分类号		密级	
UDC		学位论文编号	

重庆邮电大学硕士学位论文

中文题目	面向社交网络的信息流行度预测研究
英文题目	**Research on Prediction of Information Popularity for Social Network**
学　　号	
姓　　名	
学位类别	工学硕士
学科专业	信息与通信工程
指导教师	
完成日期	**2019 年 3 月 18 日**

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含他人已经发表或撰写过的研究成果，也不包含为获得 重庆邮电大学 或其他单位的学位或证书而使用过的材料。与我一同工作的人员对本文研究做出的贡献均已在论文中作了明确的说明并致以谢意。

作者签名：　　日期：　　年　月　日

学位论文版权使用授权书

本人完全了解 重庆邮电大学 有权保留、使用学位论文纸质版和电子版的规定，即学校有权向国家有关部门或机构送交论文，允许论文被查阅和借阅等。本人授权 重庆邮电大学 可以公布本学位论文的全部或部分内容，可编入有关数据库或信息系统进行检索、分析或评价，可以采用影印、缩印、扫描或拷贝等复制手段保存、汇编本学位论文。

（注：保密的学位论文在解密后适用本授权书。）

作者签名：　　导师签名：

日期：　　年　月　日　　日期：　　年　月　日

摘要

随着互联网技术的普及，在线社交网络已经成为当今社会人们信息交流的重要渠道和载体。在线社交网络中信息流行度预测至关重要，具有重要的研究和应用价值。但是，由于信息传播形式的多样性、网络结构的复杂性以及用户特征的多维性使信息流行度预测错综复杂且难以准确把控。如何深入剖析流行度态势变化的规律，感知信息流行态势走向，建立高效的管控措施是待研究和解决的问题。

本论文关于信息流行度预测主要涵盖两个方面的内容：宏观信息流行度态势预测，深入剖析信息流行度传播过程中的非线性动力学机制，构建跨平台信息流行度融合和预测模型；微观信息转发热度预测，分析影响用户转发的微观因素，基于传染病模型预测信息未来的转发情况。本文的详细工作总结如下：

1. 在宏观层面，旨在深入探究社交信息传播的混沌特性，提出一种基于贝叶斯估计理论的跨平台信息流行度融合预测模型。首先，定义跨平台信息流行度时间序列，利用主成分分析(Principal Component Analysis, PCA)量化和获取影响流行度的主成分。其次，基于混沌理论，探究流行度趋势变化的混沌特性，对量化后的序列实施相空间重构，在高维相空间中恢复复杂系统的演化规律和特征。同时，利用贝叶斯估计理论将多个流行度变量在同一高维空间中进行相点的最优融合，得到新的融合相空间。最后，考虑到神经网络在现实应用中有较强的逼近非线性函数的能力，通过其对融合流行度实施优化预测。

2. 在微观层面，旨在深入探究影响用户转发的多维属性，量化改进 SIR (Susceptible-Infected-Recovered)模型中的感染率，提出一种感知信息流行度的用户行为演化策略。首先，提取用户个人和社交维度的转发驱动力，利用多元线性回归量化多维转发感染率。其次，为了让模型更贴近真实网络传播架构，对传统传染病 SIR 模型中状态 S 进行改进，重新定义 SIR 模型的传播规则。最后，通过时间切片技术提取改进 SIR 模型的各个状态值，利用最小二乘法(Least Square, LS)结合量化后的感染率拟合真实模型，得到一种基于用户转发行为和改进 SIR 模型的信息流行度预测方法。

为了验证提出方法的有效性和可行性，本文基于真实社交网络数据集对模型实施验证实验。实验表明，本文提出的宏微观信息流行度预测模型，能够有效的感知信息流行度传播态势，为宏观的态势分析以及微观用户行为分析提供理论依据。

关键词：社交网络，信息传播，流行度预测，混沌时间序列，用户群体行为

Abstract

With the popularization of internet technology, online social network has become an important channel and carrier for people to exchange information in today's society. Information popularity prediction plays a vital role in online social networks, and there is extremely important research and application value. However, due to the diversity of information dissemination forms, the complexity of network structure and the multidimensional characteristics of user features, it is complicated and difficult to accurately control the information popularity. How to thoroughly analyze the changing regularities the popularity trend, how to perceive the trend of the topic popularity situation in advance, and how to establish an efficient public opinion control measures are the problems needed to be solved.

The thesis mainly covers two aspects about information popularity prediction: macro information popularity prediction deeply analyzes the nonlinear dynamic mechanism for information popularity transmission and constructs a cross-platform information popularity fusion and prediction model. Micro information retweeting popularity prediction analyzes the micro factors affecting user retweeting and predicts the future retweeting information based on epidemic model. The detailed works of this thesis can be summed up as follows:

1. At the macro level, the chaotic characteristics of social information dissemination are explored in depth, and the cross-platform information popularity fusion prediction model based on Bayesian estimation is proposed. Firstly, the information popularity time series from different social platforms are defined, and the principal components affecting popularity are quantified and obtained by Principal Component Analysis(PCA). Secondly, the chaotic characteristics of the popularity trend are explored, and the quantized sequences are reconstructed in phase space based on chaos theory. The changing regularities and properties of complex systems are restored in high-dimensional phase space. At the same time, the novel and fused phase space is obtained by using the Bayesian estimation to optimally fused multi-variable phase points in the same high-dimensional space. Finally, considering that the neural network has strong ability to approximate the nonlinear function, the neural network algorithm is applied to optimize and predict the fused information popularity.

2. At the micro level, the multi-dimensional attributes affecting user retweeting are explored, and the infection rate in improved Susceptible-Infected-Recovered(SIR) model is quantized. A user behavior evolution strategy is proposed to perceive information popularity. Firstly, the retweeting driving force of user's personal and social dimensions is extracted. Further, the retweeting driving force is quantized by multiple linear regression. Secondly, in order to make the model closer to the real network communication architecture, the state S in traditional SIR is improved, and the propagation rules of SIR model are redefined. Finally, the state values of SIR model are extracted by time slicing technology, and Least Square(LS) method and quantified infection rate were used to fit the real model. A prediction method of information popularity based on user retweeting behavior and improved SIR model is obtained.

In order to verify the validity and feasibility of the proposed method, the thesis validates the model based on real social network data set. Experiments show that the macro and micro information popularity prediction model proposed in the thesis can effectively perceive the information popularity transmission situation, and provide theoretical basis for macro situation analysis and micro user behavior analysis.

Keywords: social networks, information dissemination, popularity prediction, chaotic time series, user group behavior

目录

图录

表录

第 1 章 绪论

1.1 研究背景与意义

社交网络是以人类社交为核心的网络服务结构。具体说来，社交网络服务是基于六度分隔理论[1]，以用户之间共同的兴趣、爱好、活动或者真实的好友关系为基础，以实名或者非实名的方式在网络平台上构建的一种社交关系网络结构。在这样的传播结构下，信息的传播方式表现出速度快、覆盖范围广和影响力深等特点，给人们的生活方式带来了巨大的变革。同时，也推动着以 Sina、Twitter、Facebook 为典型代表的社交平台的用户使用量以及信息存储量呈现出爆炸式的增长模式。在 2018 年 8 月 20 日中国互联网络信息中心公布的第 42 次《中国互联网络发展状况统计报告》[2]指出，截至 2018 年 6 月，我国互联网用户已经达到 8.02 亿，超过我国总人口的一半，互联网普及率高达 55.8%。2018 年上半年新增网民 2968 万，相对于 2017 年末增长 3.8%，其中，手机网民规模已达到 7.88 亿，占总体用户规模的 98.3%。具体地，中国网民和互联网使用率走势如图 1.1 所示：

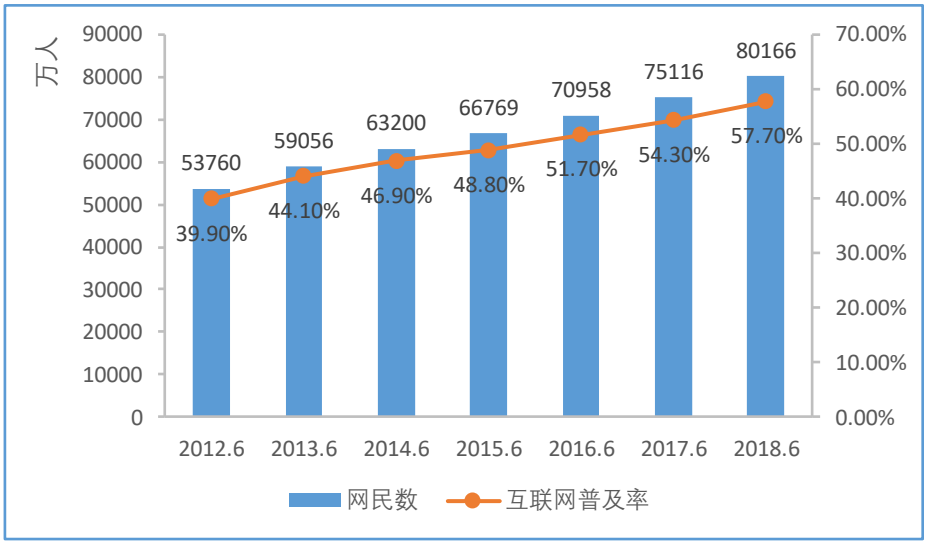


图 1.1 中国网民量以及互联网使用率统计图

总之，随着互联网技术的蓬勃发展，社交网络以空前的发展速度和规模不断壮大，给社会发展带来变革性的影响[3, 4]。一方面，社交网络逐渐成为人们获取信息、分享信息的重要社会媒体，深刻影响并改变了信息的传播方式。另一方面，

人与信息之间的互联高度融合，人人参与到信息的产生与传播过程，使信息呈现爆炸式的增长趋势，这种传播方式导致一条信息能够在短时间内传播到数百万计的用户，诸如消息过载、虚假信息泛滥等问题也随之而来。所以，开展针对社交网络的信息流行度预测[5, 6]，深入探究社交网络中信息流行度趋势的传播规律以及用户转发行为的特征，感知信息流行的变化趋势是我们关注的焦点。

就研究意义而言，本文从宏微观两个角度出发构建信息流行度预测模型。在宏观层面，本文全面分析影响信息流行度的跨平台因素，深入剖析话题信息流行度宏观态势的传播规律和特性，构建一种跨平台信息流行度融合和预测模型。在微观层面，针对推动信息传播的用户群体，深入分析和量化影响用户转发的多维属性，实现对未来时刻参与信息用户量的预测，感知信息未来流行度变化趋势。以此，为学者们深入分析和研究信息流行度的传播规律提供一定的帮助。

就现实价值而言，鉴于在线社交网络信息传播对人们的生活、社会发展的影响，在线社交网络信息传播分析引起了学术界和工业界的广泛关注。人们尝试捕获、理解以及预测在线社交网络中信息传播。而关于信息流行度预测的研究就有助于人们更深入的理解信息传播内在传播机制和规律，为舆情监控[7]、话题影响力评估[8]、网络营销[9]提供支持。在舆情监控上，预测信息热度态势变化有利于支持网络安全预警和辅助决策；在网络营销方面，准确估计信息的流行度可以帮助商家合理的进行商品推荐和广告投放；在话题影响力评估上，信息流行度预测能够实时分析热度趋势，应用于热搜话题排序。

1.2 研究现状

1.2.1 社交网络发展现状

在线社交网络是指由数以万计的网络用户自组织、自连接的集合，是真实物理世界到虚拟网络的一种映射，本质上是个体与个体直接的关系网络[10]。根据欧盟关于社会计算机的报告中，社交网络分为以下四类：(1) 即时应用网络，这是通过网络实现即时通信的社交平台，国外知名的平台有 MSN、AIM 等，国内有 QQ、微信等；(2) 在线社交关系平台，用户通过双向认证建立社交关系，进而将自己感兴趣的照片、文字信息分享给网络朋友，以国外 Facebook 和 Google，国内 QQ 空

间和人人为典型代表；(3) 微博应用平台，是一种支持用户单向关注建立信息传播网络的社交平台，用于分享、传播、发布简短实时话题信息，比较著名的社交平台包括国外的 Twitter，国内的新浪以及腾讯；(4) 可共享类应用，将信息共享成为可能，用户在相应的应用平台发布信息，其它用户通过登录平台账号来浏览已经发布的信息，以论坛、BBS、博客为典型代表。

从一定意义上看，在线社交网络的存在实质上源于人们对网络交流的需求[10]。早在 1838 年，美国科学家萨缪尔·摩尔斯发明了摩斯电码，通过电报的形式来发送信息，实现了社会群体的远距离通话。紧接着，电话的出现进一步加快了人与人之间的通信速度。随着科技的快速发展，迎来了全新的 Web2.0 互联网时代，它是主动性、社会学的典型应用代表，滋养了一大批社交平台，国外的 Facebook、Twitter 以及国内的 Sina 等都是 Web2.0 时代的应用产物。Facebook 自 2004 年成立，截止到目前已经有超过 20 亿的用户，成为全球最具影响力的社交平台，每月活跃用户达到 13 亿。2018 年 6 月为止，我国的微博用户量也达到 3.37 亿，相对于 2017 年末增加 2140 万，在整体网民中微博用户的比例达到 42.1%。毋庸置疑，在线社交网络已经成为连接物理社交世界和虚拟网络空间的桥梁。网络中信息和信息之间、用户和信息之间、用户和用户之间的各种交互行为，都促进了数据的传播和分享，成为人们的交流的重要途径。

从上面的分析看出，社交网络不论是在平台数还是平台用户量上都呈现明显的上升趋势，它当之无愧是最受欢迎的通信方式。这是因为社交网络与传统信息媒体比较具有以下一些特点：(1) 迅速性，用户可以不受时间地点的限制随时随地的分享、发布、接收信息，而且规则简单，分享接收门槛低，给人们获取信息节省了更多的时间；(2) 蔓延性，用户通过主动关注建立关系网络，信息沿关注关系层级的传递。同时，信息的传播呈现出“裂变式”的扩散态势，为人们获得网络舆情提供了支持；(3) 平等性，与传统非对等社交媒体相比，在线社交网络中所有用户之间是平等的，人人都有权利发表自己对事物的观点和认识。一旦触及人们现实生活中感兴趣的话题，就有可能引发人们的激烈讨论，从而信息发布者或转发者成为该事件的意见领袖；(4) 自组织性，社交网络中的个体呈现出自媒体的形态，每个用户都是一个有独立意识的个体，能够很快形成新的社区。总之，社交网络是社会进化的必然产物，它的发展对人们的生活和工作产生了巨大的影响。

1.2.2 信息流行度预测概述

社交网络中信息流行度预测是信息传播预测的一个重要分支，具有重要的研究和应用价值[11]。目前，针对信息流行度预测的相关研究已经取得了一些成果，本文从宏观信息传播态势和微观影响因素分析两个方面出发，概述目前信息流行度预测的相关工作。

在微观层面，通过分析影响信息流行度的关键因素来建立预测模型，其中，影响因素包括发布前影响因素和发布后影响因素。针对发布前影响因素的信息流行度预测，主要考虑信息本身固有传播影响力，在信息发布之前预测信息的受欢迎程度，为控制信息合理发布提供支持。考虑的早期因素包括发布者影响力、文本内容等。Bandari 等人对新闻的内在多维属性进行分析建模，从新闻类别、文章语言特色、用户是否认证体等多个影响因素出发，通过回归算法结合新闻早期的多维属性，利用分类算法预测未来该本文内容的流行度，预测准确度达到 84%[12]。Agarwal 等人深入研究博客内容的内在属性，即内容的雄辩力和新颖性对信息流行度的影响，用链接数量来度量新颖性，内容长短度量雄辩力。Agarwal 等人的研究表明博客引用链接越少、内容越长越有机会在后期获取更高的流行度[13]。Tsakias 等人研究一些早期直觉属性对新闻流行热度的影响，主要包括影响力和文本内容。Tsakias 等人先是使用一个分类器预测新闻是否会被评论，然后对于那些预测会被评论的新闻，再使用另一个分类器去预测评论数的多少[14]。针对发布后影响因素的分析预测，主要集中于对信息发布后的参与用户的属性以及时序因素建立预测模型。Lerman 等人通过实验发现，在 Digg 帖子上信息后续的流行度多数都来自前面点赞者的粉丝[15]。Liu 等人发现微博信息也有类似的现象，提取转发用户自身特征以及转发用户和粉丝之间的交互特征，通过时间切片处理，用分类模型预测转发用户的粉丝会不会在下一时刻参与到话题中来，进而感知信息的传播态势[16]。He 等人通过评论内容挖掘信息深层次的社交影响力，通过评论时间挖掘信息热度的时间因素，提出基于二部图和正则化的排序模型，预测信息未来时刻的流行度[17]。陈等人用转发数衡量微博信息的流行度，提取用户转发兴趣、用户活跃度等多维转发因素，并将当前背景下的热点信息作为影响信息流行度的可考虑因素，利用分类模型解决用户是否转发的问题[18]。Xiao 等人结合博弈论提取外部和内部属性

量化 SIR 的感染率，用 SIR 构建信息传播的状态模式，通过状态 I 来感知信息传播的流行度变化趋势[19]。除此之外，也有人综合考虑发布前后影响因素，将流行度预测转换为转发数的预测，通过提取属性来训练分类模型实现流行度的预测。刘和贺等人针对用户转发、信息内容的情感以及用户的兴趣三个属性，构建动态特性的用户转发行为预测模型，用来预测话题信息未来时刻转发数[20]。

综上所述，基于微观角度的信息流行度预测取得了丰硕的成果，但由于流行度受多种因素影响，演化传播模式具有高度的动态特性[21]。从微观角度出发预测时，预测难点表现在用户行为[22]的多维性和复杂性、文本内容的复杂情感[23]、网站机制的差异性[11]等，这些因素交织在一起共同作用于流行度，很难将它们分开和量化。因此，如何从微观层面提取影响流行度的主要因素，如何量化这些因素，成为微观流行度预测研究的重点和难点。

在宏观层面，一般忽略参与传播用户的个体特性，通过组织社交网络流行度态势传播规则，从统计学和模型两个角度对信息流行度传播范围和传播态势进行预测。在统计学信息流行度研究方面，主要利用统计学方法探究信息流行度变化的一些规律性并实现信息流行度的预测。Leskovec 等人研究在线信息流行度随时间的增长和衰减趋势，通过一个 K-SC(K-Spectral Centroid)算法对流行度进行时间相似度聚类，将流行度的态势演化分成 6 类[24]。Cheng 等人又在 Leskovec 的基础上，进一步研究 Facebook 上信息流行度的演化规律，研究发现流行度趋势变化和时间跨度选择有关系，除以上 6 种流行度演化模式以外，流行度态势还存在复杂的重复和起伏现象[25]。Szabo 等人通过分析 Youtube 视频和 Digg 故事的信息流行度增长模式发现，早早期流行度和未来流行度之前存在很强的对数关系[26]。除此以外还有一些研究是基于泊松过程的统计预测，他们遵循两个定律：(1) 富者更富的现象，某网络信息获得新的流行度的概率正比于该信息已有的流行度；(2) 兴趣衰减特性，认为内容对用户的吸引力会随时间变化而变化的时间衰减效应。综合两者，定义信息流行度的速率函数，然后通过生存分析理论[27]实现参数估计进而实现流行度的预测。申等人[28]提出的 RPP(Reinforced Poisson Process)和 Gao 等人[29]提出的 ERPP(Extended Reinforced Poisson Process)就是通过构建不同速率函数的增强泊松过程流行度预测模型。在模型角度的信息流行度预测方面，往往深入探究流行度态势变化的规律和规则，构建信息流行度趋势传播的时序模型，根据前面

时刻流行热度值，预测未来时刻信息热度，常用模型有传染病模型、分类回归模型、混沌时间序列模型等。王等人考虑外部非直接关注用户带来的转发量，基于传统 SIS(Susceptible-Infected-Susceptible)模型，新加入外部访问状态 E(External)，构建传染病 SISE 传播模型，通过提取各个时序下相应状态的值，训练模型求感染率，实现对转发数的预测，感知信息流行度态势变化[30]。Feng 等人同样也是基于传统 SIS 模型加入状态 E。不同的是，Feng 等人认为不同人群带来的感染率不同，用 KNN(K-Nearest Neighbor)算法对人群进行分类，定义多人群感染率[31]。Tatar 等人用信息发布后一段时间后获得的评论数作为因子，提出一个简单有效的线性回归模型，预测信息后期的流行度[32]。Cheng 等人使用时间序列中常用的自回归平均模型，感知网站上帖子流行度相对于时间的变化趋势[33]。孙等人通过遗传算法优化 BP(Back Propagation)神经网络的初始值和阈值，构建基于 BP 神经网络和遗传算法的网络舆情危机预警模型[34]。除此以外，也有一些研究学者发现网络舆情的走向具有混沌特性，将混沌理论[35]引入到信息流行度预测的研究课题中来。魏等人用百度搜索指数定义网络话题舆情传播趋势的时间序列，发现信息流行度的传播具有混沌特性，利用混沌理论求重构参数，实现对话题的流行度的预测[36]。黄等人统计话题下各个时刻相关的帖子数对应的信息热度，考虑到信息热度走势的混沌特性，将混沌理论结合支持向量回归机用于信息热度的建模[37] 。

同样地，在宏观层面的信息流行度预测也取得不错的成果，但是仍然存在一些挑战。主要表现在信息流行度存在平台差异性以及多因素交互影响，导致流行度传播趋势在一维空间中表现出复杂性和非线性等特征[36]，使信息流行度难以预测。因此，量化跨平台信息流行度及探究流行态势传播特征具有重要意义，可为进一步深入研究信息流行度规律提供支持。

1.3 主要研究内容

本论文以国家重点基金科研项目为背景，以宏观信息传播态势规律及微观群体事件的用户转发因素两个研究内容为切入点，从宏微观两个角度对信息流行度进行探究和预测。具体地，本论文的总体研究思路如图 1.2 所示：

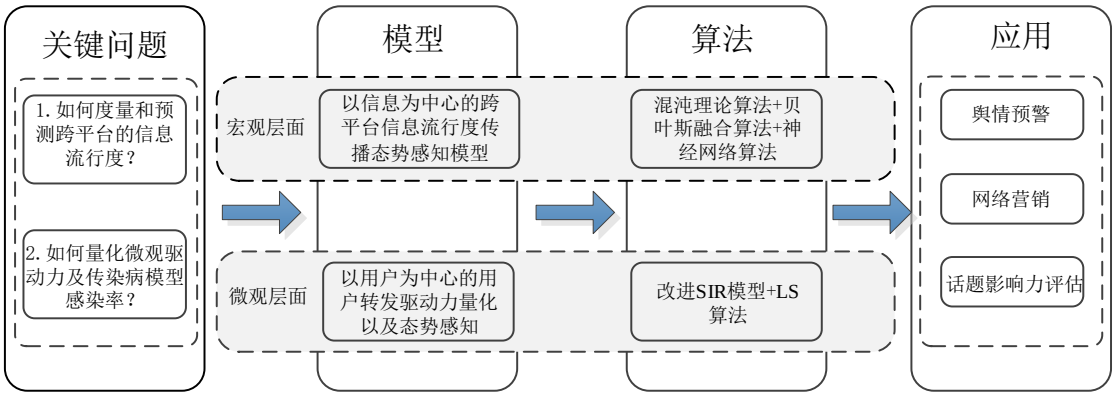


图 1.2 总体研究思路

本论文从待解决关键问题出发，在宏观层面，集中于跨平台信息流行度态势变化规律研究。在微观层面，探究影响用户转发的多维度因素，量化改进 SIR 模型中的感染率，提出一种感知信息流行度的用户行为演化策略。具体的创新点和贡献如下：

1. 针对信息传播态势变化的复杂特征以及流行度的平台差异性，提出了一种基于混沌理论和贝叶斯估计的跨平台多变量信息流行度融合预测模型。为了深入研究话题信息传播的规律，引入小数据量法发现社交信息流行度传播态势的混沌特征。与此同时，考虑到话题信息流行度态势变化存在平台差异性，首先利用 PCA 量化跨平台因素，然后基于混沌和贝叶斯估计理论，在同一重构高维相空间中实现多平台流行度序列的最优融合。在预测模型方面，考虑到神经网络在拟合非线性函数方面具有较强的能力，通过其预测融合后的流行度。

2. 针对用户行为的多样性和复杂性，提出了一种基于用户行为和改进 SIR 模型的微观信息流行度预测方法。首先，基于传统 SIR 和网络真实传播机制，重构社交网络信息传播规则。其次，从微观用户转发行为出发，综合考虑影响用户转发的多维度、多角度因素，以此作为用户状态群体量改变的理论依据，量化 SIR 模型的转发驱动力以及模型感染率。最后，利用时间切片和 LS 训练量化后的模型参量，用于感知流行度态势变化。

1.4 论文组织结构

第 1 章，主要介绍论文的研究背景及现状，首先对社交网络发展现状和信息流行度预测进行概述，其次，介绍本文的主要研究内容、研究意义、组织架构。

第 2 章，主要介绍信息流行度预测的相关技术和基础理论，首先介绍影响信息流行度的主要因素，然后总结信息流行度预测常用方法和模型。

第 3 章，从宏观信息流行度预测角度出发，介绍本文提出的一种基于混沌理论的跨平台话题流行度预测模型的实施步骤，包括模型构建、学习算法以及仿真实验。

第 4 章，从影响用户转发的微观因素出发，介绍本文提出的基于用户转发行为的信息流行度预测模型的实现过程，包括模型的构建、学习算法以及仿真验证。

第 5 章，总结和未来展望。具体地，首先对本文的研究工作进行总结，进一步地，对工作中的不足以及未来可能的研究方向进行概述。

第 2 章 相关技术和基础理论概述

本论文主要的研究对象是社交网络中信息的流行度。具体地，社交网络可以看成由用户以及用户关注关系组成的网络点边结构，信息沿着这些关系进行层级传播[38]。而信息的传播存在各种差异性，包括平台差异性、信息内容差异性、用户节点传播力差异性等，让信息流行度演化方式和预测成为一项有挑战性的工作和研究的热点。目前关于信息流行度预测的相关技术也较为成熟且应用广泛，本章将结合相关研究，以影响因素及常用模型为主线，介绍信息流行度预测研究的相关理论和技术方法。

2.1 信息流行度影响因素

研究指出信息的流行度呈现幂律分布，即仅有少部分信息能够被大量用户转发，而大部分信息只有很少的人会关注，不会变得流行。因此，研究信息流行度的相关因素尤为重要，就目前的研究看来，可以将其概括为两大类，即内部因素和外部因素。

2.1.1 内部因素

内部因素指信息固有的流行度传播因素，总结为信息内容内在影响力和发布用户内在影响力。针对信息内容影响力，研究学者认为信息内容是影响信息流行度的一个决定性的因素，包括信息是否与热点内容相关，是否@其它用户，是否包含标签、视频、图片等附加信息等。对于发布用户影响力也是内部影响因素研究的一个重点工作，多数学者集中于发布信息用户的注册年限、粉丝数、身份、活跃度等内在属性的研究。

2.1.2 外部因素

外部因素指信息发布后影响信息热度的一些因素，主要包括社交影响力和网站机制。社交影响力通过信息传播的网络结构表现出来，具体表现在参与用户之间的社交活动以及用户的从众心理特性。目前关于社交影响力的研究和量化很多，

主要集中在邻居节点转发率、内容相似度等。网站机制是指不同社交网站设置的诸如点赞、热点信息推送置顶的网站特色功能。网站机制也是影响信息流行的一个重要因素，例如：在社交平台上，高热度话题会被推荐至热点话题榜单，从而提高了信息的可见度以及关注度，这是“富者更富”现象的体现。

2.2 流行度预测常用分类和回归模型

2.2.1 决策树 C4.5 模型

C4.5 于 1993 年被 Ross Quinlan 提出，是对 ID3 算法的一个改进，常用于分类决策[39]。考虑到 ID3 算法在用信息增益选取特征时，容易偏向选择值较多的特征，C4.5 改用增益比选择特征。假设，用 A 标记特征， D 标记训练数据集，则信息增益比的计算公式可表示为：

$$g_r(D,A)=\frac{g(D,A)}{H_A(D)} \tag{2.1}$$

其中，信息增益 $g(D,A)$ 记为：

$$g(D,A)=H(D)-H(D/A) \tag{2.2}$$

式中， $H_A(D)$ ——数据集 D 关于特征 A 的熵

$H(D)$ ——数据集 D 经验熵

$H(D/A)$ ——数据集 D 关于特征 A 的经验条件熵

C4.5 决策树的构建过程包括以下三个模块：(1) 特征的选择；(2) 决策树的生成；(3) 决策树的剪枝。

2.2.2 支持向量机模型

支持向量机[40](Support Vector Machine, SVM)属于监督学习，一般用于解决二分类情况，它通过训练模型，求解能够将原始数据集正确区分的几何间隔最大的分离超平面。如图 2.1 是 SVM 解决二维特征空间中线性分类问题的图示，其中，● 表示正样本，▲表示负样本。当待训练的数据集线性可划分时，有无数条直线能使正负样本正确划分，如图 2.1 中虚线所示。然而，SVM 的解为将正负样本正确划分的直线中间隔最大的分离直线，其解具有唯一性，如图 2.1 中实线所示。

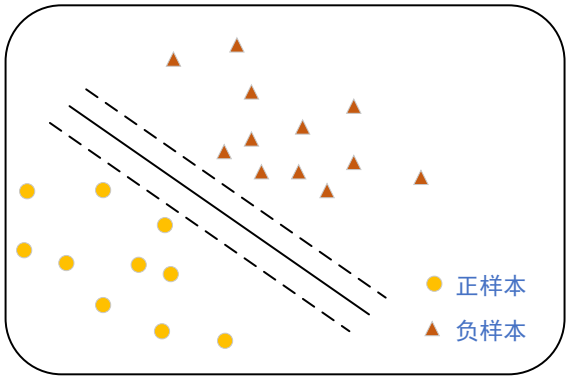


图 2.1 SVM 划分数据集图

关于 SVM 的模型由简至繁可以概况为三类：(1) 用于解决线性可分问题的硬间隔支持向量机。(2) 当大多数样本数据线性可划分时，通过加入松弛因子放宽条件的软间隔支持向量机。(3) 非线性支持向量机，其选择引用特定核函数将原始样本空间映射到可构造出最优分类决策超平面的高维特征空间，然后通过求解线性可分问题来求解原始非线性函数的解。相应地，常用核函数有如下几种形式：

(1) 线性核函数：$K(x,z)=x\cdot z$；

(2) 多项式核函数：$K(x,z)=(x\cdot z+1)^p$；

(3) 高斯核函数：$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$。

2.2.3 自回归移动平均模型

自回归移动平均模型[41] (Autoregressive Integrated Moving Average Model, ARIMA)，简记为 ARIMA(p, d, q)。ARIMA 是一种经典的线性时间序列预测模，关于三个参数 p， d， q 的相关定义如下所述：

p：称为自回归项数(Auto-Regressive, AR)，指待预测时序样本的滞后数(lags)。

d：是 Integrated 项，代表使待预测时序样本转化为平稳序列需要进行的差分阶数，即 d 阶差分后，数据才能达到稳定序列。

q：称为移动平均项(Moving Average, MA)，指待预测数据序列误差的滞后数。

当给三个参数设置不同数值时，可以得到 ARIMA 模型的几种变形，具体如下：

设置 $d=0$ 的情况，对应的 ARIMA$(p, 0, p)$ 模型可以表示为 ARMA(p, q)。

设置 $p=0$ 的情况，对应的 ARIMA$(0, d, p)$ 模型可以表示为 IMA(d, q)。

设置 $q=0$ 的情况，对应的 ARIMA$(p, d, 0)$ 模型可以记为 ARI(d, q)。

设置 $d=1$， $p=q=0$ 的情况，相应的 ARIMA$(0, 1, 0)$ 被叫做游走模型[42]，这种形式是 ARIMA 中最基本的模型，常使用于经济学的建模预测范畴。

2.2.4 多元线性回归模型

线性回归模型是统计学中一种较为常用的回归分析法，模型是一个或多个变量的线性组合函数，未知参数通常利用最小二乘法[43](Least Square, LS)求取。对一个变量建模的回归函数称为一元线性回归，大于一个变量的情况叫多元线性回归[44]。假设给定 d 个属性表述的特征空间，对给定数据集 $D=\{(\mathbf{x}_1, y_1),(\mathbf{x}_2, y_2),\ldots,(\mathbf{x}_m, y_m)\}$ 进行多元线性回归，将估计的向量函数表示为：

$$f(\mathbf{x})=\mathbf{w}^{\mathrm{T}}\mathbf{x}+b \tag{2.3}$$

其中：

$$\mathbf{w}=(w_1; w_1; \ldots w_d) \tag{2.4}$$

相应地，$\mathbf{w}$ 和 b 由 LS 估计得到。为了方便把 $\mathbf{w}$ 和 b 写成 $\hat{\mathbf{w}}=(\mathbf{w};b)$ 的向量形式，把输出向量记为 $\mathbf{y}=(y_1; y_2; \cdots y_m)$，将将给定的样本数据集合 D 用大小为 $m\times(d+1)$ 的矩阵 $\mathbf{X}$ 表示。其中，每一行代表一个数据样本，前 d 个元素表示样本的 d 个属性值，最后一个元素置 1，则有：

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{bmatrix} \tag{2.5}$$

基于 LS 让均方误差最小化有：

$$\hat{\mathbf{w}}^{*}=\arg\min_{\hat{\mathbf{w}}}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}}) \tag{2.6}$$

令 $E_{\hat{\mathbf{w}}}=(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})$，对 $\hat{\mathbf{w}}$ 求导得到：

$$\frac{\partial E_{\hat{\mathbf{w}}}}{\hat{\mathbf{w}}}=2\mathbf{X}(\mathbf{X}\hat{\mathbf{w}}-\mathbf{y}) \tag{2.7}$$

令上式为零即可得到 $\hat{\mathbf{w}}$ 的最优解，即为所求 w 和 b。

2.3 基于传染病的信息传播模型

在线社交网络中，以生物数学领域中的传染病模型为基础，构建新的传播规则和模型是信息流行度预测的一种重要手段。传染病模型的主要思想是：将人群中的个体按照其所处的状态进行分类，关注每类状态下个体数量比例的演化，处在每个状态的个体比例通过微分方程求解。目前，关于社交网络中传染病模型[45]有很多种，最经典的模型包括 SI[46]、SIS[47]、SIR[48]三种。

2.3.1 传染病 SI 模型

传统 SI(Susceptible-Infected)模型中，网络人群都处于易感染 S(Susceptible)和感染 I(Infected)两种状态。其中，易感染用户指的是有可能接触到信息并成为感染状态的用户，感染用户是收到信息并转发信息的用户群体。同时易感染用户 S 会以一个固定概率 λ 被感染群体感染转化为感染用户 I。具体地，SI 模型两个状态的转移情况如图 2.2 所示：

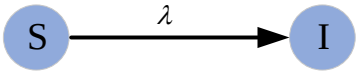


图 2.2 SI 状态转移图

在任意 t 时刻，感染率已固定为 λ，处于易感染和感染状态的用户量值分别为 $S(t)$ 和 $I(t)$，则 SI 模型的状态转换微分方程可以表示成如下形式：

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = -\lambda I(t) S(t) \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \lambda I(t) S(t) \end{cases} \tag{2.8}$$

2.3.2 传染病 SIS 模型

SIS(Susceptible-Infected-Susceptible)模型与 SI 模型有相同的状态类型，包括易感染 S(Susceptible)和感染 I(Infected)两类用户人群。但是，两者的传播和状态转换规则又有所不同。针对 SIS 模型，感染状态的用户可能又会转变成易感染人群并且有可能再次转发成为感染用户，其中，由感染状态再次变为易感染状态的速率

为 μ，感染率仍假设为 λ，其状态转移情况如图 2.3 所示：

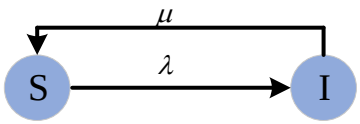


图 2.3 SIS 状态转移图

同样地，在任意 t 时刻，传播用户处于易感和感染态的用户群体比重分别为 $S(t)$、$I(t)$。当两种用户群体充分接触时，感染用户对应的增长速率为 $\lambda I(t)S(t)-\mu I(t)$，易感染用户对应的下降速率为 $\lambda I(t)S(t)+\mu I(t)$。则 SIS 模型的状态转换微分方程可以表示成如下形式：

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = -\lambda I(t)S(t) + \mu I(t) \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \lambda I(t)S(t) - \mu I(t) \end{cases} \tag{2.9}$$

2.3.3 传染病 SIR 模型

SIR(Susceptible-Infected-Recovered)模型中用户群体除了上述两种状态外，还有第三种状态，称为免疫状态 R(Recovered)。在该模型中，易感染用户 S 会以固定概率 λ 被感染群体感染转化为感染用户 I，而感染用户将会以固定概率 μ 恢复为免疫用户，并且对信息失去兴趣，不可能被再次感染。其状态转移图如图 2.4 所示：

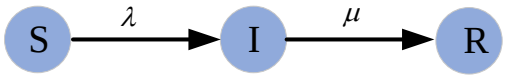


图 2.4 SIR 状态转移图

假设，在任意 t 时刻，传播用户处于易感、感染、恢复态的用户群体比重分别为 $S(t)$、$I(t)$、$R(t)$。当三种状态充分接触传播时，感染用户群体的增长率为 $\lambda I(t)S(t)-\mu I(t)$，易感用户的下降率为 $\lambda I(t)S(t)$，免疫用户的增长率为 $\mu I(t)$。则 SIR 模型的状态转换微分方程可以表示成如下形式：

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = -\lambda I(t) S(t) \\ \dfrac{\mathrm{d}I(t)}{dt} = \lambda I(t) S(t) - \mu I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \mu I(t) \end{cases} \tag{2.10}$$

2.4 基于时间特性的预测模型及方法

2.4.1 基于混沌时间序列流行度预测模型

1. 小数据量法的混沌特性判断分析

对一个时间序列应用混沌方法进行预测时，首先需要判断待预测系统是否具有混沌特性。同样地，在对网络舆情流行度建模时也需要该步骤，对于量化后的流行度序列通过小数据量法获取最大 Lyapunov 指数 λ。若 $\lambda > 0$，则证明舆情流行度序列具有混沌特性[49, 50]。具体方法归结如下：

(1) 对待判别流行度序列 $\{z(i), i = 1, 2, \cdots n\}$ 实施傅里叶变换(Fast Fourier Transformation, FFT)，计算系统轨道的平均周期；

(2) 求得延迟时间 τ 以及嵌入维数 m，并进行相空间重构得：$\{\mathrm{Z}_j, j = 1, 2, \cdots M\}$，$M = n - (m-1)\tau$；

(3) 对每一个相点 Z_j 的最近临近点 Z_j 进行短暂分离限制，进而求两者经过 i 个离散时间步长后的距离 $d_j(i)$，即：

$$d_j(i) = \left| Z_{j+i} - Z_{j+i} \right|, \qquad i = 1, 2, \cdots, \min(M - j, M - j) \tag{2.11}$$

(4) 考虑每一个离散步长 i，计算每个相点对应的 $\ln d_j(i)$ 的均值 $z(i)$：

$$z(i) = \frac{1}{q\Delta t} \sum_{j=1}^{q} \ln d_j(i) \tag{2.12}$$

式中，q ——非零 $d_j(i)$ 数量

对于求得的 $z(i)$ 用 LS 做回归直线，所求直线的斜率即为流行度时间序列的 λ。

2. 相空间重构理论

Takens 定理证明，对于一个给定的时间序列$\{z(i), i=1,2,\cdots,n\}$，选取适当的嵌入维数m和延迟时间τ，能够重构一个与其原动力系统在拓扑意义下等价的相空间，从而把握混沌时间序列的性质与规律[51, 52]。因此，针对舆情流行度的时间趋势预测，同样需要利用相空间重构理论将流行度时间序列从一维空间映射到高维相空间，恢复信息流行度传播的规律和特征，提高流行度预测的精确度。假设求得的流行度态势变化的延迟时间为τ[53, 54]和嵌入维数为m[55, 56]，则重构的相空间可以表示为：

$$\mathbf{Z}=[Z_1,Z_2,\cdots Z_M]^{\mathrm{T}}=\begin{bmatrix} z_1 & z_{1+\tau} & \cdots & z_{1+(m-1)\tau} \\ z_2 & z_{2+\tau} & \cdots & z_{2+(m-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ z_M & z_{M+\tau} & \cdots & z_{M+(m-1)\tau} \end{bmatrix} \tag{2.13}$$

式中，$M=(n-(m-1)\tau)$。

2.4.2 基于增强泊松过程的信息流行度预测方法

在线社交网络信息流行度预测中，关于增强泊松过程的流行度预测主要关注三个影响信息流行度的因素：(1) 适者生存，也就说信息本身固有吸引力对未来信息热度起至关重要的作用；(2) 富者更富现象，即若当前获得的流行度值越多，那么以后就会获得更多关注；(3) 兴趣衰减特性，认为内容对用户的吸引力会随着时间逐渐削弱[57]。根据上述信息流行度随时间演化的三个特性，构建基于增强泊松过程的信息流行度预测模型。关于增强泊松过程的建模过程如下：

首先，定义速率函数。对于一条给定的信息m，其在时刻t的速率函数为：

$$x_m(t)=\lambda_m f_m(t) i_m(t) \tag{2.14}$$

式中，λ_m ——内容m的固有吸引力

$f_m(t)$——关注度衰减函数

$i_m(t)$ ——累计关注度

其次，假设第$k-1$次转发发生在时刻t_{k-1}，则第k次转发会发生在时间t_k的概率为：

$$p(t_k/t_{k-1}) = e^{-\int_{t_{k-1}}^{t_k} x_m(t)dt} x(t_k) \tag{2.15}$$

相应地，t_n 至 T 时间段内无转发行为的概率为：

$$p(T/t_n) = e^{-\int_{T_n}^{T} x_m(t)dt} \tag{2.16}$$

在增强泊松过程中，假设各个时刻的转发事件是相互独立的统计变量，则发生这一系列转发的概率为：

$$\ell = p(T/t_n)\prod_{i=1}^{n} p(t_k/t_{k-1}) \tag{2.17}$$

最后，通过最大似然估计求上述方程式的未知解，获取流行度预测模型。

2.5 本章小结

本章主要介绍信息流行度预测的相关技术以及基础理论。首先，介绍影响信息流行度的相关因素。其次，介绍信息流行度预测的相关方法，包括：流行度预测常用的分类和回归模型、传染病模型以及时间特性模型。关于时间特性的方法重点讲解基于混沌特性的流行度预测模型以及基于增强泊松过程的流行度预测模型。本章为后续流行度预测研究奠定了理论基础。

第 3 章 基于混沌理论的跨平台信息流行度预测模型

3.1 引言

随着在线社交应用和媒体的迅速扩散，在线社交网络已将我们的日常生活与网络信息空间连接起来。而关于信息传播预测是一个具有很大科学价值和应用价值的研究课题，对应的主要研究任务包括信息流行度预测、用户传播行为预测和信息传播路径预测，其中，信息流行度预测起着至关重要的作用。

目前，针对信息流行度预测的研究取得了丰硕的果实，但仍然存在一些挑战和突破点：

1. 社交话题信息热度度量指标多来自单一社交平台，不能全面反映信息热度。信息热度具有平台差异性，不同的社交平台上热度变化不尽相同，在定义信息流行度时，多数研究将话题热度定义为某一社交平台上信息的某种数量。单一社交平台上流行度的变化趋势大体上能够反映信息传播态势变化，但是不能细致和全面地刻画和研究其变化规律。

2. 社交信息传播态势变化受多种因素交互影响，在一维空间中表现出复杂性和非线性等特征。当线性模型用于信息流行度进行预测，无法捕捉信息时间传播态势的非线性特征。而针对非线性的预测模型，往往忽略了信息流行度态势变化的一些内在特征，从而影响信息流行度的感知能力。

3. 基于混沌理论的单变量预测难以保证完备的重构原系统。在单一社交平台上，信息流行度被多种特征交互影响，而在不同社交平台上，信息流行度态势则存在平台差异性。因此，一种可以融合多个变量的混沌预测显的尤为重要。

本文围绕以上三个信息流行度预测的挑战和突破，提出一种跨平台的多变量混沌时间序列社交信息流行度预测模型。并基于真实数据实施实验验证，实验表明，本预测方法不仅能够融合多个社交平台的信息流行度，并且能在一定程度上提高流行度的预测精确度，以更细粒度视角感知热点话题发展态势。

3.2 问题形式化及相关定义

3.2.1 问题定义

本文通过一个时间切片技术序列化来自多个社交网络平台的同一话题信息，假设有 p 个影响信息流行度的网络平台，则可以将其列化为：$X_1, X_2, \cdots, X_p$。具体地，第 i 个社交平台的信息流行度时间序列 X_i 定义如下：

$num\left[mes\left(t_j\right)\right]$ 表示话题开始时刻到 t_j 时刻该社交平台上发表的与话题相关的信息数量。类似地，$num\left[mes\left(t_{j-1}\right)\right]$ 表示话题开始到 t_{j-1} 时刻与话题相关的信息数，则定义话题在第 j 个时刻来自该平台的话题热度为：

$$x_{ij} = num\left[mes\left(t_j\right)\right] - num\left[mes\left(t_{j-1}\right)\right] \tag{3.1}$$

那么，从话题开始起，每隔恒定值 $t_j - t_{j-1}$ 时间间隔采样一点，得到来自第 i 个社交平台的话题流行度时间序列 $X_i = \{x_{i1}, x_{i2}, \cdots x_{in}\}$，其中，$n$ 为采样点数。

3.2.2 问题形式化

为了形式化地描述本章的研究工作，将问题概述展示如图 3.1，问题具体的输入输出如下：

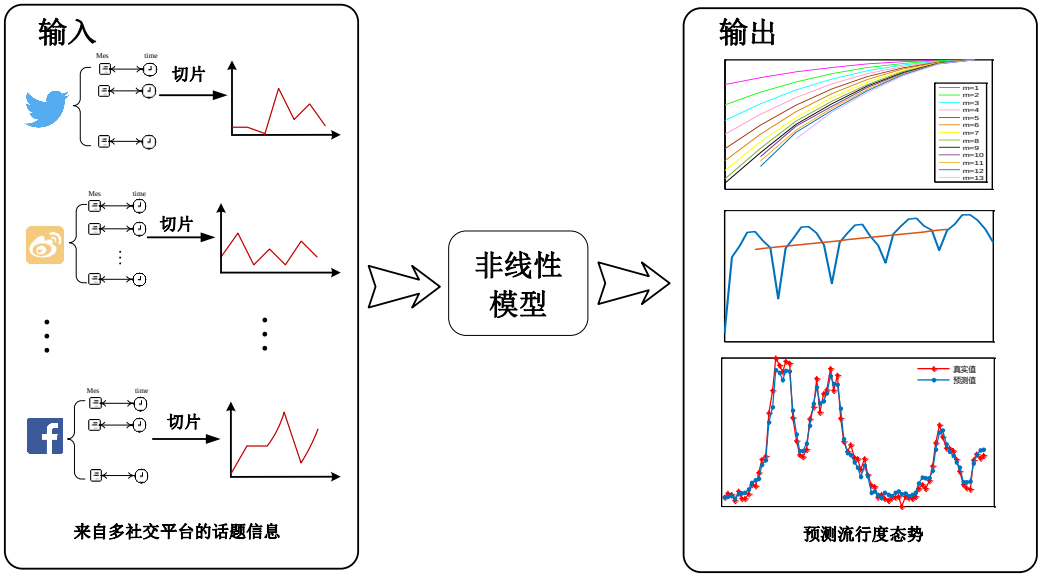


图 3.1　问题概述图

基于上述定义，将问题的输入描述如下：

从多个社交网络平台抓取信息和时间的原始数据，经过一个时间切片处理，原始数据被时间序列化为 X_1, X_2, …, X_p，并将其作为模型的输入。

根据问题输入，有以下输出问题需要解决：

1. 如何量化多社交平台的话题信息流行度时间序列？

本章中，通过 PCA 来处理这一问题，根据公式 $A_i = b_{i1}X_1 + b_{i2}X_2 + \cdots + b_{ip}X_p$ 提取影响话题流行度的 r 个主成分，进而实现量化处理。

2. 如何对量化后的主成分进行相空间的融合？

利用贝叶斯估计理论将影响流行度的主成分在高维空间进行最优融合，依据公式 $Z_k = \hat{z}_k = \dfrac{\sum_{h=1}^{r} \dfrac{\mathbf{a}_h}{\sigma_h^2} + \dfrac{z_0}{\sigma_0^2}}{\sum_{h=1}^{r} \dfrac{1}{\sigma_h^2} + \dfrac{1}{\sigma_0^2}}$，估计融合后的流行度时间序列： Z 。

3. 如何对新的融合后的序列 Z 进行流行度预测？

用神经网络对重构后的序列 Z 进行流行度预测，通过训练历史数据找到输入和输出之间的非线性映射函数 f，则预测未来时刻话题流行度为：$Z_{n+1} = f(Z)_{n,\cdots,1}$。

3.3 模型

为了实现本章提出的信息流行度预测方法，将具体的模型细化成三个模块：流行度量化模块、混沌相空间重构模块、神经网络流行度预测模块，如图 3.2 所示。其中，第一个模块用来量化多平台的信息流行度时间序列，提取影响信息热度的主要成分；第二个模块对量化后的主成分时间序列进行理论分析以及相空间重构；第三个模块是流行度预测模块，基于前面时刻流行度热度，利用径向基函数(Radial Basis Function, RBF)神经网络预测未来时刻流行度热度。

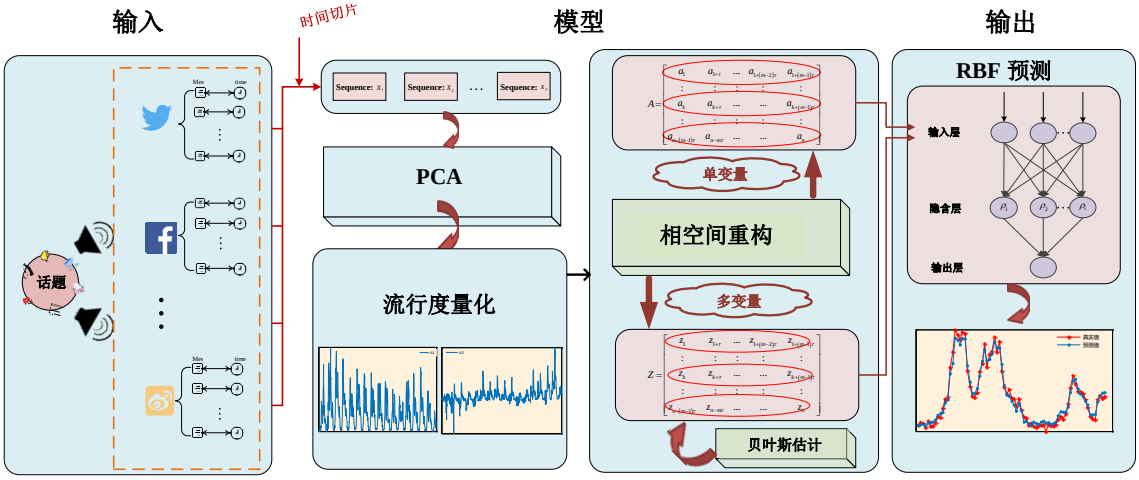


图 3.2　整体框图

最后，为了方便读者阅读和查阅，归纳文章中常用符号，详细归纳如表 3.1 展示：

表 3.1　符号及描述

符号	描述	符号	描述
X_i	任一平台流行度序列	$\mathbf{X}$	多平台流行度矩阵
A_i	任一主成分	$\mathbf{A}$	主成分矩阵
m,τ,λ	相空间重构参数	$\mathbf{Q_i}$	A_i 的重构相空间
Z	融合流行度序列	$\mathbf{Z}$	Z 的重构相空间
p	社交平台数	n	采样点数
r	主成分数	l	预测样本数

3.3.1 PCA 量化流行度

PCA 的目的是将多平台的流行度时间序列量化为少数几个综合指标，发现影响社交话题流行度的主要成分，降低多变量融合预测的计算复杂度。对话题流行度主成分分析的具体步骤如下：

(1)　p 个平台的流行度时间序列可以写成矩阵形式：

$$\mathbf{X}=\left[X_1,X_2,\cdots,X_p\right]=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{3.2}$$

(2) 将 **X** 标准化。$E(X_k)=\mu_k$ 及 $\mathrm{Var}(X_k)=\sigma_k^2$ 代表变量 X_k 的均值和方差，则标准后的变量为：

$$\mathbf{X}=\left[X_1,X_2,\cdots X_p\right]=\left[x_{ij}\right]_{n\times p},\quad X_k=\frac{X_k-\mu_k}{\sqrt{\sigma_k^2}} \tag{3.3}$$

(3) 按照公式(3.4)计算 **X** 的相关矩阵 **R**，即：

$$\mathbf{R}=\frac{1}{n}\mathbf{X}^T\mathbf{X} \tag{3.4}$$

(4) 求 **R** 对应的特征值和特征向量，确定影响信息流行度的主要成分。通常选取累计贡献率在 85%~95%之间的特征根 θ_1，θ_2，…，θ_r，来对应第 1、第 2、第 r 个主成分。将任意一个特征值对应的特征向量记为 $\mathbf{B}_i=\left[b_{i1},b_{i1},\cdots b_{ip}\right]^{\mathrm{T}}$，那么第 i 个主成分可以表示如下：

$$A_i=b_{i1}X_1+b_{i2}X_2+\cdots+b_{ip}X_p \tag{3.5}$$

式中，A_1，A_2，…，A_r ——影响信息流行度的主成分，用于跨平台因素的多变量的融合。

3.3.2 流行度时间序列的相空间重构

1. 多变量相空间重构

首先，将上述影响流行度主成分的多变量 A_1，A_2，…，A_r 写成如下形式：

$$\mathbf{A}=\left[A_1,A_2,\cdots A_r\right]^{\mathrm{T}}=\begin{bmatrix}a_{1,1} & a_{1,2} & \cdots & a_{1,n}\\ a_{2,1} & a_{2,2} & \cdots & a_{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{r,1} & a_{r,2} & \cdots & a_{r,n}\end{bmatrix} \tag{3.6}$$

其次，考虑到 C-C 和 G-P 方法在求系统重构参数的实际应用中取得很好的效果，且计算简单。本文利用 C-C 计算流行度序列的延迟时间，G-P 计算嵌入维数。假设 r 个主成分对应的延迟时间和嵌入维数分别为 τ_1，τ_2，…，τ_r 和 m_1，m_2，…，m_r。为了保证各个变量重构后都能完全展开，选择 r 个参数中最小延迟时间和最大嵌入维数作为多变量流行度时间序列相空间融合的重构参数[58]。即：

$$m=\max(m_i)\qquad \tau=\min(\tau_i) \tag{3.7}$$

式中，$i=1,2,\cdots r$ 。

进一步地，分别对 r 个影响流行度的主成分实施相空间重构，任意一个主成分的重构相空间为：

$$\mathbf{Q}_i=\begin{bmatrix} a_{i,1} & a_{i,1+\tau} & \cdots & a_{i,1+(m-1)\tau} \\ a_{i,2} & a_{i,2+\tau} & \cdots & a_{i,2+(m-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i,M} & a_{i,M+\tau} & \cdots & a_{i,M+(m-1)\tau} \end{bmatrix} \tag{3.8}$$

2. 基于贝叶斯估计的流行度相点融合

理论上根据嵌入延时定理单变量流行度时间序列就能恢复动态系统的特性。但是考虑到社交话题流行度的变化趋势系统受多平台因素的影响，单变量不足以反映流行度趋势变化的实际情况，而多变量建模能够融合多个社交平台流行趋势信息，且包含流行度态势系统更多的动态特性。所以利用贝叶斯估计理论融合影响流行度的多个变量，弥补单变量预测不足的问题。

上述步骤 1 已经对 r 个影响流行度主成分的相空间进行重构。基于此，将第 i 个主成分的任意一个相点表示如下：

$$\mathbf{a}_i=(a_{i,k},a_{i,k+\tau},\cdots a_{i,k+(m-1)\tau}) \qquad (k=1,2,\cdots,M) \tag{3.9}$$

则可以把融合后的相点集合记为：

$$D_k=[\mathbf{a_1};\mathbf{a}_2;\cdots;\mathbf{a}_r] \tag{3.10}$$

用 z_k 表示融合后的相点，则依据贝叶斯估计理论有：

$$p(z_k/\mathbf{a_1},\mathbf{a}_2,\cdots,\mathbf{a}_r)=\frac{p(z_k;\mathbf{a_1},\mathbf{a}_2,\cdots,\mathbf{a}_r)}{p(\mathbf{a_1},\mathbf{a}_2,\cdots,\mathbf{a}_r)} \tag{3.11}$$

假设 z_k 和 D_k 分布服从参数为 $N(z_0,\sigma_0^2)$ 和 $N(z_k,\sigma_h^2)$ 的正态分布，假设 $\alpha=\dfrac{1}{p(\mathbf{a_1},\mathbf{a}_2,\cdots,\mathbf{a}_r)}$，则：

$$\begin{aligned} p(z_k/\mathbf{a_1},\mathbf{a}_2,\cdots,\mathbf{a}_r)&=\alpha\prod_{h=1}^{r}\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{h}}}\exp\left[-\frac{1}{2}(\frac{\mathbf{a}_h-z_k}{\sigma_h})^2\right]\times\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}(\frac{z_k-z_0}{\sigma_0})^2\right] \\ &=\alpha\prod_{h=1}^{r}\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{h}}}\times\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\sum_{h=1}^{r}(\frac{\mathbf{a}_h-z_k}{\sigma_h})^2-\frac{1}{2}(\frac{z_k-z_0}{\sigma_0})^2\right] \\ &=\alpha"\exp\left\{\left[\left(\sum_{h=1}^{r}\frac{1}{\sigma_h^2}+\frac{1}{\sigma_0^2}\right)z_k^2-2\left(\sum_{h=1}^{r}\frac{\mathbf{a}_h}{\sigma_h^2}+\frac{z_0}{\sigma_0^{\,2}}\right)z_k\right]\right\} \end{aligned} \tag{3.12}$$

其中：

$$\alpha'' = \alpha \prod_{h=1}^{r} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{h}}} \times \frac{1}{\sqrt{2\pi_0}\sigma} \exp\left[-\frac{1}{2}\left(\sum_{h=1}^{r} \frac{\mathbf{a}_h^2}{\sigma_h^2} + \frac{z_0^2}{\sigma_0^2} \right) \right] \tag{3.13}$$

因为公式(3.12)的指数部分是关于 z_k 的二次函数，所以 $p(z_k/\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_r)$ 仍然服从正态分布 $N(z,\sigma^2)$，则：

$$p(z_k/\mathbf{a}_1,\mathbf{a}_2...,\mathbf{a}_r) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}(\frac{z_k - z}{\sigma})^2 \right] \tag{3.14}$$

根据式(3.12)和式(3.14)可得：

$$\begin{cases} \dfrac{1}{\sigma^2} = \displaystyle\sum_{h=1}^{r} \frac{1}{\sigma_h{}^2} + \frac{1}{\sigma_0{}^2} \\ \dfrac{z}{\sigma^2} = \displaystyle\sum_{h=1}^{r} \frac{\mathbf{a}_h}{\sigma_h{}^2} + \frac{z_0}{\sigma_0{}^2} \end{cases} \tag{3.15}$$

求解式(3.15)有：

$$z = \frac{\displaystyle\sum_{h=1}^{r} \frac{\mathbf{a}_h}{\sigma_h^2} + \frac{z_0}{\sigma_0^2}}{\displaystyle\sum_{h=1}^{r} \frac{1}{\sigma_h^2} + \frac{1}{\sigma_0^2}} \tag{3.16}$$

则通过贝叶斯估计获得的最优融合相点 z_k 的估计值 $\hat{z}_k$ 如下：

$$\hat{z}_k = \int_{\Omega} z_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}(\frac{z_k - z}{\sigma})^2 \right] dz_k = z \tag{3.17}$$

则新的相空间点 $Z_k = \hat{z}_k$，进一步地，新的 m 维重构相空间表示为：

$$\mathbf{Z} = \left[Z_1, Z_2, ..., Z_M \right]^T \tag{3.18}$$

其中：

$$Z_k = \left[z_k, z_{k+\tau}, ..., z_{k+(m-1)\tau} \right],\ (k = 1,2,...,n-(m-1)\tau) \tag{3.19}$$

式中，k ——新融合变量的任一相点坐标

3.3.3 流行度预测模型

考虑到神经网络具有很强的逼近非线性函数的能力，可用其对流行度进行预测。同时，介于 RBF 具有网络结构简单、逼近能力强和学习速度快等优点，把其应用到本文预测中。RBF 结合混沌理论用于流行度预测时，其中输入为重构数据

的每个相点：$Z_n=\left[z_{n-(m-1)\tau},...,z_{n-1},z_n\right]$，输出为每个相点对应的下一时刻的值：$z_{n+1}$。相应地，RBF 流行度预测框图如图 3.3 所示：

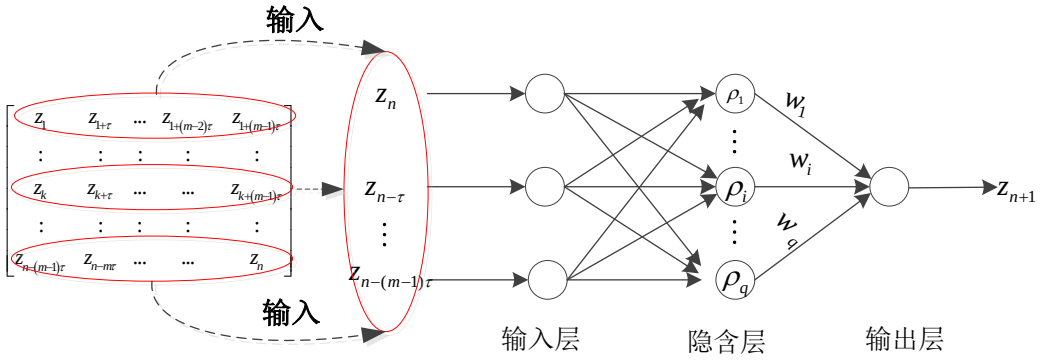


图 3.3 RBF 流行度预测模型

基于混沌理论的 RBF 流行度预测网络表示为：

$$z_{n+1}=\sum_{i=1}^{q}w_i\rho(z,z_i) \tag{3.20}$$

其中，$\rho(z,z_i)$ 是高斯径向基函数，形式如下：

$$\rho(z,z_i)=\exp(-\frac{1}{2\sigma_i^2}\left\|z-z_i\right\|^2) \tag{3.21}$$

式中，q ——隐含层神经元个数

z_i ——第 i 个隐藏神经元中心

w_i ——第 i 个隐藏神经元权重

σ_i ——第 i 个隐含层神经元宽度向量

在上述流行度预测模型中，z_i 利用 K-Means 聚类来确定，而 w_i 和 σ_i 通过 BP 神经网络训练获得。

3.3.4 模型算法设计及分析

遵循模型框图，模型学习算法的实现细节如表 3.2 所示：

表 3.2 模型算法表

Input: time series of each platform: $X_1, X_2, \cdots X_p$

Output: principal components: $A_1, A_2, \ldots, A_r$; fusion variable: Z; mapping function: $Z_{n+1} = f(Z)_{n,\cdots,1}$

Part 1: quantification popularity

//extract main components parameters: $A_1, A_2, \ldots, A_r$

using PCA to get the main components $A_1, A_2, \ldots, A_r$ from Eqs.(3.2)-(3.5)

Part 2: Phase space reconstruction and RBF prediction

//2.1 reconstruction phase space

for each variable **do**

compute τ_i by the method of C-C

compute m_i by the method of G-P

compute the largest Lyapunov exponent λ by the method of small-data

if $\lambda > 0$ continue **do**

reconstruction phase space for the parameters of $A_1, A_2, \ldots, A_r$ from Eqs.(3.6)-(3.8)

end for

//2.2: the theory of Bayesian estimation

obtain the optimal fusion phase space from Eqs.(3.9)-(3.19)

repeat 2.1 to obtain new multivariable reconstruction phase space: Z

//2.3: the popularity prediction

train the RBF model and obtain the nonlinear mapping function: $Z_{n+1} = f(Z)_{n,\cdots,1}$

test the mapping function and obtain the prediction value

分析模型学习算法复杂度，利用 PCA 对多平台信息流行度序列进行量化的时间复杂度为$O(N)$。对量化后对应的主成分实施重构相空间，确定其对应的嵌入维数和延迟时间所需要的复杂度为$O(r \times N)$。最后，对经过重构的流行度序列利用 RBF 进行预测所需要的时间复杂度为$O(N^2)$。通过以上分析，整个算法的复杂度为：$O(N)+O(r \times N)+O(N^2) \sim O(N^2)$。

3.4 仿真实验与结果讨论

3.4.1 实验数据

本文实验数据选自中国的社交网络媒体(Platform 1)、微信(Platform 2)、微博(Platform 3)、天涯论坛(Platform 4)作为获取社交信息的渠道平台。这些平台上拥有上亿的用户或新闻媒体，包含了丰富的数据信息，是促进话题信息传播和演化的

重要平台，使用这四个平台的数据来研究话题信息的一系列问题具有真实性、可靠性，表 3.3 为具体数据介绍。

Topic A：话题 A 研究 2016 中国召开的两会会议流行度趋势。以上述四个社交平台为种子平台，抓取从 2016.03.02-2016.03.22 时间段内话题 A 相关的信息。其中，四个平台相关的信息数量分别为 49981、50000、37443、34319。

Topic B：话题 B 指的是 2017 年的一件社会性羞辱话题事件。同样，从上述四个社交平台抓取从 2017.03.24-2017.04.05 时间段内话题 B 相关的信息。其中，四个平台相关的信息数量分别为 5168、9163、3289、3837。

表 3.3 时间间隔及话题信息表

数据集	时间间隔	平台 1	平台 2	平台 3	平台 4
Topic A	2016.03.02-2016.03.22	49981	50000	37443	34319
Topic B	2017.03.24-2017.04.05	5168	9163	3289	3837

3.4.2 基础方法

为了评估本文提出的社交话题流行度预测的性能，将以下经典方法与本文模型进行对比：

C-RBFNN 模型[16]：C-RBFNN(Cloud-RBFNN)是一种改进的 RBF 模型，考虑到信息传播过程中的不确定性，引入模糊数学中的云理论对 RBF 中的隐含层的激活函数进行优化，把其应用到热点话题转发数的预测，目的是为了提高对非线性类问题的预测性能。

LWLP 模型[59]：LWLP(Local Weighted Linear Prediction)全称为局域加权线性预测模型，是一种经典的混沌时间序列线性预测模型。文献[59]将该模型应用于 Lorenz 系统预测中，并取得不错的预测效果。

LS-SVM 模型[60]：LS-SVM(Least Squares-SVM)是一种改进的 SVM 模型，模型将标准 SVM 中的不等式约束改成等式约束，并将经验风险由误差的一范数改为二范数，由此将求解二次优化问题转化成解一次线性方程组，将其应用于混沌系统的预测，加快学习算法的收敛速度。

3.4.3 评估指标

为了全面衡量模型的预测性能，本文采用绝对误差(Absolute Error, ERR)、平均绝对误差(Mean Absolute Error, MAE)、平均绝对百分误差(Mean Absolute Percentage Error, MAPE)、均等系数(Equality Coefficient, EC)作为衡量预测性能的技术指标。具体定义如下：

$$\mathrm{ERR} = \hat{z}_i - z_i \tag{3.22}$$

$$\mathrm{MAE} = \frac{1}{l}\sum_{i=1}^{l}\left|z_i - \hat{z}_i\right| \tag{3.23}$$

$$\mathrm{MAPE} = \frac{1}{l}\sum_{i=1}^{l}\left|\frac{z_i - \hat{z}_i}{z_i}\right| \tag{3.24}$$

$$\mathrm{EC}=1-\frac{\sqrt{\sum_{i=1}^{l}\left(z_i - \hat{z}_i\right)^2}}{\sqrt{\sum_{i=1}^{l} z_i^{\,2}} + \sqrt{\sum_{i=1}^{l} \hat{z}_i^{\,2}}} \tag{3.25}$$

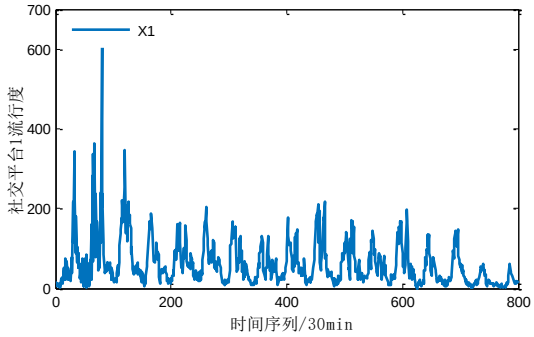
其中，z_i代表真实值，$\hat{z}_i$代表预测值，l为预测样本长度，ERR 是预测值和真实值的残差序列，MAE 用来衡量预测值偏离真实值的偏离程度，MAPE 代表预测值与真实值实际偏差绝对值占观测值百分比的均值，EC 用来衡量预测值和真实值之间拟合的准确度。

3.4.4 预测性能分析

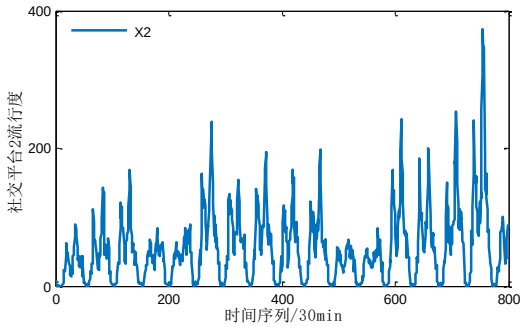
依据前面数据介绍，采用 Topic A 和 Topic B 两个话题来进行本文的实验验证，其中，具体介绍 Topic A 实施步骤，而 Topic B 通过表 3.7 来展示预测结果。

1. 各平台流行度时间序列采样

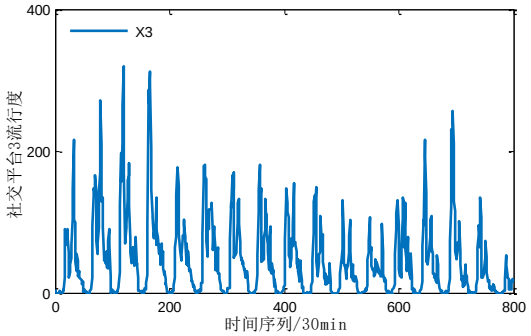
本文研究话题信息整个生命周期的流行度演化态势，从话题开始到话题消亡，每 30 分钟为一个时间切片进行采样，共采样 800 个时刻点。具体地，采样的上述四大社交平台的流行度时间序列如图 3.4 所示：



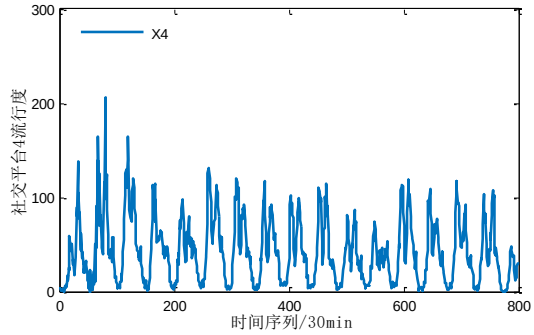
(a) Platform1 流行度时间序列



(b) Platform 2 流行度时间序列



(c) Platform 3 流行度时间序列



(d) Platform 4 流行度时间序列

图 3.4 多社交平台流行度时间序列

对比来自 4 个平台关于两会话题的流行度时间序列发现，单个社交平台数据能够大体反映流行度的态势变化，但是仍然存在很明显的平台差异，仅仅通过单平台的数据来衡量话题热度不够细致全面，在反映前后时刻话题流行度热度差距的时候不够精确。

2. PCA 量化流行度

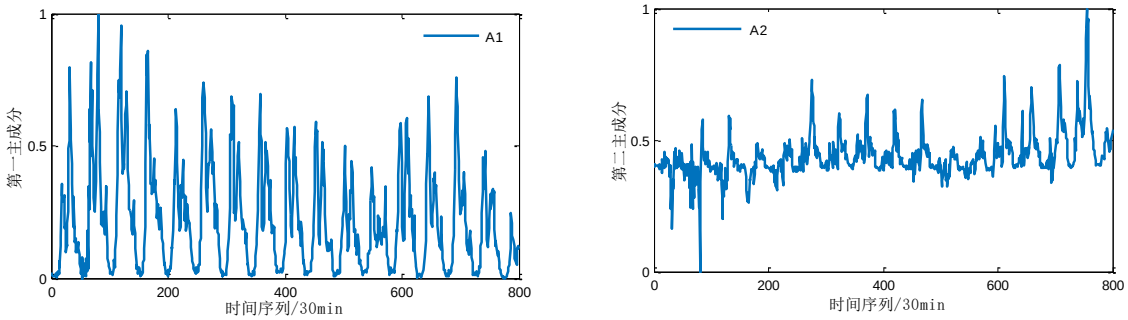
对来自上述 4 个平台的流行度时间序列进行主成分分析，其方差贡献率如表 3.4 所示：

表 3.4 主成分贡献率

成分	特征值	方差(%)	累积方差(%)
1	2.8325	70.81	70.8130
2	0.7814	19.54	90.35
3	0.3501	8.75	99.10
4	0.0360	0.90	100

观测表 3.4 发现，前两个主成分 A_1、A_2 的累计贡献率值达到 90.35%，包含原数据绝大部分信息，是影响流行度态势变化的主要成分，可将其用于社交话题多变量流行度融合预测。对两个主成分时间序列进行归一化处理，使其处于区间[0, 1]

之间，其主成分时间序列如图 3.5 所示。



(a) 第一主成分　　(b) 第二主成分

图 3.5 社交话题流行度主成分分析图

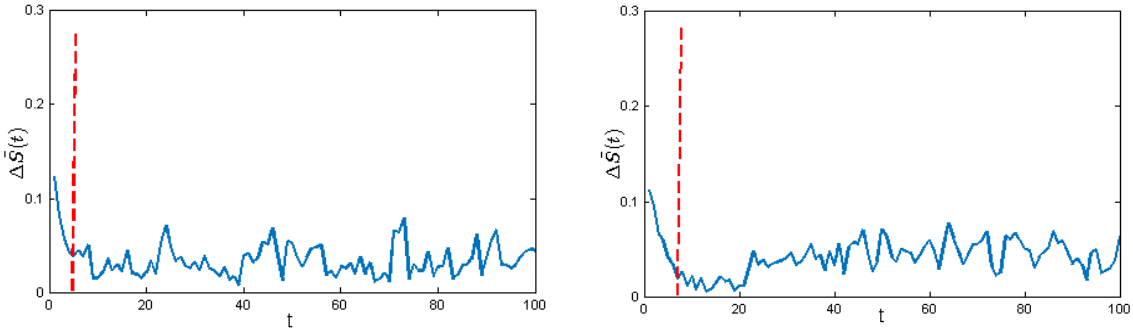
3. 混沌特性和重构参数分析

分别对 A_1、A_2 两个主成分进行混沌特性分析和重构参数的求取，具体分析结果如表 3.5。其中，τ 表示延迟时间，m 表示嵌入维数，λ 表示最大 Lyapunov 指数。

表 3.5 相空间重构参数

主成分	τ	m	λ
第一主成分（A_1）	5	8	0.0087
第二主成分（A_2）	7	7	0.0082

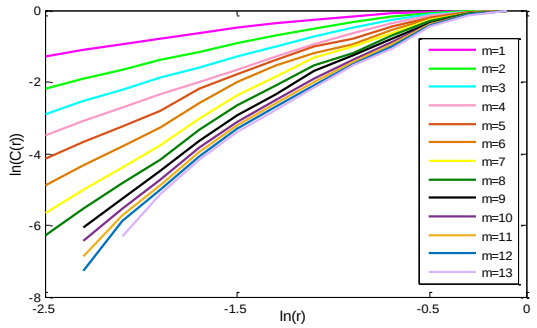
首先，利用 C-C 算法求延迟时间。延迟时间为 $\Delta\bar{S}(t)$ 的第一个极小值，而 $\Delta\bar{S}(t)$ 是通过关联积分 $C(r)$ 计算统计变量偏差获得。具体地，各个变量对应的 $\Delta\bar{S}(t)$ 如图 3.6 展示，不难求得主成分 A_1、A_2 对应的延迟时间 τ 分别为 5，7。



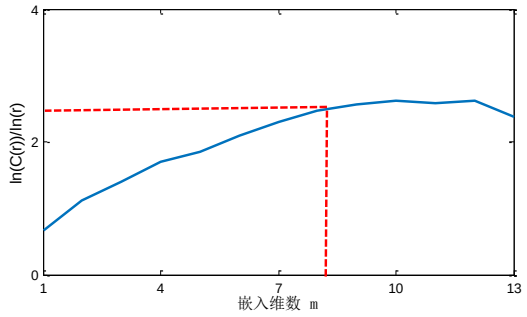
(a) 第一主成分的 $\Delta\bar{S}(t)$　　(b) 第二主成分的 $\Delta\bar{S}(t)$

图 3.6 两个主成分的延迟时间图

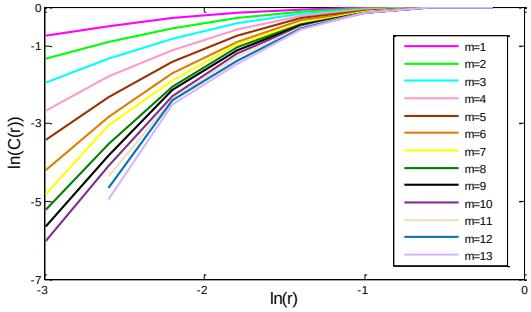
其次，利用 G-P 算法分析 A_1、A_2 的关联积分，进而求嵌入维数。图 3.7(a)、(c)绘制了在任意半径内，不同嵌入维数下，关联积分 $C(r)$ 的 $\ln C(r)-\ln r$ 图线。当 r 取某个适当的范围时，关联维数相对于 $C(r)$ 存在对数线性关系，则能通过图 3.7(a)、(c)拟合出关联维数随嵌入维数变化的图例，如图 3.7(b)、(d)所示。当关联维数随嵌入维数的增大而不再发生变化时所对应的最小嵌入维数即为所求 m，也显示了量化后 A_1、A_2 序列的混沌特性。通过观察容易求得 A_1、A_2 的嵌入维数分别为 8，7。



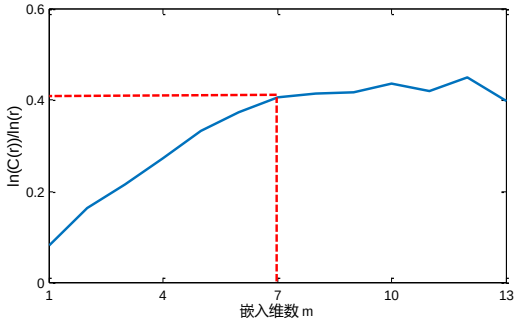
(a) A_1 的 $\ln C(r)-\ln r$ 曲线



(b) A_1 拟合(a)的曲线图



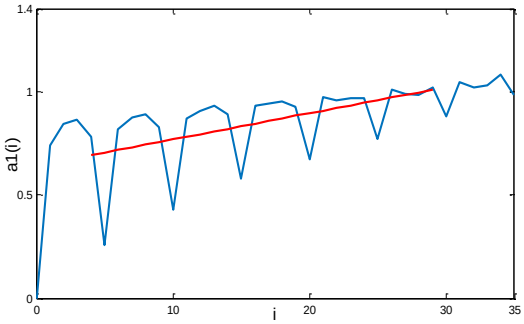
(c) A_2 的 $\ln C(r)-\ln r$ 曲线



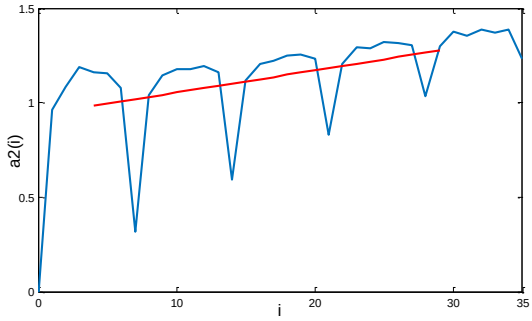
(d) A_2 拟合(c)的曲线图

图 3.7 两个主成分的嵌入维数图

最后，利用小数据量法求取最大 Lyapunov 指数，对重构后的时间序列做进一步的混沌特性分析。若流行度序列对应的最大 Lyapunov 指数大于 0，则说明社交话题流行度趋势具有混沌特性，可用混沌理论对流行度序列进行预测。小数据量法求两个主成分的最大 Lyapunov 指数如图 3.8 所示，图中蓝线为对数平均值随离散步长的变化图。进一步，可以从离散步长图中找出最佳拟合区间，用 LS 做回归直线，该直线的斜率即为最大 Lyapunov 指数，图 3.8 中的红线为 LS 拟合线段。A_1、A_2 两主成分对应的最大 Lyapunov 指数分别为对应 0.0087、0.0082，结果均大于零。



(a) 第一主成分 A_1 的对数均值图

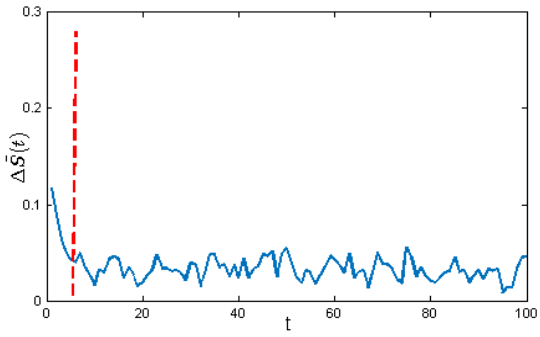


(b) 第二主成分 A_2 的对数均值图

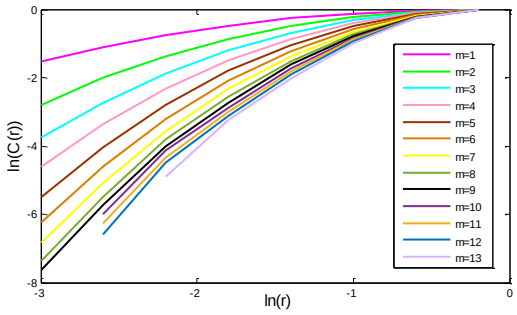
图 3.8 两个主成分的最大 Lyapunov 指数图

4. 多变量相点融合

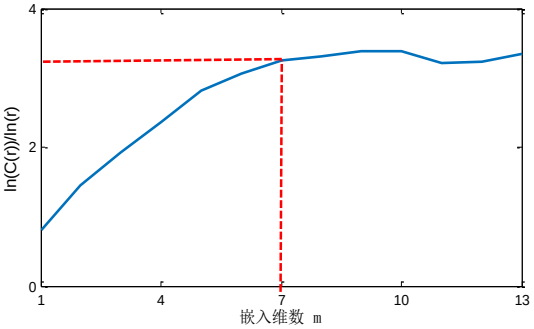
对影响流行度两个主要成分 A_1、A_2 进行最优相点融合。首先，按照公式(3.7)选择两个主成分变量的最大嵌入维数和最小延迟时间，依据公式(3.8)对两个主成分分别进行相空间重构。然后，依据公式(3.9)-(3.19)在同一重构高维相空间中实现影响话题流行度主成分的最优融合。对融合后的变量 Z 求嵌入维数及延迟时间分别对应 7、6。利用小数据量法检验融合后新的状态的最大 Lyapunov 指数为 0.0073，说明融合后变量仍具有混沌特性。具体地，图 3.9 为融合后流行度的求参图。



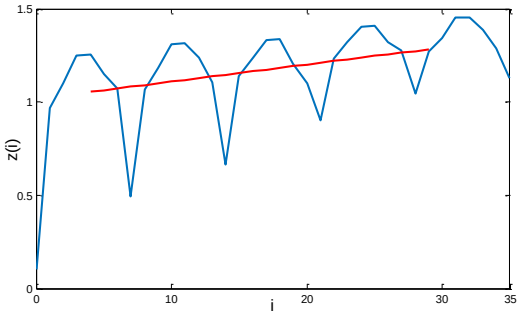
(a) 融合后话题流行度 $\Delta\bar{S}(t)$



(b) 融合后话题流行度 $\ln C(r)-\ln r$ 曲线



(c) 融合后话题流行度拟合(b)的曲线图

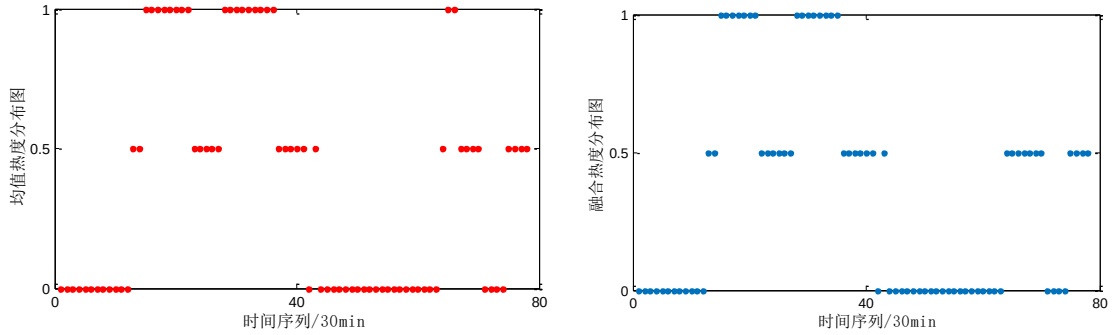


(d) 融合后话题流行度的对数均值

图 3.9 融合流行度参数图

5. RBFNN 流行度预测结果

本文的预测范围为流行度时间序列的后 77 个值，其余时刻点的值用作训练集，在预测之前，对融合的合理性进行分析。为了量化直接用均值热度和融合流行度比较带来的偏差，将话题热度分为三个等级，即：1 级超热度(0.6-1.0)，用 1 表示；2 级中热度(0.3-0.6)，用 0.5 表示；3 级微热度(0.0-0.3)，用 0 表示。将四个平台在各个时刻的热度均值和融合后的序列的热度值按等级统计比较发现，两者能够很好的拟合，其中，正确拟合的点数达 730 个，占总采样点数的 92%。进一步地，将预测部分的热度分布图展示如图 3.10(a)和(b)。



(a) 均值流行度热度分布图　　(b) 融合流行度热度分布图

图 3.10 热度分布图

进一步，Topic A 的预测性能对比详见表 3.6，Topic B 预测性能对比详见表 3.7。融合后流行度时间序列 Z 预测分析如图 3.11 所展示，其中，(a)为预测值和真实值对比图，(b)预测误差图。

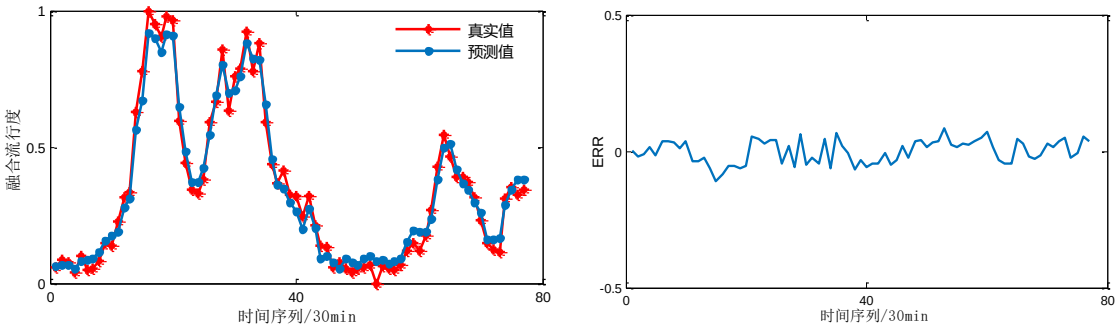
表 3.6 Topic A 预测性能对比表

误差指标	C-RBF	LWLP	LS-SVM	$A_1(t)$	$A_2(t)$	$Z(t)$
MAE	0.0720	0.0727	0.0558	0.0660	0.0411	0.0374
MAPE	0.8685	0.2645	0.2494	0.2933	0.1287	0.2014
EC	0.8619	0.8841	0.9222	0.9069	0.9280	0.9434

表 3.7 Topic B 预测性能对比表

误差指标	C-RBF	LWLP	LS-SVM	$A_1(t)$	$A_2(t)$	$Z(t)$
MAE	0.1391	0.1013	0.0871	0.0640	0.0799	0.0617
MAPE	0.7220	0.5790	0.5300	0.434	0.5211	0.4006
EC	0.8746	0.8917	0.9033	0.9019	0.8950	0.9187

通过图 3.11 发现，融合流行度时间序列的预测值和真实值能很好地贴近，预测误差平稳且较低，这证明了混沌理论在流行度预测中的可实施性。与此同时，通过预测结果，也可以感知到话题的涨落情况，及时掌控话题的有用时刻点，如：话题什么时候到达流行度峰值，什么时候流行度明显下降。这为舆情预警、潜在用户发现、广告推荐等领域提供了实用性的帮助。

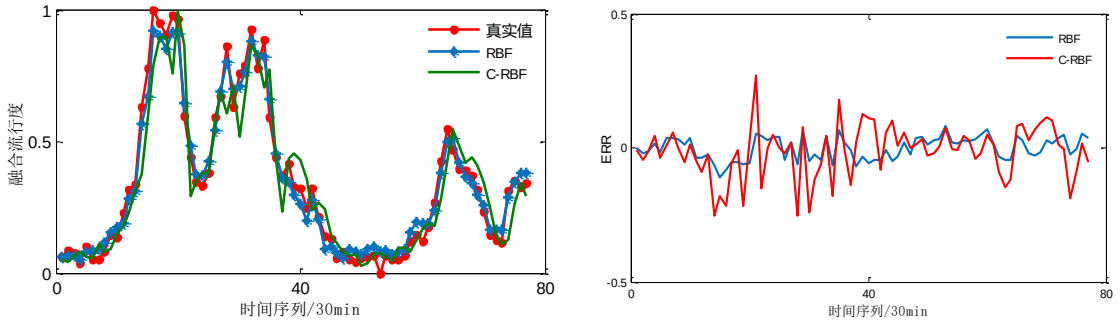


(a) 融合序列 Z 真实值和预测值对比图　　(b) 融合序列 Z 的预测绝对误差图

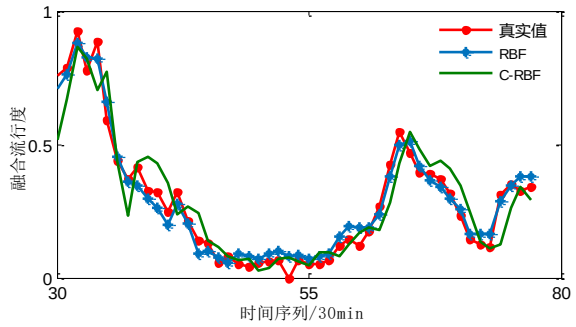
图 3.11 融合流行度 Z 预测结果

进一步地，为了证明本文提出模型的高效性，通过与一种应用于热点话题转发数预测的 C-RBFNN 模型、一种应用混沌理论的线性 LWLP 模型以及应用混沌理论的非线性 LS-SVM 模型进行对比实验。

图 3.12 为本文与 C-RBFNN 预测效果对比图，(a)为总体趋势对比图，(b)为两者的预测误差对比图，(c)为细节对比图。通过对比发现，假设不考虑社交话题态势变化的混沌特性，即使改进和优化算法也难以大幅度提高预测性能。



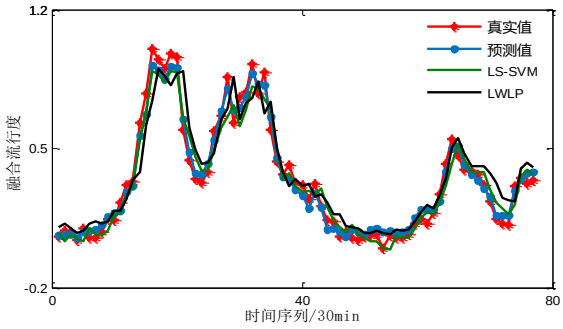
(a) 整体趋势对比图　　(b) 预测误差对比



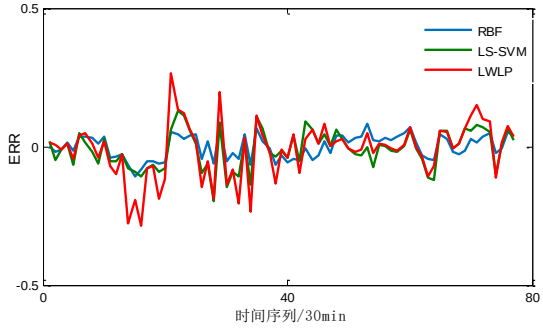
(c) 细节对比图

图 3.12 本文混沌 RBF 与 C-RBF 预测对比图

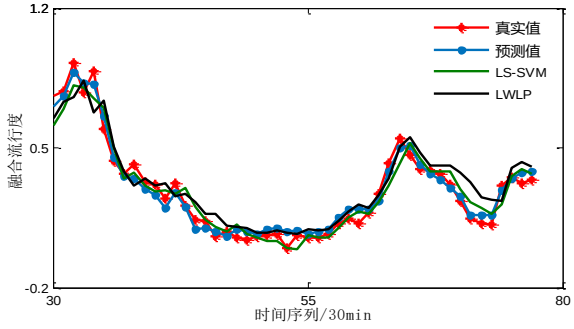
图 3.13 为 LWLP、LS-SVM 与本文的对比图，其中，(a)为的总体态势对比图，(b)为预测误差图，(c)为对比细节图。通过对比发现，LWLP、LS-SVM 模型也能感知到融合后的流行度的态势变化，且取得不错的预测效果，但 RBF 相对于其它模型在话题流行度的预测准确率更突出一点。



(a) 整体趋势对比图



(b) 预测误差对比



(c) 细节对比图

图 3.13 本文混沌 RBF 和两种混沌模型的预测对比图

3.5 本章小结

本章通过分析多个平台的话题信息，构建话题流行度态势传播的时间序列，进而提供一种基于混沌理论的社交话题传播流行度预测模型，用于感知社交话题在未来时刻流行度态势变化，预测话题信息热度值。首先，利用 PCA 分析影响流行度态势变化的主要成分，量化跨平台信息流行度。其次，基于混沌理论对影响流行度的主成分实施相空间重构。与此同时，在多变量相空间融合时，利用贝叶斯估计理论在同一重构高维相空间中实现多平台流行度序列的最优融合，弥补了单变量不足以反应流行度态势变化问题。最后，通过 RBF 预测融合后的多变量流行度序列。本章通过四个社交信息渠道的数据来实施相关实验。实验结果表明，该模型不仅能够很好地适用于社交话题流行度预测，且能够动态地感知话题态势变化，在网络安全和经济商业领域都有十分重要的研究和现实意义。

第 4 章 基于 F-SIR 和用户转发行为的信息流行度预测模型

4.1 引言

目前关于信息流行度的预测中，多数研究者将流行度定义为某种数量，用其衡量话题热度，如话题信息的分享[61]、评论[62]、转发数[63]等。关于传染病模型的流行度预测也不例外，它对疾病在人群中的表现和分布形式进行数学建模，求解模型对应的感染人群量，其中，感染人群量即参与转发人群热度。可见，基于传染病模型的方法是分析信息的传播过程和动力学研究的基础，是信息流行度预测的一个重要工具。

目前，基于传染病的信息流行度预测也存在一些挑战和突破点：

1. 信息流行度态势变化受多种驱动机制影响。在宏观上表现出信息传播流行度态势的多样化；在微观上表现在用户行为复杂性引起的用户转发行为的不确定性。因此，如何从微观上分析用户行为引起的转发参与过程以及分析影响转发的因素，进而预测转发流行度也成为研究的难点。

2. 信息传播的过程类似传染病的过程，但是仍然存在不同的地方。一方面，由于社交网络平台具有公开特性，传染病模型中的群体量为动态量，而传统传染病模型中将群体量设置为静态常量。另一方面，社交网络中信息沿着关注关系传播，只有用户关注者转发了这条信息，用户才有机会以粉丝的身份接收到信息，成为易感染者。所以说，网络中的易感染者大多数来自感染者的粉丝。

3. 传统传染病模型针对参数训练的问题，往往人为的设定固定的群体状态转换概率来构建信息传播网络。然而人为设定的参数具有随机性且缺乏理论依据，忽略了信息传播过程中时间特性引起的转换概率的动态变化，使模型的真实值和预测值有较大的差量。

针对以上传染病信息流行度的存在的挑战，本文构建基于用户行为和改进 SIR 模型的信息流行度预测模型。并基于腾讯微博数据集实施模型的验证实验，实验结果表明，本文提出的模型为用户转发行为分析提供了依据，且能较好地感知信息流行度趋势，预测信息未来时刻流行热度值。

4.2 问题形式化及相关定义

4.2.1 问题定义

1. 社交网络背景知识定义

为了清晰表述信息传播的过程，在这里定义网络的基本传播关系。

首先，定义信息传播网络$G=\{U,E\}$。其中，$U=\{u_1,u_2,u_3\cdots\}$代表时间T内信息传播的用户集，$E\subseteq U\times U$为用户之间的关注关系集合。若存在边$e_{ij}=<u_i,u_j>$，代表用户u_i是u_j用户的关注者即粉丝，信息可以沿着边e_{ij}从用户u_i传递到u_j。

接着，令$A=\{(b,u_i,t)\}$表示用户的过往信息集，其中，(b,u_i,t)用于描述用户u_i在t时刻采取的行为动作b。

最后，定义 SIR 模型的用户状态集合$Status_T=\{(u_i,t)\}$，代表在T时间的所有用户所处的相应状态。其中，$(u_i,t)\in\{\mathrm{S/\ I/\ R}\}$表示用户在$t$时刻可能处于 S(易感染态)、I(感染态)、R(恢复态)三个状态中的任一状态。

2. 转发特征提取与定义

在网络信息传播中，驱动用户参与转发传播的因素多种多样。不同用户的转发概率不同，例如：经常转发或者发布信息的用户参与新话题的可能性就越大；用户的好友参与越多的用户下一时刻越有可能参与到话题中来。本文将转发驱动力归结为两个方面的因素，即个人转发驱动力和社交转发驱动力，定义用户转发多维属性。具体如下：

(1) 个人转发驱动力

在用户个人转发驱动力方面，认为用户的转发具有延续性，这源于用户个人兴趣的记忆效应，与用户自身的属性息息相关。因此，从用户的个人属性出发，定义用户个人转发驱动力，相关属性如表 4.1 所示，具体定义如下：

表 4.1 个人转发驱动力符号及描述

序号	符号	描述
1	$value[attention(u_i)]$	用户 u_i 的关注度
2	$value[retweetRate(u_i)]$	用户 u_i 的历史转发率
3	$value[activity(u_i)]$	用户 u_i 的活跃度

定义 1：个人关注度

个人关注度定义为用户粉丝数和用户偶像数的比值，关注度越大的用户越有可能通过参与话题讨论来吸引粉丝。在线社交网络可以看成一个有向图，若用户 u_i 关注用户 u_j，则存在一条边 $u_i \to u_j$，认为 u_j 是用户 u_i 的偶像，用户的偶像总和为 $num\left[idol(u_i)\right]$。同理，若用户 u_i 被用户 u_j 关注，则存在一条边 $u_j \to u_i$，认为 u_j 是 u_i 的粉丝，用户的粉丝总和记为 $num\left[fans(u_i)\right]$。所以有：

$$value\left[attention(u_i)\right]=\frac{num\left[fans(u_i)\right]}{num\left[idol(u_i)\right]} \tag{4.1}$$

定义 2：个人历史转发率

历史转发率定义为话题信息开始前一个月用户转发信息占总信息的比值。根据用户的记忆特性，有转发过往行为的用户有更大的可能再参与到新话题的转发中，具体公式如下：

$$value\left[retweetRate(u_i)\right]=\frac{num\left[retweet(u_i)\right]}{num\left[totalTweet(u_i)\right]} \tag{4.2}$$

式中，$num\left[retweet(u_i)\right]$ ——话题开始前一个月用户转发信息的数量

$num\left[totalTweet(u_i)\right]$——话题开始前一个月用户所有信息数量

定义 3：个人活跃度

活跃度是用来衡量用户过往参与信息以及使用社交应用的主动性。相比活跃度较低的用户，高活跃度的用户在转发上起到的作用更大，定义如下：

$$value\left[activity(u_i)\right]=\rho num[orig(u_i)]+num[retw(u_i)] \tag{4.3}$$

式中，$\rho \in [0,1]$ ——弱化系数

$num[orig(u_i)]$——话题开始前一个月用户 u_i 发布信息数量

$num[retw(u_i)]$——话题开始前一个月用户 u_i 转发信息数量

(2) 社交转发驱动力

在用户社交转发驱动力方面，主要考虑用户的从众心理。这种从众心理主要来自于用户的社交行为的影响，相关属性如表 4.2 所示，具体定义如下：

表 4.2 社交转发驱动力符号及描述

序号	符号	描述
1	$value[interSimilarity(u_i, w)]$	用户 u_i 和信息 w 的相似度
2	$value[socInfRate(u_i, u_j)]$	用户 u_i 社交感染率
3	$value[socEffect(u_i, v_j)]$	用户 u_i 的社交影响力

定义 4：兴趣相似度

兴趣相似度用于衡量社交信息内容和用户兴趣的相似程度，认为两者相似度越大用户越容易被话题吸引，越容易转发信息。分别提取用户个人兴趣和信息的关键字，利用 Jaccard 系数计算两者相似度，计算公式如下：

$$value\left[interSimilarity(u_i, w)\right] = \frac{userInterest(u_i) \bigcup topicInterest(w)}{userInterest(u_i) \bigcap topicInterest(w)} \tag{4.4}$$

式中，$userInterest(u_i)$——用户行为兴趣标签

$topicInterest(w)$——信息关键字标签

定义 5：社交感染率

社交感染率考虑的是邻居节点参与情况对当前用户参与情况的影响。将其定义为参与到信息的关注节点占所有关注节点的比例，具体定义如下：

$$value\left[socInfRate(u_i, u_j)\right] = \frac{num[neigRetw(u_j)]}{num[neig(u_i)]} \tag{4.5}$$

式中，$num[neigRetw(u_i)]$——用户的关注节点参与转发数量

$num[neig(u_i)]$　　——用户关注节点总数。

定义 6：社交影响力

社交影响力是衡量当前用户对应的邻居好友具有的信息传播带动。社交影响力越大说明邻居好友的信息传播带动力越大，越有可能吸引用户参与话题信息。用 $\overline{num}[read(v_j)]$、$\overline{num}[retw(v_j)]$、$\overline{num}[comt(v_j)]$ 分别表示当前话题开始的前一个月关注节点发表信息的平均阅读数、转发数、评论数。则社交影响力如下：

$$value\left[socEffect(u_i, v_j)\right] = \beta\overline{num}[read(v_j)] + \overline{num}[retw(v_j)] + \overline{num}[comt(v_j)] \tag{4.6}$$

式中，β——弱化系数

4.2.2 问题形式化

为了形式化的描述本章的研究工作，将问题概述展示如图 4.1 所示，具体输入输出描述如下：

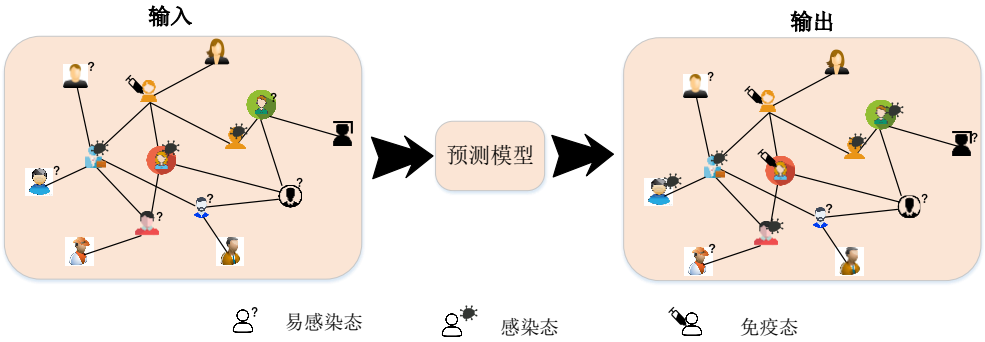


图 4.1 问题概述图

针对上述定义，将问题输入总结如下：

1. 某个话题下，信息传播的关系网络 $G_{T_k}=\{U,E\}$ 以及用户集合 U 的过往历史行为集 $A=\{(b,u_i,t)\}$。

2. 时间 T_k 内用户的状态集合 $Status_{T_k}=\{(u_i,t)\}$， k 为训练窗口长度。

根据问题输入，有以下问题输出需要解决：

1. 如何量化微观转发驱动力以及模型感染率？

从微观用户出发，考虑影响用户转发的个人转发驱动力 $P_{individual}(u_i)$ 和社交转发驱动力 $P_{social}(u_i)$，利用多元线性回归量化转发驱动力，以此获得改进 SIR 中转发感染率。

2. 如何基于传统 SIR 模型构建新的类传染病传播过程？

考虑到在线社交网络中信息是以发布者为源头沿着关注网络传播，即：当一个用户被感染后，他/她的粉丝则具有接触到信息的可能，并转变为易感染用户。重新定义易感染用户，将感染用户的粉丝作为易感染用户群体量，更加真实地模拟网络中信息传播的规律。

3. 如何描绘信息传播趋势？

将量化的用户转发驱动力和改进的传染病模型相结合，为状态转换的感染率提供理论依据。与此同时，通过时间切片技术提取各个状态量值，利用 LS 拟合模

型真实参量，提前感知话题信息流行度传播态势，预测话题信息参与趋势。

4.3 模型

为了解决上述问题，将模型的整体框架展示如图 4.2 所示。首先，提取社交网络结构和用户历史行为数据集，从个人和社交两个角度出发，分析影响用户转发的多维度驱动力，利用多元线性回归量化多维度用户转发驱动力，进而量化感染率。其次，考虑到真实社交网络中话题信息传播情况，对易感染用户进行改进，基于 SIR 构建新的社交传染病传播规则，获得新的状态传播方程。最后，通过时间切片处理量化各个时刻三个状态量值，利用 LS 模型训练得到状态方程预测模型，进而动态感知信息流行度态势变化，在微观层面深入分析个人和社交转发驱动力对传播态势的影响。

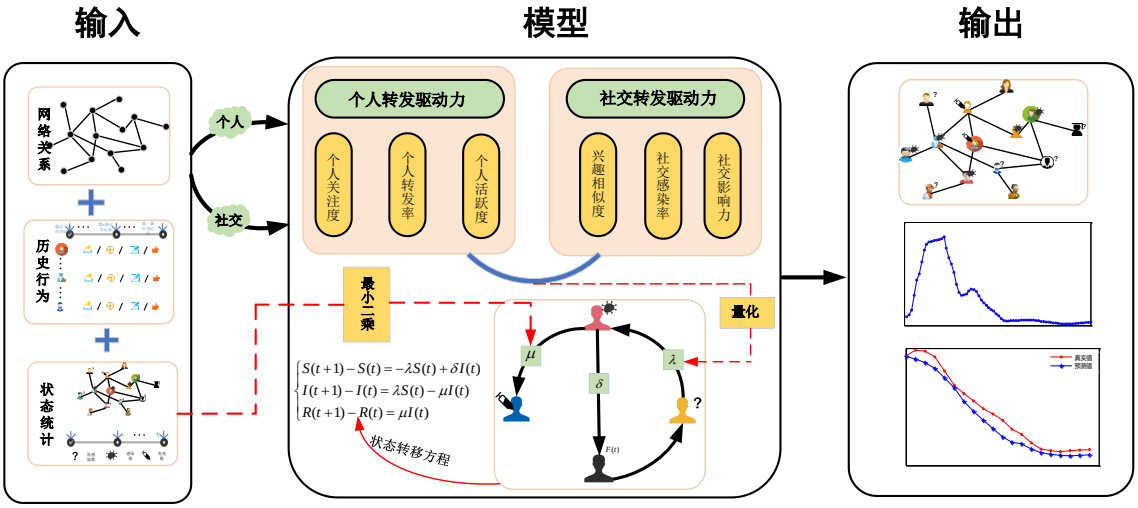


图 4.2 整体框图

4.3.1 转发驱动力量化

用户是否参与信息不仅仅与用户自身属性有关，如：关注度、转发率、活跃度，还与用户周围的事物有关，包括兴趣相似度、社交感染率、社交影响力，即个人和社交两个角度的转发驱动力。针对前面已经提取的个人和社交转发驱动力，利用多元线性回归量化多维度用户转发概率，在微观层面量化用户参与信息的细粒度动力学因素。具体地，转发驱动力如式(4.7)所示：

$$P(u_i)=\theta_0+\theta_1 P_{individual}(u_i)+\theta_2 P_{social}(u_i) \tag{4.7}$$

式中，θ_0、θ_1、θ_2——偏回归系数，由多元线性回归训练得到。θ_1、θ_2 分别是个人转发驱动力和社交转发驱动力相对于总体转发驱动概率的权重。

量化后的个人转发驱动力 $P_{individual}(u_i)$ 为：

$$P_{individual}(u_i)=\sum_{m=1}^{m}\frac{\kappa_{im}}{\max\limits_{u\in U}(\kappa_{u_i m})}\quad m=1,2,3 \tag{4.8}$$

式中，κ_{im}　——用户个人转发驱动驱动力，可以提取个人关注度、个人转发率、个人活跃度三个属性

$\max\limits_{u\in U}(\kappa_{im})$——不同属性下的最大值，以此实现个人转发驱动力的归一化处理

信息传播过程中用户因邻居节点而转发的能力会随着话题时间的弥散而越来越低。也就是说，信息的传播时间和邻居节点的带动力成反比，这类似于物理学中元素的半衰减性质。因此，引入半衰减因子函数 $(1/2)^{\frac{t-t'}{w}}$，用来刻画信息兴趣度随时间的衰减过程。其中，t、t'分别代表话题当前时刻和开始时刻，w为正则化因子。在此前提下定义用户社交转发驱动力 $P_{social}(u_i)$：

$$P_{social}(u_i)=\frac{\chi_{i1}}{\max\limits_{u\in U}(\chi_{u1})}+\sum_{n=1}^{n}\chi_{in}\times(1/2)^{\frac{t-t'}{w}}\quad n=2,3 \tag{4.9}$$

式中，χ_{i1}——兴趣相似度

χ_{in}——社交感染率和社交影响力

4.3.2 信息传播的传染病模型

在线社交网络中，信息的传播过程和传染病的传播过程类似，可以利用传染病中的 SIR 模型对信息传播过程实施建模。但是，在真实社交网络中两者的传播机理又略有不同。在生物种群中，当一个个体感染某种疾病后，除了一些有抗体免疫个体外，网络中其余个体是易感染者，都有机会接触到感染个体。然而，在线社交网络中，信息是沿着关注网络层级传播，只有用户的关注者转发了这条信息，用户才有机会以粉丝的身份接收到信息，成为易感染者，受个人和社交特性

的影响，以一定的转发感染率变为感染人群。所以说，网络中的易感染者大多数来自感染者的粉丝。在此基础上，构建基于 SIR 的改进 F-SIR 模型，其中，F 表示粉丝，是感染者的粉丝到易感染的一个过度状态。具体地，F-SIR 的状态转移图如图 4.3 所示：

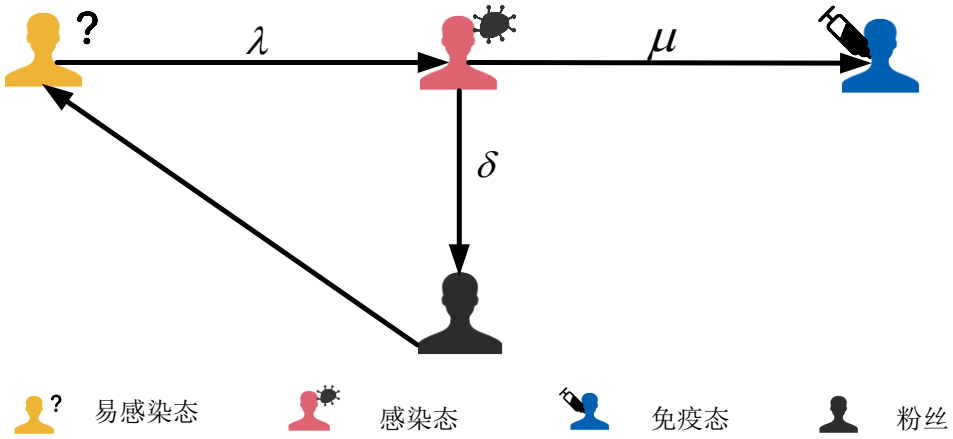


图 4.3 状态转移图

该模型建立如下假设：

1. 当一个易感染用户被感染而发布或者转发一条信息时，该用户的状态则由易感染态转换为感染态，而用户是否转发由用户的个人和社交特性决定，经多元线性回归量化求得转发概率。假设在t时刻，感染人数是$I(t)$，用户对应的平均转发感染率记为$\lambda=\overline{P(u_i)(t)}$。

2. 当一个用户转发了一条信息，则该用户状态为感染态，他/她的粉丝立马转换易感染态。在t时刻，假设感染用户的平均粉丝数为$\overline{F(t)}$，则由感染用户带来易感染用户的速率为$\delta=\overline{F(t)}$。

3. 假设用户的话题参与有效时间为12小时。12小时后，如果用户没有重新参与话题，则会由感染状态变为免疫状态。t时刻，免疫人数为$R(t)$，免疫率μ由LS训练获得。

根据以上的传播规则，可以得到如下的动力学方程：

$$\begin{cases} S(t+1)-S(t)=-\overline{P(u_i)(t)}S(t)+\overline{F(t)}I(t) \\ I(t+1)-I(t)=\overline{P(u_i)(t)}S(t)-\mu I(t) \\ R(t+1)-R(t)=\mu I(t) \end{cases} \tag{4.10}$$

4.3.3 构建流行度预测模型

为了完整构建 F-SIR 的传播模型，深入探究感染免疫状况，利用 LS 训练误差函数，使训练模型的参量在拟合真实值时误差最小，具体步骤如下：

通过时间切片技术，设定切片窗口大小，统计各个时间窗口下对应的三个状态的真实值记为$S(t+1)$、$I(t+1)$、$R(t+1)$，预测值用$S'(t+1)$、$I'(t+1)$、$R'(t+1)$表示，预测值可以通过状态转移方程估计获得。那么，预测值和真实值的误差函数可以表示如下：

$$f=\sum_{t=1}^{k}[(S(t+1)-S'(t+1))^2+(I(t+1)-I'(t+1))^2+(R(t+1)-R'(t+1))^2] \tag{4.11}$$

式中，k——训练窗口长度，而：

$$S'(t+1)=S(t)-\overline{P(u_i)(t)}S(t)+\overline{F(t)}I(t) \tag{4.12}$$

$$I'(t+1)=I(t)+\overline{P(u_i)(t)}S(t)-\mu(t)I(t) \tag{4.13}$$

$$R'(t+1)=R(t)+\mu(t)I(t) \tag{4.14}$$

为了最小化误差函数 f，利用 LS 求未知参量，令偏导数为 0，结果即为所求。

4.3.4 模型算法设计及分析

传统 SIR 模型参数训练中，往往人为的设定固定的群体状态转换概率来构建完整的 SIR 预测传播网络。按照上述方法，一方面，人为设定的参数具有随机性且缺乏理论依据；另一方面，忽略了话题传播过程中时间特性引起的转换概率的动态变化，使得预测值和真实值有较大的差量。考虑到以上情况，本文构建 F-SIR 模型，从微观用户角度出发，多维度、细粒度的研究微观用户转发特性引起的宏观群体状态量变化。与此同时，引入时间衰减函数和切片处理共同刻画时间特性引起的状态转移概率的动态变化过程。具体地，相应模型的学习算法可以大致分为三个部分：(1) 构建 F-SIR 模型传播规则；(2) 通过多元线性回归，挖掘微观用户转发特征量，量化转发驱动力；(3) 用 LS 调节学习参数，获得不同时刻信息传播的扩散情况，预测信息流行度随时间变化值。具体的算法如表 4.3 所示。

表 4.3 模型算法表

Input: diffusion network: $G_{T_k}=\{U,E\}$; Behavior Set of history: $A=\{(b,u_i,t)\}$; Status set: $Status_{T_k}=\{(u_i,t)\}$

Output: retweeting driving force: $P(u_i)$; infection rate: λ; fans rate: δ; immunization rate: μ; the status set of prediction: $\{S(t)\}$, $\{I(t)\}$, $\{R(t)\}$, $t\in[k+1,n]$

constructing propagation rules of improved SIR model: F-SIR
initialize S_0, I_0, R_0
for each user u_i in topic t_j **do**
compute dynamic retweeting driving force: $\{P_{individual}(u_i), P_{social}(u_i)\}$ from Eqs.(4.1)-(4.6)
perform multiple linear regression to calculate influence factors from $P(u_i)$ Eqs.(4.7)-(4.9)
obtain residual, confidence interval, parameter θ_0, θ_1, θ_2
end for
perform least square method fit data from Eqs.(4.10)-(4.14)
obtain parameter $\lambda=\overline{P(u_i)(t)}$, $\delta=\overline{F(t)}$, μ
executing state transition equation from Eqs.(4.10);
result: retweeting driving force: $P(u_i)$; the status set of prediction: $\{S(t)\}$, $\{I(t)\}$, $\{R(t)\}$, $t\in[k+1,n]$

4.4 仿真实验与结果讨论

4.4.1 实验数据

本文的实验数据集来自腾讯微博平台，该平台为一种提供微型博客服务的类 Twitter 网站，是促进信息传播的重要平台。据统计截至 2011 年 9 月 30 日，腾讯微博平台上的注册用户数超过 3.1 亿，日活跃用户数将近 5000 万人。所以，用腾讯微博的数据来研究话题信息具有真实性、可靠性。为了建立热点信息流行度预测模型，从腾讯社交平台抓取近 10 个月的信息传播数据，经筛选选取其中三个话题信息作为本文信息流行度预测模型的实验数据，包括“私人订制”(Topic A)，“爸爸去哪儿 2”(Topic B)以及“熊猫血女孩救助”(Topic C)，介绍如表 4.4 所示：

表 4.4 相关数据统计表

数据集	时间间隔	用户数	粉丝数	网络边数	用户行为数
Topic A	2013.12.19-2014.01.04	4359	898126	1399436	26417807
Topic B	2014.05.14-2014.09.04	7041	879780	1075051	40680234
Topic C	2014.02.25-2014.09.07	9581	534459	582529	21449323

4.4.2 基础方法

为了评估本文构建的基于用户行为的信息流行度预测模型的性能，将本文 F-SIR 模型与以下经典模型进行对比。

SIR[64]：SIR 模型是传染病模型中最经典的模型。模型中信息传播用户处于三个状态，易感染态 S、感染态 I 和免疫态 R。已知一条信息的传播网络，易感染用户指信息未知者；感染用户指已经接收到信息，并且有感染其它用户能力的群体；免疫用户指已经接触到信息，但不会继续传播信息的群体。

SISE 模型[30]：SISE 模型是一种改进的 SIS 传染病模型，它考虑到话题信息的易感染用户群体除可能来自感染用户的粉丝群外，还有可能是直接通过外部开放平台间接感染的用户。基于这种情况，在传统 SIS 模型的基础上，加入外部状态量 E 状态，构建基于 SISE 的流行度预测模型。

4.4.3 评估指标

为了对模型性能做进一步地评估，本文采用传染病模型中信息流行度预测常用指标绝对误差(Absolute Error, ERR)、平均绝对误差(Mean Absolute Error, MAE)、平均绝对百分误差(Mean Absolute Percentage Error, MAPE)作为预测性能评价指标，具体定义如下：

$$\mathrm{ERR} = \hat{I}(t) - I(t) \tag{4.15}$$

$$\mathrm{MAE} = \frac{1}{n-k}\sum_{t=k+1}^{n}\left|I(t) - \hat{I}(t)\right| \tag{4.16}$$

$$\mathrm{MAPE} = \frac{1}{n-k}\sum_{i=k+1}^{n}\left|\frac{I(t) - \hat{I}(t)}{I(t)}\right| \tag{4.17}$$

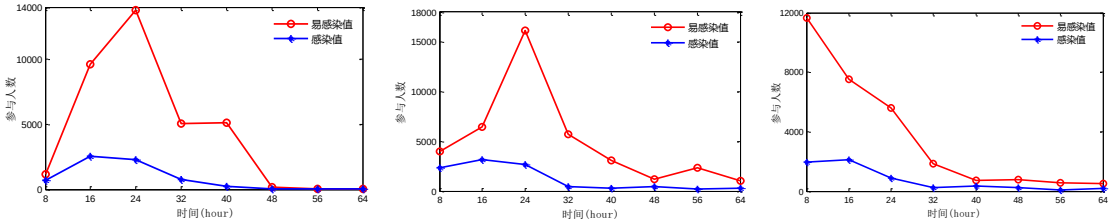
其中，$I(t)$ 代表真实值，$\hat{I}(t)$ 代表预测值，k 为训练窗口长度，n 为样本总数。

4.4.4 预测性能分析

1. 观察分析

首先，以小时为时间单位进行时间切片处理，简单统计不同时刻易感染用户量以及下一时刻易感染用户参与话题的讨论量，统计结果如图 4.4 所示。在图 4.4

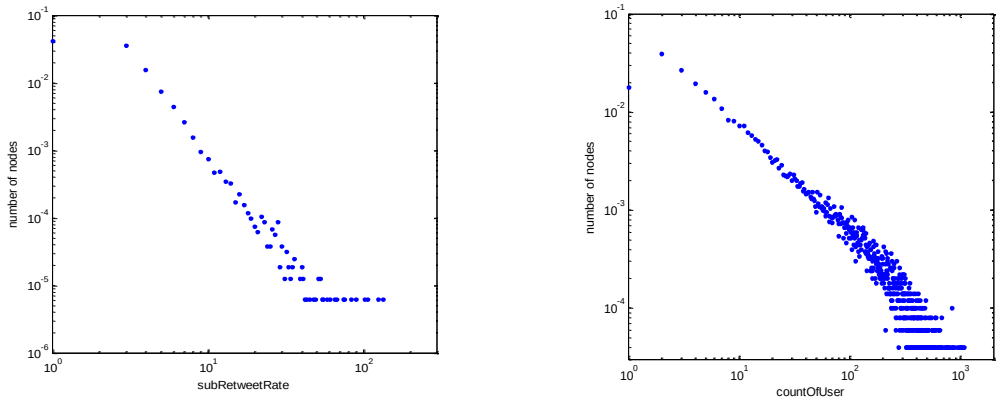
中，红线是当前时刻易感染用户数量，蓝线是下一时刻易感染用户中被感染的用户数量。通过比较发现，易感染用户的在不同时刻被感染的感染率 $P(u_i)(t)$ 会随时间波动，是一个动态变量。



(a) Topic A 潜在用户趋势图　(b) Topic B 潜在用户趋势图　(c) Topic C 潜在用户趋势图

图 4.4 潜在用户感染趋势图

进一步地，分析用户的转发属性，如图 4.5 所示。在图 4.5 中，(a)代表潜在感染人群的转发率分布形式图，(b)代表潜在感染人群关注的参与用户分布形式图。通过观察图 4.5 发现，潜在感染人群的转发率以及其关注的参与用户数量均呈现幂率分布形式。这间接说明绝大多数人群的过往行为具有一定的规律，受其对应的关注人群的参与情况以及历史转发比率的影响，可以通过潜在人群的个人和社交特性量化感染率。



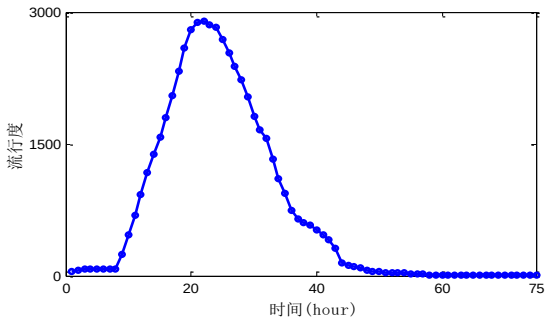
(a) 潜在感染人群转发率　(b) 潜在感染人群关注的参与人群图

图 4.5 属性幂律分布图

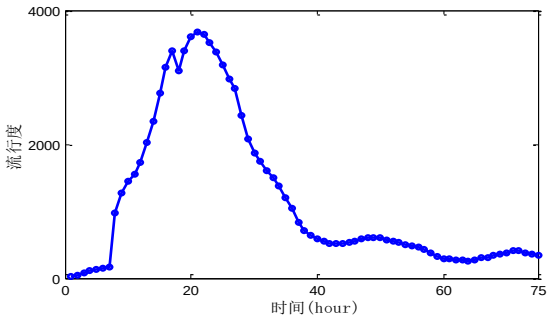
2. 预测结果分析

在 F-SIR 模型中，状态 I 代表感染状态人群，感染状态量越大，说明参与话题的人数越多。因此，可以通过时间切片处理，统计不同时刻感染用户量 $I(t)$，构建话题信息在不同时刻的流行热度 $I(t)$ 趋势图，基于前面时间点的话题流行度预测未

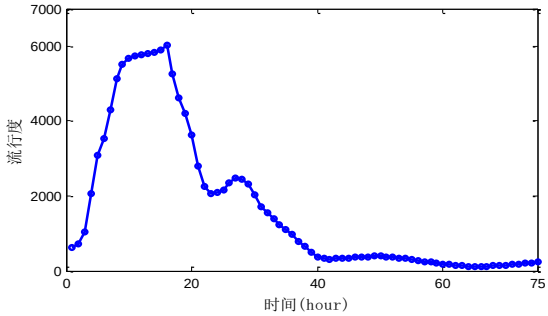
来时刻流行度。具体地，统计获得的话题流行热度趋势图如图 4.6 所示，其中，(a)为 Topic A 的流行度态势变化图，(b)为 Topic B 的流行度态势图，(c)为 Topic C 的流行度态势图。通过对比三个话题信息的流行度趋势图发现，信息的流行热度随时间大体呈现出先增后减的变化趋势，即一个话题信息存在热度爆发期以及热度消亡期，在热度达到最高点后，话题流行度会逐渐衰弱，渐渐地走向死亡。通过观察统计发现，Topic A 的热度活跃区间为 9-43 小时，Topic B 的活跃度区间为 8-81 小时，Topic C 的热度活跃区间为 1-59 小时。进一步地发现，不同的话题消亡的方式会有所不同，对于 Topic A 在热度达到最高热度点以后，热度值一直呈现下降趋势，而 Topic B 和 Topic C 仍然会存在余热波峰。可见，针对不同的信息或者话题其流行度变化趋势不同，信息流行度预测至关重要，能够及时捕获话题未来热度变化趋势，为网络安全管控、舆情预警以及提供帮助。



(a) Topic A 流行度趋势图



(b) Topic B 流行度趋势图



(c) Topic C 流行度趋势图

图 4.6 流行度趋势图

本文将预测时间长度设置为 20 个时间切片。首先，调节训练集合的长度 k，进一步地，获取不同训练长度下对应的预测误差，选择合适的训练长度。具体地，三个话题信息在不同训练长度下的 MAPE 误差如表 4.5 所示，误差整体趋势如图 4.7 所示。通过表 4.5 以及图 4.7 总结发现，三个话题的预测误差呈现出先减少后

增加的趋势。这是因为刚开始随着训练长度地增加，能够捕捉更多话题趋势变化的细节信息，从而使训练误差降低，而当训练长度过长时，预测值接近话题消亡时期，话题信息的参与人数很低，导致即使预测值与真实值相差很小也会造成 MAPE 值较大。考虑到需要捕捉话题有效时期的流行热度，因此，训练长度不易太长。本文选择三个话题预测误差下降比较明显的训练长度作为合适的训练长度，其中，Topic A、Topic B、Topic C 训练长度分别设置为 35、35、25。

表 4.5 训练长度及 MAPE 误差表

训练长度	Topic A	Topic B	Topic C
15	0.5130	0.4400	0.3729
25	0.3711	0.3530	0.2173
35	0.2845	0.2426	0.1966
45	0.2610	0.2011	0.1810
55	0.3400	0.2483	0.2219

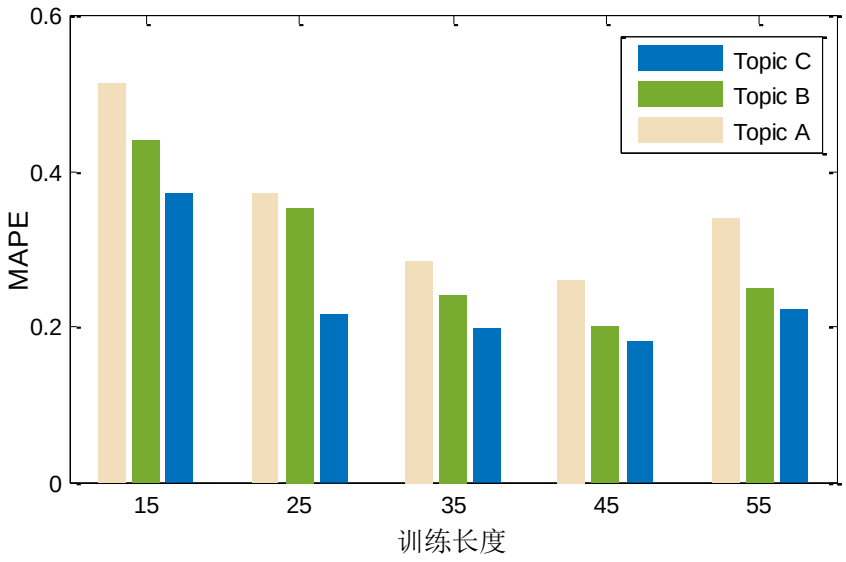
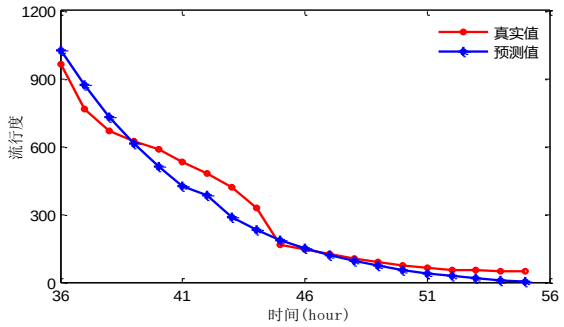
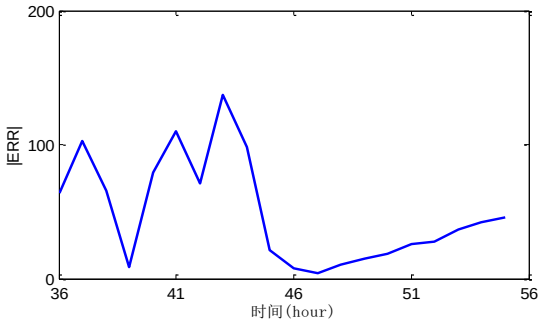


图 4.7 训练窗口测试图

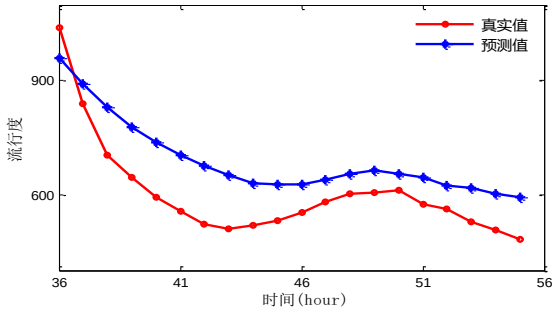
进一步地，在相对应的训练长度下，各话题流行度的预测和实际值的对比及绝对误差结果通过图 4.8 来展示。在图 4.8 中能够发现，本文所提出的方法能够较好地预测话题信息的发展趋势，有较低的绝对误差值。相对于 Topic A 和 Topic C 来讲，Topic B 的变化趋势中存在余波峰，由于 F-SIR 模型动态调整感染率，能够在一定程度上感知波峰趋势的变化。



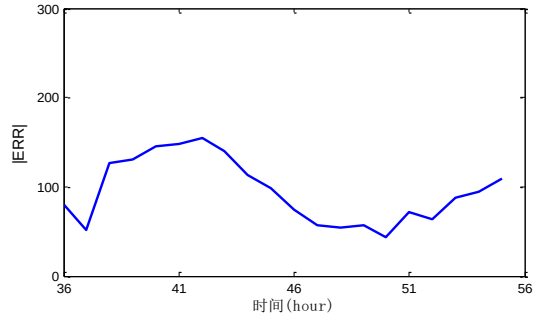
(a) Topic A 真实与预测对比图



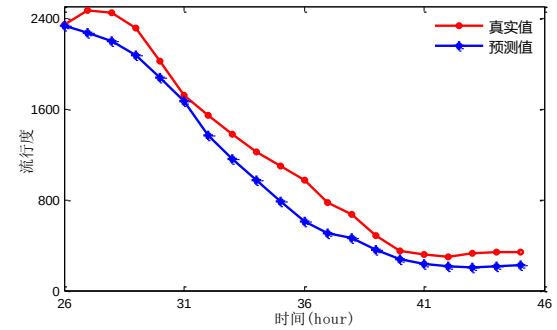
(b) Topic A 预测误差图



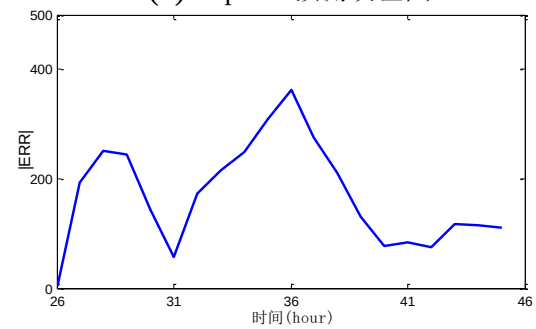
(c) Topic B 真实与预测对比图



(d) Topic B 预测误差图



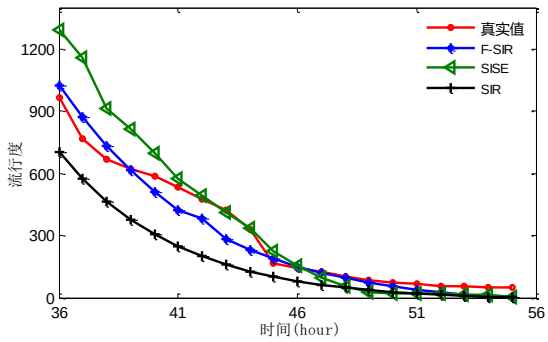
(e) Topic C 真实与预测对比



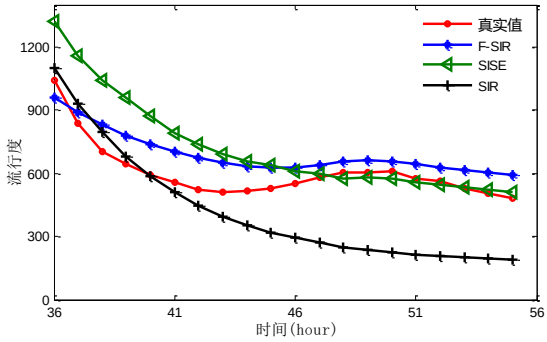
(f) Topic C 预测误差图

图 4.8 F-SIR 模型预测结果

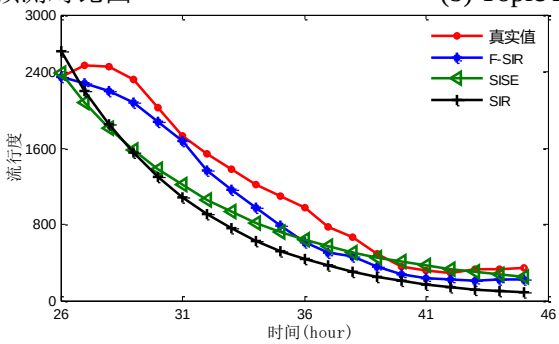
进一步地，为了证明本文提出模型的有效性和可靠性，将本文 F-SIR 模型与传统 SIR 和 SISE 模型实施误差对比分析，具体的预测趋势对比图如图 4.9 所示，预测对比误差如表 4.6 所示。通过对比发现，相对于传统 SIR 以及加入外部访问状态 E 的 SISE 模型，改进的 F-SIR 模型都显示了较好的预测效果。可见，该模型能够预测话题热度，用于感知和预测话题态势变化趋势。



(a) Topic A 预测对比图



(b) Topic B 预测对比图



(c) Topic C 预测对比图

图 4.9 F-SIR 模型预测方法对比图

表 4.6 预测性能对比表

Topic	Error Index	F-SIR	SISE	SIR
Topic A	MAE	50.6259	91.1500	138.9216
	MAPE	0.2845	0.3552	0.4800
Topic B	MAE	94.9697	132.400	211.3840
	MAPE	0.2426	0.3288	0.3738
Topic C	MAE	169.9500	286.3292	392.6176
	MAPE	0.2173	0.2416	0.4487

4.5 本章小结

本章分析影响用户转发的多维属性，量化改进 SIR 模型中的感染率，提出一种基于用户行为的信息流行度趋势预测模型，用于感知话题未来时刻流行度热度值。首先，从用户的个人和社交两个层面出发，提取影响用户转发的多维属性，利用多元线性回归量化用户转发驱动力，获得转发感染率。其次，基于传统 SIR 模型以及社交网络中信息传播的特点，引入感染者粉丝转化为易感染群体用户的过渡状态 F，重构信息传播规则和状态转移方程。最后，综合考虑影响量化后的用户驱动力以及时间特性，结合时间切片技术提取改进 SIR 模型的各个状态值，利

用最 LS 拟合真实模型，训练模型参量，获得 F-SIR 模型的状态转移方程，基于此模型对未来时刻的信息热度实施预测。本文使用从腾讯微博抓取的三个热点话题传播网络的相关数据来实施相关模型的实验验证。实验结果表明，本章提出的模型能够较好地感知信息流行度趋势，预测信息未来时刻流行热度值。

第 5 章 总结及展望

5.1 研究工作总结

随着科技蓬勃迅速的发展，线上交流和分享信息已经成为人们的日常活动，在生活和工作中扮演着重要的角色。面对多样化的社交平台以及错综复杂的用户群体，信息流行度传播和演化规律也变的更加复杂和不确定。因此研究和把握社交平台下话题信息的传播流行态势，有助于及时掌控网络舆情的走向，有助于更好的把握网民的上网行为，有助于构建更健康的社交环境。本论文以科研项目为背景，从宏观和微观两个角度构建流行度传播态势。在宏观上，研究信息流行度态势变化的规律；在微观上，探究用户的转发行为，预测时序转发流行度。将本文研究工作总结如下：

1. 在宏观上基于话题信息量构建流行度时序态势，研究流行度趋势传播规律。本文从具体话题入手，构建多社交平台流行度时序变化趋势图，利用小数据量法计算最大 Lyapunov 指数，发现社交话题流行度时间序列存在混沌特性，旨在深入剖析话题传播的非线性动力学机制。

2. 基于贝叶斯估计理论构建跨平台流行度融合模型。考虑到社交话题存在平台差异性，不同平台下的话题信息流行度变化趋势不尽相同，本文对跨平台时序流行度实施相空间重构，利用贝叶斯估计方法，在同一重构高维相空间中实现多平台流行度序列的最优融合，获得跨平台流行度序列。

3. 从微观用户转发属性出发，综合考虑影响用户转发的多维度因素，以此作为用户状态群体量改变的理论依据，量化 SIR 模型的转发驱动力以及模型感染率。以传染病 SIR 模型为基础，从用户个人和社交维度探究转发驱动力，提出一种多维度的基于多元线性回归的转发感染驱动力量化和度量方法，从而为细粒度研究微观层面用户转发驱动力提供理论依据。

4. 基于真实数据验证模型的准确性和有效性，为信息传播以及流行度态势预测的后续研究提供一定的支持。

5.2 未来工作展望

本文仅仅是从宏观的传播规律和微观用户行为为信息流行度预测提供一些支持，而针对信息流行度预测的相关研究，本文的研究工作是微不足道的。目前，针对信息流行度的预测很多，但信息流行度预测也存在诸多挑战，如：时间跨度的选择，内外影响因素的量化，以及用户行为复杂性引起的流行度态势传播的不确定性等。由于本人知识储备量以及时间问题，本文的工作还存在一些不足，需要进一步提高和充实。具体总结如下：

1. 如何划分流行度时间粒度。信息流行度是与时间有关的热度量，在不同时间粒度下，流行度的演化趋势不尽相同。时间粒度划分的太细，可能导致工作量过大，同时造成大量时间段无数据而无法进行后续研究；若时间粒度选取过大，容易超出信息的生命周期，使预测结果没有意义。而本文针对信息流行度的预测时，通过选区适合话题传播的预测粒度实施研究。但是，若改变话题传播的粒度，话题流行度传播态势如何以及预测结果如何仍然需要进一步的实验。因此，合理划分时间粒度是信息流行度预测工作的一个重要问题。

2. 用户转发驱动力的定义和量化问题。本文构建的用户多维转发驱动机制，仅仅是常用的用户转发属性，在实际场景下用户行为错综复杂，如何定义和发现有效的转发驱动力成为是否能够准确预测用户行为的关键问题。同时，除了用户自身和外部邻居引起的转发行为，用户的转发与文本的内容及情感特征有关。因此，如何结合信息的上下文情境深入分析文本语义流行度以及用户情感信息，也成为流行度预测待研究和探讨的关键问题。

3. 在线社交网络中的异构数据和海量数据的处理问题。相对来说，目前大部分信息流行度的预测研究可实施的数据集普遍较小，但在线社交网络中信息传播包含海量数据，而基于小数据集的预测模型可能不适合大规模数据预测或因为时间原因难以在大数据集上部署。所以，在以后的任务中，可以探究如何构建高效的数据处理模型，提高预测模型的执行广度和宽度。具体的包括：借助大数据平台 Hadoop 部署并行数据处理方案；探究基于网络论坛、新浪、Twitter 等多社交平台的数据，实现流行度全方位预测。

社交网络中信息流行度传播态势感知对舆情管控、市场营销以及为网民营造

健康的上网环境都有重要的研究价值。大数据平台的发展为解决和研究信息流行度预测提供了支持，但这项工作任重道远，未来的研究工作可以从以上几方面进一步探讨。

参考文献

[1] Milgram S. The small world problem[J]. Psycholgy Today, 1967, 2(1): 60-67.

[2] CNNIC: 2018 年第 42 次中国互联网络发展状况统计报告[EB/OL]. URL: [2018-08-20] https://cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201808/t20180820_70488.htm.

[3] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks[C]// Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2007: 29-42.

[4] Zhao D, Rosson M B. How and why people Twitter: the role that micro-blogging plays in informal communication at work[C]// Proceedings of the ACM 2009 International Conference on Supporting Group Work. New York: ACM Press, 2009: 243-252.

[5] 李洋, 陈毅恒, 刘挺. 微博信息传播预测研究综述[J]. 软件学报, 2016, 27(2): 247-263.

[6] Davoudi A, Chatterjee M. Prediction of information diffusion in social networks using dynamic carrying capacity[C]// IEEE International Conference on Big Data. Washington: IEEE Press, 2016: 2466-2469.

[7] Masuda N. Opinion control in complex networks[J]. New Journal of Physics, 2015, 17(2): 33031-33041.

[8] Ye Shaozhi, Wu F. Measuring message propagation and social influence on twitter.com[C]// Proceedings of International Conference on Social Informatics. Berlin: IEEE Press, 2010: 216-231.

[9] Doerr C, Blenn N, Van Mieghem P. Lognormal infection times of online information spread[J]. PloS one, 2013, 8(5): 245-252.

[10] Emrouznejad A , Marra M. The state of the art development of AHP (1979–2017): a literature review with a social network analysis[J]. International Journal of Production Research, 2017, 55(22): 1-23.

[11] 胡颖, 胡长军, 傅树深, 等. 流行度演化分析与预测综述[J]. 电子与信息学报, 2017, 39(4): 805-816.

[12] Bandari R, Asur S, Huberman B A. The pulse of news in social media: forecasting popularity[C]// Proceedings of Association for the Advancement of Artificial Intelligence. New York: IEEE Press, 2012: 231-241.

[13] Agrawal N. Identifying the influential bloggers in a community[C]// ACM International Conference on Web Search & Data Mining. New York: ACM Press, 2008: 207-217.

[14] Tsagkias M, Weerkamp W, De Rijke M. Predicting the volume of comments on online news stories[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2009: 1765-1768.

[15] Lerman K, Galstyan A. Analysis of social voting patterns on digg[C]// Proceedings of the First Workshop on Online Social Networks. New York: ACM Press, 2008: 7-12.

[16] Liu Yanbing , Zhao Jinzhe, Xiao Yunpeng. C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network[J]. Neurocomputing, 2018, 275(7): 733-746.

[17] He Xiangnan, Gao Ming, Kan M Y, et al. Predicting the popularity of web 2.0 items based on user comments[C]// Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2014: 233-242.

[18] 陈江, 刘玮, 巢文涵, 等. 融合热点话题的微博转发预测研究[J]. 中文信息学报, 2015, 29(6): 150-158.

[19] Xiao Yunpeng, Song Chenguang, Liu Yanbing. Social hotspot propagation dynamics model based on multidimensional attributes and evolutionary games[J]. Communications in Nonlinear Science and Numerical Simulation, 2019, 67(3): 13-25.

[20] 刘玮, 贺敏, 王丽宏, 等. 基于用户行为特征的微博转发预测研究[J]. 计算机学报, 2016, 39(10): 1992-2006.

[21] Bae Y, Ryu P M, Kim H. Predicting the lifespan and retweet times of tweets based on multiple feature analysis[J]. ETRI Journal, 2014, 36(3): 418-428.

[22] Sun Wanlong, Zheng Dequan, Hu Xinchen, et al. Microblog-oriented backbone nodes identification in public opinion diffusion[C]// Proceedings of 2014

International Conference on Audio, Language and Image Processing (ICALIP). Piscataway: IEEE Press, 2014: 570-573.

[23] Tan Chenhao, Lee L, Pang Bo. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter[C]// Proceedings of the Association for Computational Linguistics. Baltimore: IEEE Press, 2014: 175−185.

[24] Yang J, Leskovec J. Patterns of temporal variation in online media[C]// Proceedings of ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 177-186.

[25] Cheng J, Adamic L A, Kleinberg J M, et al. Do cascades recur?[C]// Proceedings of International Conference on World Wide Web. Montreal: IEEE Press, 2016: 671-681.

[26] Szabo G, Huberman B A. Predicting the popularity of online content[J]. Communications of the ACM, 2010, 53(8): 80-88.

[27] Oakes D. Survival analysis[J]. European Journal of Operational Research, 2000, 95(449): 282-285.

[28] Shen Huawei, Wang Dashun, Song Chaoming, et al. Modeling and predicting popularity dynamics via reinforced poisson processes[C]// Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Québec: AAAI Press, 2014: 291-297.

[29] Gao Shuai, Ma Jun, Chen Zhumin. Modeling and predicting retweeting dynamics on microblogging platforms[C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2015: 107-116.

[30] Wang Hao, Li Yiping, Feng Zhuonan, et al. Retweeting analysis and prediction in microblogs: An epidemic inspired approach[J]. China Communications, 2013, 10(3): 13-24.

[31] Feng Zhuonan, Li Yiping, Li Jin, et al. A cluster-based epidemic model for retweeting trend prediction on micro-blog[C]// Proceedings of International Conference on Database and Expert Systems Applications. New York: IEEE Press, 2015: 558-573.

[32] Tatar A, Leguay J, Antoniadis P, et al. Predicting the popularity of online articles based on user comments[C]// Proceedings of the International Conference on Web Intelligence. New York: IEEE Press, 2011: 671-678.

[33] Cheng Hui, Liu Yun. An online public opinion forecast model based on time series[J]. Journal of Internet Technology, 2008, 9(5): 429-432.

[34] Sun Lingfang, Zhou Jiabo, Lin Weijian, et al. On network public opinion crisis early warning based on the BP neural network and genetic algorithm[J]. Journal of Intelligence, 2014, 11(3): 18-24

[35] Hajiloo R, Salarieh H, Alasty A. Chaos control in delayed phase space constructed by the takens embedding theory[J]. Communications in Nonlinear Science and Numerical Simulation, 2018, 54(2): 453–465.

[36] 魏德志, 陈福集, 郑小雪. 基于混沌理论和改进径向基函数神经网络的网络舆情预测方法[J]. 物理学报, 2015, 64(11): 110503.

[37] 黄 敏, 胡学钢. 基于支持向量机的网络舆情混沌预测[J]. 计算机工程与应用, 2013, 49(24): 130-134.

[38] Chatzopoulou G, Sheng Chen, Faloutsos M. A first step towards understanding popularity in YouTube[C]// Proceedings of 2010 INFOCOM IEEE Conference on Computer Communications Workshops. New York: IEEE Press, 2010: 1-6.

[39] 徐鹏, 林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报, 2009, 20(10): 2692-2704.

[40] Nguyen-Thi A T, Nguyen P Q, Ngo T D, et al. Transfer adaboost SVM for link prediction in newly signed social networks using explicit and PNR features[J]. Procedia Computer Science, 2015, 60(1): 332-341.

[41] Saito K, Nakano R, Kimura M. Prediction of Information diffusion probabilities for independent cascade model[C]// Proceedings of International Conference on Knowledge-Based Intelligent Information and Engineering Systems. Berlin: IEEE Press, 2008: 67-75.

[42] Li Xiaolong, Pan Gang, Wu Zhaohui, et al. Prediction of urban human mobility using large-scale taxi traces and its applications[J]. Frontiers of Computer Science, 2012, 6(1): 111-121.

[43] Qiu Xueheng, Zhang Le, Suganthan P N, et al. Oblique random forest ensemble via least square estimation for time series forecasting[J]. Information Sciences, 2017, 420: 249-262.

[44] Eberly L E. Multiple linear regression[J]. Methods in Molecular Biology, 2007, 404(2): 165-168.

[45] Pastor-Satorras R, Castellano C, Van Mieghem P, et al. Epidemic processes in complex networks[J]. Reviews of Modern Physics, 2015, 87(3): 925-979.

[46] Lee C, Garbett A, Wilkinson D J. A network epidemic model for online community commissioning data[J]. Statistics and Computing, 2018, 28(4): 891-904.

[47] Zhao Yanan, Jiang Daqing. The threshold of a stochastic SIS epidemic model with vaccination[J]. Applied Mathematics and Computation, 2014, 243(1): 718-727.

[48] Zhu Kai, Ying Lei. Information source detection in the SIR model: A sample path based approach[J]. IEEE/ACM Transactions on Networking, 2012, 24(1): 408-421.

[49] Rosenstein M T, Collins J J, Deluca C J. A pratical method for calculating largest layapunov exponents from small data sets[J]. Physica D: Nonlinear Phenomena, 1993, 65: 117-134.

[50] Maron G, Dante A C, Ramos E A. Spatial cognition degree of development classication using articial neural networks and largest Lyapunov exponents[C]// Proceedings of Brazilian Congress on Computational Intelligence. New York: IEEE Press, 2013: 495-500.

[51] Hajiloo R, Salarieh H, Alasty A. Chaos control in delayed phase space constructed by the takens embedding theory[J]. Communications in Nonlinear Science and Numerical Simulation, 2018, 54(11): 453-465.

[52] Cheng Anyu, Jiang Xiao, Li Yongfu, et al. Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 466(8): 422-434.

[53] Kim H S, Eykholt R, Salas J D. Delay time window and plateau onset of the correlation dimension for small data sets[J]. Physical Review E, 1998, 58(5): 5676-5682.

[54] Chen Zhuo, Lu Chen, Zhang Wenjin, et al. A chaotic time series prediction method based on fuzzy neural network and its application[C]// Proceedings of International Workshop on Chaos Fractals Theories and Applications. Kunming: IEEE Press, 2010: 355-359.

[55] Zhang Zeyin, Wang Ting, Liu Xinggao. Melt index prediction by aggregated RBF neural networks trained with chaotic theory[J]. Neurocomputing, 2014, 131(9): 368-376.

[56] Grassberger P, Procaccia I. Measuring the strangeness of strange attractors[J]. Physica D: Nonlinear Phenomena, 2004, 9(1): 189-208.

[57] Gao Shuai, Ma Jun, Chen Zhumin. Modeling and predicting retweeting dynamics on microblogging platforms[C]// Proceedings of ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2015: 107−116.

[58] 丛蕊, 刘树林, 马锐. 基于数据融合的多变量相空间重构方法[J]. 物理学报, 2008, 57(12): 7487-7497.

[59] Qu Jianling, Wu Xiaofei, Qiao Yuchuan, et al. An improved local weighted linear prediction model for chaotic time series[J]. Chinese Physics Letters, 2014, 31(2): 020503.

[60] 田中大, 高宪文, 石彤. 用于混沌时间序列预测的组合核函数最小二乘支持向量机[J]. 物理学报, 2014, 63(16): 66-76.

[61] Liebig J, Rao A. Predicting item popularity: Analysing local clustering behaviour of users[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 442(13): 523-531.

[62] Figueiredo F, Almeida J M, Goncalves M A. Trendlearner: Early prediction of popularity trends of user generated content[J]. Information Sciences, 2016, 350(18): 172-187.

[63] Liu Wei, He Min, Wang Lihong, et al. Research on microblog retweeting prediction based on user behavior features[J]. Chinese Journal of Computers, 2016, 39(10): 1992-2006.

[64] 肖云鹏, 李松阳, 刘宴兵. 一种基于社交影响力和平均场理论的信息传播动力学模型[J]. 物理学报, 2017, 66(3): 233-245.

致谢

就像是流星划过天空一般，三年的研究生生活转瞬即逝，已经接近尾声，好想抓住毕业的尾巴，让时间静止在这一刻。在这秀丽的南山之上，樱花盛开的重邮，有刚进校园的迷茫、有课题项目下的压力、有朋友争吵的沮丧，也有温馨暖人的鼓励、有论文发表的兴奋、有获得奖学金的欣喜，这里是我播种和收获的热土。想借此论文，像研究生三年给予我关心和支持的可爱人儿致以最诚挚的感谢。

感谢我的导师 XXX 教授，他是我学术研究的敲门人，助我开启学习路上另一扇别样的大门。同时，特别感谢我的指导老师 XXX 教授，他对于我是老师更是家人。还记得，刚进实验室，由于技术水平跟不上而沮丧时，XXX 及时给予我暖心的鼓励以及学习方面的建议，此刻他是亲切的。还记得，小论文撰写时，XXX 跟我一起讨论和解决问题，为论文提出很多珍贵的建议，此刻他是严谨的、学识渊博的。在研究生学习，XXX 老师是跟我讨论学术问题最多的人，我的成长以及取得的成果，都离不开他的潜心指导。由衷的感谢 XXX 老师，您辛苦了。

感谢实验室的小伙伴，他们让我感受到了实验室大家庭的温暖，是他们陪我一起为项目、小论文、工作而努力奋斗，也是他们陪我一起收获和庆祝胜利的果实，感谢研究生生活一路有他们。感谢我的室友 XXX，每一次促膝长谈，都能让我豁然开朗。感谢 XXX 同学，是他在我考研和读研期间一直陪着我，不论是在生活还是学习中对我都无微不至，他也是我学习的榜样，希望以后继续和你携手共进步，平安简单快乐。

感谢我的家人，每次从家赶往千里外的重庆时，在离别的车站，总是让我想到陆游的《游子吟》，他们在我求学路上默默地给予我最伟大的关怀和支持，他们是我努力和前进的动力。

最后，感谢评阅本论文以及答辩组的专家和老师，向您们致以崇高的敬意和真挚的感谢。

攻读硕士学位期间从事的科研工作及取得的成果

参与科研项目：

[1] XXX，重庆市基础科学与前沿技术研究项目, 2017-2020

[2] XXX，重庆市教委科学计划项目, 2015-2017

[3] XXX，重庆市科委重点研发项目, 2017-2019

[4] XXX，重庆市科委重点研发项目, 2017-2019

[5] XXX，重庆市研究生教改项目, 2018-2021

发表及完成论文：

[1] XXX, **XXX**, XXX, et al. XXXXX[J]. Physica A: Statistical Mechanics and its Applications, 2019. (SCI期刊，已录用)

[2] XXX, **XXX**, XXX, et al. XXXXX[P]. 中国，受理号: XXXX, 2017.12.20.

[3] **XXX**, XXX, XXX, et al. XXXXX[P]. 中国，受理号: XXXX, 2019.02.14.

[4] XXX, XXX, XXX, **XXX**, et al. XXXXX[P]. 中国，受理号: XXXX, 2017.12.20.

[5] XXX, XXX, XXX, **XXX**, et al. XXXXX[P]. 中国，受理号: XXXX, 2019.02.14.

[6] XXX，XXX，XXX，XXX，**XXX**，et al. XXXXX[P]. 中国，受理号：XXXX, 2019.02.14.

获奖：

[1] 重庆邮电大学2017年度二等奖学金