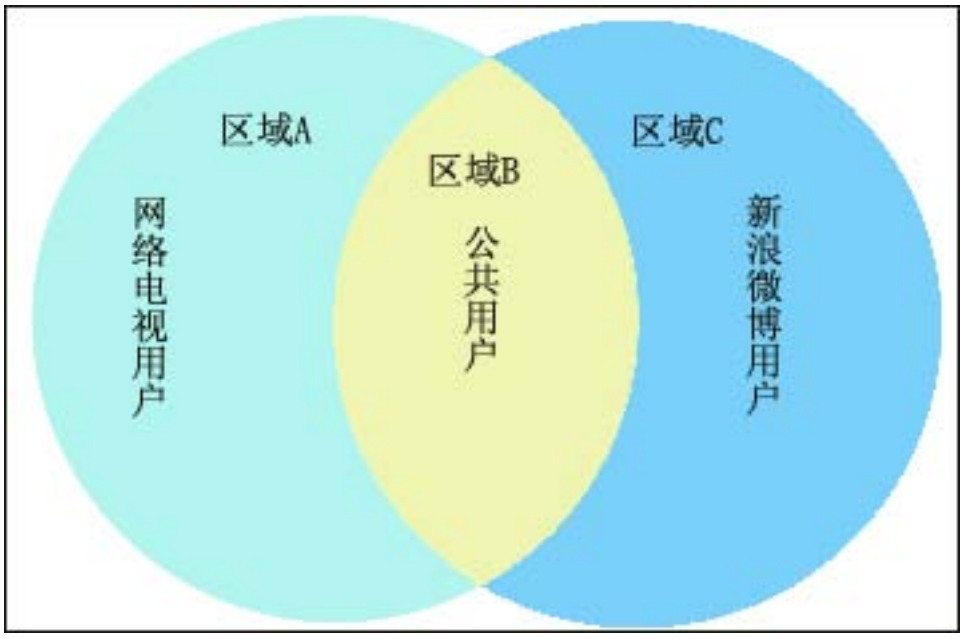

图 1.2 社会媒体用户分布

1.3 研究现状

本文以基于在线社会媒体和网络电视的社会电视为研究对象，测量、分析社会电视受众行为模式和行为特征，并进一步基于受众行为研究社会电视影响力。由于社会电视是一个新兴媒体，目前还较为少见系统的有针对性的研究，下面分别就几个方面的相关研究动向进行综述。

1.3.1 社会电视受众行为测量与研究

由于社会电视兴起的时间不长，目前针对社会电视受众的行为测量研究还处于刚刚起步的阶段。而作为社会电视受众测量的基础，对于网络电视用户的测量研究，以及在线社会媒体为用户的测量研究则得到了广泛的关注，并取得了不少有意义的成果。

网络电视的测量方法分为主分为主动测量和被动测量两种：（1） 主动测量方法是使用网络爬虫，模仿节点主动加入到P2P网络，获取相应的网络特性和用户节点信息。主动测量方法一般通过分析P2P客户端产生的流量进行，寻找tracker特征，使得网络爬虫能像普通节点一样加入到P2P系统，然后尽可能的收集如邻居节点列表包括节点的IP、端口号等信息。主动测量方法主要用于测量P2P网络的拓扑、延迟、内容可用性等行为特性。（2） 被动测量方法主要是根据测量目的的不同在网络的不同位置部署一定数量的测量点，使用特定的软硬件设备捕获网络流量，通过分析流量被动测量相关的P2P流量信息。根据测量目的的不同，测量点可以位于核心路由器或某个ISP网络的边缘出口，也可以位于中端机上。被动测量方法主要用于测量P2P网络的流量上下行数据流量、节点数目、连接持续时间等宏观流量信息。网络电视的测量研究主要以大规模运营的P2P TV系统平台，CoolStreaming[10]是最早在互联网上大规模部署的基于P2P技术的网络电视，Zhang[11]等人通过系统的日志数据，对CoolStreaming的在线人数、用户行为和用户的地理分布等进行了分析和研究。Li等人[12]通过CoolStreaming日志数据，进一步分析研究了其用户类型和分布，以及在线人数的时间演化。Wu等人[13]与悠视网合作，对UUSee系统的在线人数、用户行为、用户分布等进行了统计分析，认为UUSee的在线人数具有明显“日模式”。Liu[14]研究了UUSee冷门频道中用户会话长度及其影响因素。对于不公开协议的P2P TV系统的用户行为测量，主要还是通过主动爬行和被动嗅探两种方式实现。Hei[15][16]和Vu[17][18]分别设计了PPLive协议爬行器，测量了PPLive的用户行为，包括频道在线人数和系统在线人数变化，用户动态性，用户会话长度等特征。Silverston等人[19]通过嗅探的方式测量和比较了PPLive、PPStream、SOPCast和TVAnts等四个P2P TV系统，采用威布尔曲线对节点会话长度分布进行了拟合。

Qiu等人[20]针对美国国内基于机顶盒的IPTV系统进行的用户行为分析和建模研究。这些网络电视的测量和用户行为分析，多是从通信协议改进与复杂网络分析的角度，而非传播学受众分析的角度来展开的。由于研究目的不同，其用户研究内容多属于受众结构性分析，尚无法做到进一步把握受众的态度、情感和需求等更深层次的受众分析。

在线社会媒体的测量方面，Mislove[21]等采集了四种不同的在线社会媒体数据：Flickr、YouTube、LiveJournal和Orkut，对比分析这四种社会网络的拓扑特性。Guo[22]等针对三种不同类型的知识共享网络（博客网络、书签共享网络、知识问答网络），对其用户时长以及用户对网络的贡献度进行了测量研究。Cha[23]、Cheng[24]和Biel[25][26]则分别研究了YouTube的用户行为特征、视频内容的属性（类别、长度、大小等）特征、社会网络的拓扑结构和动态演化特征等，得到了YouTube用户的统计行为模式。Cha[27]还在Mislove的基础上对Flickr网络中的图片拓扑分布、时间演化分布以及信息传播过程进行了分析。Java[28]等对Twitter的网络拓扑、地理分布进行了研究，并利用文本处理技术分析了用户发博的兴趣和动机。Huberman[29]则对Twitter中潜在的朋友关系进行了挖掘，发现Twitter网络是由密度很大的相互关注网络与稀疏的真实朋友网络组成的。Kwak[30]利用爬行技术采集了比先前研究更为精确的Twitter数据，从Twitter网络的拓扑结构、用户排序方法、热点话题传播模式等多个角度展开研究，实现了Twitter网络及其信息分发的量化分析。Donghee等人[31]选择奥巴马接受诺贝尔和平奖演讲的现场直播，和娱乐节目“美国木偶”两个不同类型电视节目为背景，提出了AEIOU模型（Attention, Emotion, Information, Opinion, Utility），测量和分析了Twitter用户对于电视内容的情感和态度，验证了从在线社交媒体采集电视受众对内容的认知行为的可行性。上述这些研究仍然重点集中于拓扑结构测量以及社会网络分析、社区结构发现等相关领域，对于用户行为特征和模式的挖掘仍然有待深入。

综上所述，目前的研究在很大程度上仍然将网络电视和在线社会媒体视作单纯信息交流或者通信网络，而不是一种大众媒体形式。对于大众媒体进行精确受众分析，不仅需要获取受众的媒体接触行为，同时还有必要采集受众的心理认知行为信息，并进一步地进行行为信息融合，以得到相对完整的可用的用户复合行为，用于深层次的受众分析研究和影响力研究，从已知文献看，目前在这方面的理论和方法研究都较为少见。

1.3.2 社会电视受众行为分析与建模

传统大众媒体的受众分析主流还是采用定性的方法，但是也有一些研究采用了分析建模的定量研究方式，如Li等人[32]提出了一个电视受众收视率模型，并在此

模型的基础上研究了多电视台的收视竞争问题。Zheng[33]则提出了基于灰色GM模型的电视受众收视率预测方法，以解决小样本和信息不完全条件下的电视受众收视率预测。Chen等人[34]和Shin[35]都通过研究Web TV的用户特征和内容评价行为模式，用于实现视频节目的推荐。Takama[36]采用情感分析和模糊推理研究电视用户的兴趣模式。Gopalakrishnan等人通过分析交互式IPTV点播用户的行为实例，提出了基于semi-Markov模型的用户行为描述方法。这些分析和建模研究主要目的在于理解或预测受众的媒体接触行为，改进电视或网络电视的节目推荐或内容调度等服务。

为了深入研究社会电视受众行为的特征和发生机理，要研究受众复合行为的各种统计特性，定量的了解受众复合行为，进一步地研究社会电视受众复合行为中蕴含的人类行为动力学特征，并给出解释模型。人类动力学的研究始于Barabosi[37][38]对电子邮件和普通邮件的发送与回复行为的时间间隔的统计研究，发现人类行为同时具有长时间的静默与短期的高频率爆发，相邻两个事件的时间间隔分布满足负幂次分布。这些行为的统计特性不能用传统的泊松过程进行描述，说明人类的个体行为可能存在复杂的动力学机制。目前人类动力学的研究已经吸引了越来越多的关注。大量的人类行为的实证研究表明，时间统计的非泊松特性可能是在人类行为中普遍存在的一种现象。人类动力学研究的一个重要方面是探索这种非泊松行为特征的动力学机制与来源。目前的几种重要的解释包括任务队列理论模型[39][40]、自适应兴趣模型[41]和修改泊松模型[42]，人类动力学的另外一个重要的研究方面是在人类行为特性对社会系统的影响。综上所述，目前国内外关于人类行为动力学的研究主要分为两个层面，一是有大量研究针对某具体人类行为实例展开，统计和发现其行为动力学规律，本文的部分研究与此思路相吻合，即基于社会电视复合行为这一新兴的人类行为实例，研究其固有的潜在动力学特征。二是针对实例中表现出来的非泊松行为特征等的动力学机制与来源开展研究，提出解释性模型并进行验证。但是目前的研究是针对宏观意义上的，通用的人类行为统计规律而言的，并没有考虑到不同目的的人群（如社会电视中可能存在的推手型用户），以及其在行为特征和解释模型上的差异性。

1.3.3 社会电视影响力分析与评价方法

对于新兴的社会电视，目前还未见到有关其影响力评价的研究，也未见到相对应的影响力分析和评价方法的研究。相关的研究大量见于传统电视，网络媒体等媒体领域中。在指标体系上，发行量、收视率、收听率和点击率来衡量媒体影响力的最基本指标，这些指标反映了媒体覆盖和影响的受众数量的大小。在媒体影响力评估方面也有较为成熟的方法体系，常见的评估方法包括[43]：媒介资源评

价法，客体归类法，专业威望评价法，二级传播评价法和综合评价法等，这些方法各自有自己的指标体系偏好，但是大多沿用了对传统媒体的认识研究思路，即将媒体传播过程视作封闭过程，忽略了社会电视中大量的受众参与行为，尤其是推手型用户的参与行为可能对影响力造成的干扰。

我们认为，对社会电视这一类强调交互性和用户参与的大众媒体，其中的信息传播过程是开放的，可能收到某些受众干扰和影响，因而要相对精确评估其影响力，必须不考虑某些特定行为模式（如推手）的参与程度。

1.4 论文主要工作

社会电视从结构上看包括网络电视和关联的在线社会媒体两个功能模块，本文主要结合网络电视和新浪微博两个平台。通过网络电视爬行器技术对前者实施测量，可以获取受众对社会电视视频内容进行实时接收和分发等媒体接触行为，例如访问时间、停留时长、访问频度等，这些行为数据都具有严格实时性和高度动态性。而通过 Web 爬行技术对后者实施测量，可以获取受众对社会电视内容的心理认知行为，如推荐强度、情感喜好和态度等，这些数据可能是不完整的和不确定的。为了提高受众行为数据的完整性和可用性，需要对来自于上述两个不同数据源的，包含多个目标受众的行为数据进行关联和融合。本文基本研究思路是在测量方法上综合网络电视测量技术和在线社会媒体测量技术，提出面向社会电视受众媒体接触行为和收看内容认知行为的同步测量方法，测量过程中尽可能获取和保留受众行为数据的时空一致信息，建立受众复合行为的统一形式化描述结构，研究基于统一时空框架下的，具有多源、异类特点的受众复合行为信息的组织管理、关联分析和融合技术，以便获得尽可能丰富、完备和语义一致的社会电视受众行为数据。

本文采集2011年湖南卫视快乐女生播出时段PPlive[44]、UUsee[45]和PPStream[46]三个网络电视平台的节点数据和新浪微博快乐女生用户的的博文及粉丝评论数据为基础，并在此基础上进行了一些分析。具体工作如下：

（1）设计并实现 P2P TV 用户节点采集器和新浪微博采集器

为获取实验所需的数据集，本文设计并实现了两个采集系统，分别是 P2P TV 用户节点采集器和新浪微博采集器，其中 P2P TV 采集器采用主动测量技术，获取用户节点信息包括 IP 和端口号；新浪微博采集器，采集快乐女生播放期间微博用户讨论该事件的数据，包括用户信息、微博内容、评论内容等。

（2）焦点人物分析

以此次采集的快女数据为例，本文主要分析了快乐女生前三强的选手段林希、刘忻和洪辰三个焦点人物的微博数据，主要包括其粉丝的地理分布、粉丝关注度

和支持度等信息。

（3）P2P TV 用户地理可视化

将 P2P TV 采集器采集到的快照文件中的 IP 信息，利用 GEO 数据库转化为经纬度信息，并利用 GoogleMap API 将经纬度信息和用户数量，以图标的方式画在地图上，以可视化的方式展现出来。

（4）P2P TV 在线用户行为分析

主要研究用户在线时长分布、根据在线长的用户分类、用户在线时长演化分析及用户到达率分析。

社会电视受众行为分析整体框架如图 1.1 所示。

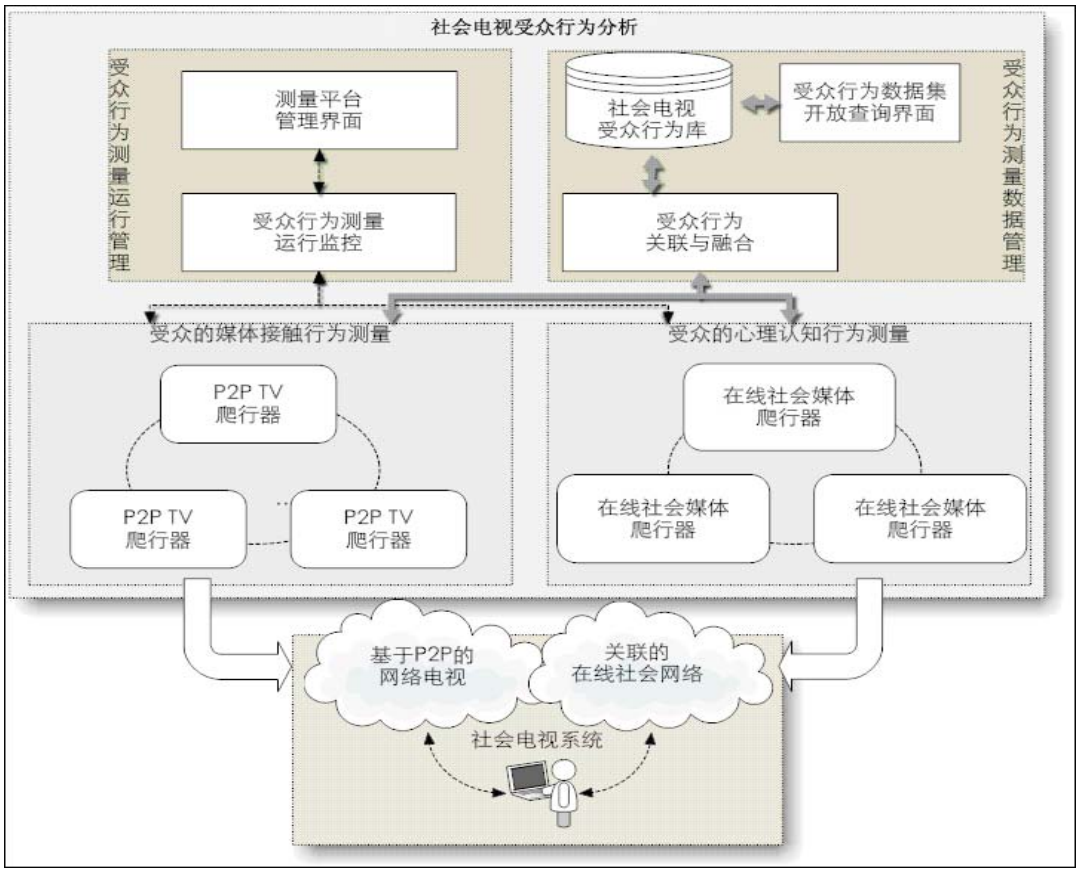

图 1.3 社会电视受众行为分析框架

1.5 论文组织结构

本文共分为五章：

第一章：绪论。本章首先介绍论文的背景、研究意义及社会电视用户行为研究现状，以及新浪微博和网络电视用户的特点，然后介绍论文面临的问题以及论文主要工作。

第二章：相关理论知识和关键技术。本章主要介绍论文相关的理论知识和系

统设计和实验过程中使用到的关键技术。主要包括网络爬虫技术、网页信息抽取技术、Spring 技术在数据库中的应用、P2P 网络测量技术的种类及本文采用的测量手段和 IP 地址解析技术。

第三章：基于微博数据的用户行为分析，主要介绍新浪微博数据采集系统、焦点人物分析和热点事件中的用户行为分析。主要包括焦点人物的定义、粉丝地理分布分析、粉丝活跃度分析和关注度分析。重点分析 2011 年湖南卫视“快乐女生“事件中微博用户行为。

第四章：基于网络电视数据的用户行为分析，主要包括网络电视数据采集系统的设计、用户地理分布、用户平均在现实时长等分析。本章以 PPlive、PPStream 和 UUSee 三个网络电视平台的湖南卫视频道快乐女生播出期间的数据为基础展开分析。

第五章：对比分析微博用户行为和网络电视用户行为。就用户在线数量、用户在线时长分布、用户在线时间演化和用户地理位置分布这四个方面展开对比分析。

第六章：结束语。本章总结了全文工作及创新点，并对下一步研究工作进行展望。

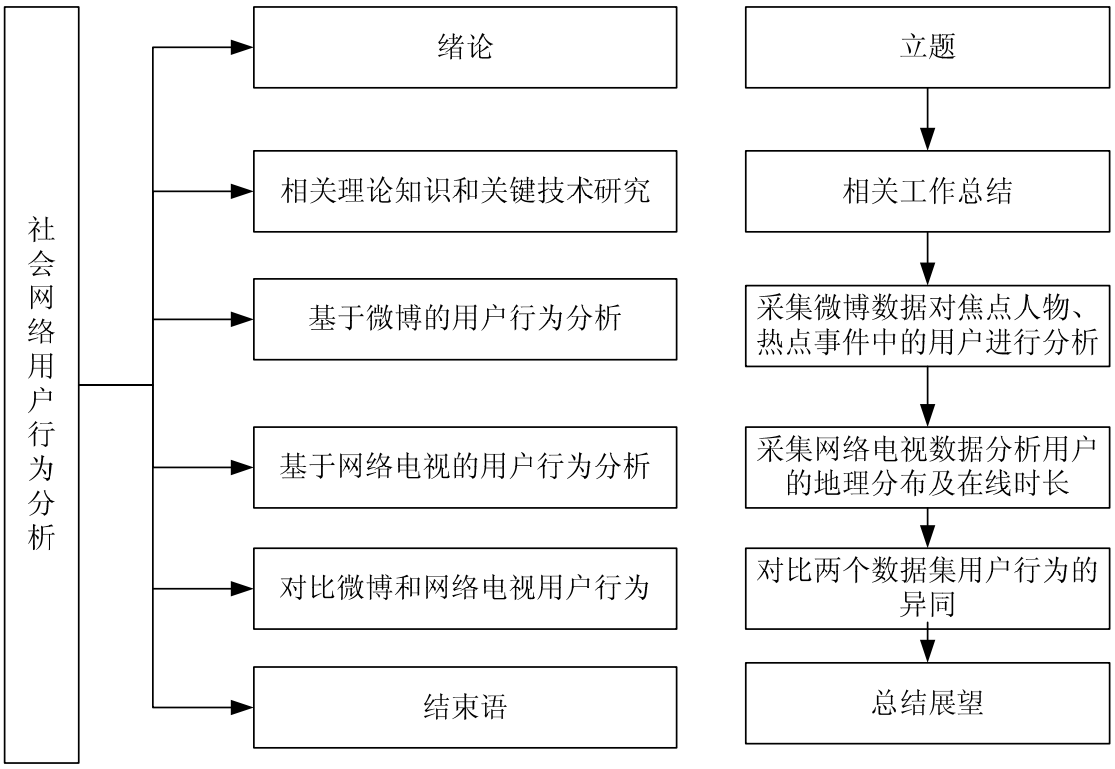

图 1.4　论文组织结构图

第二章　相关理论知识和关键技术

本章主要介绍与本文密切相关的六种关键技术：网络爬虫技术、网络电视测量技术、Spring 技术、Google Maps API 技术及 IP 地址解析技术。网络爬虫技术和网络电视测量技术分别是研究微博用户行为和网络电视用户行为过程中获取数据的基础；Google Maps API 是在用户地理位置可视化过程中直观展示用户地理分布的接口。IP 地址解析技术和 Spring 技术分别使用在用户地理可视化之前将 IP 地址转化为经纬度和解决数据库存储慢的问题。

2.1　网络爬虫技术

网络爬虫（又被称为网页蜘蛛，网络机器人），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。按照网络爬虫系统结构和实现技术大致可以分为以下三种：通用网络爬虫、聚焦网络爬虫和增量式网络爬虫。

（1）通用网络爬虫[47]

通用网络爬虫，爬行对象从给定的初始种子节点开始，扩充到整个 Web，采集全网数据。这类爬虫的爬行范围和爬行数量都十分巨大，对于爬行速度和存储空间要求较高，对于爬行页面的顺序要求相对较低，完成一次爬行，需要较长时间。所以为了提高爬行效率，通用网络爬虫通常采取一定的爬行策略来提高爬行效率，通常采用的爬行策略有深度优先策略和广度优先策略。其中深度优先策略基本思想是按照深度由低到高的顺序，依次访问下一级链接，直到不能再深入为止，在完成一个爬行分支后再回到上一个链接节点进行搜索。这类爬行策略比较适合垂直搜索或站内搜索。广度优先策略的基本思想是按照网页内容目录层级深浅来爬行页面，处于较浅目录层次的页面首先被爬行，当同一层次中的页面爬行完毕后，爬虫进入下一层次继续爬行。

（2）聚焦网络爬虫[48][49]

聚焦网络爬虫又称主题爬虫，是指有选择性的爬取那些与预先定义好的主题相关页面的网络爬虫。与通用网络爬虫相比，聚焦爬虫只需要爬行与主题相关的页面，极大的节省了硬件和网络资源。

（3）增量式网络爬虫[50]

增量式网络爬虫是指对已下载网页采取增量式更新，只爬取新产生的或发生变化的网页的爬虫，在一定程度上所爬取的页面是新的页面。和周期性爬行的网络爬虫相比，增量式爬虫智慧在需要的时候爬取新的页面，不重复下载爬取过的内容，这样有效见小了数据的下载量，减少了时间和存储空间上的消耗。

2.2 网页信息抽取技术

网页爬行器向服务器提交请求后获得到的是网页的 html 代码，需要将其中的信息如用户姓名、省份、博文等，抽取出来存入数据库。所以需要对获得的 html 进行抽取，本文使用 XPATH 技术对网页进行抽取。XPATH 是一种专门针对 XML 的结构化查询语言， HTML 是一种特殊的 XML 语言。

基于字符串表达式的 XPATH 路径语言可以非常高效的定位 XML 数据，从而灵活地选择文档的组成成份。XPATH 将所有的页面信息全都看成各种不同类型的树状节点，而每一个页面只有唯一的一个根节，从这个根节点树状的展开了各种节点。节点之间存在着许多关系，当 A 节点直接从属与 B 节点时，分别称 B，A 为对方的父，子节点；当 A，B 同属于一个父节点时，A，B 互为对方的兄弟节点。根据类型不同，节点还可以分为，元素节点，属性节点，正文节点。XPATH 定义了一种方法来计算每类的节点的字串值，可以通过节点的名字，属性，以及父子节点或兄弟节点来唯一的定位。

而本文采用比较灵活的正则表达式进行新浪微博信息的抽取，采用 javabean 技术，只需把写好的正则表达式写入配置文件，在用到时只需加载即可。

2.3 Hibernate 和 Spring 技术在数据库中的应用

在传统的 JDBC 编程来访问数据库，需要在 Java 应用程序中嵌入大量琐碎的且价值不高的代码，使得编程设计人员不能把精力集中在业务逻辑的设计上，如何能够在复杂的大型系统中，使编程人员不过多的关注数据库的操作上，屏蔽数据库访问的细节，ORM 框架应运而生。

在应用系统的设计中，涉及的实体一般都会使用对象和关系数据库这两种表现形式，实体在内存中表现为对象，而在数据库中表现为关系数据，对象之间会有关联、继承和多态等面向对象的特性，而在数据库的数据之间却无法表达这些特性。因而需要一种神奇的技术，可以实现对象与关系数据库中数据的自由转换，而对象/关系映射(Object/Relation Mapping ，ORM)的框架的功能正是如此。而 Hibernate[51]是一种强大的对象/关系映射框架[12]。

在应用系统中使用ORM[52]框架可以避免直接使用SQL语句对关系数据库中的数据进行操作，而是借助ORM框架把数据库中的数据转换为对象，通过操作这些对象实现对数据库中的CRUD（Creat、Retrieve、Update和Delete）的操作。同时把面向对象的思想贯彻在一个系统的分析、设计、编程等方面。

在 ORM 框架中为了将数据库的数据操作转换成为对对象的操作，需要实现数据到对象的映射：就是把数据库中的表映射为面向对象思想中的类，表中的字段

映射成类中的属性。

Spring[53]是个轻量级的Java EE的开发框架，内部由Core、AOP、DAO、ORM、JEE和Web共 6 个核心子模块组成，其中DAO子模块提供封装JDBC操作的工具类，简化直接使用JDBC操作数据库的繁琐度，并支持通过声明方式管理事物。ORM子模块提供对其他ORM框架整合的支持，进而降低了使用ORM框架的繁琐度。

2.4 P2P 网络测量技术

目前对于P2P网络的测量研究主要包括P2P文件共享系统[54]的测量和P2P TV系统的测量。就方法而言，目前P2P网络测量主要采用三种方法[55][56][57]：一是基于嗅探的被动测量方法[58]（见图 2.2），即利用Wireshark[59]等网络嗅探软件来捕捉特定环境下的P2P系统客户端的通信流量，并对通信流量进行分析和统计。基于嗅探的被动测量方法通常以流量相关的系统局部特征为侧重点，包括本地流量统计、上传/下载带宽使用情况和数据包大小分布等，相关研究一般是在特定实验环境下搭建测试平台，虽然得到的统计数据是基于真实流量，但是其实验规模限制了结论的普适性。二是P2P应用平台提供商参与的测量方法（见图 2.3），即研究机构与P2P 应用平台提供商展开合作，平台提供商在其客户端软件中增加数据采集接口，或在整个系统中部署数据采集方案，供研究者进行数据的收集和分析使用。这一类研究方法由于要求比较好的合作条件，因此相对比较困难。三是基于通信协议的主动测量方法[60]（见图 2.4），即通过对被测系统的通信协议进行分析和理解，设计一个协议爬行器来主动探测系统和收集信息。

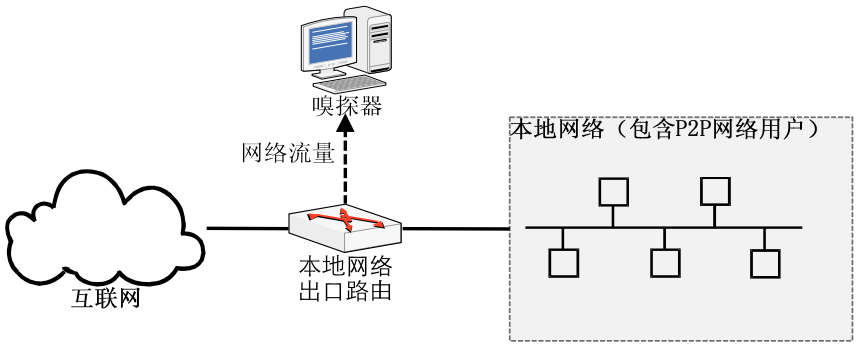

图 2.1 P2P 网络基于嗅探的被动测量方式

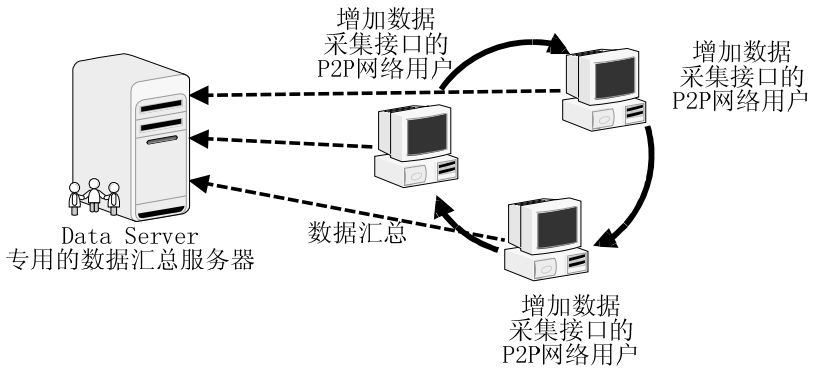

图 2.2 P2P 网络基于平台提供商参与的测量方式

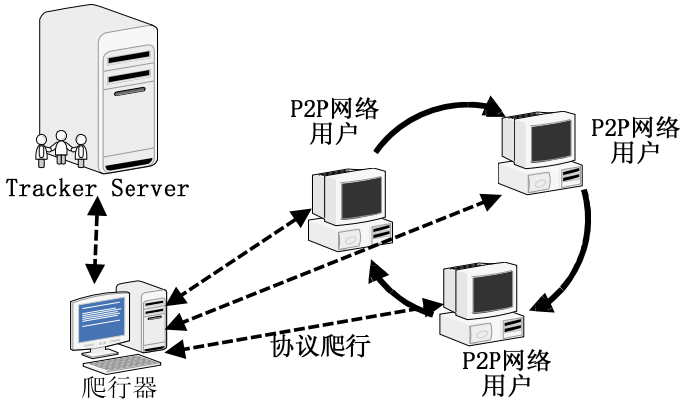

图 2.3 P2P 网络基于爬行器的测量方式

从 P2P 网络技术的发展趋势和应用来看，目前对于 P2P 网络的测量的研究内容主要包括：

1. 测量框架研究：研究 P2P 网络应用的体系结构，建立规范、完整的测量系统框架；研究快速测量方法和测量数据获取策略；研究主被动相结合的 P2P 网络测量方案；利用形式化的方法研究测量结果的完备性与正确性；研究测量系统的评价指标，包括稳定性、资源消耗代价、结果可靠性、准确性等；开发测量系统的仿真验证平台等；

2. 测量测度研究：定义统一的、具体的、可重复的测量测度，从而解决目前测量目标分散，结果各异的问题，满足研究者、用户、运营商、应用运行人员之间知识传递以及进行不同系统比较的需求；

3. 测量关键技术：包括面向运行规律的测量和面向运行效果的测量。从具体的测量内容来看，面向运行规律的测量包括用户行为的测量与建模、网络拓扑的测量与建模、网络流量的测量与建模等内容，需要解决的技术难点包括用户行为特征的发现与采集、网络流量的识别等；面向运行效果的测量主要是从用户的角度对互联网应用系统进行测量研究，包括系统可用性测量、系统效率的测量等内

容，需要解决的技术难点是测量结果的可信性验证等问题；

4. P2P 网络应用的设计与改进：基于网络应用的研究成果，研究和设计更符合新型网络运行特征的协议或系统。

5. P2P 网络应用的信息传播监管：这是以网络舆论传播管控为直接背景需求，结合社会计算和情报与安全信息学（Intelligence and Security Informatics，ISI）理论的新型研究领域。研究基于 P2P 应用的信息传播对国家和社会安全的可能影响，并对这些网络信息进行实时有效地了解和掌控，分析 P2P 网络应用用户的构成和分布特征，建立网上各种利益和兴趣团体的动态在线档案，利用合法的情报收集和分析手段，主动及时地采取安全措施，保障社会的运行和发展。

2.5 Google Maps API 技术

谷歌地图[61]（Google Maps）是 Google 公司提供的电子地图服务，包括局部详细的卫星照片。能提供三种视图：一是矢量地图（传统地图），可提供政区和交通以及商业信息；二是不同分辨率的卫星照片（俯视图，跟 Google Earth 上的卫星照片基本一样）；三是地形视图，可以用以显示地形和等高线。谷歌地图还提供了免费的地图应用接口服务Google Maps API，主要提供了地图的显示、标点、划线等操作包含 59 个类，空间、叠加层等包含 18 个类。可以实现地图的加载和显示、添加控件、添加标注、经纬度编码、地址解析和逆向解析等操作，使用Google Maps API可以在JavaScript和XML的基础上灵活开发各种地理信息应用。Google Maps API采用申请密钥的方式授权使用，每个密钥只能对应一个网站，当然，任何一个密钥都可以在Localhost上使用。Google Maps API主要有三种API调用形式分别为“JavaScript Maps API”、“Maps API for Flash”和“HTTP Service”。本文采用的是JavaScript Maps API，其基本调用流程如图 2.4 所示。

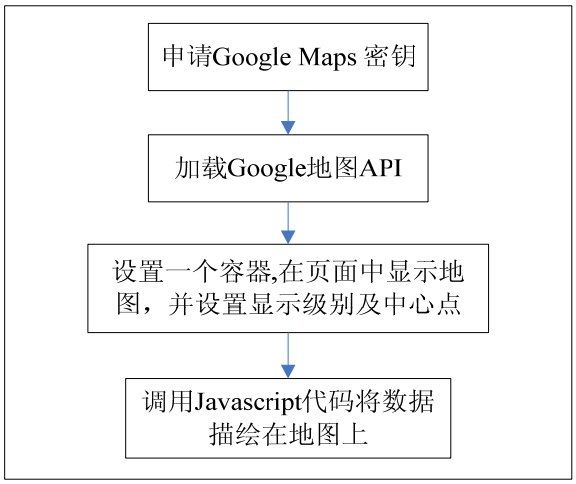

图 2.4 Google Maps API 调用流程

2.6 IP 地址解析技术

主动测量获取的IP地址信息，在用户地理可视化时需要将其转化为地理信息，本文使用MaxMind[62]公司的GeoIP数据库[63]来获取用户IP精确的获得用户地理位置信息，目前流行的火狐浏览器就是使用它来统计下载用户所在区域的，精确率在 99.5%以上。GeoIP通过来访者的IP，可定位经纬度，国家/地区，省市，甚至街道等位置信息。本文主要是利用其来将IP地址转化为经纬度。MaxMind公司提供了包含C语言在内的多种调用GeoIP数据库的API接口，用户可以方便的根据IP地址查询到所需要的信息。

2.7 本章小结

本章主要介绍了论文中所涉及到的相关理论知识，包括数据挖掘相关概念和方法、网络爬虫的分类及特点和 P2P 网络测量等技术，论文中所涉及的网络爬虫是集成了通用网络爬虫、聚焦爬虫和增量式网络爬虫的特点设计而成，可以针对微博数据有选择的爬取用户信息。P2P 测量技术本文主要采取主动测量的方式，模拟用户节点加入 P2P 网络中，从而获取邻居节点信息。

第三章　微博用户行为分析

针对新浪微博采集用户数据，分析用户行为，主要包括：焦点人物分析、活跃用户行为分析。其中新浪微博数据采集模块采用垂直搜索方式，从预先的种子节点开始，按照粉丝和关注关系一级一级爬取用户数据，主要包括用户的博文、评论及用户基本信息。采取的爬行技术主要有两种一是模拟数据包发送方式[64]，二是基于新浪微博自身的API获取数据，两种方式各有优劣。焦点人物分析部分主要包括焦点人物的定义、粉丝地理分布、活跃粉丝用户的分析和焦点人物关注度分析；活跃用户行为分析部分主要从用户发帖量、用户回帖行为及用户网络拓扑结构三个方面展开分析。

3.1　微博数据采集

3.1.1　新浪微博数据采集系统

要分析新浪微博用户行为，首先需要获取新浪微博的用户及事件数据，为此我们设计了一套微博数据采集系统SinaCrawler[65]。SinaCrawler的基本思想是从给定的初始节点开始，爬行用户的粉丝，获取粉丝数据后再爬取粉丝的粉丝，如此循环下去，直到满足给定的爬行深度自动停止爬行。SinaCrawler系统框架如图 3.1 所示。

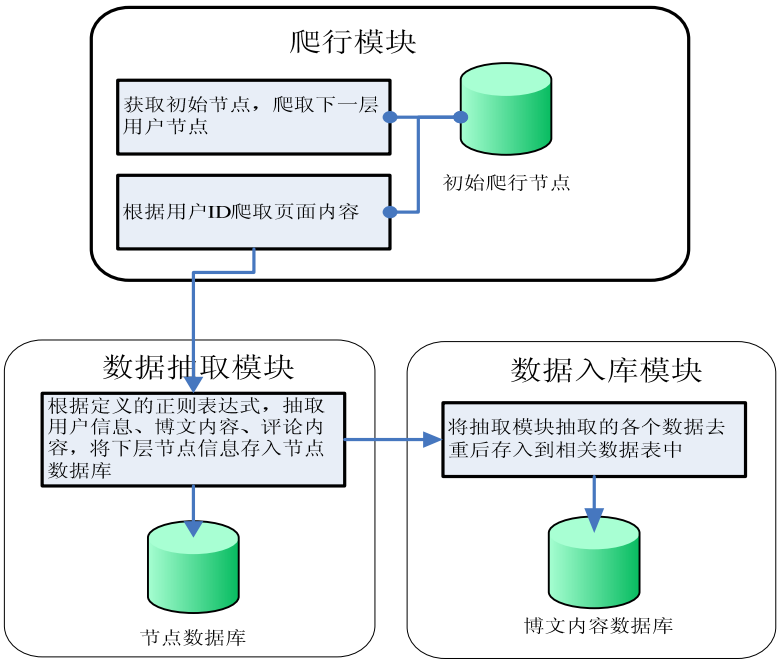

图 3.1　SinaCrawler 系统框架

（1）爬行模块：该模块包括四个子模块分别是登陆子模块、控制子模块、配置子模块和爬虫引擎子模块。其中登陆子模块主要是模拟浏览器登录新浪微博以获得相应的Cookie[66]，以便爬取所需的信息；控制子模块主要是对程序的与运行进行控制，包含对登录用户名密码、网页解析配置文件的、初始种子结点的、爬行参数等的配置；配置子模块主要包含种子结点配置子模块、网页解析配置文件子模块，前者主要是对初始的种子结点进行配置，后者主要是对是对抽取网页信息的正则表达式[67]文件进行配置；爬虫引擎子模块是爬行模块的核心，主要是根据种子结点构造的URL，下载相应URL的源文件。爬虫引擎设计的健壮性、鲁棒性对整个爬虫系统能否持续高效的运行具有很重要的作用，通过对SinaCrawler爬行系统进行测试，达到 7×24 小时爬行的要求。

（2）数据抽取模块：主要是对下载后的网页源文件进行运用正则表达式进行解析获得相应的信息，包括用户 ID、省份、教育程度、兴趣爱好、博文、评论、粉丝和关注数等信息。

（3）数据入库模块：主要是对解析出的微博信息进行存储，SinaCrawler爬行系统采用MySql[68]数据库，数据库轻便且操作维护简单。

本文以湖南卫视 2011 年“快乐女生“事件为基础，爬行快乐女生前 37 强选手 2011 年 6 月至 9 月的新浪微博数据并进行分析。在爬行中，首先将这 37 个快乐女生选手的微博地址及湖南卫视官方地址列为初始节点，将这些微博的 ID 存入初始节点数据库中，爬行系统从数据库中读取初始节点，首先爬取这些初始节点的粉丝用户，将其粉丝用户的 ID 作为第二层的爬行节点，并存入到节点数据库中，爬行系统设置爬行深度为 3，即表示爬取完第三层的用户之后，爬行系统自动停止。本爬行系统采用 Java 语言编写，通过构造数据包的方式向 sina 微博服务器提交请求，并获取服务器返回的 html 代码，通过正则表达式抽取页面上的用户名、粉丝用户、关注用户、地理信息、近期发表的博文、用户的评论等信息，将这些信息存入到相应的 MYSQL 数据表中，供下步用户行为分析调用。SinaClawer 爬行系统的基本工作流程如图 3.2 所示。

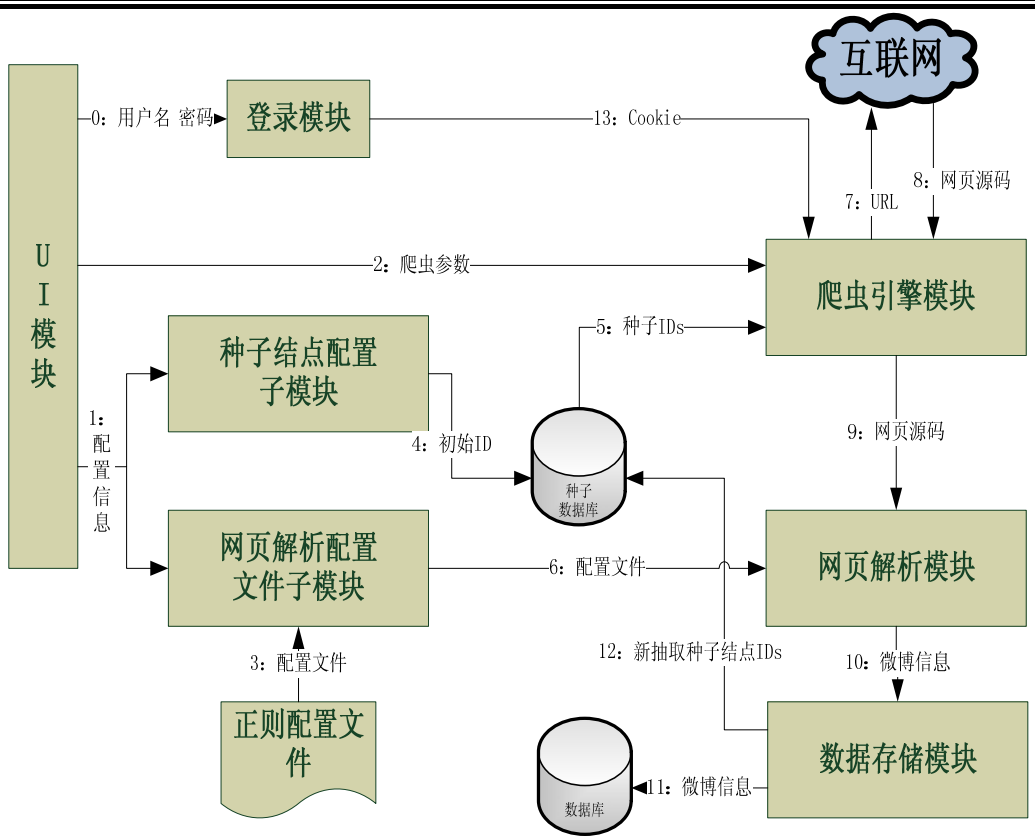

图 3.2 SinaClawer 运行流程

3.1.2 基于新浪微博 API 的数据采集

由于新浪微博自身的特点，SinaClawer不能完全满足需求，如：当用户粉丝超过 5000 条时SinaClawer不能爬全用户的粉丝ID，同样用户评论和博文较多时也面临同样的问题。由此本文采取调用微博API[69]的方式，用以获取粉丝数、评论数和微博数超过 5000 的内容，以弥补SinaClawer爬行系统的不足。新浪微博API主要提供了用户基础数据接口和微博地理位置信息接口，本系统主要集成了微博用户基础数据接口[70]，主要包括下行数据接口、用户接口、关注接口、话题接口、Social Graph接口和登录接口，其中获取下行数据接口主要提供获取公共微博消息、用户微博消息列表、用户评论列表、博文转发数等服务；用户接口主要提供获取用户资料、用户关注列表以及与该用户有相同兴趣的用户列表等服务；关注接口主要提供获取两个用户之间是否存在关注与被关注关系的服务；话题接口主要提供按小时、当天和当周获取热门话题服务；Social Graph接口主要提供获取用户粉丝和关注用户ID服务；登录接口主要提供获取登录授权服务；与SinaClawer相比使用API方式有着速度快、数据全的特点。本文调用新浪微博Java版API，主要用于获取用户的粉丝ID，粉丝的地理位置等下行数据。其基本流程如下：

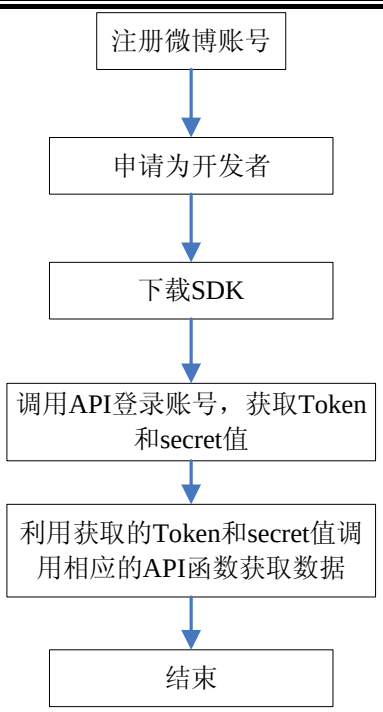

图 3.3 新浪 API 运行流程

3.2 焦点人物分析

至 2011 年 10 月新浪微博用户已达到 2 亿，这些用户通过关注和粉丝组成一个巨大的有联系的网络，在这个网络中用户大致分为两类：名人和草根。其中名人用户在该网络中的连接度较之草根用户要大很多，其显著特点就是粉丝数量多。这就决定了名人用户在网络中推送消息能得到更多的关注，所以不论是进行网络营销还是舆论宣传，利用名人用户能收到更好的效果。

3.2.1 焦点人物的定义

焦点人物即收到关注数目众多，在某一事件有重要影响力的人物，其显著特点就是粉丝数量众多。以 2011 年湖南卫视的快女事件为例，比赛期间湖南卫视官方和刘忻、段林希、洪辰等快女选手的微博被关注数量达到几十万，针对快女事件这些人物即成为焦点人物，其一言一行受到众多微博用户的关注，其影响力和号召力远远大于其他用户。本文定义焦点人物使用两个指标粉丝数和博文的评论量。以快乐女生前三强选手刘忻为例其粉丝数量达到 100 万之多，其微博自 4 月份至 10 月份的用户平均评论数超过 6000 条，我们将其微博与普通用户的微博做了对比如图 3.4 所示。从图中我们可以看出焦点人物博文的平均评论数大大超过普通用户。另外在 4 月份至 5 月份之间刘欣博文的评论数走势与普通用户相似，这与快乐女生比赛安排相符合，说明在 6 月份之前刘欣在微博中所受到的关注并不

是很多，此阶段我们将其定义为普通用户，随着比赛进程的进行其知名度越来越高，所受到的关注也越来越多，此阶段刘欣已经发展为了焦点人物。

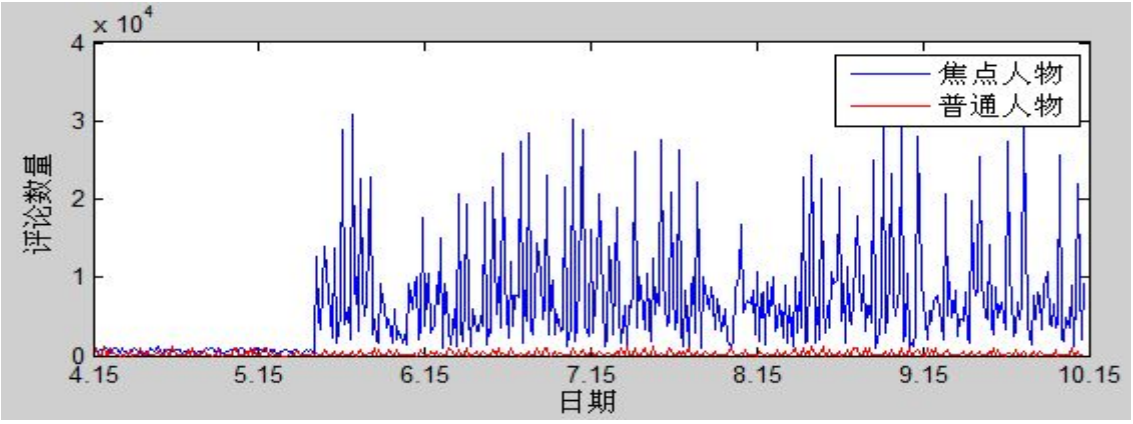

图 3.4 用户评论数量分布

3.2.2 粉丝地理分布分析

针对某一焦点人物分析其最新活跃的 5000 个粉丝，统计其地理位置，可以得出该焦点人物在哪几个省份获得的支持程度更高、号召力更大，通过该地理位置分析可以辅助用户进行下步决策，如：对于舆论宣传来说，可以重点选择宣传的省份。本文以 2011 年快女数据为基础分析前三强的选手刘忻、段林希和洪辰其粉丝用户的地理分布。首先爬取三个焦点用户的前 5000 个活跃粉丝的地理位置，然后选取每个用户粉丝分布最多的三个省份，进行对比分析，如图 3.5 所示。从图中我们可以清晰的看出洪辰在浙江省获得的支持最大，刘忻在湖南省的支持度最大，段林希则在广东省获得的支持度最大。获取粉丝地理分布特性能够帮助用户有更有效的展开舆论宣传，做到有的放矢，从而节约宣传成本。

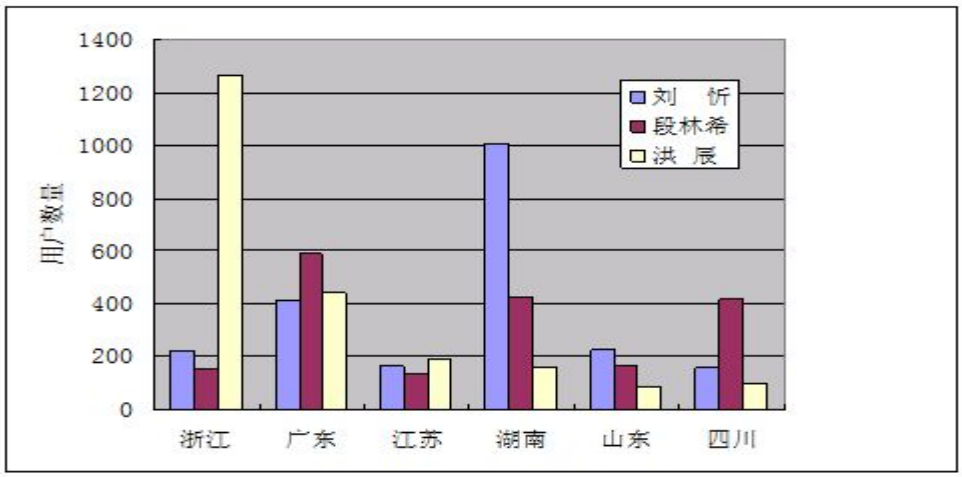

图 3.5 粉丝地理位置分布

3.2.3 基于活跃度的粉丝用户分类

根据粉丝参与评论数目的多少我们可以将粉丝分为三类：潜水粉丝、忠实粉丝和较忠诚粉丝。潜水用户即参与博主话题评论数为 0 的用户，该类用户关注微博主要是获取信息，参与度为 0；忠实分析定义为评论数大于 10 的用户，表明该类用户比较乐于参与博主发表的各种言论；较忠实粉丝定义为评论数在 1 到 10 之间的用户。针对 2011 年快乐女生前三强用户刘忻、段林希和洪辰进行分析，如图 3.6 所示。具体数据如表 3.1 所示。

表 3.1 用户活跃度分布

姓　名	潜水粉丝	忠实用户	较忠实用户
刘　忻	90%	0.82%	9.18%
段林希	84.8%	0.72%	14.48%
洪　辰	79.64%	1.66%	18.7%

图 3.6 粉丝活跃度分布

3.2.4 关注度分析

关注度反应了博主发表博文后收到的关注程度。本文在时间轴上对用户发表的博文数和用户评论数和转发数的比例来比较快乐女生前三强用户刘忻、段林希和洪辰所受到的关注。从而比较三个用户在一定时期内的影响力。分析结果如图 3.7 所示。从图中我们可以看出这三个用户在比赛阶段最受关注的时段各不相同，并且随着比赛的结束，三人收关注的程度较比赛前期趋于稳定。用户关注度分析可以帮助管理者在一定时段内对关注度较高的用户进行重点监控，防止负面舆论的传播。这一点特别是对涉及国家经济、政治和军事的名人的监控尤为重要，他们所受到的关注程度与消息的传播快慢有着密切联系，对于这类受关注程度高的名人对其微博内容进行监控是防止危害国家安全，防止煽动网民负面情绪的重要途径。

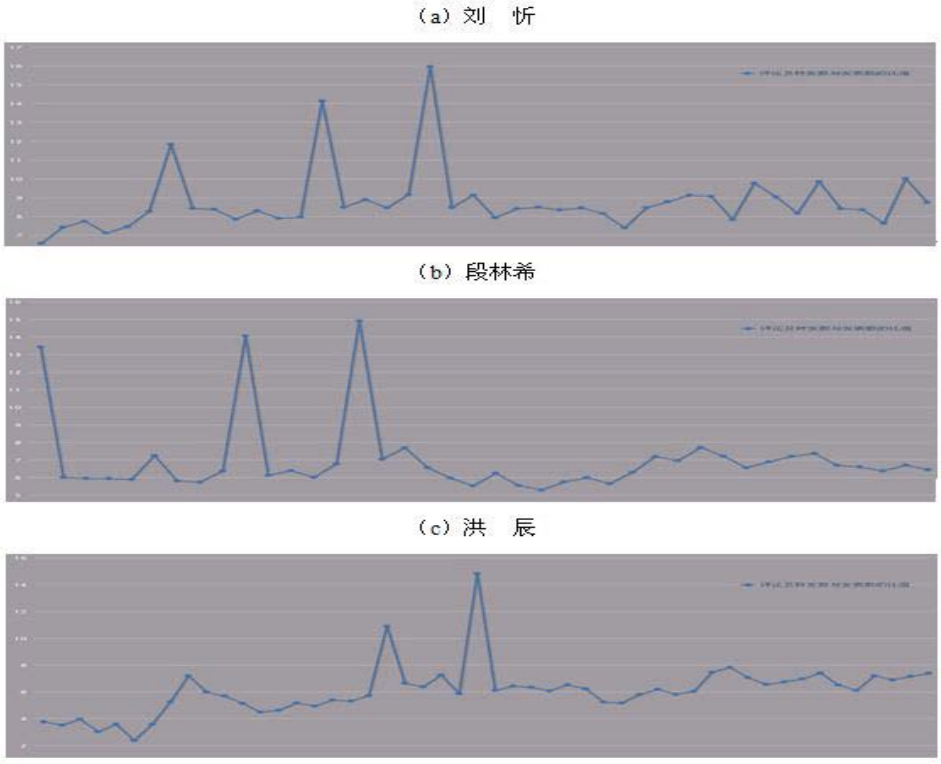

图 3.7 关注度分析

图中横轴为日期，纵轴用户关注度，其计算公式为：$GZD = \lg(\frac{ZF + PL}{NB})$，其中 GZD 为用户关注度；NB 为博主发表博文的数量；ZF 为博文被转发的数量；PL 为博文被评论的数量。用户关注度越高表明用户受到的关注越大。

3.3 活跃用户行为分析

本节从快乐女生相关的用户中选取发帖量、评论和转发量最高的 5%作为研究对象，我们称之为活跃用户。主要就总的用户发帖量的分布、单用户发帖行为和用户拓扑结构进行分析。

3.3.1 用户发帖量分布

（1）用户发帖量宏观分析

用户在一定时期内的发帖量反映了一个事件所受到的关注程度和热点程度，有利于从宏观上把握事件的走势。用户就某一事件的发帖量越多，则表明该事件受关注程度越高。本文以 2011 年快乐女生事件为基础，采集了与快乐女生事件相关用户的新浪微博 7 月份的博文信息并分析用户关注快乐女生事件的走势情况。

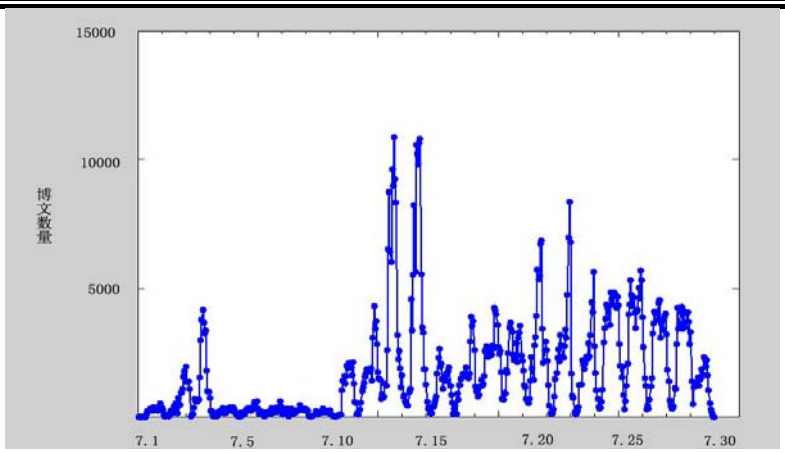

图 3.8 用户发帖量分布

从图中我们可以看到 7 月 10 号至 7 月 30 之间，微博上用户发表有关快乐女生的帖子数目较高，可以得出该时间段内用户对“快乐女生事件”较为关注，这与快乐女生总决赛的时间安排相符合，用户对“快乐女生事件”的关注时段正好是总决赛的时段。

（2）用户发帖量排序分析

将用户的发帖数量进行排序，如图 3.9 所示，我们发现随着排序靠后的用户发帖量呈现骤减趋势，其中排序靠前的 20%用户发表的博文数量占了博文总量的 80%。

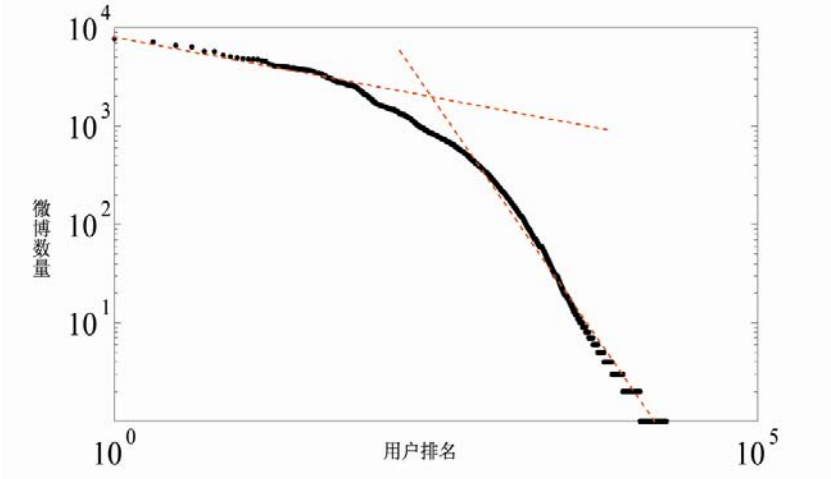

图 3.9 用户发帖排序

3.3.2 用户回帖行为分析

用户的回帖行为主要是针对某焦点人物发表的博文，分析用户的评论行为，主要包括用户回帖量分析和用户在线时长分析。

（1）用户回帖量分布

焦点人物发表的博文一个重要特点就是用户回帖量较多，某个用户对博文的回帖量越多，说明该用户忠诚度越高。本文以刘忻的微博为例分析用户的回帖行为。如图 3.10 所示。

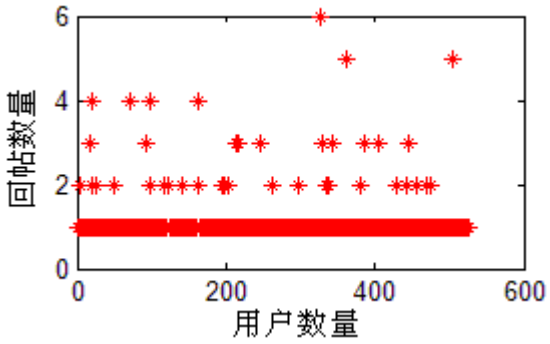

图 3.10 用户回帖分布

从图中我们可以看出，绝大多数的用户回帖数都是 1，这类用户对于该事件表现出了一定的关注性，但是热衷程度并不是很高，在线观看帖子的时间也很短，回帖数大于 1 的用户表现出的行为是有大量的在线时间观看帖子，并对帖子内容进行多次回复，该类用户对于自己所关注的事件热衷程度高，对于事件的传播和影响较大。在信息的传播和舆论控制时对此类用户的监管显得尤为重要。

（2）用户在线时长分析

针对回帖量大于 1 的用户，分析其在线时长分布，对于某一事件用户第一次评论的时间作为用户关注该事件的初始时间点，用户对事件的多次回复定义为用户的在线时间，用户的最后一次回帖作为用户离开事件。若用户对事件 10 分钟未回复，10 分钟之后再次回复，我们认定其为第二次进入该事件进行评论，将其作为新的用户看待。以刘忻的微博为例，分析结果如图 3.11 所示。

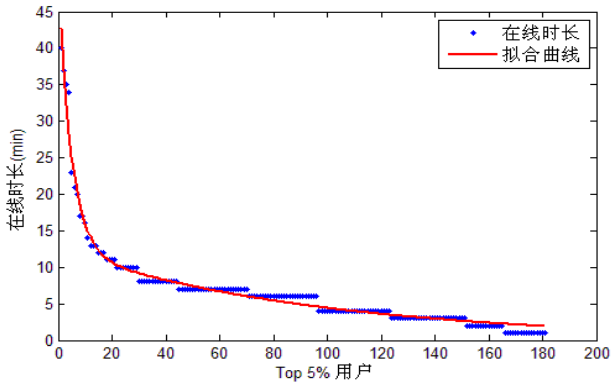

图 3.11 用户在线时长分布

从图中我们可以看到用户在线时长 80%集中在 5 分钟到 10 分钟，平均在线时长大约为 6 分钟。用户在线时长符合负指数分布。在线时长在 5 分钟以下的用户

我们定义其为正常用户，该类用户对事件表现出一定的热衷，乐于参与事件讨论，并发表自己的观点，此类用户的行为不足以主导舆论走势，在舆论监管时此类用户不需花费大量精力。在线时长在 5 分钟至 10 分钟之间的用户，用户数量众多，在舆论监管时应给予一定的重视，该类用户如果观点相似则极易引导舆论走向。在线时长在 10 分钟以上的用户属于高度热衷用户，此类用户行为不同于普通用户，其在线时间之长，回帖量之多，足以引导舆论走向，故对于此类用户的监管应予以高度重视，此类用户中极有可能存在网络推手。

3.3.3 用户网络拓扑结构分析

分析完单用户行为特征后，本文针对发帖量在前 50 名的用户进行了用户拓扑结构的分析。主要是通过用户的关注联系、工作、毕业院校和兴趣爱好，分析用户之间的联系。我们将用户定义为节点，用户之间的关注关系、具有相同工作和毕业院校以及含有“快乐女生”标签的用户之间的联系定义为边。由此我们找到了跟这 50 名用户产生关系的用户，并将其之间的联系用边连接，黄色的点表示前 50 名用户；红色的点表示其他快乐女生选手；绿色的点表示湖南卫视官方微博；黑色的点表示其他用户，如图 3.12 所示。从图中我们发现 21，24，51 和 62 节点为关键节点，连接数较多。如 62 号节点，我们发现其与多数前 50 名的节点有联系，说明该选手收到了众多网友的支持和关注；24 号节点与其他快女参赛选手之间有广泛联系，可以推测出 24 号用户对“快乐女生“事件十分热衷。

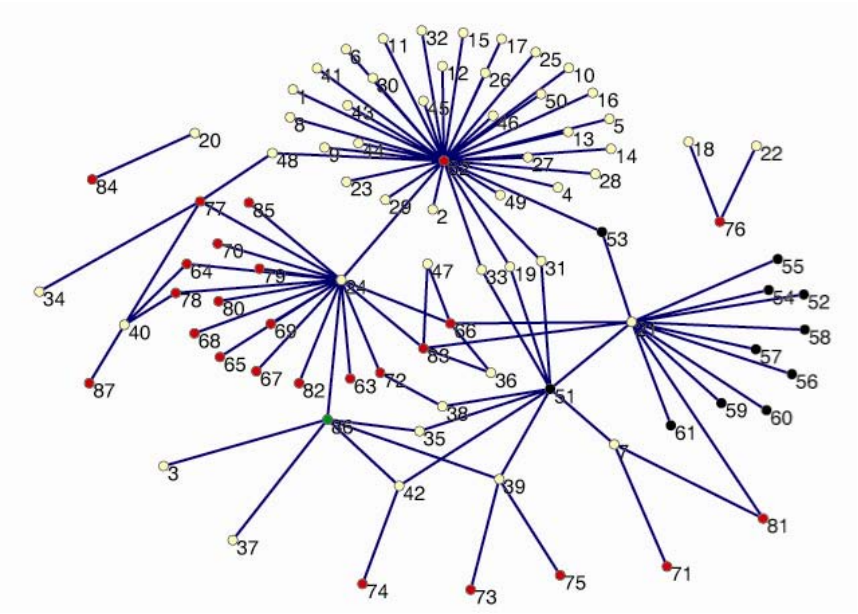

图 3.12 用户拓扑结构图

3.5 本章小结

本章主要针对微博用户的行为展开分析，主要包括焦点人物分析和活跃用户行为分析。首先，焦点人物以其被关注的数量众多，在网络中的影响力较大，在消息的传播中起到了关键作用，对于这些用户我们就其粉丝地理分布、活跃度及其所受到的关注度对比分析各焦点人物的特性。其次，针对快乐女生事件分析了关注此事件的活跃用户在该事件发生过程中的一些行为，包括用户发帖量的分布，单用户发帖行为分析和用户拓扑结构分析。

第四章　P2P TV 用户行为分析

4.1　爬行器设计与实现

目前网络电视多是基于 P2P 结构，每个节点用户既是客户端也是服务器，为了获取整个网络的拓扑结构，需要伪造用户节点，接受邻居节点数据包，为此我们利用先前在研究 P2P 网络电视课题中设计的网络电视爬行器 TVCrawler 来爬行拓扑结构，下面我们详细介绍 TVCrawler 的设计和实现过程。

4.1.1　系统框架设计

TVCrawler是一个具有分布式结构的P2P TV网络爬行器，其基本思想是遵循P2P TV系统的协议流程和语义[71]，模拟客户端与Tracker Server[72]及其他节点通信，从而获取节点邻接关系并加入本地节点队列，反复迭代直至获取到足够的P2P覆盖网络信息。

由于目前主要的 P2P TV 系统都是私有协议，为了获得爬行器所必需的协议语义，需要对多个 P2P TV 系统的网络协议进行逆向工程。通过网络嗅探软件 WireShark，对 PPLive，PPStream 和 UUSee 三个 P2P TV 系统，分别捕获了大量的实际通信流量，重点分析其中与节点邻接关系相关的部分通信接口，获得了邻居节点集的查询接口的格式和语义。

与以往对P2P文件共享系统进行网络测量的一个根本不同在于，上述三种P2P TV系统大量基于无连接的UDP通信方式，而不是TCP连接方式来实现协议交互。其中PPLive和UUSee两个系统都采用Pull[73][74]的方式交换邻居信息，即节点主动向邻居节点查询，邻居节点则产生一个本地邻接关系列表作为应答返回；而PPStream则采用了Push[75][76]的方式，即由节点定期产生一个本地邻接关系列表，发送给随机选择的邻居节点。

为了加快爬行数据的收敛速度，TVCrawler 采用主从结构分布式部署，分为爬行控制器（Controller）和爬行器终端（Crawler）两个部分，其系统框架见图 4.1. 图 4.3 和图 4.4 分别为爬行控制器和爬行器终端的系统界面图。其中，作为主控端的爬行控制器，根据爬行策略控制全部爬行器客户端的同步启动和终止爬行，并通过接收报文的方式监控爬行终端的爬行过程。同时同步接收多个爬行器上报的拓扑数据，合并生成覆盖网络拓扑总表。爬行器终端是系统的数据采集子系统，负责对 P2P TV 覆盖网络拓扑信息的采集，单个爬行器终端通过运行 P2P TV 爬行引擎，获取拓扑的局域视图，并将局域视图以增量方式发送给控制器。多个爬行器终端可同时运行在不同网络终端上，在控制器的控制管理下，对同一个 P2P TV

覆盖网络并行地进行爬行。

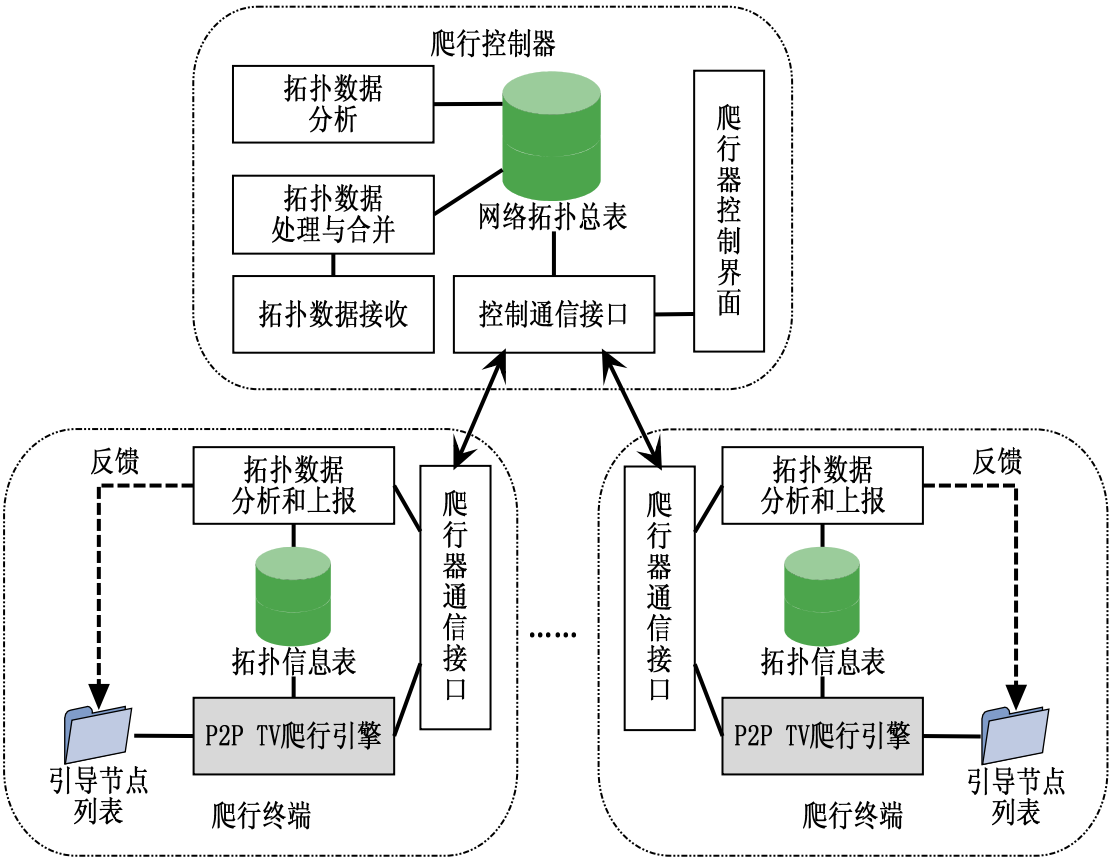

图 4.1 TVCrawler 系统框架结构

TVCrawler 的核心模块是 TV 爬行器引擎（Crawler Engine），用于实现多 P2P TV 系统爬行测量，爬行器引擎的结构见图 4.2。P2P TV 协议知识库通过前期的逆向工程工作分析得到，包括网络协议的格式、编码等语义信息以及预定义的报文模板。覆盖网络拓扑数据库则存储爬行的结果，包括 P2P TV 节点和连接信息。P2P TV 协议报文生成与解析在协议知识库的基础上生成爬行所需要的协议报文（如邻居节点集查询报文），并且对爬行过程接收到的结果报文进行解析，以便获取邻居节点信息。待爬行节点生成模块是迭代爬行的基础，反复从拓扑数据库中取出新的未爬行节点作为下次爬行的目标节点。邻居查询结果过滤对爬行得到的节点集进行去重、过滤等处理。邻居节点集查询模块完成查询报文的发送和查询结果报文的接收。

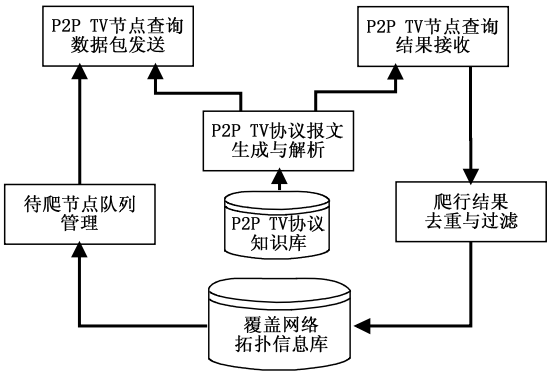

图 4.2 TVCrawler 爬行引擎结构

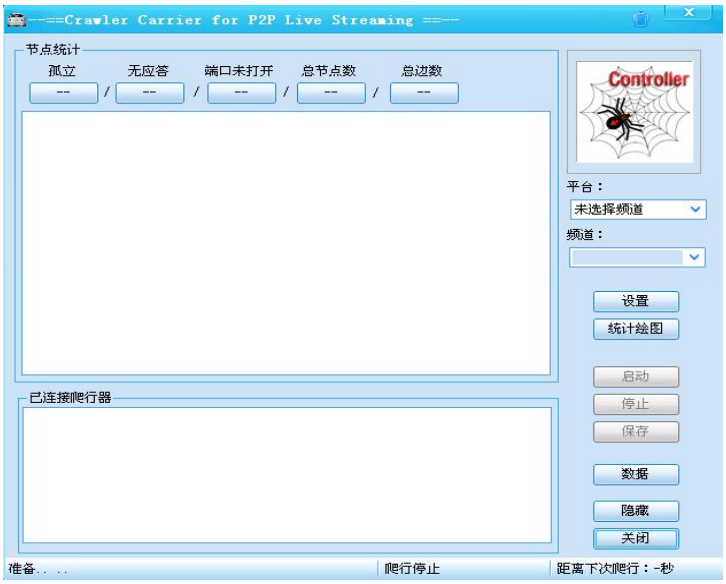

图 4.3 爬行控制器界面图

图 4.4 爬行终端界面图

4.1.2 引导节点构造

爬行器的爬行启动需要给定初始的引导节点集，作为爬行的初始邻居节点。TVCrawler 在爬行引导节点集的构造上采用以下三种方式：

（1）P2P TV 网络中通常提供若干固定地址的成员服务器，TVCrawler 模拟新加入节点的注册过程，向成员服务器请求获取引导节点集。

（2）在测量网络中运行 P2P TV 客户端并加入指定的直播频道，将该客户端作为初始引导节点。

（3）采用反馈方式更新和构造初始引导节点集[77][78]。在连续多次地爬行测量中，TVCrawler对上一次爬行测量得到的拓扑数据进行分析，将节点按度数从大到小排序，选取排名前 n 位的节点作为反馈，加入或替换引导节点集。

4.2 在线用户分布模型

4.2.1 用户在线时长分布

用户在线时长是指用户节点一次加入频道至离开频道的时间，反应了节点一次停留在频道中的时间长度。本研究的测量数据是一系列在时间上连续的快照集 $S=\{s_0,s_1,...,s_N\}$，s_i 表示第 i 张快照。对测量得到的每一个节点 p，查找对应的快照子序列 $S_p=\{s_n,s_{n+1},...,s_{n+i}\}$，表示在第 n 张快照中首次发现节点 p，并在随后的连续 i 张快照中都包含节点 p，直至快照 s_{n+i+1} 该节点消失。将快照序列 S_p 的总时长 T_s 作为节点的会话长度。对于退出频道后再次加入的节点，同样当作新节点加入处理。实际上，由于存在快照测量时长和测量间隔时长，T_s 只是节点实际会话长度的近似值，偏差最大时约为连续两张快照形成的间隔时间。图 4.5 为 PPlive、PPStream 和 UUsee 平台用户在线时长分布。

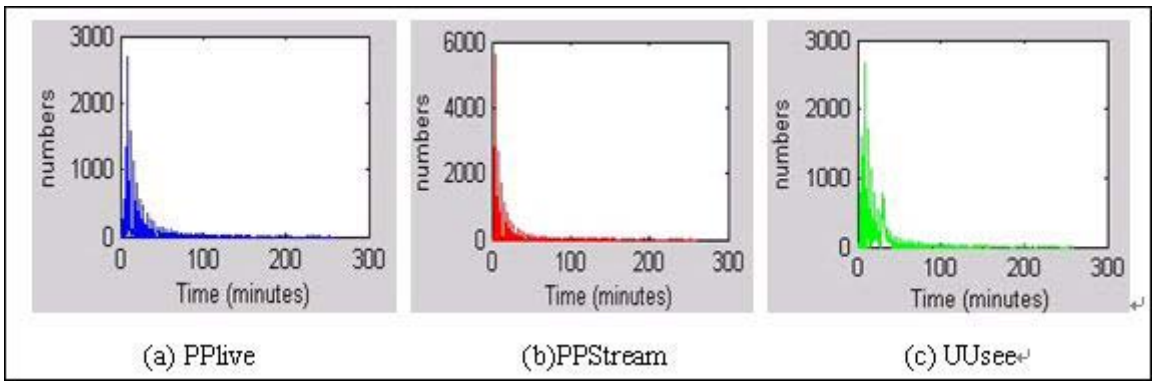

图 4.5 用户在线时长分布

从图 4.5 中可以看出，PPLive 和 PPStream 平台有 70%的节点的会话长度小于 10

分钟，其平均会话长度在 15 分钟左右。UUsee 平台的会话长度相较 PPlive 和 PPStream 平台稍长，其中有 80%的节点会话长度小于 50 分钟，平均会话长度为 40 分钟左右。

4.2.2 基于在线时长的用户分类

根据用户的在线会话时长可以将用户分为三类：轻度收看者、中度收看者和重度收看者。根据用户平均会话分布，我们将会话时长在 10 分钟以内的用户定义为轻度收看者，该类用户行为一般表现为浏览电视频道，短时间内不断换台，查找自己喜欢看的节目；10 分钟至 100 分钟之间的用户定义为中度手看者，该类用户为正常收看者，多数拥有自己固定的喜好，对自己喜爱的节目热衷程度较高；100 分钟以上的定义为重度手看者，该类用户长时间在线，此类用户一般为学生、无业青年或特殊工作群体如网管，有大量时间收看电视节目。图 4.6 为快女播放期间，PPlive、PPStream 和 UUSee 三个平台中湖南卫视频道三类用户会话时长分布。

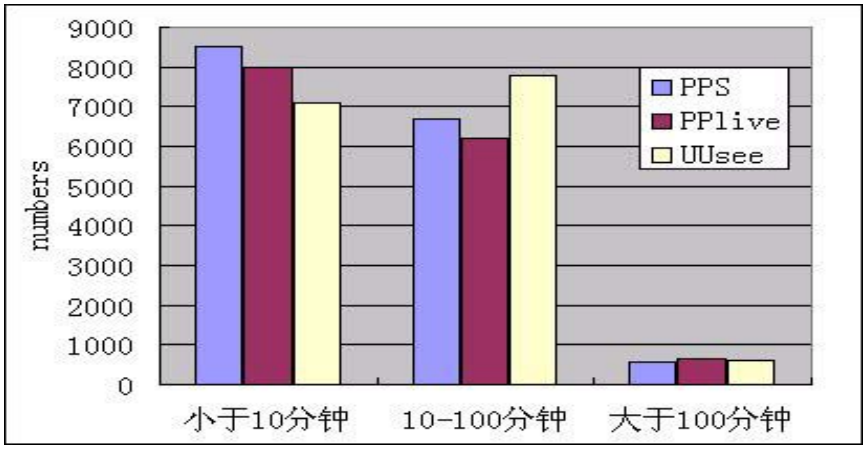

图 4.6 三类用户平均在线时长比较

4.2.3 分类用户的在线时间演化

为研究用户关注内容及关注时间，本文进行了用户在线时长演化分析，依然以 2011 年湖南卫视快女播放期间，PPLive、PPStream 和 UUSee 三个平台的数据进行。分析三个平台湖南卫视频道在快女播放当天用户的行为和平常用户观看湖南卫视的行为。我们以 24 小时内每个时间点，用户的数量为特征描述了如图 4.7、图 4.8 和图 4.9 所示的用户在线时长演化情况。

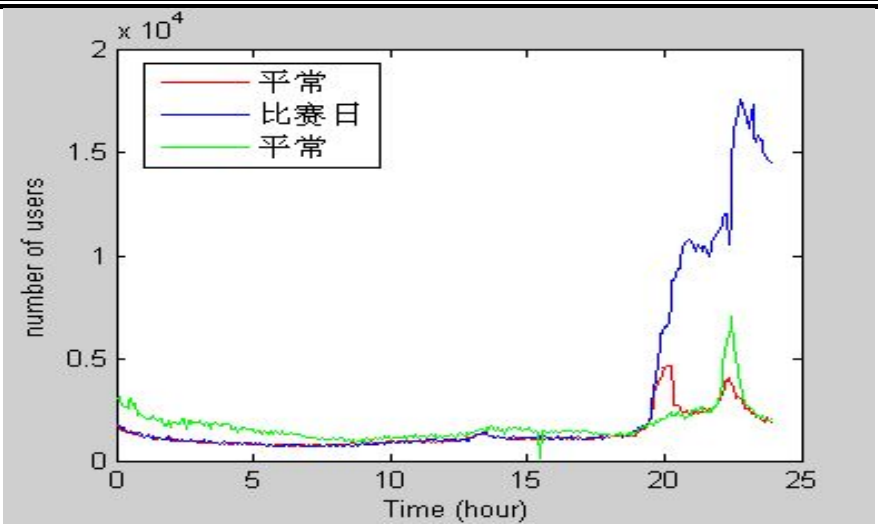

图 4.7 PPlive 用户在线时长演化

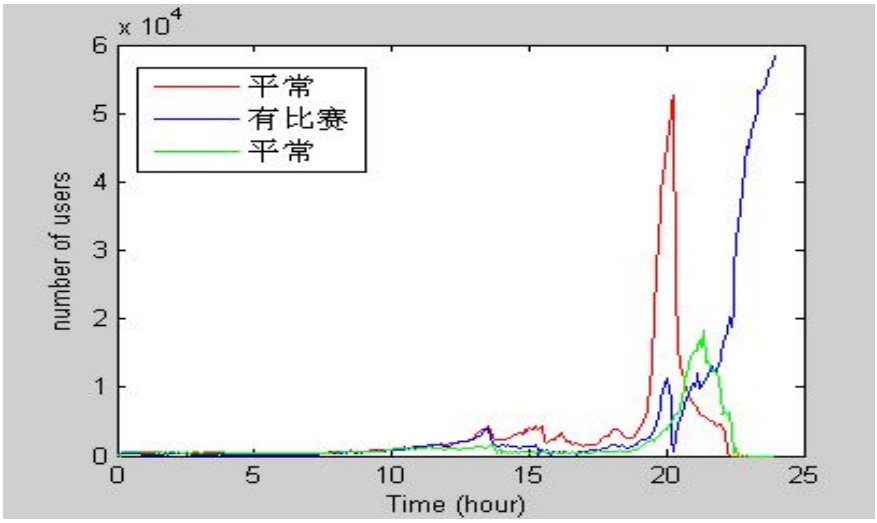

图 4.8 PPStream 用户在线时长演化

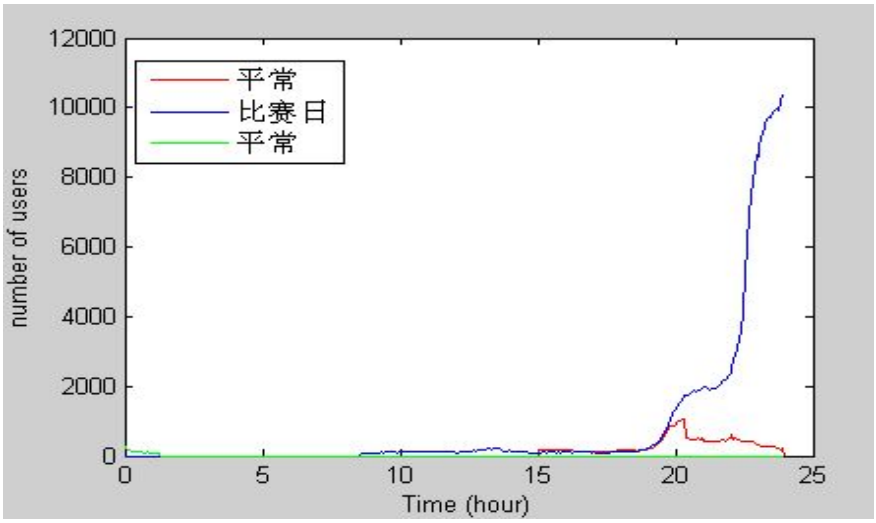

图 4.9 UUsee 用户在线时长演化

从图中可以看出无论在平时还是在比赛日在晚 20:00 时左右，用户数量有个小的突起，这反映了 20:00 左右为湖南卫视的黄金时间。一般电视台在新闻联播播出后会播出各种收视率较高的电视节目，这与电视台安排的黄金时间相符合。但是在快乐女生比赛当天我们可以看出在晚 22:00 至 24:00 之间用户数量会剧增，这说明了在比赛当天，有众多用户在观看湖南卫视的快乐女生，且多为重度收看者，在两个小时之内用户数量一直呈增长状态。在试验中我们发现在比赛第二天的在线用户人数会比平常的要少，图中绿色的线表示比赛第二天用户的在线时长分布。这跟用户个人的行为习惯息息相关，前一天熬夜看快乐女生比赛，第二天习惯于早点休息，以恢复精力。通过用户在线时长的演化分析可以辅助决策者调整战略。如电视台可以了解到什么电视内容比较受欢迎，什么时间投放广告效果最好，在该例中我们可以看出在平常 20:00 左右投放广告效果比较好，而在快乐女生播放当天在 22:00 至 24:00 投放广告效果更佳，而在比赛第二天的广告效果较平常会减弱。

4.2.4 用户到达率

用户到达率(arrival rate)是指在 P2P TV 网络中，单位时间内加入 P2P 网络的数量。用户到达率是研究 P2P 网络动态性和用户行为的重要指标。定义 t 时刻的用户到达率如公式(4.1)：

$$A_t = \frac{N_{t+1} - N_t}{T_{t+1} - T_t} \tag{4.1}$$

其中 T_t 表示第 t 张快照的获取时间，N_t 是第 t 张快照中的用户数。在本文研究中时间间隔以分钟为单位，即 A_t 表示每分钟内的用户到达率。图 4.15 为在三个系统被测量频道在一周范围内用户到达率的时间演化。可以看出用户到达率与频道在线人数类似，具有很强的时间相关性，而且同样具有工作日模式、周末模式和特殊事件模式等三种日模式。在 P2P TV 系统中，用户到达行为与频道的节目编排是高度相关，当频道开始播放热门节目时，通常会在短时间产生大量用户到达事件，引发了所谓的 Flash crowd 现象。比较在线用户人数演化图与图 4.14 可以看出在线用户人数与用户到达率呈现正相关性，即随着在线人数的增加或减少，到达率也呈明显上升或下降趋势。因此我们认为在 P2P TV 系统中，用户到达事件并非独立随机事件。表 4.1 给出了湖南卫视快乐女生总决赛的时刻表

表 4.1 快乐女生总决赛时刻表

日期	星期	时间
7 月 15 日	星期五	22:00-次日 01:00
7 月 22 日	星期五	22:00-次日 01:00
7 月 29 日	星期五	22:00-次日 01:00
8 月 5 日	星期五	22:00-次日 01:00
8 月 12 日	星期五	22:00-次日 01:00
8 月 19 日	星期五	22:00-次日 01:00
8 月 26 日	星期五	22:00-次日 01:00
9 月 2 日	星期五	22:00-次日 01:00
9 月 9 日	星期五	22:00-次日 01:00
9 月 16 日	星期五	22:00-次日 01:00

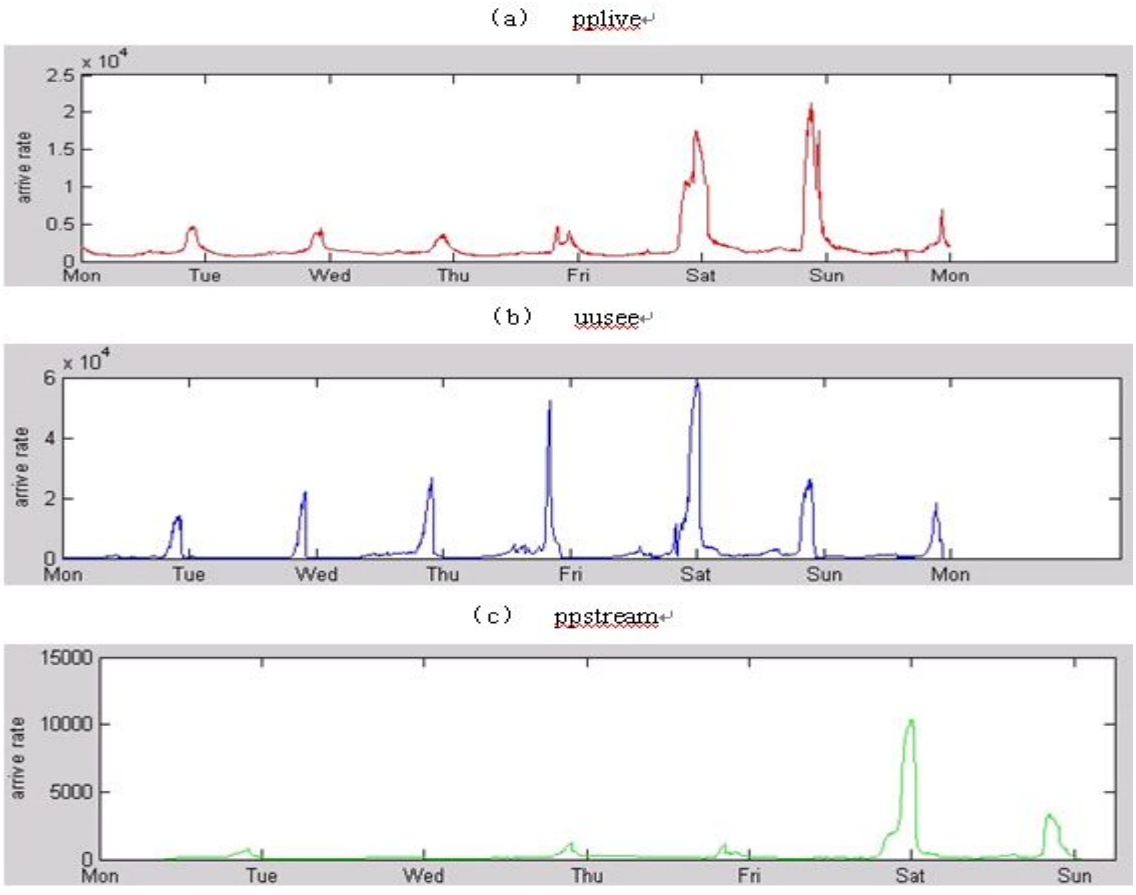

图 4.10 频道用户到达率的时间演化

对比表 4.1 和图 4.10 三个平台的用户到达率在每周五晚至周六凌晨达到一个高峰，与快乐女生播放时间相吻合。说明在快乐女生播放期间，有大量用户加入到网络中，即为特殊事件模式，在短时间内有大量用户到达事件，引发了 Flash

crowd 现象。

4.3 用户地理位置分布

4.3.1 拓扑预处理

地理空间信息可视化是信息可视化中重要的技术，涉及大多数国民经济的行业，已经有广泛的应用。P2P TV 在线用户的地理空间可视化通过强大的、有效的地图系统将复杂的空间和属性数据以地理的形式展现出来，从而挖掘数据之间的关联性和发展趋势，了解 P2P TV 传播动态和范围、预测传播趋势，发现传播行为异常，进而做出及时和正确的判断和监管决策。

本文采用了谷歌地图（Google Map）技术[79]进行频道用户的地理空间可视化。谷歌地图是Google公司向全球提供的在线电子地图服务，能提供三种视图：一是矢量地图；二是不同分辨率的卫星照片；三是可以用以显示地形和等高线的地形图，本文主要采用矢量地图运用谷歌还提供的免费的地图应用接口服务Google Map API来进行用户地理位置可视化。

由于快照中包含的节点数和覆盖网络的边数过于巨大，直接进行可视化计算和显示的复杂度高，因而需要对快照中的覆盖网络进行面向地理信息的简化。简化算法的基本思想在于合并地理上接近的节点，并同时对关联的边进行合并处理。

记快照对应的覆盖网络为 $G=(V,E)$，其中 $V=\{v_i,w_i\},i=1,2,\ldots,N$ 为覆盖网络的节点集，w_i 为节点权重，每个节点的初始权重为 1。每一个节点都有对应的经纬度坐标属性。e_{ij} 表示节点 v_i 到节点 v_j 的边，$E=\{e_{ij}\}$ 为覆盖网络的边集。D 表示节点间的距离算子，K 表示合并阈值。简化算法如图 4.11 所示：

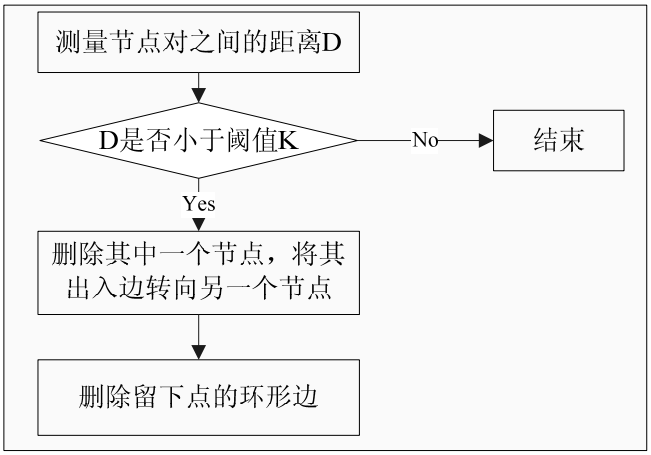

图 4.11 拓扑预处理流程

4.3.2 用户分布可视化

本文以 2011 年湖南卫视快乐女生举办期间，采集 PPLive、UUsee 和 PPstream 三个网络电视平台的数据进行分析。TVCrawler 爬取到用户的数据格式为一个二元组<IP 地址:端口号>定义了测量时刻的一个在线用户 IP 地址、本地端口号等信息。为了得到用户的地理信息，需要将用户 IP 地址转换为对应的地理位置信息，目前国内常用的 IP 数据库有 QQIP 数据库、纯真 IP 数据库等，但这些数据库针对国外的 IP 数据存在严重不足的缺陷，由此本文采用国外较为著名的 MaxMind 公司的 GeoIP 数据库，GeoIP 数据库不仅提供了 IP 地址到国家、省市、街道的映射和查询，同时还提供了 IP 地址到经纬度的定位信息查询。用户分布可视化的流程如图 4.12 所示。由于 TVCrawler 爬行器在爬取用户信息时采用了简化算法，故在同一经纬度上可能有多个用户。本文在将用户 IP 转化为经纬度后，采取合并策略，统计同一经纬度上的用户数目，根据用户数目的多少我们定义了三个级别：小于 10 个用户、10 个用户至 100 个用户和大于 100 个用户。分别采用三种不同大小的图标来表示用户群体的大小，描绘在 GoogleMap 上，并且点击图标后即可显示该经纬度具体的用户数量。

以 2011 年 9 月 16 日 8 点至 9 点采集的 PPLive、UUsee 和 PPstream 三个平台的湖南卫视的数据进行试验。我们可以看出大多数用户来自于中国大陆，且多分布在大中城市，乡村和城镇的用户数量较少。结果如图 4.13、图 4.14 和图 4.15 所示

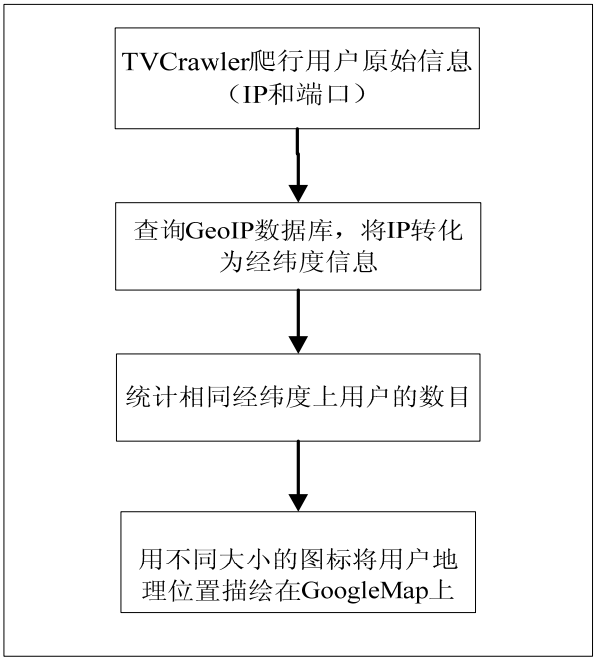

图 4.12　用户地理位置可视化流程图

图 4.13 PPlive 用户地理分布

图 4.14 PPstream 用户地理分布

图 4.15 UUSee 用户地理分布

4.4 本章小结

本章利用 TVClawer 爬行器，爬取 PPLive、PPStream 和 UUsee 三个平台湖南卫视快乐女生播放期间的数据，进行了用户行为分析，包括用户地理位置分布、用户在线时长分布、在线用户分类分析和用户在线时长分析。通过这些分析让我们可以局部了解用户在某事件中观看习惯、用户所关心的事件内容、针对某一事件哪个地区的用户较为关注等等。通过这些用户行为的分析，可以辅助我们调整战略。

第五章　微博和 P2P TV 用户行为对比分析

本文第三、四章分别就微博用户和 P2P TV 用户的行为做了分析，本章将从用户在线数量分布、在线时长分布和地理位置分布这三个方面就快乐女生事件对比分析两个不同数据集的用户行为的异同。

5.1　用户在线数量对比分析

用户在线数量的分布表明了事件所受的关注程度。本文对比分析 2011 年 7 月 1 日至 2011 年 9 月 20 日，新浪微博用户在线数量分布和 PPlive 湖南卫视用户在线数量分布情况。针对微博数据，我们将该期间内新浪微博上发表有关快乐女生事件用户的数量，用户参与快乐女生选手及湖南卫视官方微博，关于快乐女生博文的评论和转发的用户数量，以天为单位进行统计，从而获取用户在该段时间内每天用户的参与量。针对 PPlive 网络电视数据，按天统计爬行下来的快照数据，统计不重复 IP 的数量即为该天用户的在线量，如图 5.1 所示。

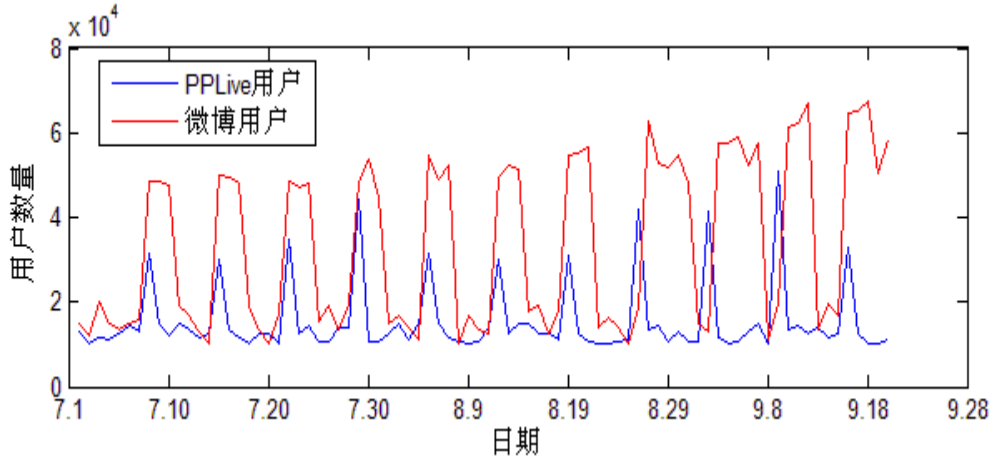

图 5.1 用户在线人数分布

图中蓝色的曲线为该段时间内收看 PPlive 湖南卫视用户的分布情况，红色的曲线为该段时间内新浪微博上发表有关快乐女生博文或参与转发和评论快乐女生微博的用户数量。从图中我们可以看出从 7 月 1 日至 9 月 20 日，快乐女生节目播出期间用户的在线数量相交平常都要多出许多，表现为曲线中的尖峰。另外微博用户在线数量分布与 P2P TV 用户在线数量分布上也有一些微小的差别，主要表现在 P2P TV 的用户在线数量突起的时刻为快乐女生节目播出当天，第二天就会迅速减退回到平常状态，而微博用户在线数量的突起时间一般会持续一到两天，随后缓慢减少。

5.2 用户在线时长对比分析

P2P TV 的用户在线时长是指用户节点一次加入频道至离开频道的时间，反应了节点一次停留在频道中的时间长度。微博用户在线时长是指用户发表博文或参与评论的时间长度，当用户 10 分钟未发表博文或评论定义用户下线。对比两个平台用户在线时长我们发现，用户在线时长分布都呈现负指数分布，如图 5.2 所示。

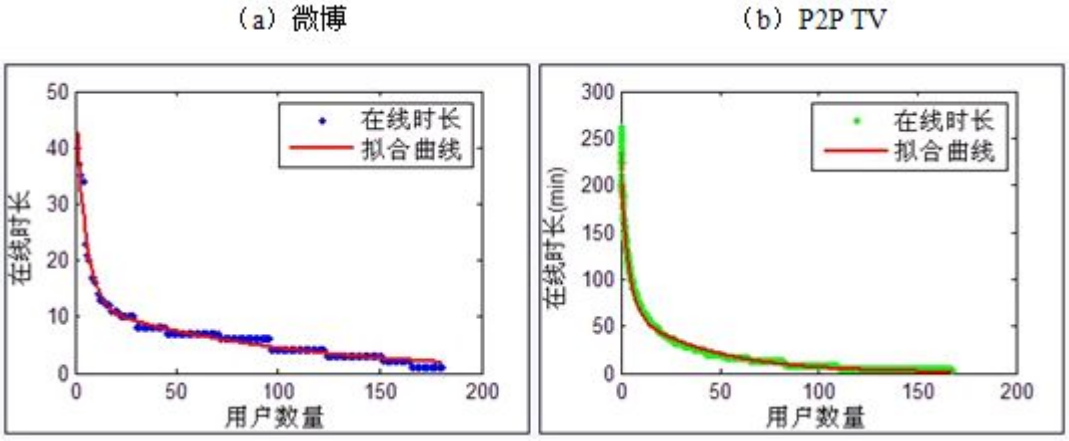

图 5.2 用户在线时长分布

P2P TV 用户和微博用户在线时长都符合负指数函数 $f(x)=a\times e^{(bx)}+c\times e^{(dx)}$ 其中 X 表示用户数量，$f(x)$ 表示用户在线时长，a,b,c,d 为参数，图中红色的曲线为 Matlab 拟合曲线。表 5.1 给出了对应两个平台参数 a,b,c,d 的值。

表 5.1 拟合曲线参数

参 数	微博用户	P2P TV 用户
a	37.35	163.8
b	-0.2051	-0.2772
c	12.38	70.91
d	-0.01032	-0.02485

通过计算我们得出微博用户平均在线时长大约为 6 分钟，而 P2P TV 用户平均在线时长大约为 15 分钟。

5.3 用户在线时间演化分析对比

本部分对比分析一个工作日内两个平台用户在线时间演化情况，了解两个平台用户上网行为习惯，两个平台用户各自趋向在什么时间段上网。P2P TV 平台以 PPlive 湖南卫视用户为例，研究 24 小时内用户在线时间走势；微博平台以 24 小时内用户发布有关快乐女生博文及转发和评论快乐女生相关博文的数据为基础，研

究用户 24 小时内各个时间段内用户数量分布。图 5.3 和图 5.4 分别为 PPlive 平台用户在线时间分布和微博用户在线时间分布。

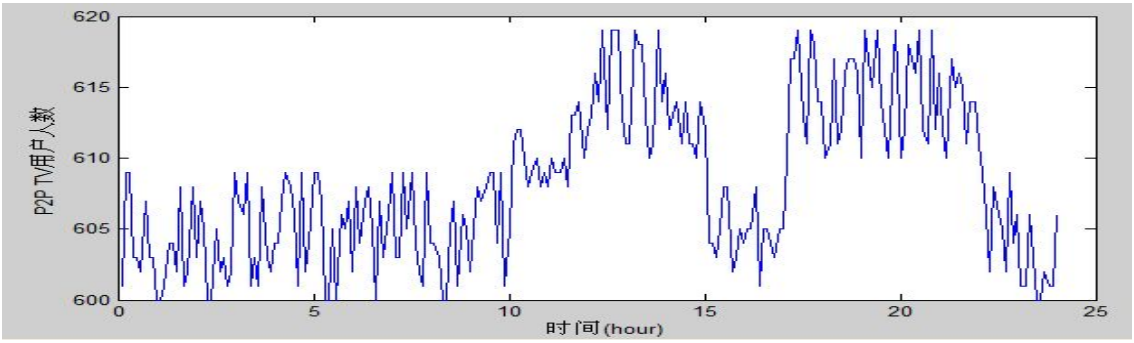

图 5.3 P2P TV 用户在线时间分布

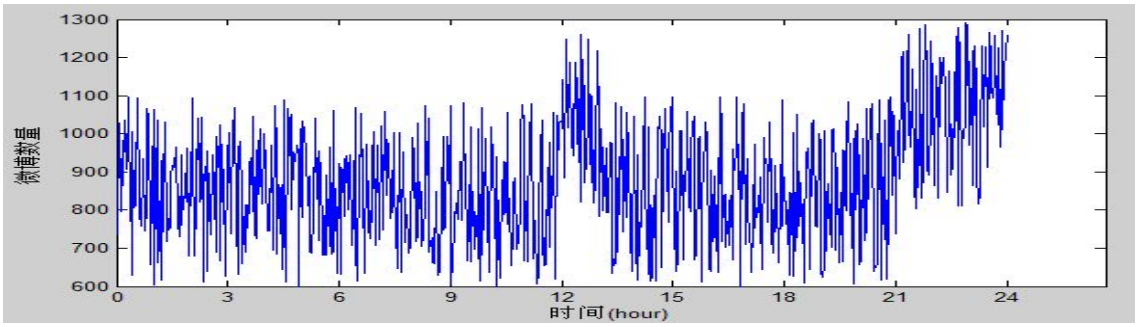

图 5.4 微博用户在线时间分布

从图中可以看出，P2P TV 用户在线时间比较固定，在中午 12 点至下午 15 点及晚上 20 点至 22 点 P2P TV 在线人数较其他时间段多，而微博用户在线时间则分布比较平均，在中午 12 点至下午 15 点及晚上 20 点至 22 点用户人数虽然有一定的增加，但增加幅度并不大，造成这一现象的原因在于微博作为消息传递平台，使用简单，用户可以随时随地发表博文和评论，发送手段不仅仅限于 PC 机，还包括手机、PDA、IPad 等其他移动设备。P2P TV 平台和微博平台用户数量在在中午 12 点至下午 15 点及晚上 20 点至 22 点都有一个跃起，这段时间主要是 PC 机用户，一般表现为用户下班后，时候电脑登陆互联网观看节目和发表博文的。

5.4 用户地理分布对比分析

以 2011 年 9 月 16 日快乐女生总决赛数据为基础，对比分析当晚观看快乐女生湖南卫视频道用户的地理分布和总决赛三位选手微博粉丝各自前排在前三的省份。总决赛三位选手粉丝中排在前三名的省份如表 5.2 所示。

表 5.2 总决赛选手微博粉丝分布

姓　名	数量最多的省份	数量第二的省份	数量第三的省份
刘　忻	湖南省	广东省	山东省
段林希	广东省	湖南省	四川省
洪　辰	浙江省	广东省	江苏省

从表中我们可以看出，三位选手粉丝数量最多的三个省份总共分布在湖南省、广东省、山东省、浙江省、江苏省和四川省这 6 个省份。同样我们分析 9 月 16 日晚观看湖南卫视快乐女生总决赛用户数量排在前 6 名的省份。如表 5.3 所示。

表 5.3 P2P TV 用户省份分布

第一名	广东省	第四名	湖南省
第二名	浙江省	第五名	湖北省
第三名	江苏省	第六名	福建省

对比表 5.2 和表 5.3 我们发现两个平台用户地理分布呈现一定的规律，微博平台用户地理分布前 6 名的省份中包含 P2P TV 用户数量前 4 名的省份，即广东省、浙江省、江苏省和湖南省。说明这四个省份用户对“快乐女生“事件关注程度最高，在两个平台上都有所体现。

第六章　结束语

6.1　本文主要工作及创新点

本文以 2011 年湖南卫视快乐女生这个事件为基础，采集新浪微博数据和 P2P TV 数据，进行了用户行为分析。得出了很多有意思的结论。文章涉及两个数据集，针对这两个数据集分别进行了不同的用户行为分析。

（1） 针对新浪微博用户行为分析

本文从焦点人物和活跃用户行为两个方面出发，分析了在网络中处于消息传播的关键节点。其中焦点人物分析分三个方面展开：粉丝地理位置分布、活跃用户分类和关注度分析。通过该三个方面的分析，能够对焦点人物有个宏观的了解，有利于辅助消息传播和舆论宣传，做到有的放矢。另外对于某一事件中的活跃用户，我们对其行为展开了初步分析，主要包括用户参与事件整体态势分析、单个用户发帖行为分析和用户网络拓扑结构分析，活跃用户以其高度的参与量，易于主导舆论方向，故对其研究具有重要意义。

（2） 针对 P2P TV 用户行为分析

本文从 P2P TV 主动测量的角度着手，重点研究 P2P TV 系统中用户分布以及用户行为模式等内容的测量和分析，从而为 P2P TV 系统的用户监测、异常监管、传播预测等内容提供数据基础和模型支持。主要内容为 P2P TV 频道用户的在线行为，提出了一个基于在线时长分布特征的直播频道用户分类方法，并对分类用户的到达率和动态性进行了比较研究。本文以 2011 年湖南卫视快乐女生播放期间的数据为基础，对 PPLive、PPStream 和 UUSee 三个系统的湖南卫视频道进行测量分析，重点研究了 P2P TV 频道中用户行为。其中按照 P2P TV 用户的在线时长的分布，将用户分为轻度收看者、中度收看者和重度收看者三类，对应的平均在线时长分别为 3~10 分钟，10~100 分钟和 100~200 分钟，并进一步分析了在被测频道中，三类用户各自的在线人数和到达率。

论文的创新点包括三个方面：

（1）设计并实现了，基于垂直搜索的 SinaClawer 爬行器，可根据所关注的内容，设定初始爬行节点，是针对性的为某一特定领域、某一特定人群或某一特定需求提供的有一定价值数据采集系统。垂直搜索爬行器只搜索特定的主题信息，按预先已定义好的专题有选择地收集相关的网页。这样大大降低了收集信息的难度，提高了信息的质量。

（2）设计并实现了，基于主动测量的 TVClawer 爬行器，基本思想仍然是遵循 P2P TV 系统的协议流程和语义，模拟客户端与 Tracker Server 及其他节点通信

从而获取节点邻接关系并加入本地节点队列，反复迭代直至获取到足够的 P2P 覆盖网络信息，该爬行器采用 C/S 结构，采用分布式爬行方式，分有爬行控制端和爬行客户端，增加了爬行器的灵活性和爬行效率。

（3）研究了 P2P TV 在线用户行为，提出了基于在线时长演化的用户行为分析，从直观的图表中，我们可以清楚明了的了解用户在各个时段收看节目的习惯，所关心的播放内容等等，这对于电台调整其播放内容和播放时间、用户行为的监控有着重要的参考意义。

6.2 未来工作展望

本文的研究工作是针对新浪微博用户及 P2P TV 用户进行网络用户行为分析，研究用户参与网络事件的规律，寻找异常用户，为事件监控提供依据。

本文的研究工作在以下三个方面还有待深入研究：

（1）大事件预测

结合博文内容和用户行为，分析近期内讨论比较热烈的热点话题，并分析其未来走势，预测其是否可能演化为网络大事件，并分析网民对该事件的态度，分析讨论该事件网民的行为，为大事件的引导和控制提供依据。

（2）基于博文内容的用户行为分析

本文主要从统计学的角度对用户行为进行研究和分析，在下一步工作中，拟采用自然语言处理的方法，从博文内容出发，研究用户情感和事件发展趋势、寻找网络推手等等。针对博文内容进行深度挖掘。

（3）基于 P2P TV 的网络舆论传播影响力研究与建模

作为视频信息传播的重要渠道，P2P TV 在传播特点、传播规律以及传播影响力上都有着与论坛、博客和微博等不同的方式和特点。本文拟对 P2P TV 的传播规模和范围进行了定量的测量和分析，并结合一些用户反馈调研和测量手段，如 P2P TV 用户抽样调查、P2P TV 相关网站的 Web 挖掘等，对 P2P TV 的信息传播的过程和舆论影响力进行模型化的研究，以用于舆论预测和引导等应用。目前主流的 P2P TV 系统的信息传播过程还是以单向流动为主，但是已经出现若干新的客户端功能，如用户评论、推荐等，表现出 P2P TV 系统增强用户互动性的趋势，这些用户互动信息将是舆论传播影响力研究的又一个重要数据来源。

致　谢

时光飞逝，两年半的硕士生活即将结束；蓦然回首，两年半的硕士学习收获良多。军训的汗水锻炼了我的毅力、上课的充实丰富了我的知识、科研的艰辛培养了我不怕困难、奋勇前进、志在高峰的品质。一次次战胜困难和挑战的背后，有自己的不懈努力，也有各位老师、同学和亲友的热心关怀和无私帮助。在此我要向他们表示我最衷心的感谢和最诚挚的敬意。

首先我要感谢我的恩师王晖教授。两年多来，恩师在学习和生活方面都给予我无微不至的关心，对我在学习和生活上碰到的许多问题都给予很大帮助和精心指导。从课程学习到论文发表，从课题开题到毕业设计，从设备配置到出差活动，从社交礼仪到人生道路，恩师的悉心教导和帮助均历历在目，记忆犹新。恩师的指导方法灵活，善于因材施教，为我们创造宽松舒适的学习环境，让我们能够专心从事自己喜欢的科研方向。恩师严谨的治学态度，渊博的理论知识，深厚的学术造诣，对研究热点的敏锐洞察力是我今后做学问的学习榜样；恩师忘我的工作作风，积极的人生态度，豁达的人生哲学和和蔼可亲的言行举止更是我今后做人的学习榜样。恩师不仅教会了我如何做学问，更教会了我如何做人，在此向恩师致以深深的谢意和诚挚的祝福。

感谢实验室的姜志宏老师、张鑫老师和李沛老师。他们给我介绍许多关于科研方法的宝贵经验，对课题研究提出了许多好的思路和想法，对我的论文写作提出了许多建议，为我的课题出谋划策，帮助我锻炼动手能力，并为建立课题实验环境做了许多工作。

感谢实验室中一起奋斗的各位师兄弟。首先感谢高超、樊鹏翼和李伟师兄，他们在我进实验室初期给我许多帮助，为我介绍课题的研究现状以及一些做课题的经验，使我能够顺利的进入课题研究，与我朝夕相处，我们一起学习，一起讨论，共同度过一段非常美好的时光，课题上给我许多建议和帮助，为人处事上给我许多经验和启发，生活上给我许多关心和照顾，为我搜寻许多关于课题的最新文献，使我能够掌握课题的国际前沿水平，找到课题研究方向和思路。感谢郑亮、李进，他们陪我度过一段非常快乐的时光，在平时的学习和生活中给我许多帮助。

学员队方面，首先要感谢队长李龙美，政委朱志伟、熊启贵，正是他们的悉心关怀和辛勤的工作为我创造了良好的学习环境。感谢队里所有同学，他们在三年的同窗生活中给我时刻有家的感觉，带来的无边乐趣与无数的帮助。

感谢我的妈妈，对她我始终怀着深挚的爱。一路风雨历程，我前进的每一步都凝聚着她无私的付出，是她用辛勤的汗水和超乎寻常的努力为我创造了优良的学习和成长环境，是她不屈不挠、艰苦奋斗的精神时时刻刻感染着我，激励着我，

振奋着我，是她让我在拼搏奋进中拥有不竭的动力。我要以丰硕的成绩和无尽的孝心来报答她。

感谢国防科技大学。在这里我完成了本科和硕士学习，六年半的学习生活使我成长了许多，提高了许多，进步了许多。这里所经历的一切，所认识的老师和同学，将是我一生的财富！

最后向所有关心和帮助我的人们致以衷心感谢！

参考文献

[1] Coppens T, Trappeniers L, Godon M. AmigoTV: towards a social TV experience [C]. In Proceedings of the 2nd European Conference on Interactive Television. 2004: 1-4.

[2] Philo. Philo [EB/OL]. http://?www.playphilo.com/?, 2010-10.

[3] Regan T, Todd I. Media Center Buddies: Instant Messaging Around a Media Center [C]. In NordiCHI ’04: Proceedings of the third Nordic conference on Human-computer interaction. 2004: 141-144.

[4] Boertjes E. ConnecTV: Share the Experience [C]. In EuroITV 2007: Proceedings of the 5th European Conference on Interactive TV: a Shared Experience. Amsterdam, The Netherlands, 2007: 139-140.

[5] Harboe G, Massey N, Metcalf C, et al. Perceptions of value: The uses of social television [C] // Pablo César J F J, Konstantinos Chorianopoulos. In EuroITV 2007: Proceedings of the 5th European Conference on Interactive TV: a Shared Experience. Amsterdam, The Netherlands, 2007: 116-125.

[6] FanTalkTV. FanTalkTV [EB/OL]. http://?www.fantalk.tv/?, 2010-10.

[7] tvClickr. tvClickr [EB/OL]. http://?www.tvclickr.com/?, 2010-10.

[8] http://?www.exmachinagames.com/playtotv-live-social- gaming, 2010-10.

[9] http://www.iresearch.com.cn

[10] CoolStreaming website [EB/OL]. http://?www.coolstreaming.us/?, 2009-05.

[11] Zhang X, Liu J, Li B. On Large Scale Peer-To-Peer Live Video Distribution: CoolStreaming and Its preliminary Experimental Results [C]. In MMSP2005: Proceedings of the IEEE 7th International Workshop on Multimedia Signal Processing. Piscataway, NJ, USA, 2005.

[12] Li B, Xie S, Qu Y, et al. Inside the New Coolstreaming: Principles, Measurements and Performance Implications [C]. In Proceeding of the 27th IEEE Conference on Computer Communications. Piscataway, NJ, USA, 2008: 1031-1039.

[13] Wu C, Li B, Zhao S. Exploring Large-Scale Peer-to-Peer Live Streaming Topologies [J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP). 2008, 4 (3): 19.

[14] Liu Z, Wu C, Li B, et al. Why are Peers Less Stable in Unpopular P2P Streaming Channels [J].

[15] Hei X, Liang C, Liang J, et al. Insights into PPLive: A Measurement Study of a

Large-Scale P2P IPTV System [C]. In Proceedings of the IPTV Workshop, International World Wide Web Conference. New York, May 22-26 2006.

[16] Hei X, Liang C, Liang J, et al. A Measurement Study of a Large-Scale P2P IPTV System [J]. IEEE Transactions on Multimedia. 2007, 9 (8): 1672-1687.

[17] Vu L, Gupta I, Liang J, et al. Measurement of a Large-scale Overlay for Multimedia Streaming [C]. In Proceedings of the 16th international symposium on High performance distributed computing. New York, June 2007: 241-242.

[18] Vu L, Gupta I, Nahrstedt K, et al. Understanding Overlay Characteristics of a Large-scale Peer-to-Peer IPTV System [J]. ACM Transactions on Multimedia Computing, Communications and Applications. 2009, 6 (4): 2010.

[19] Silverston T, Fourmaux O. Measuring P2P IPTV Systems [C].In NOSSDAV' 2007: Proceedings of the 17th International workshop on Network and Operating Systems Support for Digital Audio & Video. New York, 2007.

[20] Qiu T, Ge Z, Lee S, et al. Modeling user activities in a large IPTV system [C]. In IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. New York, NY, USA, 2009: 430-441.

[21] M A, M M, PG K, et al. Measurement and analysis of online social networks [C]. In Proceedings of Internet Measurement Conference'07. California, 2007.

[22] G L, T E, C S, et al. Analyzing patterns of user content generation in online social network [C]. In Proceedings of SIGKDD,. Paris, 2009.

[23] Meeyoung C, Haewoon K, Pablo R, et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system [C]. In IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. New York, NY, USA, 2007: 1-14.

[24] Cheng X, Dale C, Liu J. Statistics and social network of YouTube videos [C]. In Proceedings of the 16th IWQoS'08. Enschede, 2008.

[25] Biel J. subscribe to me! Analyzing the structure and dynamics of the Youtube network [C]. 2009.

[26] Biel J-I, Gatica-Perez D. Wearing a YouTube hat: directors, comedians, gurus, and user aggregated behavior [C]. In MM '09: Proceedings of the seventeen ACM international conference on Multimedia. New York, NY, USA, 2009: 833-836.

[27] C M, M A, PG K. A measurement-driven analysis of information propagation in Flickr social network [C]. In in proceedings of WWW'09. Madrid, 2009.

[28] Java A, Song X. Why we twitter: understanding microblogging usage and communities [C]. In Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD workshop. California, 2007.

[29] Huberman B, Romero D, Wu F. Social networks that matter: twitter under microsope [J]. First Monday. 2008, 14 (1).

[30] Kawk H, Lee C, Park H, et al. What is twitter, a social network or a news media [C]. In Proceedings of the WWW'10. Raleigh, 2010.

[31] Donghee W, Eun-Kyung N. Tweeting About TV: An AEIOU Model of Sociable Television Behavior [C]. In presented at the annual meeting of the International Communication Association. Singapore, 2010.

[32] Li X, Shang G, Zhao Z, et al. A typical TV audience rating model and its competition dynamics analysis [C]. In Control and Decision Conference (CCDC), 2010 Chinese. 2010: 1840 - 1844.

[33] Zheng Y. Audience Rating Prediction of New TV Programs Based on GM (1.1) Envelopment Model [C]. In Proceedings of 2009 IEEE International Conference on Grey Systems and Intelligent Services. Nanjing, China, November 2009.

[34] Chen K-C, Teng W-G. Adopting User Profiles and Behavior Patterns in a Web-TV Recommendation System [C]. In Proceedings of The 13th IEEE International Symposium on Consumer Electronics (ISCE2009). 2009: 320-324.

[35] Shin H, Lee M, Kim E Y. Personalized Digital TV Content Recommendation with Integration of User Behavior Profiling and Multimodal Content Rating [J]. IEEE Transactions on Consumer Electronics. 2009, 55 (3): 1417-1423.

[36] Takama Y, Muto Y. Profile Generation from TV Watching Behavior Using Sentiment Analysis [C]. In Proceedings of 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2007: 191-194.

[37] Barabási A-L. The origin of bursts and heavy tails in human dynamics [J]. Nature. 2005, 435: 207-211.

[38] Oliveira J G, Barabási A-L. Human Dynamics: The Correspondence Patterns of Darwin and Einstein [J]. Nature. 2005, 437: 1251.

[39] Vázquez A, Oliveira J G, Dezs? Z, et al. Modeling bursts and heavy tails in human dynamics [J]. Physical Review E. 2006, 73: 036127.

[40] Vázquez A. Exact results for the Barabasi model of human dynamics [J]. Physical Review Letters. 2005, 95: 248701.

[41] 韩筱璞, 周涛, 汪秉宏. 基于自适应调节的人类动力学模型 [J]. 复杂系统与复杂性科学. 2007, 4 (4): 1-5.

[42] Malmgren R D, Stouffer D B, Campanharo A S L O, et al. On Universality in Human Correspondence Activity [J]. Science. 2009, 325 (5948): 1696-1700.

[43] 郑丽勇, 郑丹妮, 赵纯. 媒介影响力评价指标体系研究 [J]. 新闻大学. 2010

(1): 121-125.
[44] PPLive. http://www.pplive.com
[45] UUSee. http://www.uusee.com
[46] PPStream. http://www.ppstream.com
[47] Brian Pinkerton.Finding What People Want: Experiences with the WebCrawler. Second International WWW Conference[C]. Chicago,Illinois,1994.
[48] 刘玮玮.搜索引擎中主题爬虫的研究与实现[D]. 南京: 南京理工大学，2006.
[49] 周立柱,林玲. 聚焦爬虫技术研究综述[J]. 计算机应用,2005,25(9): 1965-1989.
[50] 王辉. 基于质心具有增量性质的主题爬行[D]. 长春: 吉林大学. 2007.
[51] http://www.hibernate.com/
[52] Torkvist L, Stahlberg L, Bohman L, et al. LMWH for the treatment of mild to moderately active steroid refractory/dependent UC – an independent, multicentre, randomised controlled study. Gastroenterology 2001; 120: A1435.
[53] Spring Java/J2EE Application Framework, Reference Documentation, 2004.
[54] I.Stoica, R.Morris, D.Karger, F.Kaashoek and H.Balakrishnan,"Chord: A scalable content-addressable network," in Proceedings of the ACM SIGCOMM 2001 Technical Conference, (San Diego, CA, USA), August 2001.
[55] Gummadi KP,Dunn RJ,Saroiu S,Gribble SD, Levy HM,Zahorjan J. Measurement ,modeling,and analysis of a peer-to-peer file-sharing workload.In:Proc. of the 19th ACM Symp. on Operating Systems Principles(SOSP-19). 2003.314-329.
[56] Liang J, Kumar R, Ross KW. The KaZaA overlay: A measurement study. In: Proc. of the 19th IEEE Annual Computer Communications Workshop.2004.
[57] Ripeanu M, Foster I, Iamnitchi A. Mapping the gnutella network: properties of large-scale peer-to-peer systems and implications for system design. IEEE Internet Computing Journal, 2002,6(1):50-57.
[58] A new method for analyzing feedback-based protocols with applications to engineering Web traffic over the Internet Heyman D.P.;Lakshman T.V.;Neidhardt A.L Computer Communications Volume 26 Number 8,20 May 2003,pp.785-803.
[59] Ethereal Labs: Supplement to Computer Networking: A Top-Down Approach Featuring the Internet, 3rd ed. J. Kurose and K. Ross, http://www-net.cs.umass.edu/ethereal-labs/, accessed 9/5/2007.
[60] Kalidindi S, Zekauskas MJ. Surveyor: an infrastructure for Internet performance measurements. In: Proceedings of the INET'99. San Jose, 1999. http://www.isoc.org/inet99/proceedings/4h/4h_2.htm.
[61] http://code.google.com/intl/zh-CN/apis/maps/signup.html
[62] http://www.maxmind.com/
[63] http://www.maxmind.com/app/ip-location

[64] Golgher PB, Laender AHF, Silva AS da,et al.An example-based environment for wrapper generation[A]. Proceedings of the 2nd International Workshop on The World Wide Web and Conceptual Modeling[C]. USA:Salt Lake City,2000.152-164.

[65] JeffHeaton. 网络机器人 Java 编程指南[M]. 童兆丰. 北京: 电子工业出版社,2002.

[66] http://archive.cnblogs.com/a/2217350/

[67] http://blog.csdn.net/allwefantasy/article/details/3136570

[68] http://www.mysql.com/

[69] http://open.weibo.com/wiki/index.php

[70] http://open.weibo.com/wiki/index.php/API%E6%96%87%E6%A1A3#.E8.8E.B7.E5.8F.96.E4.B8.8B.E8.A1.8C.E6.95.B0.E6.8D.AE.E9.9B.86.28timeline.29.E6.8E.A5.E5.8F.A3

[71] RATNASAMY S, HANDLEY M,KARP R,et al. Topologically-aware overlay construction and server selection[C]//Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies.New York, USA:IEEE,2002:1190-1199.

[72] K.Gummadi, R.Dunn, S.Saroiu, Gribble, H.Levy and J.Za-horjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In SOSP, 2003.

[73] X. Zhang, J. Liu, B. Li, et al. 2005. CoolStreaming/DONet: A Data-driven Overlay Network for Peer-to-Peer Live Media Streaming[C], Infocom 2005, Miami, FL, USA, March 13-17, vol. 3, 2102 - 2111.

[74] M. Zhang, Q. Zhang, L.f. Sun, et al. 2007. Understanding the Power of Pull-based P2P Streaming Protocol: We Can Do Even Better [J], IEEE journal on selected areas in communications, Vol. 25, Issue 9, 1678-1694.

[75] S. Asaduzzaman, Y. Qiao, G. Bochmann. 2008. CliqueStream: An efficient and fault-resilient live streaming network on a clustered peer-to-peer overlay[C], in Proceedings of the Eighth International Conference on Peer-to-Peer Computing, P2P'08, pp.269-278.

[76] F. Wang, Y. Q. Xiong, J. C. Liu. 2007. mTreebone: A Hybrid Tree/Mesh Overlay for Application-Layer Live Video Multicast[C]. In Proceedings of the 27th international Conference on Distributed Computing Systems (ICDCS). IEEE Computer Society, Washington, DC, 49.

[77] 王勇，云晓春，李奕飞，等. 一种基于正反馈的对等网络拓扑获取方法[J]. 计算机研究与发展. 2007，44 (9): 1550-1556.

[78] 刘刚，方滨兴，胡铭曾，等. 类 Gnutella 的对等网络的测量方法研究 [J]. 计

算机应用研究. 2006 (6): 230-232.

[79] Google Map [EB/OL]. http://ditu.google.com/, 2010-10.

作者在学期间取得的学术成果

[1] Su He, Zhihong Jiang, Xin Zhang, Hao Kang, Hui Wang, A Traffic Recognition System of P2P IPTV. 2011 IEEE International Conference on Multimedia Information Networking and Security(MINES 2011), 2011.11. 已发表（EI Compendex 检索，ISTP 检索）.