**中图分类号：TP391**

**论文编号：10006ZY1306311**



硕士学位论文

基于背景热点的微博转发预测研究

作者姓名 陈 江

学科专业 计算机技术

指导教师 巢文涵 讲师

王丽宏 教授级高工

培养院系 计算机学院

**Research on Microblog Forwarding Prediction**

**Based on Hot Topics**

A Dissertation Submitted for the Degree of Master

**Candidate：Chen Jiang**

**Supervisor：Lecturer Chao WenHan**

School of Computer Science & Engineering

Beihang University, Beijing, China

**­­­­中图分类号：TP391**

**论文编号：10006ZY1306311**

硕 士 学 位 论 文

基于背景热点的微博转发预测研究

作者姓名 陈 江 申请学位级别 工学硕士学位

指导教师姓名 巢文涵 职 称 讲师

王丽宏 教授级高工

学科专业 计算机技术 研究方向 信息安全

学习时间自 2013年 09月 05日 起至 2016年 03月 01日止

论文提交日期 2015年 11月 30日 论文答辩日期 2015年 12月 21日

学位授予单位 北京航空航天大学 学位授予日期 年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名： 日期： 年 月 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘 要

从2009年发布至今，微博已经演化为一个成熟的被广泛接受的自媒体平台，是人们在移动互联网时代获取信息的最及时、最便捷的渠道。微博是人们发布资讯、讨论社会热点、发表个人观点最集中的网络平台，研究微博用户的转发行为对舆情监测、商业营销、行为预测等具有重要的意义。用户转发行为受多方面的影响，包括用户自身的行为特点、微博的内容特征、所处的社交网络环境等，此外，在热点事件背景下，用户转发行为也会受到热点事件的发生及发展趋势的影响。本文基于背景热点话题研究用户转发行为并提出基于背景热点话题的微博转发预测方法。本文的主要工作包括以下几方面：

1. 基于背景热点话题的微博用户转发行为影响因素分析。微博是否会被转发与用户个体行为和用户对微博背景知识的认知具有紧密相关性；用户所掌握的微博背景知识一方面由历史微博获取，一方面由用户对微博内容的综合认知程度决定；背景热点话题是影响用户对微博综合认知的重要因素。本文从以下几方面分析用户转发行为：a. 基于背景热点话题的用户转发兴趣对用户转发行为的影响；b. 基于背景热点话题的用户行为特征对用户转发行为的影响；c. 基于背景热点话题的微博内容特征对用户转发行为的影响。实验表明，基于背景热点的微博转发特征与用户转发行为之间具有密切关系。

2. 基于背景热点话题的微博转发特征提取。根据背景热点话题对微博用户转发行为的影响，提取能有效区分用户转发行为的转发特征。本文基于背景热点话题提取了如下几个特征：a. 基于背景热点话题的用户转发兴趣特征；b. 基于背景热点话题的用户转发行为特征；c. 基于背景热点话题的微博内容特征。

3. 提出基于背景热点话题的微博转发预测方法。通过融合基于背景热点话题的用户转发兴趣特征、基于背景热点话题的用户转发行为特征和基于背景热点话题的微博内容特征等三类微博转发特征，对用户转发行为进行预测。实验表明，将基于背景热点话题的微博转发特征用于微博转发行为预测，能有效提高微博转发预测的效果。

最后，本文基于以上工作设计和实现了微博转发预测原型系统，系统包括数据采集、特征提取、微博转发预测等主要功能模块，实现了微博数据采集、背景热点数据采集、背景热点提取、微博转发特征提取、微博转发预测等功能。

关键词：背景热点、转发行为、转发预测

Abstract

After released in 2009, microblog has evolved into a widely accepted self-media, now it’s the most timely and convenient channels for people to get information. Besides, microblogging is the most centralized network platform where people distribute information, discuss social hot spots and express their personal point of view. Microblog is the network platform on which people most intensively post information, discuss social hotspot, and express their personal point of view. Studies on microblog users' forwarding behavior are of great significance in public opinion monitoring, business marketing, behavior prediction, etc. User’s forwarding behavior is influenced by a wide range of facts, including their own behavior characteristics, feature of content, and the network environment where there are placed in. And it’s also proved that people’s forwarding behavior can be influenced with the background hot topics (BHT). We study the forwarding behavior based on the BHT, and proposed a method to forward prediction based on the BHT.

The main work of this paper is listed as follow:

1. We studied the affecting factors on users’ forwarding behavior based on BHT. Whether a microblog will be forwarded or not depends not only on individual user characteristic, but also on the users’ knowledge about the background involved in microblog content. The knowledge about background of a microblog people have in mind was acquired by historical microblog on one hand, on the other hand, it also comes from the users’ comprehensive awareness of a microblog content. BHT is one of the factors that matters a lot to people’s integrated cognition about microblog. In this article, we study the users’ forward behavior from the aspects as follow: (a) the impact of users’ forwarding interest based on BHT on users’ forwarding behavior. (b) The influence on users’ forwarding behavior caused by the characteristics of users’. (c) The impact of content feature based on BHT on users’ forwarding behavior. Our experiments show that users’ forwarding behavior has a close relationship to the features based on BHT.

2. Microblog forwarding feature extraction base on BHT. We extracted BHT based features feature that can effectively distinguish users’ forwarding behavior according to the impact of BHT on users’ forwarding behavior. These features are listed as follow: (a) users’ forwarding interest based on BHT. (b) The characteristic of users’ forwarding behavior base on BHT. (c) Content feature based on BHT.

3. We proposed BHT based microblog forwarding prediction method. Our method predicts the users’ forwarding behavior based on BHT based forwarding interest, BHT based characteristic of users’ forwarding behavior, and BHT based content features. Our experiments show that our method performs well and effectively improved the microblog forwarding prediction results.

Finally, we designed and implemented a prototype system to predict whether a microblog will be forwarded or not based on all of our works above. The system works with functions like data crawling, feature extraction, microblog forwarding prediction.

**Keywords**: Background hot topics, Forwarding prediction, Microblog forwarding

目 录

[第一章 绪论 1](#_Toc437504500)

[1.1 选题背景及研究意义 1](#_Toc437504501)

[1.2 国内外研究现状 2](#_Toc437504502)

[1.2.1 基于特征提取的微博转发预测研究 2](#_Toc437504503)

[1.2.2 基于社交网络结构的微博转发预测研究 3](#_Toc437504504)

[1.3 本文的研究内容 4](#_Toc437504505)

[1.4 本文结构安排 5](#_Toc437504506)

[1.5 本章小结 6](#_Toc437504507)

[第二章 微博转发预测研究的相关技术和方法 7](#_Toc437504508)

[2.1 概述 7](#_Toc437504509)

[2.2 背景热点话题相关技术 7](#_Toc437504510)

[2.2.1 背景热点获取 8](#_Toc437504511)

[2.2.2 背景热点聚类 10](#_Toc437504512)

[2.2.3 背景热点表示 11](#_Toc437504513)

[2.3 微博转发预测相关技术和方法 13](#_Toc437504514)

[2.3.1 基于用户特征的微博转发预测 13](#_Toc437504515)

[2.3.2 基于内容特征的微博转发预测 16](#_Toc437504516)

[2.3.3 基于结构特征的微博转发预测 18](#_Toc437504517)

[2.3.4 微博转发预测中的分类方法 19](#_Toc437504518)

[2.4 本章小结 25](#_Toc437504519)

[第三章 基于背景热点的微博转发行为影响因素分析 26](#_Toc437504520)

[3.1 概述 26](#_Toc437504521)

[3.2 背景热点话题相关定义 27](#_Toc437504522)

[3.3 背景热点话题对微博用户转发行为的影响分析 28](#_Toc437504523)

[3.4 基于热点话题的影响因素分析 29](#_Toc437504524)

[3.4.1 基于背景热点的用户转发兴趣对转发行为的影响 29](#_Toc437504525)

[3.4.2 融合背景热点话题的用户行为一致性对转发行为的影响 31](#_Toc437504526)

[3.4.3 基于背景热点的微博内容对转发行为的影响 33](#_Toc437504527)

[3.5 本章小结 34](#_Toc437504528)

[第四章 基于背景热点的微博转发特征提取及转发预测 35](#_Toc437504529)

[4.1 概述 35](#_Toc437504530)

[4.2 数据集构建 35](#_Toc437504531)

[4.3 微博转发预测特征选取 36](#_Toc437504532)

[4.3.1 基于用户和微博内容的转发特征 36](#_Toc437504533)

[4.3.2 基于背景热点话题的转发特征 38](#_Toc437504534)

[4.4 微博转发预测 40](#_Toc437504535)

[4.4.1 微博转发预测基本思路及评价指标 40](#_Toc437504536)

[4.4.2 微博转发预测 41](#_Toc437504537)

[4.5 实验与结果分析 43](#_Toc437504538)

[4.6 本章小结 47](#_Toc437504539)

[第五章 基于背景热点微博转发系统设计与实现 48](#_Toc437504540)

[5.1 概述 48](#_Toc437504541)

[5.2 总体系统设计 48](#_Toc437504542)

[5.3 数据采集 49](#_Toc437504543)

[5.4 数据存储 50](#_Toc437504544)

[5.5 微博转发预测模型 51](#_Toc437504545)

[5.5.1 背景热点提取 52](#_Toc437504546)

[5.5.2 提取微博转发特征 53](#_Toc437504547)

[5.6 微博转发预测 55](#_Toc437504548)

[5.7 本章小结 55](#_Toc437504549)

[总结与展望 56](#_Toc437504550)

[参考文献 58](#_Toc437504551)

[攻读学位期间取得的成果 63](#_Toc437504552)

[致谢 64](#_Toc437504553)

**图目录**

[图 1 微博转发预测概述图 7](#_Toc437504554)

[图 2 网络资源表示图 8](#_Toc437504555)

[图 3 网络爬虫数据采集过程 9](#_Toc437504556)

[图 4 互联网网页关系图 15](#_Toc437504557)

[图 5 微博用户关系图 15](#_Toc437504558)

[图 6 一个线性可分数据集上的可能决策边界 20](#_Toc437504559)

[图 7 决策边界的边缘 20](#_Toc437504560)

[图 8 线性不可分情况下SVM的决策边界 22](#_Toc437504561)

[图 9 背景热点话题期间微博转发量变化趋势 28](#_Toc437504562)

[图 10 背景热点话题对各类用户微博转发量的影响 29](#_Toc437504563)

[图 11 微博转发量对应融合背景热点话题的用户转发兴趣的关系图 31](#_Toc437504564)

[图 12 微博转发量对应行为一致性特征的关系图 33](#_Toc437504565)

[图 13 微博转发对应融合热点话题的微博内容特征对应的关系图 34](#_Toc437504566)

[图 14 基本微博转发预测流程图 40](#_Toc437504567)

[图 15 基于基本特征的对比实验流程图 41](#_Toc437504568)

[图 16 基于背景热点微博转发特征的转发预测实验流程图 42](#_Toc437504569)

[图 17 多特征融合的转发预测实验流程图 43](#_Toc437504570)

[图 18 基于背景热点微博转发预测总体系统设计图 48](#_Toc437504571)

[图 19 数据采集模块流程图 49](#_Toc437504572)

[图 20 微博转发预测模型流程图 52](#_Toc437504573)

**表目录**

[表 1 协同过滤评分矩阵 18](#_Toc437504574)

[表 2 不同微博用户转发同一条微博的被转发情况 36](#_Toc437504575)

[表 3 基于用户特征和内容特征转发预测实验结果 44](#_Toc437504576)

[表 4 基于背景热点的微博转发预测实验结果 44](#_Toc437504577)

[表 5 增加单个特征变量对比实验结果(Precision) 45](#_Toc437504578)

[表 6 增加单个特征变量对比实验结果(Recall) 45](#_Toc437504579)

[表 7 增加单个特征变量对比实验结果(F-Measure) 45](#_Toc437504580)

[表 8 逐个增加特征变量对比实验结果(Precision) 46](#_Toc437504581)

[表 9 逐个增加特征变量对比实验结果(Recall) 46](#_Toc437504582)

[表 10 逐个增加特征变量对比实验结果(F-Measure) 46](#_Toc437504583)

[表 11 背景热点新闻表 50](#_Toc437504584)

[表 12 背景热点表 50](#_Toc437504585)

[表 13 微博消息存储表 51](#_Toc437504586)

[表 14 用户信息存储表 51](#_Toc437504587)

# 绪论

## 选题背景及研究意义

微博作为一个社交媒体从2009年发布至今，已经演化成一个重要的自媒体平台。微博是人们获取信息的重要渠道，也是人们讨论社会热点、发表观点的最具代表性的网络平台。根据2015年7月《CNNIC：2015年第36次中国互联网络发展状况统计报告》，截止2015年6月，我国微博的用户量达到2.04亿，其中手机移动端的用户数为1.62亿，使用率为27.3%，用户之间通过关注形成复杂的关系网络。

在微博平台中，用户之间通过关注关系构成错综复杂的网络结构，用户通过转发微博传播信息，这种传播方式具有传播快、覆盖广的特点，这使得某些信息能够在微博中快速扩散，短时间内获得极大的关注。因此，微博转发研究对话题检测、热点跟踪、舆情监控以及商业营销具有重要价值。

目前针对微博转发的研究主要基于网络结构或基于微博特征，前者通过分析微博网络中信息传播的特点，研究微博转发问题，但该方法局限于微博网络复杂而庞大，难以获得完整的网络结构，而基于部分网络结构数据往往造成较大的偏差。后者通过分析影响微博转发的因素，构建微博转发模型，该类主要针对用户静态属性或消息特征来预测消息是否会被转发，没有充分考虑待预测用户的个体差异和背景知识对转发决策的影响。

成立之初，微博是一个社交平台，经过多年的发展与演化，微博已经成为一个成熟的被广泛认可的自媒体平台。人们可以通过微博及时地获取信息、讨论社会热点、发表自己的观点，微博已经成为人们表达观点最集中的网络平台。这种观点表达方式，由于受微博字数少，篇幅小的限制，加上上网的便捷性，使得一些具有代表性和区分性的观点能够在微博平台迅速传播，形成不同的舆论阵营。新闻报道或者事件的发生以简明概要的信息表示方式在微博平台传播，能够很快引导人们对事件的关注，这使得事件更容易发展成为集聚各方舆论观点的热点事件，而热点事件的不同观点反过来也会促使人们更多地关注事件，影响微博用户的转发行为。所以热点事件的发展和用户转发行为是一个交替促进、交替影响的过程。

事实上，用户阅读到一条微博时，会根据自己已有知识对微博价值和新颖性进行判断，然后决定是否转发。微博是否会被转发与用户个体行为和用户对微博的背景知识具有紧密相关性，用户所掌握的微博背景知识一方面由历史微博获取，一方面由用户对微博内容的综合认知程度决定，而用户对微博内容的综合认知程度受多种复杂因素影响，现实生活中正在发生的或者已经发生的热点话题是其中极其重要的影响因素。

基于背景热点事件用户转发行为，能够更好地了解用户行为动机，更好地了解事件发展趋势。这对话题检测、热点跟踪、舆情监控以及商业营销具有重要价值。

## 国内外研究现状

### 基于特征提取的微博转发预测研究

微博从社交平台发展成为被广泛认可的自媒体平台，是消息传播的重要渠道，已经成为人们表达观点最集中的网络平台。为了体现微博用户的个性化特点，微博平台允许用户设置昵称、头像、教育信息、职业信息、个人标签、个性域名、个人简介等信息，当用户通过发布和转发微博参与到信息流通过程中，用户又会获得粉丝数、关注数、关注列表、粉丝列表、评论列表、转发数量、点赞数量等属性，这些特征微博用户在微博平台的社交关系，形成不同的社区，构成虚拟的网络社会。传统互联网的普及，移动互联网的快速发展，使得人们访问互联网越来越便捷，现在已经是一个全新的自媒体时代，每个人都是媒体。微博在信息传播、公众生活、商业营销、经济发展、政治影响等众多与全体社会人息息相关的众多领域扮演着重要角色。基于微博基本特征的研究包括微博转发预测、营销推荐分析、人物关系挖掘、用户行为分析、意见领袖等。国内的研究主要基于新浪平台，而国外的研究主要基于Twitter、Facebook平台。

在微博转发预测的研究中，Petrovic S[[[1]](#endnote-2)]等人提取tweet作者相关的粉丝数、关注数、收藏数等属性以及和tweet相关的标签、URL数、tweet长度等属性，基于tweeter平台，采用passive-aggressive算法，研究tweet的转发预测，得出社交属性在微博转发预测研究中是一个具有良好区分度的特征的结论。李英乐[[[2]](#endnote-3)]和曹玖新[[[3]](#endnote-4)]等人提取用户的粉丝数量属性、用户影响力、用户转发活跃度及用户亲密程度等作为转发特征，同时考虑用户历史微博的内容特征以提取用户转发兴趣，并且基于TF-IDF考虑微博内容的重要性，研究微博用户的转发预测行为。Kanavos A[[[4]](#endnote-5)]等人提取tweet用户的粉丝数、微博转发数量和原创数量、用户活跃度及tweet中标签数量等特征，并且结合Perikos[[[5]](#endnote-6)]等人提出的情感分析模型，基于Ekman P[[[6]](#endnote-7)]等人的六类情感体系进行情感分析，研究微博传播的广度和深度的预测问题。Yang Z[[[7]](#endnote-8)]等人分析用户转发兴趣、转发活跃度以及微博重要性等因素对用户转发决策的影响，构建因子图[[[8]](#endnote-9),[[9]](#endnote-10),[[10]](#endnote-11)]模型，对用户转发行为进行预测。Boyd D[[[11]](#endnote-12)]等人以对话形式看待微博转发，对用户为何转发微博以及转发什么样的微博进行了研究分析。Suh B[[[12]](#endnote-13)]等人量化比较分析微博转发的影响因素，他们的研究结果表明内容特征、是否含URL和标签以及关注数和粉丝数等与微博转发可能性有极大的关系，基于这些量化的影响因子，利用广义线性模型他们构建微博转发预测模型。张旸[[[13]](#endnote-14)]等人根据用户粉丝数量、用户被提及次数、微博长度、微博被转发比例以及微博是否包含标签等因素对微博转发影响程度的不同，构建了基于特征加权的微博转发预测模型。

Hong L[[[14]](#endnote-15)]等人提取微博内容特征、微博用户特征以及用户所在网络结构特征等微博转发影响因子，构建多分类模型，研究微博流行度预测问题。Bandari R[[[15]](#endnote-16)]等人则通过提取新闻所属类别、新闻的语言特性、是否包含实体以及新闻来源等特征构建多维度向量，研究新闻在微博网络中的流行度问题。Ma H[[[16]](#endnote-17)]等人以微博用户为出发点构建数据集，研究分析微博及标签的转发频度和发展趋势，对微博流行度进行预测，他们的研究表明外部热点事件的发生会对微博的流行程度造成很大的影响。Sanli C[[[17]](#endnote-18)]等人通过研究社交平台用户之间的沟通规律来研究社交网络中谣言的传播规律。Zhao Q[[[18]](#endnote-19)]等人构建无须训练和特征提取的自激点过程统计模型，准确而高效低预测Twitter中推文的转发规模。Gao S[[[19]](#endnote-20)]等人构建加强的泊松过程模型研究微博转发在时间上的动态变化过程，有效地预测了信息在微博平台传播的流行度。Ma Z[[[20]](#endnote-21)]等人通过提取热点标签的内容特征及上下文特征，采用朴素贝叶斯[[[21]](#endnote-22)]、KNN[[[22]](#endnote-23)]、决策树[[[23]](#endnote-24)]、SVM[[[24]](#endnote-25),[[25]](#endnote-26)]以及逻辑回归等分类算法，提出了预测热点标签流行度的方法。

### 基于社交网络结构的微博转发预测研究

转发行为是促使信息在微博网络中病毒式传播[[[26]](#endnote-27),[[27]](#endnote-28)]的主要原因，信息的传播将用户和用户、用户与事件关联在一起，加上用户之间通过关注与被关在构建起来的联系，是微博成为一个具有复制现实社会功能性质的虚拟社交网络。

Galuba W[[[28]](#endnote-29)]等人研究微博平台的用户网络中路径的长度、用户的活跃度、URL的关注度等问题，探索URL在微博网络中的传播情况，从而对URL未来在微博网络中的传播情况及流行度做出预测。Yang J[[[29]](#endnote-30)]等人以“@微博ID”作为用户之间的联系，构建网络结构图，研究话题信息在微博网络中的传播规律，并对信息传播的速度、规模以及范围进行预测。Cheng J[[[30]](#endnote-31)]等人根据微博客的转发路径构建转发关系图，同时考虑内容等其他特征，对信息传播链以及传播规模进行预测。Goldenberg J[[[31]](#endnote-32)]和Kempe D[[[32]](#endnote-33)]等人基于包含活跃结点和非活跃结点的有向图研究信息传播，分别构建了Independent Cascades(IC)模型和Linear Threshold(LT)模型。Bakshy E[[[33]](#endnote-34)]等人基于关注关系构建图模型，基于图模型研究信息传播规律，他们的研究结果表明具有粉丝多的用户和名人在信息传播过程中具有更大的影响力。Jin F[[[34]](#endnote-35)]等人研究信息的传播方式和传播路径，基于SEIZ流行病模型，研究社交网络中有害信息和谣言的传播。除了信息传播领域外，基于社交网络的研究还大量应用于意见领袖、人物关系挖掘以及社区发现[[[35]](#endnote-36),[[36]](#endnote-37)]等领域。

综合国内外研究现状，现有研究的出发点主要有两个方面：基于微博和用户的基本特征、基于社交网络结构特征，这些研究工作将微博平台视作一个独立系统，不受其他渠道信息影响。事实上，Yang Z [7]以及Ma h16]等人的研究工作表明，当有突发话题发生时，微博传播很大程度上会收到外界信息的影响。微博是否会被转发与用户个体行为和用户对微博的知识背景具有紧密相关性。用户所掌握的微博背景知识一方面由历史微博获取，一方面由用户对微博内容的综合认知程度决定，而用户的认知是一个综合个人知识和个人社会背景认识事物的复杂过程，影响因素众多。其次，通过微博内容与用户兴趣相似度判断用户转发的方法，往往因为微博内容非常短，所含内容特征有限，使得微博与用户兴趣之间的相似度计算准确性低，转发行为预测准确性低。

## 本文的研究内容

本文研究融合背景热点话题的微博转发预测方法。现有关于微博转发预测的研究，仅是将特征提取限于微博平台范围内，没有考虑背景热点话题对微博转发行为的影响。本文基于背景热点研究微博转发行为，提取转发特征，提出基于背景热点的微博转发预测方法。本文主要研究内容包括：

1. 微博转发行为影响因素分析。用户转发行为受多种因素影响，包括微博用户特点、微博特征、用户所处社交网络环境等，此外，Ma H等人的研究也表明外部热点事件的发生会对微博的流行程度造成较大影响。本文研究基于背景热点的微博用户转发行为影响因素，主要包括以下几方面：a. 基于背景热点话题的用户转发兴趣对用户转发行为的影响；b. 基于背景热点话题的用户行为特征对用户转发行为的影响；c. 基于背景热点话题的微博内容特征对用户转发行为的影响。

2. 微博转发特征提取。根据背景热点话题对微博用户转发行为的影响，提取能有效区分用户转发行为的转发特征。本文提取基本的用户特征和微博内容特征，同时，基于背景热点提取微博转发特征，包括：a. 基于背景热点话题的用户转发兴趣特征；b. 基于背景热点话题的用户转发行为特征；c. 基于背景热点话题的微博内容特征。

3. 微博转发预测。根据背景热点话题对微博转发行为的影响，提取基于背景热点话题的用户转发兴趣特征、基于背景热点话题的用户转发行为特征和基于背景热点话题的微博内容特征等微博转发特征，结合支持向量机、贝叶斯、决策树的等机器学习方法，对用户转发行为进行预测，并分析基于背景热点的微博转发特征对转发行为预测能力的影响。

最后，本文基于以上研究内容设计和实现了微博转发预测原型系统，系统包括数据采集、特征提取、微博转发预测等主要功能模块，实现了微博数据采集、背景热点数据采集、背景热点提取、微博转发特征提取、微博转发预测等功能。

## 本文结构安排

本文共有六章，结构安排如下：

第一章 绪论。本章阐述论文选题背景及意义、国内外研究现状和本文研究内容。

第二章 微博转发预测研究的相关技术和方法。本章分两部分介绍微博转发预测研究相关的技术和方法：2.2 背景热点话题相关技术和2.3 微博转发预测相关技术和方法。其中，2.2从背景热点获取、背景热点聚类、背景热点表示以及等方面加以阐述；2.3从基于用户特征的微博转发预测、基于内容特征的微博转发预测、基于结构特征的微博转发预测以及微博转发预测中的分类方法等方面加以阐述。

第三章 基于背景热点的微博转发行为影响因素分析。本章首先通过具体事例研究背景热点话题对微博转发预测的影响，然后从基于背景热点的用户转发兴趣、融合背景热点话题的用户行为一致性、基于背景热点的微博内容三个不同角度研究用户转发行为与背景热点事件的关系。

第四章 基于背景热点的微博转发特征提取及转发预测。本章描述的数据集的构建过程，并从基于用户、微博基本特征的转发预测实验、基于背景热点话题的转发预测实验、多特征融合的转发预测实验三个不同的层面进行实验，研究微博转发预测问题。

第五章 基于背景热点微博转发系统设计与实现。本章的主要内容是系统的设计与实现，包括数据采集、数据存储、微博转发预测模型、转发预测以及界面设计等内容。

总结与展望。总结和分析论文工作，明确下一步研究工作的方向。

论文最后部分是参考文献、攻读学位期间取得的成果和致谢。

## 本章小结

本章中，本文首先阐述了微博转发预测研究的背景及本文选题的意义，然后，对国内外现有关于微博转发预测研究工作进行总结，分类阐述了国内外研究现状，接着，阐明了本文的主要研究内容，最后，梳理了本篇论文的结构。

# 微博转发预测研究的相关技术和方法

## 概述

微博转发预测研究的基本思路是首先提取能有效区分转发行为的转发特征，然后转发特征训练转发模型，最后将转发模型用户转发行为预测。微博转发预测的过程可以概述如图 1。转发特征主要有用户相关特征、微博内容相关特征以及网络结构相关特征等，转发预测模型通常采用贝叶斯、决策树、支持向量机等机器学习方法。

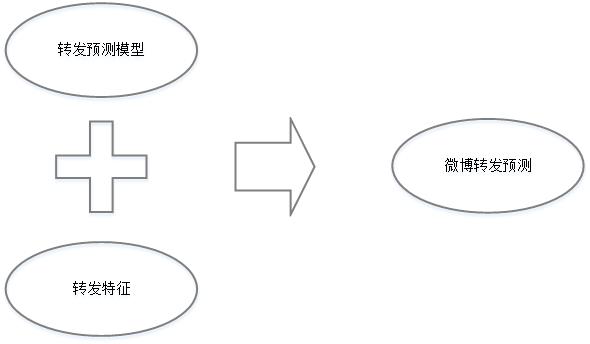


图 1 微博转发预测概述图

## 背景热点话题相关技术

微博平台中用户对微博的转发通常和热点事件联系在一起，或者用户更偏向于转发热点事件相关的微博，或者一些传播路径中对信息传播具有显著促进作用的用户对微博的转发会使微博相关事件热度升级。所以，本文提出基于背景热点研究微博转发预测问题。背景热点是指从微博平台以外的新闻媒体获取的热点话题数据，背景热点数据经过处理后表示为关键词向量，以此表示用户能够从其他渠道获知的微博内容相关的背景知识。本文的背景热点来源主要包括凤凰新闻网热点新闻频道、网易新闻网新闻排行榜频道和新浪网新闻排行频道。

获得背景热点话题主要包括背景热点的获取、背景热点聚类、背景热点提取三个过程。

### 背景热点获取

背景热点获取是指从凤凰网、网易和新浪相关热点新闻频道获取热点新闻的过程。本节将从背景热点获取的基本技术和背景热点获取的相关工具开始，阐述背景热点获取的相关技术。

#### 背景热点获取的基本技术

从新闻热点排行榜频道获取热点新闻时，可以将新闻站点看成是一张图，新闻网页的抓取是一个遍历节点并将网页信息保存到本地的过程。新闻站点构成图的过程是将每个网页作为图的一个节点，页面中的超链接看作是连接图中结点的边。爬取新闻站点的过程可以通过深度优先遍历和广度优先遍历的方式爬取所有站点数据，但是，如果采用深度优先方式，容易因为图中存在的循环而陷入一个“黑洞”。因此，大多数的数据采集都采用广度优先的方式。

宽度优先采集数据往往是从多个种子链接开始的，相对于背景热点获取则是背景热点的多个来源，比如网易新闻网、新浪新闻网和凤凰新闻网。互联网上的每个HTML页面或者文件都可以看作是待爬取的一个结点，而指向该页面或文件的链接可以看做是结点之间的有向边。如图 2所示，表示种子链接，或都表示待爬取的热点新闻频道，或表示具体的热点新闻。

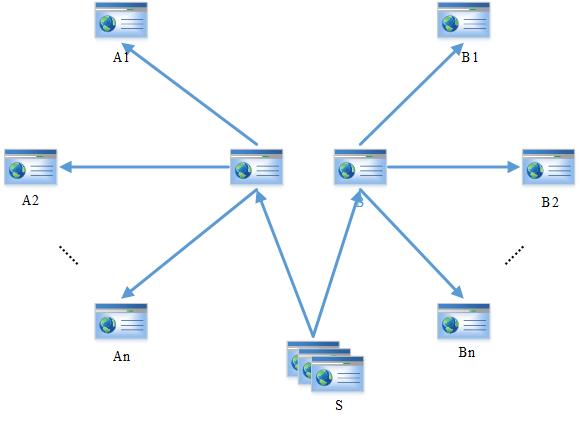


图 2 网络资源表示图

宽度优先采集数据的过程中，对于已经采集过的站点，可以将其URL 放入一张VISITED表，每次需要爬取一个新的链接时，需要通过VISITED表判断链接是否已经被爬取，如果没有爬取，则按照处理页面提取子结点的方式处理该链接，否则，不做处理。

网络爬虫采集数据的过程如图 3所示。以指定的背景热点数据源站点作为初始URL，数据采集的过程中，解析出来的超链接先和VISITED表中已经采集过的URL进行对比，如果已存在于VISITED中，则对其该链接地址，否则，作为新的待爬取的URL地址放入TODO表中，每次都将刚刚采集过的网页对应的URL地址放入VISITED表。

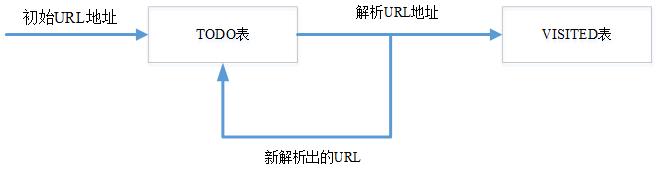


图 3 网络爬虫数据采集过程

#### 背景热点获取的相关工具

背景热点的获取主要基于现有的网络爬虫技术，因此，要获取背景热点，既可以根据网络爬虫的基本原理实现数据爬取，也可以借助于成熟的第三方工具。自己实现背景热点数据获取的技术基础是上一小节阐述的背景热点数据获取基本技术，第三方工具则主要有Java语言实现的和Python语言实现的。

基于Java语言实现的背景热点获取工具：

**Nutch**

Nutch是一个搜索引擎项目，先已在github上开源，遵守Apache开源协议，随着版本的不断更新与改进，最新版本的Nutch，已经不止是一个搜索引擎，它已经具有强大的数据采集功能。同时，Nutch已是一个具有极高性能的数据采集工具，支持大规模的数据采集，并且支持对已采集到数据的高校检索。

**WebCollector**

WebCollector无须配置、提供精简的API，是一个可以用少量代码完成强大的数据采集功能的爬虫框架。

此外，基于Java语言实现的数据采集工具还有WebMagic、JSpider、SpiderMan、Clawler4J等。

基于Python语言实现的背景热点获取工具：

**PySpider**

PySpider提供强大的API，支持Python 2和Python 3，提供强大的WebUI和脚本编辑器，并具有任务监控、项目管理和结果查看等功能。它不局限于静态HTML页面的抓取，对JavaScript页面也具有很好的支持，同时支持任务优先级、任务重启及任务定期抓取等功能，并且采用分布式架构，支持大规模抓取。

**Scrapy**

Scrapy是基于Twisted实现的异步处理框架，用户可以通过模块定制来实现自己的数据采集系统。

此外，基于Python语言实现的数据采集工具还有urlwatch、pyrailgun、gcrawler等。

### 背景热点聚类

对于同一件热点事件，不同的新闻频道会从不同的角度去报道，不同的新闻频道对同一热点事件的报道可能会存在一定程度上的重合。因此，对从多个热点新闻频道获取到的背景热点数据进行聚类处理，既能去除重复的热点，也能从不同的角度描述一个热点事件。背景热点聚类过程中的主要技术是K均值聚类。

#### 基于K-Means算法的背景热点聚类

首先以VSM（向量空间模型）表示获取到的背景热点数据的每篇文档，然后再随机选取K篇文档作为初始质心，然后每篇文档指派到最近的质心，被指派到同一个质心的点集合称为一个簇。然后，重新计算簇的质心，重复以上过程，直到文档簇的质心不再发生变化或簇文档不再发生变化。基本的K-Means算法如下：

**算法1（基本K-Means算法）**

1：选择K个点作为初始质心

2：do:

3： 将每个点指派到最近的质心，形成K个簇

4： 重新计算每个簇的质心

5：while(质心不发生变化)

将文档合并到质心离它最近的文档簇时需要计算文档之间的距离。计算文档之间距离或比较文档之间相似性的方式有欧式距离、用Jaccard距离、余弦距离、编辑距离和海明距离等。

在背景热点聚类的过程中需要多次计算文档簇的质心，所以文档簇的质心是不断变化的。聚类时常使用误差的平方和（SSE）作为聚类的目标函数：



是文档与文档簇的质心的欧式距离。以SSE作为聚类目标函数，则文档向量各维度均值作为质心文档对应维度的值，即可得到文档簇的质心。即，第个簇心的质心可以表示为：



为了减少K均值相似度的计算量，可以采用二分K均值算法来改进K均值算法。二分K均值的基本思路是：首先，所有文档构成一个簇，然后将簇按一定规则分裂成两个簇，然后按此规则一直分裂，直到产生K个簇。

### 背景热点表示

背景热点的表示通常有两种方式：(1)选择背景热点文档集中最具代表性的文档，用一个文档来表示热点；(2)综合考虑背景热点文档集，选择多个词，以多个词表示热点。第一种方式相对简单，可以通过文档包含关键词个数等方式对文档进行打分，选择打分最高的文档作为表示热点的文档即可。第二种方式，选择最能代表这些文档集主要思想的多个词，用以表示热点。选择能表示文档集主题的词的技术包括TF-IDF、互信息、信息增益等多种方式，对于微博数据，由于微博文本短，词稀疏等问题，所以对微博进行分词、去除停用词以及噪音数据等处理之后，将词向量直接作为热点的表示的方式，简单而高效。

#### 基于TF-IDF选择主题的词表示

TF和IDF是用来提取文档特征的常用方法，通常和向量空间模型结合使用。TF(Tern Frequency)表示词在文档中出现的频率，IDF(Inverse Document Frequency)表示特征提取过程中更加重视在其他文档中出现比较少的特征，是文档总数与包含该特征的文档数的比值。TF和IDF的计算方式如下：



其中，是词语在文档中出现的次数。



其中，表示文档总数，表示包含词语的文档数。通常以TF\*IDF作为标准来选择能表示主题的关键词。

#### 基于互信息的主题词选择

互信息能够用来衡量词语主题之间的统计关系，度量词语对于主题的区分度，互信息用于主题词选择的含义是某个词在一个主题中出现的频率高，但在其他主题中出现的频率低，则该词与该主题的互信息较大。即互信息越大，词语属于一个主题的概率就越大。主题词与主题之间的互信息可以表示为：



其中，表示文档总数，表示主题中包含词的文档数，表示主题以外的包含词的文档数，表示主题中不包含词的文档数。选择主题词的方式表示如下：



即选取互信息大于一定阈值的作为表示话题的词集合。

#### 基于信息增益主题词选择

信息增益可以用来衡量主题词为主题提供的信息量的多少，并依此来衡量主题词的重要性。通过信息增益可以将主题词从主体词文档集和非主题词文档集中过滤出来。信息增益基于熵来计算，它表示不考虑主题词时文档集的熵与考虑主题词时文档集的熵的差值：



其中，表示主题出现的概率，表示包含词语的文档出现的概率，表示包含词语的文档包含在主题的文档集合中的概率，表示不包含词语的文档出现的概率，表示不包含词语的文档包含在主题的文档集合中的概率，表示主题个数。选择主题词时，可以先删除信息增益很小的词，其他的词按信息增益从大到小排序，选择信息增益较大的词作为表示主题的词集合。

## 微博转发预测相关技术和方法

微博转发预测基本的方法和思路是提取有效特征，然后利用机器学习算法来预测用户的转发行为。根据特征选取对转发预测技术进行分类，主要有基于用户特征的微博转发预测方法、基于内容特征的微博转发预测方法和基于结构特征的微博转发预测方法。微博转发预测中常用的机器学习方法包括：支持向量机、决策树、贝叶斯分类等。

### 基于用户特征的微博转发预测

基于用户特征的微博转发预测方法主要提取与微博用户相关的用户特征，包括用户粉丝数、关注数、用户影响力、用户活跃度、用户转发活跃的等。其中，用户粉丝数、关注数等式能够直接获取的特征，他们直接或间接地作为转发预测特征。用户影响力、用户活跃度和用户转发活跃度等特征一般通过间接计算获得。一般以用户发布微博的多少或者单位时间内用户发布微博的多少作为用户活跃度，而用户转发活跃度一般以用户转发微博的频繁程度作为度量。用户影响力的计算一般基于PageRank[[[37]](#endnote-38)]算法、HITS[[[38]](#endnote-39)]算法及Influence-passivity[[[39]](#endnote-40)]算法等。

#### PageRank算法用于用户影响力计算

PageRank将网页之间的超链接作为网页权威性的一种体现，它将指向某一个网页的一个超链接作为对该网页的一次代表其重要程度的投票，一个页面的重要性由拥有指向该页面超链接的其他所有页面决定。通过迭代地投票，最后决定页面之间的相对重要性。

用户的鼠标点击是根据用户个人的兴趣爱好和浏览目的决定的，用户浏览某个页面具有随机性。假设有共个页面，则初始化每个页面的PR值为，某个页面，假设含有指向该页面链接的页面集合为，则其PR值可以表示为



其中，是页面中含有的指向其他页面的链接的数量，表示页面的PR值。

Larry Page提出PageRank算法时认为，互联网中用户浏览信息是一个随个人喜好变化的随机过程，用户在有超链接的网页继续点击浏览是一个概率事件，同时，用户从一个没有超链接的网页跳出，跳到其他某一个页面继续浏览也是一个概率事件。为了模拟这种情况，引入了阻尼系数，其含义是，用户在某一个含有链出链接的网页中，点击网页中的链接继续浏览的概率为，停止点击，随机跳出到另一页面开始浏览的概率为。引入阻尼系数之后的PR值计算方式如下：



我们在互联网中浏览网页时，可以通过点击网页中的链接跳转到另一个网页，因此可以用一个有向图表示互联网，如图 4所示。P(i)表示页面，箭头指向表示页面之间的链接关系，有一条P(i)指向P(j)的有向边，则表示P(i)含有P(j)的链接地址。

微博中用户的关系也可以用类似于网页的方式来表示。一个用户用一个结点来表示，用户之间的关注关系用一条有向边来表示，如果用户关注了用户，则在图中添加一条有用户指向用户的有向边。 如图 5所示。

对应于用于网页排名的PageRank算法，可以将每个用户的PR值初始化为用户的粉丝量或者微博转发能力等，用户之间的关注关系类比于网页中的超链接，考虑到即使一个用户关注了另一个用户，用户也并不会百分之百地转发其关注者的微博，如果没有微博转发的存在，这种影响力的传播事实上是没有意义的，所以同样可以引入阻尼系数。

引入阻尼系数后的PR值计算方式如下：



其中，和分别表示用户和用户的影响力，表示关注用户的用户数（即用户的粉丝数），表示用户关注的用户。

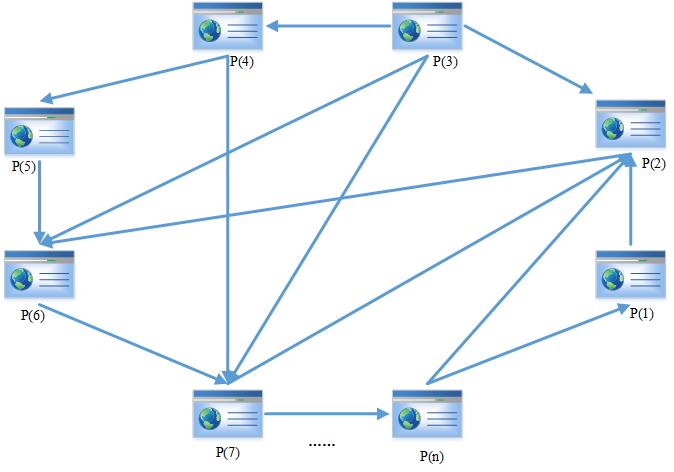


图 4 互联网网页关系图

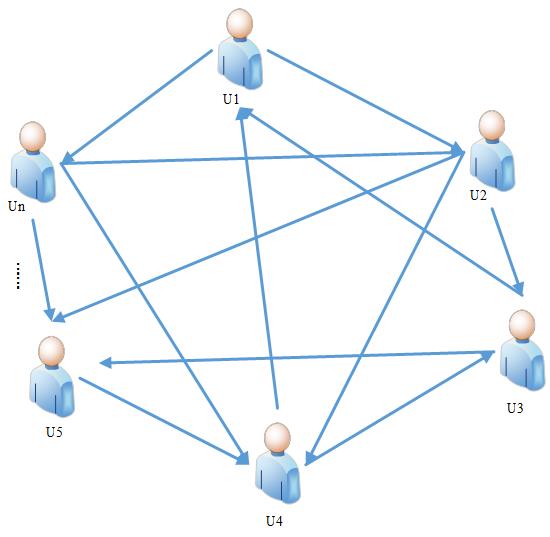


图 5 微博用户关系图

#### HITS算法用于用户影响力计算

HITS即超链接诱导主题搜索算法也是一种与PageRank算法类似的计算网页重要性的算法。该算法利用网页内容的重要程度和网页中所含链接的重要程度即内容权威度(Authority)和链接权威度(Hub)两个指标来衡量一个网页的重要性。某个网页的Authority值是根据含有该网页链接的网页的Hub值计算的，而这个网页的Hub值是根据该网页中所包含超链接对应网页的Authority值计算的，即：





其中，表示引用网页的网页集合，表示网页所包含的链接集合。

Authority值和Hub值的计算过程如下：

1. 初始化所有结点的Authority值和Hub值为1；

2. 更新所有结点的Hub值；

3. 更新所有结点的Authority值；

4. 按结点的Hub值（或Authority值）除以所有结点的Hub值（或Authority值）的方式，对所有结点的Hub值和Authority值进行归一化；

5. 重复步骤2~步骤4直至收敛。

对应到微博用户影响力的计算，用户对应网页，用一个结点表示，用户之间的关注关系对应网页链接，微博用户的影响力可以按如下公式计算：





其中，表示用户关注的用户集合，表示的粉丝用户集合。

### 基于内容特征的微博转发预测

基于内容特征的微博转发预测方法主要选取与内容相关的一些特征，包括是否含有标签、是否含有图片、是否包含URL链接地址、是否参与某个话题、微博长度、提及他人的比例、以及用户兴趣等。其中，除用户兴趣和微博内容重要性之外，其他特征都可以根据微博用户信息和微博内容信息很容易就获得。用户兴趣挖掘包括向量空间模型、协同过滤挖掘用户兴趣[[[40]](#endnote-41),[[41]](#endnote-42)]等。

#### 向量空间模型表示微博用户兴趣

向量空间模型（VSM）的目的是将文档或者文档集表示成一个向量，以便于对非结构化的文档数据进行计算。其基本思路是提取文档或文档集的特征项，然后计算特征项的特征权重，以特征项和特征项权重共同表示文档或文档集。向量空间模型被广泛应用于信息检索、文档相似性计算、话题发现、舆情监测等众多领域。

以向量空间模型表示微博平台中用户的兴趣时，通常将微博用户一段时间内的历史微博作为一个整体，从微博数据提取特征、计算特征权重，然后将用户兴趣表示成特征项和特征权重的向量。记一段时间内用户历史微博为,其中()表示用户的一条历史微博。然后对微博数据进行过滤、提取关键词等预处理，得到历史微博的词集合。针对每个词计算词的特征权重，权重的计算方式有多种，包括TF-IDF、互信息、信息增益等。最后，将历史微博表示成一个向量：，其中表示特征权重。

#### 基于用户的协同过滤的微博用户兴趣挖掘

协同过滤最典型的应用是商品推荐[[[42]](#endnote-43),[[43]](#endnote-44),[[44]](#endnote-45)]、营销推广等领域。协同过滤用于商品推荐时通常是通过计算与用户相关联的其他用户对产品的感兴趣情况，根据其他用户对商品的兴趣程度进行推荐。

协同过滤用于微博平台中用户的兴趣挖掘时，主要通过与某微博用户相关联的其他微博用户的兴趣来确定该用户的兴趣。这主要包括收集数据和寻找关联用户与相似兴趣两个步骤。收集数据主要指收集用户历史数据，包括部分用户（不包括待挖掘用户）的转发兴趣，以及用户之间通过关注等形成的关系。协同过滤的关键是构建协同过滤评分矩阵，如表 1。

所示，然后根据协同矩阵来挖掘用户兴趣。表 1中表示用户可能感兴趣的领域，表示用户,表示用户对领域是否感兴趣或感兴趣程度。

基于用户的协同过滤来挖掘的关键是寻找兴趣相似用户，然后根据相似用户的偏好来挖掘用户兴趣。在微博用户兴趣挖掘的应用中，相似用户可以通过微博用户之间的关注关系获得，用户之间的相似性可以根据用户之间的距离，即用户在关注路径上的距离来表示。如果以用户之间是否有直接关注关系来计算用户相似性，则用户之间只有相似或者不相似两种情况；如果用户之间的相似性也可以通过用户之间的关注距离来计算，例如如果用户关注了用户，用户关注了用户，但是用户没有关注用户，这时可以是用户与用户的关注距离为2。

表 1 协同过滤评分矩阵

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  | …… |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

协同过滤挖掘用户兴趣的基本思路是根据与某用户具有一定偏好相似性的用户来确定该用户的兴趣。用户对领域的兴趣爱好可以用一个能代表用户对各个领域偏好情况的向量来表示，即，其中表示用户对领域是否感兴趣或感兴趣程度。则用户与用户对各领域偏好的相似度可以计算如下：



其中，的计算方式有多种，例如欧几里得距离。通常根据来挖掘用户的兴趣，当用户与用户的兴趣相似度较高时，可以根据用户（或）的兴趣来挖掘用户（或）的兴趣。例如，用户对领域感兴趣，而根据历史信息不能知道对领域也感兴趣，当他们之间的兴趣相似度较高时，可以认为对领域也感兴趣。

### 基于结构特征的微博转发预测

基于结构特征的微博转发预测研究主要选取一些社交网络结构相关的特征，包括社交网络图结构中节点的入度、出度、用户的影响力、用户之间的网络社交关系、用户之间的关系等特征。节点的入度、出度一般以用户的关注数和粉丝数来表示。用户影响力一般采用2.3.1中的一些方法来计算。用户之间的关系有时通过一些图的特征来确定，譬如邻接关系、最大连通子图等确定。用户之间的网络社交关系可以通过用户之间的交互频繁程度来计算。

用户之间的社交交互往往通过评论、转发、私信、提及（@）等方式体现，其中私信属于个人隐私，是无法获取的数据。通常，用户之间的交互频繁程度可以按如下方式：



其中，、、分别表示用户对用户的转发数、评论数和提及次数。

### 微博转发预测中的分类方法

微博转发预测通常是选取能够有效区分微博转发预测的转发特征，然后基于机器学习分类算法训练转发预测模型，最后对用户转发行为做出预测。常用的分类方法有支持向量机、贝叶斯方法和决策树方法等。

#### 支持向量机用于微博转发预测

支持向量机的基本思想是寻找边缘最大化的决策边界,支持向量机中超平面的定义如下：



该方程的子集确定了维空间中的超平面，其中是不全为零的常数。如果存在一个超平面，将数据集所在特征空间划分不同类别的两部分，使得不同类别的数据分属于超平面的两侧，这样的超平面称为决策边界。一个数据集可以对应多个决策边界，如图 6所示，数据集中包含不同类的样本，分别用圆圈和方块表示，他们分别位于每个决策边界的不同侧。

一个数据集上存在多个决策边界，每个边界具有不同的分类能力。根据结构化风险最小理论（structural risk minimization），随着分类能力的增加，泛化误差的上界也随之提高。决策边界的边缘能够刻画分类器的好坏，如图 7所示。因此，设计的SVM分类器，需要满足：决策边界的边缘最大化。

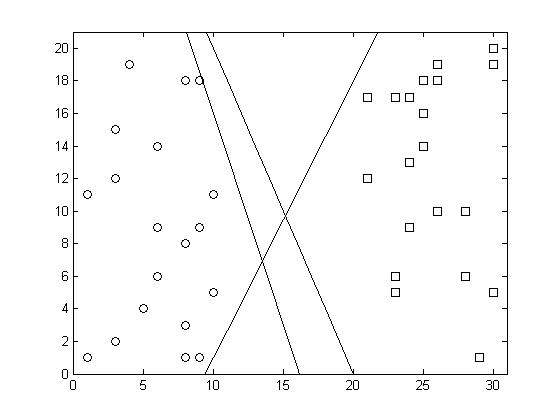


图 6 一个线性可分数据集上的可能决策边界

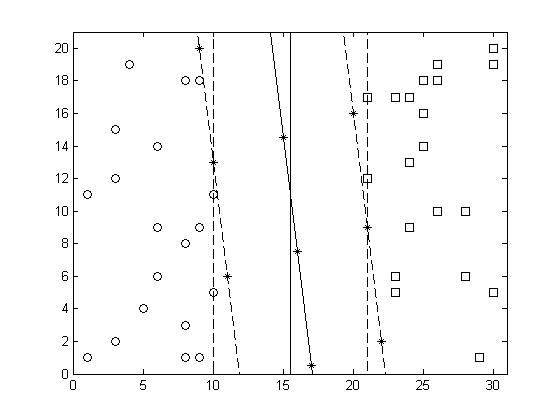


图 7 决策边界的边缘

线性SVM分类器的目标是寻找具有最大边缘的超平面，因此也被称为最大边缘分类器。一个线性分类器的决策边界表示为：



平行于决策边界的两个超平面可以表示为：





本文可以通过计算这两个超平面之间的距离来计算边缘。为了计算边缘，令是超平面上的数据点，是超平面上的数据点，将和分别代入和的表达式中，则边缘可以通过两式相减得到：







最大化边缘等价于最小化目标函数：。因此，线性可分情况下，边缘最大化等价于以下求最小值的包含约束条件的最优化问题：



这样，可以改写目标函数为：



其中，参数是拉格朗日乘子。为了最小化拉格朗日函数，对关于和求导，并令其值为零：





将约束条件进行KTT等价（Karuch-Kuhn-Tucher）变化：



公式(2.28)表明，当时，必有成立。哪些的训练实例位于超平面和上，称为支持向量，决策边界的参数和仅依赖于这些支持向量。

线性可分的支持向量机是比较理想的情况，现实生活中并不常见，实际应用中，更多的是线性不可分的支持向量机。如图 8所示，直线x=15.5左右两侧的方块和圆圈无法被正确地划分到所属的类别。

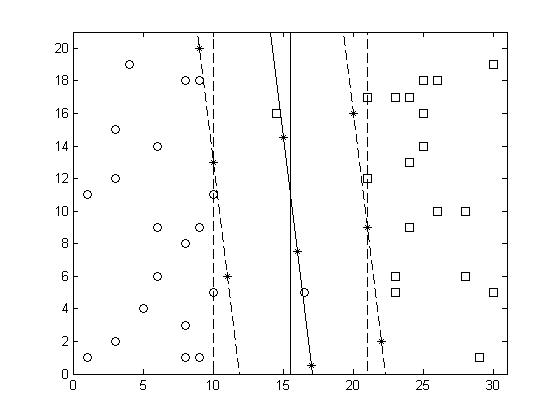


图 8 线性不可分情况下SVM的决策边界

为了解决线性不可分情况下的边缘最大化问题，可以采用软边缘的方法，训练时的决策边界允许一定的误差存在。为此，可以引入松弛变量（slack variable）来实现：



其中。提供了决策边界在训练样本上的误差。理论上，依然可以试用线性可分情况下的目标函数加上带有松弛变量的约束条件来确定决策边界。松弛变量的引入，可能使得一些样本被错误划分。为了防止因引入松弛变量而错分过多样本，导致分类器不准确的问题发生，针对松弛变量引入惩罚因子。综合可得目标函数如下：



其中和是用户指定的参数，表示对误分训练实例的惩罚。因而，被约束的最优化问题的拉格朗日函数可表示为：



利用如下的KTT条件，可以将此不等式约束问题转化为等数约束问题：



令关于，和的一阶导数为零，得到如下公式：



代入拉格朗日函数并进行对偶化，得到对偶化拉格朗日函数：



最后，利用对偶问题的求解方式求解该问题，得到拉格朗日乘子，将拉格朗日乘子代入KTT条件，从而得到决策边界的参数。

#### 贝叶斯分类方法

##### 朴素贝叶斯分类方法

朴素贝叶斯（Naïve Bayes）法是基于特征条件独立假设和贝叶斯定理的分类方法。首先，基于特征条件独立假设学习给定数据集上输入输出的联合概率分布；然后，基于此模型，利用贝叶斯定理求出给定输入后验概率最大的输出。

设输入空间为维向量集合，输出空间为类标记集合。输入特征向量为，输出类标记为。是定义在输入空间上的随机向量，是定义在输出空间上的随机变量。和的联合概率分布是，由独立同分布产生的训练集可以表示为：



朴素贝叶斯方法通过学习先验概率及条件概率，从而得到分类器的联合概率分布。朴素贝叶斯方法学习的先验概率为：



朴素贝叶斯方法学习的条件概率分布为：



条件概率分布有指数级数量的参数，实际上其计算与估计是不可行的。朴素贝叶斯假设各属性或特征之间是条件独立的，条件独立假设可以表示为：



朴素贝叶斯用于分类是一个计算后验概率的过程。假设待分类数据是，根据贝叶斯定理可得后验概率：



根据条件独立建设，上式可以表示为：



使得后验概率最大的类即为输入样本所属的类别。由于在的表达式中分母对所有的都是相同的，所以，输入所属的类别可以表示如下：



##### 贝叶斯信念网络

朴素贝叶斯分类器的条件独立假设过于严格，对于那些属性之间存在相关性的分类问题，分类准确率会有较大的损失。相比于朴素贝叶斯方法的条件独立假设，贝叶斯信念网络允许特征或属性之间具有一定的关联性。贝叶斯信念网络由一个有向无环图和概率表确定。其中，有向图中的边的方向指明了特征或属性之间的依赖关系，概率表是贝叶斯信念网络计算的依据。

贝叶斯信念网络用于分类主要包括两个步骤：a. 由具有领域专业只是的领域专家根据领域特点构建贝叶斯网络，并根据经验给出估计概率表；b. 根据贝叶斯网络和概率表中给定的先验概率计算后验概率，得到分类结果。

在贝叶斯网络中，每个节点代表一个属性，每个节点维持一个概率表，概率表给出属性取不同值的概率。当结点没有父母结点时，其对应属性取相应值的概率独立于其他属性的取值；当结点存在父母结点时，其对应属性取相应值的概率是关于其父母结点对应属性的条件概率。

#### 决策树分类方法

决策树分类是根据数据集的特点，从训练数据中选取能有效将数据分类的特征，构建分类树，对待分类数据进行分类的过程。决策树主要包括有效分类特征选取、分类树构建、优化分类树三个过程。

数据集的每个特征对应分类树的一个结点，有效分类特征的选取和分类树的构建是在同一个过程完成的，以某一个数据集的某个属性开始，构建决策分类树的根节点，不同的属性取值对应不同的决策树分支，分支的选择应该使得当前对数据集的划分达到最优，在不同的分支上，再根据未检测过的属性作为结点，重复以上过程，构建分类树。分类决策树构建完成后，为了防止过拟合现象的发生，需要对决策树进行适当的剪枝。分类决策树中，叶节点对应不同的类别，对待分类数据进行分类时，从根节点开始检测，根据待分类数据相应属性的不同取值选择不同的分支，一直到叶结点，叶结点对应的类别即是待分类数据所属的类别。

## 本章小结

本章主要阐述微博转发预测相关的技术和方法。首先，本文从背景热点获取相关技术、背景热点聚类相关技术和背景热点表示相关技术以及统计变量的相关性分析方法等方面介绍了基于背景热点的特征提取相关的技术和方法。然后，本文从基于用户特征的微博转发预测技术、基于内容特征的微博转发预测技术、基于结构特征的微博转发预测技术和微博转发预测中常用的分类方法等四个方面介绍了微博转发预测相关的方法和技术。

# 基于背景热点的微博转发行为影响因素分析

## 概述

通过对国内外关于微博转发研究现状的研究，可以发现，目前针对微博转发的研究主要基于网络结构或微博特征。前者或者根据用户之间的关注关系构建网络结构图，或者通过信息在网络中的传播路径构建网络结构图。在构建的社交网络结构的基础上研究信息在网络中的传播特点，以此研究微博转发预测问题。但这种方式往往受微博网络复杂庞大，难以获得完整的社交网络关系等局限，基于不完整的社交网络结构研究微博转发规律往往是得结果偏差较大。后者通过分析影响微博转发的因素，在影响微博转发因素的基础上提取对于研究微博转发问题有效的特征，基于机器学习方法构建微博转发模型，这种方式主要针对静态的用户特征和微博特征来预测微博是否会被转发，没有充分考虑待预测用户的个体差异和背景知识对转发决策的影响。

事实上，微博不仅是一个社交网络平台，同时也是一个自媒体平台，是网民集中表述和讨论自己观点的地方。网民往往针对现实生活中真实发生的事件发表自己的看法和观点，即网民的讨论一般情况下并不是子虚乌有的，而是来源于真实的社会事件。同时，网民对一桩社会事件的热议，也会扩大事件的影响力，引起关于事件的更多的讨论，尤其是一些意见领袖的参与，对于事件影响力的扩大具有更大的促进作用。因此，用户转发行为和社会热点事件的发生是存在一定联系的，他们之间相互促进。实际上，用户阅读到一条微博时，会根据自己已有的知识对微博价值和微博新颖性进行判断，然后决定是否转发。微博是否会被转发与用户个体行为特点和用户对与微博相关的背景知识的了解具有紧密联系。用户所掌握的微博背景知识一方面由用户历史转发微博体现，一方面由用户对微博内容的综合认知程度决定，而用户对微博内容的综合认知程度受多种复杂因素影响，社会上发生的热点话题信息是其中重要的影响因素。

综上所述，现有方法仅基于微博本身的用户特征和微博特征，无法综合利用背景热点话题特征对用户转发行为进行预测。本章基于背景热点话题基于背景热点话题研究微博用户转发行为，基于背景热点话题研究微博转发行为的影响因素。

## 背景热点话题相关定义

本文背景热点话题的来源有新浪网新闻排行榜频道[[45]](#footnote-2)注1、网易新闻排行榜频道[[46]](#footnote-3)注2和凤凰网资讯排行榜频道[[47]](#footnote-4)注3。背景热点话题相关定义如下：

**定义1：**背景热点话题内容，指从新闻网站获取的热点话题数据，经过预处理后表示为热点话题关键词集合，以此表示用户能够从其他渠道获知的微博内容相关的背景知识。

**背景热点话题内容表示：**首先采集新浪网新闻排行榜频道、网易新闻排行榜频道和凤凰网资讯排行榜频道三个新闻频道的热点新闻，一段时间内存在多个热点话题，通过聚类算法分离多个热点话题。得到一个背景热点话题文档级别的表示为，其中()表示背景热点话题文档集合中的一篇文档。根据表示背景热点话题的文档集合，提取关键词，获得背景热点话题词语级别的表示。

**定义2：**背景热点话题传播趋势，指热点话题相关报告的热度分布，以此表示热点话题热度传播趋势。

**背景热点话题传播趋势表示：**背景热点的传播趋势可以用不同时间段与背景热点话题相关的新闻报道数量或微博数量刻画。本文以一定时间内新闻报道数量变化来衡量背景热点话题的背景热点话题传播趋势，即背景热点话题传播趋势可以看作是热点话题相关新闻报道数量或微博数量关于时间的一个概率分布。

**定义3：**融合背景热点话题的微博转发预测问题可以定义为，其中：表示用户特征，表示微博特征，表示当前网络上正在发生的热点话题特征,本文称为背景热点话题，表示用户行为，即用户对微博的动作，，表示用户转发了微博，表示用户没有转发微博，用户是否转发微博的转发预测问题可以转化为二分类问题。本文基于背景热点提取微博转发特征，构建微博转发预测模型。

## 背景热点话题对微博用户转发行为的影响分析

基于微博自身属性的微博转发预测研究大多假设用户转发行为不受微博之外的因素影响。然而，用户不仅是属于微博平台的虚拟社会人，更是属于真实社会的自然人，其具有社会属性，接收信息渠道具有多元化特点，转发行为会受到微博数据以外的多种因素影响。由于微博平台已经成为网民讨论社会热点事件及发表相关观点的集散地，所以，背景热点事件的发生能在一定程度上吸引用户更多地参与到相关微博话题的讨论中，提高话题相关微博的转发量。

为了能够更有说服力的证明背景热点事件确实会在一定程度上影响用户转发行为，本文选取了一个具有代表性的具体事件，跟踪事件的发展趋势，从统计分析的角度加以说明。以“世界杯”热点话题为例，本文爬取新浪微博2014年4月12日至9月13日期间数据，以周为时间单位统计微博总量变化趋势和话题相关微博总量的变化趋势。统计结果如图 9所示。从图中可以看出，热点话题期间与背景热点话题相关的微博转发量呈现出先增加后减少的趋势，而在此期间微博转发总量整体也呈现出先增加后减少的趋势，这从宏观的角度说明了用户转发行为受到了热点话题的影响。

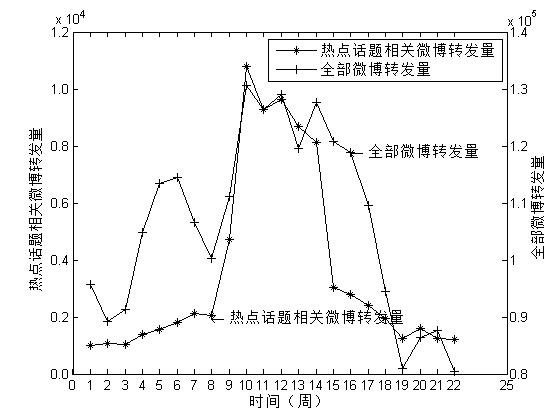


图 9 背景热点话题期间微博转发量变化趋势

进一步的，本文研究热点话题对不同类别微博用户转发行为的影响。本文针对2014年4月12日至9月13日期间微博用户，根据用户在热点话题期间是否发表过与之相关的微博判断用户是否与热点话题相关，将用户分为与热点话题相关和不相关两类，分别对用户转发量趋势进行统计。统计结果如图 10所示，三条曲线分别表示所有用户（all users）、与热点话题相关的用户(users prefer soccer)、及与热点话题不相关的用户(other users)所转发的与热点话题相关的微博量的变化趋势。本文可以看出在话题传播周期内，热点话题对各类用户的转发量都有明显的提升，与热点话题相关用户的转发量提升幅度较大。同时，历史上与热点话题不相关的用户也在热点话题期间增加了对热点话题相关微博的转发量。这些现象表明，影响用户转发行为的不仅仅是根据用户历史微博计算的用户转发兴趣，在热点事件背景下，用户的转发行为会受到当前社会热点话题的影响。

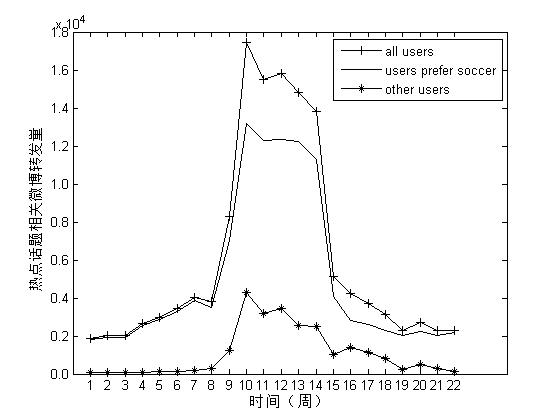


图 10 背景热点话题对各类用户微博转发量的影响

根据以上的统计实验和分析结果可知，在热点话题背景下，用户转发行为会受到背景热点话题的影响，背景热点话题影响对热点话题感兴趣的用户的转发行为，同时也会影响对背景热点话题不一定感兴趣的用户的转发行为。接下来的工作中，本文将研究基于背景热点话题提取转发特征的方法，并在此基础上研究微博转发预测问题。

## 基于热点话题的影响因素分析

### 基于背景热点的用户转发兴趣对转发行为的影响

兴趣是人们对事物喜好或关切的情绪，它表现为人们对某件事物、某项活动的选择性态度和积极的情绪反应[[[48]](#endnote-46)]。微博用户的兴趣部分地通过用户的转发行为体现出来，本文称之为用户转发兴趣。用户的转发行为受用户对微博的综合认知程度影响，而用户对微博的综合认知程度受多种复杂因素影响，用户转发兴趣及背景热点话题也是影响因素之一。兴趣作为用户的情绪反映，是用户个体行为一种相对稳定表现的形式。而在做转发决策时，个人兴趣、微博内容及当前发生的热点话题共同影响着用户转发决策。

从3.2节的分析，可以看出，背景热点话题不仅从宏观上影响用户转发行为，而且对于对背景热点话题感兴趣和不感兴趣的用户都有影响，只是影响程度不一样。事实上，用户兴趣与热点话题越相关，越容易在热点话题期间进行大量的转发。因此，本文提出融合背景热点话题的用户转发兴趣特征，研究背景热点话题下的用户转发兴趣对用户转发行为的影响。

**用户转发兴趣表示：**用户转发兴趣作为用户个体行为的一种具体表现形式，在微博转发过程中通过用户所转发的历史微博体现出来，可以根据用户历史微博来计算用户兴趣。用户历史转发微博表示为,其中()表示用户的一条历史微博。由于微博具有文本短，词稀疏的特点，所以本文选择以词集合的形式表示用户转发兴趣。对用户微博进行分词，去除停用词后，形成用户转发兴趣的词语级别的表示：



**背景热点话题内容表示：**背景热点话题文档级别的表示，其中()表示背景热点话题文档集合中的一篇文档。根据表示背景热点话题的文档集合，提取关键词，获得背景热点话题词语级别的表示：



从3.2节的实验，可以看出，背景热点话题对不同类别的用户的影响程度是不同的，原本对背景热点话题感兴趣的用户相比原本对背景热点话题不感兴趣的用户更容易受背景热点话题的影响。为此，本文通过研究用户转发兴趣与背景热点的匹配程度和用户转发行为受影响程度之间的关系来研究融合背景热点的用户转发兴趣对用户转发行为的影响。定义为用户转发兴趣与背景热点的匹配程度的计算方法如下：



统计并计算转发微博数量对应融合背景热点的用户转发兴趣的变化关系，如图 11所示，横坐标表示融合热点话题的用户转发兴趣，纵坐标表示与背景热点话题相关的微博转发量。该图反映了用户转发行为与融合背景热点话题的用户转发兴趣之间的关系。由图中可以看出，用户转发兴趣与背景热点话题之间的匹配程度越高（越大），则用户所转发与背景热点话题相关微博的数量越多，这表明融合背景热点话题的用户转发兴趣能够促进用户对与背景热点相关微博的转发，融合背景热点的用户转发兴趣特征能够作为研究微博转发预测时的有效特征。

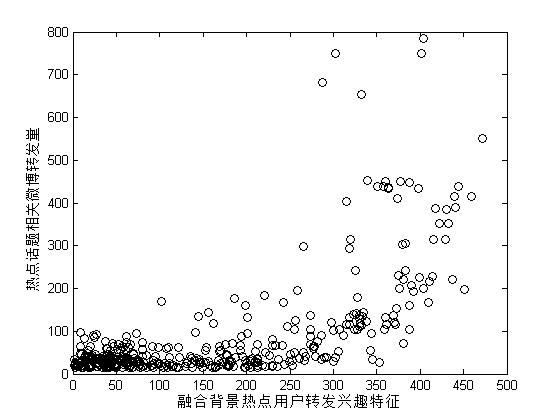


图 11 微博转发量对应融合背景热点话题的用户转发兴趣的关系图

### 融合背景热点话题的用户行为一致性对转发行为的影响

第3.3节从内容上考虑影响用户转发行为的因素，从图 9和图 10还可以看出，由于微博长度短且用户通常利用碎片时间登陆微博进行浏览和转发，大部分用户转发量在0到100条之间，转发兴趣关键词集中在0到150之间，仅从兴趣内容和累积转发量上很难全面刻画用户对热点话题的关注程度。用户转发微博的行为具有差异性，有的用户登录频繁且兴趣广泛，从累积的转发活跃度和兴趣特征上都表现出较高的转发概率，但是这类用户对热点话题相关微博的转发行为具有突发性和随机性特点。而有的用户转发活跃度较低，只是持续在自己关注的某些领域进行转发，这类用户未来转发热点相关微博的概率更大。所以考虑用户对热点话题相关微博的持续关注程度，能够有效检测用户是否是该热点话题的黏性用户，黏性用户未来转发热点话题相关微博的概率较高。

某一背景热点话题的高黏性用户通常时刻关注热点话题对应事件的发展趋势，体现在用户转发行为上，通常他们的转发行为与事件发展趋势具有高度一致性，因此，本文以用户行为与热点话题传播趋势的一致性作为用户对热点话题黏性程度的度量。一段时间内用户转发微博数量的变化趋势可以看作时间轴上的一个概率分布；本文以一定时间内新闻报道数量变化来衡量背景热点话题的热度变化趋势，也可以看做是时间轴上的一个概率分布。本文通过计算两个分布之间的相似度来计算用户行为与热点话题传播趋势的一致性特征。

KL（Kullback-Leiber divergence）散度又称相对熵（relative entropy），是用来描述两个概率分布之间差异性的一种方法[[[49]](#endnote-47)]，具有非对称性。KL距离越小表示两个分布越相似，KL距离等于0时表示两个分布完全一样。本文用KL距离来反映概率分布和概率分布之间的关系。





考虑到KL距离的非对称性，本文以



来计算概率分布之间的关系。通过上述计算方法本文计算了热点话题相关微博转发量与用户行为一致性之间的对应关系。

如图 12所示，当KL距离增大时，分布之间差异变大，用户转发行为与背景热点话题发展趋势之间一致性较小，即用户与背景热点事件之间的黏性更小，用户的转发行为具有随机性和非持续性，用户转发的与背景热点话题相关的微博较少；当KL距离减小时，分布差异较小，说明用户转发行为与背景热点话题发展趋势之间相关性较大，转发行为和热点话题趋势具有较高的一致性，即用户与背景热点事件之间的黏性更大，用户对该热点话题进行了持续关注，用户转发与背景热点话题相关微博较多。

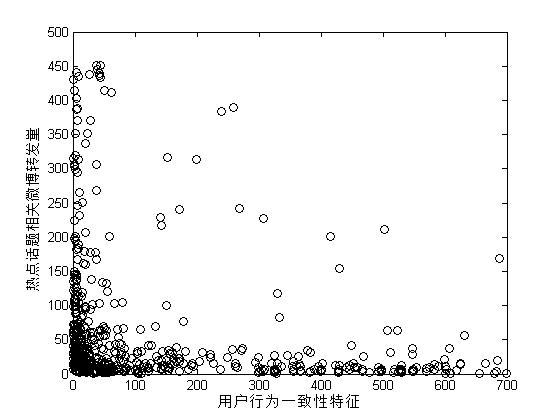


图 12 微博转发量对应行为一致性特征的关系图

这说明，持续关注某一背景热点话题的微博用户对该背景热点话题具有更高的转发兴趣，融合背景热点话题的用户行为一致性特征能够有效检测出热点话题的持续关注用户，同时避免因其总发帖量不高而被忽略的问题。因此，融合背景热点话题的用户行为一致性特征能够作为研究微博转发预测问题的有效特征。

### 基于背景热点的微博内容对转发行为的影响

当用户浏览微博时，吸引用户眼球的不仅仅是用户兴趣等用户主观因素，微博内容本身以及微博内容是否能够引起用户关注等因素也会影响用户的转发决策。所以，本节结合背景热点话题及待预测微博内容来分析热点话题对微博转发的影响。对于与背景热点话题之间具有一定黏性的微博用户，微博内容与热点话题越相关，用户转发微博的概率就越大。本文通过研究微博内容与背景热点之间的相似性和微博转发之间的关系来研究基于背景热点话题的微博内容特征对微博转发的影响。

同样地，本文用词集合来表示背景热点话题内容。本文对微博进行分词、去除停用词等预处理之后，将微博表示成一个词语级别的集合：。由于微博内容很短且都是特征词语，本文用Jaccard相似系数来表示微博内容与背景热点话题之间的相似性：



本文对融合热点话题的微博内容特征值不同的微博获得的转发总量及平均值进行了统计分析。如图 13所示,横坐标为值，纵坐标分别对应微博所获得的平均转发量和转发总量。从图中可以看出，以右侧坐标轴为标示的曲线表明微博转发总量随微博内容与背景热点话题相似性的增大而减少，这是因为大部分微博内容简短，所含内容特征较少，高相似性的微博数量大量减少，导致转发总量降低。以左侧坐标轴为标示的曲线表明，微博获得的平均转发量随微博内容与背景热点话题相似性的增大而提高，表明微博内容与热点话题越相似，越容易受到转发，融合热点话题的微博内容特征能够有效区分微博转发行为。

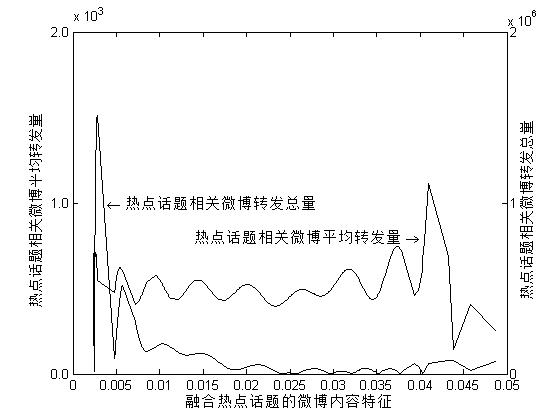


图 13 微博转发对应融合热点话题的微博内容特征对应的关系图

## 本章小结

本章主要研究微博用户转发行为受背景热点话题的影响情况，为基于背景热点的微博转发特征的选择提供理论支持。首先，通过具体事例研究了背景热点话题是否影响微博用户转发行为；然后，分别研究了基于背景热点的微博转发兴趣和微博内容特征以及用户转发行为与背景热点话题发展趋势的一致性对用户转发行为的影响。

# 基于背景热点的微博转发特征提取及转发预测

## 概述

从第三章的实验和分析可知，在热点事件背景下，微博用户转发行为会受事件发展趋势的影响。并且，受影响的不仅仅是密切关注事件发展的那一部分微博用户，原本不太关注事件发展的微博用户，也会因为背景热点事件在微博平台或者现实生活中所引起的巨大反响而参与到事件传播过程中，转发与事件相关的微博。经过第三章的实验，本文提取了融合背景热点的一些特征，包括基于背景热点的用户转发兴趣特征、基于背景热点的用户转发行为特征、基于背景热点的用户转发活跃度以及基于背景热点的微博内容特征，从实验结果及分析来看，它们都对微博用户转发行为有一定影响，本章将进一步基于这些特征，结合机器学习算法，研究微博转发预测问题。

## 数据集构建

一般地，背景热点数据能够从百度热搜榜、新浪新闻排行榜、网易新闻排行榜、凤凰资讯排行榜等各类新闻类门户网站的热点新闻频道获取，获取到的热点新闻经过聚类分析，获得当前的背景热点事件，然后对单个背景热点事件对应的新闻文档集合提取关键词，再用这些关键词通过百度搜索（或者google搜索）获得更多的关于背景热点时间的新闻报道，这样能够扩展背景热点事件的信息内容，从多个角度描述一个背景热点事件。最后，基于百度搜索获得的结果集形成背景热点时间的表示。

对应于具体的背景热点事件—巴西世界杯，本文分别从网易新闻排行榜、凤凰资讯排行榜和新浪新闻排行榜三个信息源获得背景热点事件相关的初始文档。然后基于百度新闻搜索获取热点话题数据，作为背景热点话题内容。提取热点话题关键词，利用新浪微博高级检索功能，根据检索结果，获得用户1725个，以及用户在2014年5月12日至2014年8月13日期间转发的微博，共计1,210,810条，并对用户和微博之间的转发和非转发关系进行标注。本文按时间将数据集分为训练和测试两个部分，2014年5月12日-2014年7月12日之间作为训练数据，共计895,552条，其中正样例682,324条，负样例213,228条；2014年7月13日-2014年8月13日之间作为测试数据共计315,258条，其中正样例209,999条，负样例105,259条。

在微博数据集的基础上构造矩阵：

其中：表示用户个数，表示微博个数，。标记后的数据集中每个元素为一个三元组，当时表示用户转发微博，当时表示用户没有转发微博。

## 微博转发预测特征选取

### 基于用户和微博内容的转发特征

1. **发布用户影响力**

发布用户的影响力大小会影响其微博被转发的可能性。对于同一条微博，不同的人转发之后，该条微博再被转发的可能性时不同的，通常影响力大的用户所转发的微博被再次转发的可能性更大。例如：对于微博http://weibo.com/1642591402/D2tJ8it2J，本文利用北京大学的PKUVIS微博可视化工具进行分析，得出该微博转发情况如表 2所示：

表 2 不同微博用户转发同一条微博的被转发情况

|  |  |  |  |
| --- | --- | --- | --- |
| 排名 | 应户名 | 粉丝数 | 转发数 |
| 1 | 邓超 | 37594895 | 33720 |
| 2 | 新浪娱乐 | 16795078 | 2545 |
| 3 | 刘亦菲 | 43397742 | 1826 |
| 4 | 奔跑吧兄弟 | 3291589 | 259 |
| 5 | 微博电影 | 4381673 | 46 |
| 6 | 恶棍天使 | 322290 | 41 |
| 7 | 新浪电影 | 4298982 | 19 |

由可视化分析的结果可知，该微博转发总量为39866，而表中转发量排名靠前的用户的转发总量已达38456，其余人的转发总量只有1410，只占总量的3.54%，而这些转发数比较高的用户同时也是粉丝数量很多的用户，这说明该微博的转发主要来自于粉丝数多的用户。实际上，一个用户的粉丝数越多，他所发表或转发的微博就能被更多人浏览，也就能得到更多的转发。

计算用户影响力的方式有多种，PageRank算法、HITS算法、粉丝数等都可以用于计算微博用户影响力。由于本文数据集的特点是用户数量少，缺少完整的用户之间关注关系或者转发关系，所以PageRank算法和HITS算法并不适合本文数据集的用户影响力计算，而以用户粉丝数作为用户影响力的衡量简单而有效，且用户粉丝数较容易获取，故本文以用户粉丝数来衡量用户影响力。



1. **用户活跃度**

用户拥有微博的动机各不一样，有的用户通过微博来扩大自身的社区影响力，他们会转发或发表各类微博，以获得更多关注；有的用户只是通过微博来获取信息，他们很少或从不转发或发表微博。有的用户偏向于仅仅通过微博来表达自己的观点和心情，所以他们发表的大多是原创微博，转发微博较少。有的用户为了提高自己的影响力，会很频繁地转发微博，不管微博所述是不是自己所偏爱的，这类用户转发微博相对原创微博较多。也有的用户转发和原创比较均衡。显然，对于转发较活跃的用户，更容易产生转发行为。用户发表微博的方式包括转发微博和原创微博，所以用户活跃度也包含原创和转发两个方面。

1) 发表微博活跃度：发表微博活跃度表示用户在选定时间内发表微博的频繁程度，可以用单位时间内发表微博的数量表示：



其中，为发表微博活跃度，代表在时间内发表微博的数量为。

2) 转发微博活跃度：发表微博活跃度表示用户在选定时间内转发微博的频繁程度，可以用选定时间内转发微博数量与发表微博数量的比值表示：



其中，和分别表示第天用户转发微博和发表微博的数量。

1. **用户兴趣与微博相似性**

用户对微博内容是否感兴趣，也是影响用户转发决策的一个因素。通常，用户偏爱于一类或几类微博，比如，有的用户偏爱篮球和羽毛球，而有的用户偏爱音乐和会话。用户在转发行为中体现出来的用户兴趣可以根据用户历史转发微博来计算，用户历史转发微博表示为,其中()表示用户的一条历史微博。由于微博具有文本短，词稀疏的特点，所以本文选择以词集合的形式表示用户转发兴趣。对用户微博进行分词，去除停用词后，形成用户转发兴趣的词语级别的表示：。对于一条待预测微博，同样进行分词、去除停用词等处理，得到一条微博的词语级别表示为：



本文以Jaccard相似性系数来计算用户兴趣与微博的相似度，计算方式如下：



### 基于背景热点话题的转发特征

1. **基于背景热点话题的用户转发兴趣特征**

背景热点话题对不同类别的用户的影响程度是不同的，原本对背景热点话题感兴趣的用户相比原本对背景热点话题不感兴趣的用户更容易受背景热点话题的影响。用户转发行为和用户转发兴趣与背景热点的匹配程度具有密切关系，用户转发兴趣与背景热点话题之间的匹配程度越高，用户越容易转发与背景热点相关微博。融合背景热点话题的用户转发兴趣能够促进用户对与背景热点相关微博的转发，因此，用户转发兴趣与背景热点的匹配程度能够作为微博转发预测的有效转发特征。定义为用户转发兴趣与背景热点的匹配程度的计算方法如下：



其中，和分别如公式(3.1)、(3.2)所示，分别表示用户转发兴趣的词语级表示和背景热点话题的词语级表示。

1. **基于背景热点话题的用户活跃度**

用户活跃度通常指用户发布微博的频繁程度，用户转发活跃度通常指用户转发微博的频繁程度。用户转发微博始终是用户的一种主动行为，长期不登录微博平台或者没有转发或发表微博习惯的微博用户，不太可能会转发微博，转发与背景热点相关的微博的可能性更小，用户活跃度或转发活跃度从侧面反映了用户的转发能力。融合背景热点的用户活跃的可以由用户转发与背景热点相关微博的频繁程度来刻画。用户转发行为活跃度通过用户在热点话题期间的累积转发量表示，融合背景热点话题的用户转发活跃度通过热点话题期间与热点话题相关的微博的累积转发量表示，该特征能够表明用户对热点话题的感兴趣程度。计算公式如下：

用户在一段时间t内转发的与热点话题相关的微博的频繁程度：



其中：表示用户在时间t内发布的微博i，表示对应热点话题的词语级表示，是微博是否与热点话题相关的阈值。

1. **用户转发行为与背景热点话题发展趋势一致性特征**

某一背景热点话题的高黏性用户通常时刻关注热点话题对应事件的发展趋势，体现在用户转发行为上，通常他们的转发行为与事件发展趋势具有高度一致性。用户转发行为与背景热点话题发展趋势之间相关性较大，转发行为和热点话题趋势具有较高的一致性，即用户与背景热点事件之间的黏性更大，说明用户对该热点话题进行了持续关注，用户转发与背景热点话题相关微博较多。以用户行为与热点话题传播趋势的一致性作为用户对热点话题黏性程度的度量，能够有效刻画用户转发行为与背景热点发展趋势的关系对用户转发行为的影响。一段时间内用户转发微博数量的变化趋势可以看作时间轴上的一个概率分布，记为；同理，一定时间内热点话题相关微博数量的变化也可以看做是时间轴上的一个概率分布，记为。本文通过计算两个分布之间的相似度来计算用户行为与热点话题传播趋势的一致性特征：



其中，和同公式(3.4)和(3.5)。

1. **基于背景热点话题的微博内容特征**

微博内容本身以及微博内容是否能够引起用户关注等因素是用户的转发决策的重要因素。对于与背景热点话题之间具有一定黏性的微博用户，微博内容与热点话题越相关，用户转发微博的概率就越大。用微博内容与背景热点话题相似性表示基于背景热点的微博内容特征。以词集合来表示背景热点话题内容，以词集合表示微博内容，则可用Jaccard相似系数来表示微博内容与背景热点话题之间的相似性：



## 微博转发预测

### 微博转发预测基本思路及评价指标

基本微博转发预测流程图如图 14所示，首先对训练数据和测试数据提取转发特征，本文提取的转发特征主要有基于用户的转发特征、基于微博内容的转发特征和基于背景热点的转发特征；然后，输入训练数据，结合机器学习算法，训练微博转发预测模型；最后，将转发预测模型应用于测试数据，预测转发行为。

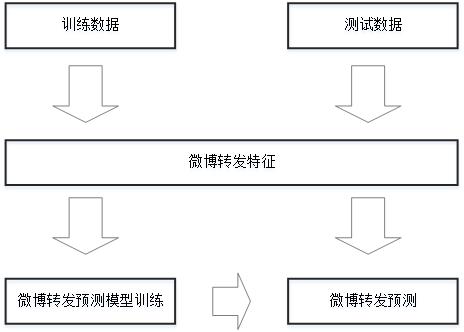


图 14 基本微博转发预测流程图

微博转发预测通常选用的评价指标为准确率（Precision）、召回率（Recall）和综合评价指标(F-Measure)：







其中，表示实际是转发微博，预测结果也为转发微博的微博数量；表示实际是非转发微博，预测结果为转发微博的微博数量；表示测试集中实际是转发微博的微博总数。

### 微博转发预测

经过前期的实验和分析，本文验证了背景热点话题对用户转发行为的影响，并提取了基于背景热点话题的微博转发特征。本节将在本文已有工作的基础上，基于所提取的转发特征进行转发预测实验，研究基于背景热点话题的转发特征对转发预测效果的影响。

#### 基于用户特征和微博特征的微博转发预测

为了验证基于背景热点话题的转发特征对转发预测效果的影响，本文先选取了不含基于背景热点话题的特征的转发特征进行转发预测实验，作为对比后续实验的对比实验。对于每条微博和发表该微博的用户，可以构建基于和的特征。其中，、、、分别对应4.3.1节描述的用户影响力特征、用户发表微博活跃度特征、用户转发微博活跃度特征和用户兴趣与微博相似度特征，，即表示用户未转发微博，表示用户转发微博。

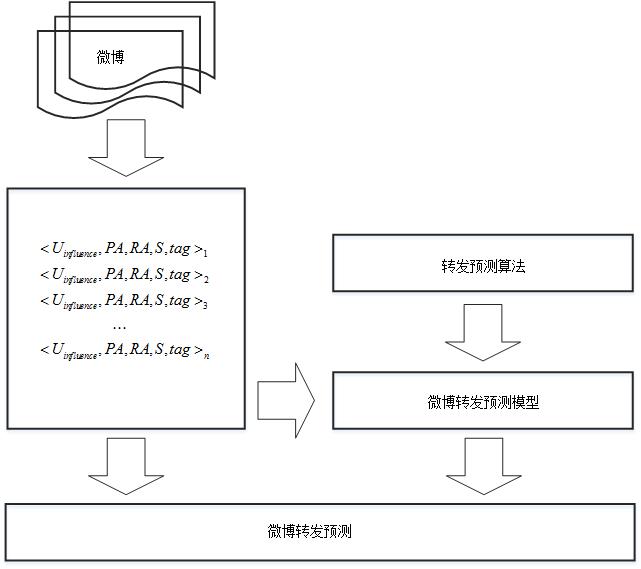


图 15 基于基本特征的对比实验流程图

本实验流程图如图 15所示，实验根据从微博提取的特征，包括用户影响力、用户发表微博活跃度、用户转发微博活跃度、用户兴趣与微博相似度，转发特征作为SVM、决策树、贝叶斯等机器学习算法的输入数据，训练微博转发预测模型，然后对待预测数据进行转发预测。

#### 基于背景热点特征的微博转发预测

本实验设计的目的在于研究基于背景热点的微博转发特征对微博转发预测效果的影响。对于每条微博和发表该微博的用户，构建基于和的特征。其中，表示基于背景热点的用户转发兴趣特征，表示用户转发与热点话题相关的微博的频繁程度，表示用户转发行为与背景热点发展趋势的一致性特征，表示融合背景热点的微博内容特征；，即表示用户未转发微博，表示用户转发微博。

本实验流程图如图 16所示，实验根据从微博提取的基于背景热点话题的特征，包括基于背景热点的用户转发兴趣特征、用户转发与热点话题相关的微博的频繁程度、用户转发行为与背景热点发展趋势的一致性特征、融合背景热点的微博内容特征，将基于背景热点话题的转发特征作为SVM、决策树、贝叶斯等机器学习算法的输入数据，训练微博转发预测模型，然后对待预测数据进行转发预测，对比基于用户和微博特征的转发预测实验效果，分析基于背景热点话题的转发特征对转发预测实验的影响。

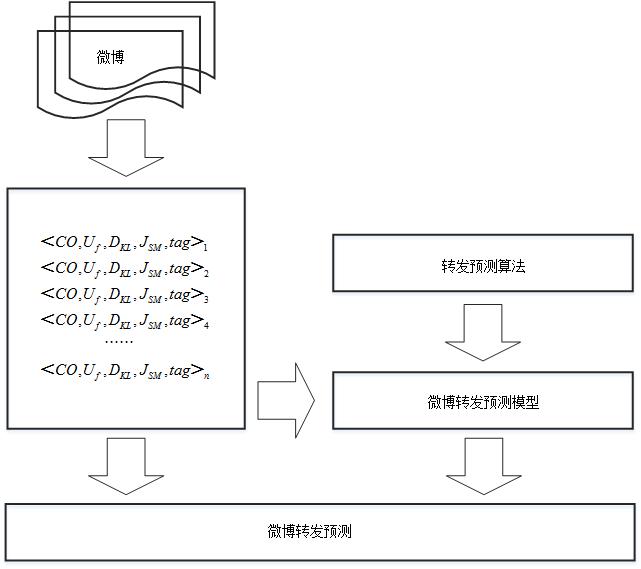


图 16 基于背景热点微博转发特征的转发预测实验流程图

#### 多特征融合的微博转发预测

设，，其中，表示原博主的用户影响力，表示微博内容与微博用户兴趣相似度；表示基于背景热点的用户转发兴趣特征，表示用户转发与热点话题相关的微博的频繁程度，表示用户转发行为与背景热点发展趋势的一致性特征，表示融合背景热点的微博内容特征。本文选取特征，进行实验，其中。

本实验流程图如图 17所示，实验结合用户、微博基本转发特征和基于背景热点话题的转发特征，进行微博转发预测实验。本实验设计了两组实验，一组实验是从基于背景热点话题的转发特征中任意选取一个特征，再融合用户、微博基本特征，构成转发特征集合，进行实验，其中，进行实验，与基于基本特征的实验对比，进行分析；另一组实验从基于背景热点话题的转发特征中选取一个到多个特征，并融合用户、微博基本特征，构成转发特征集合，进行实验，其中，进行实验，与基于基本特征的实验对比，进行分析。

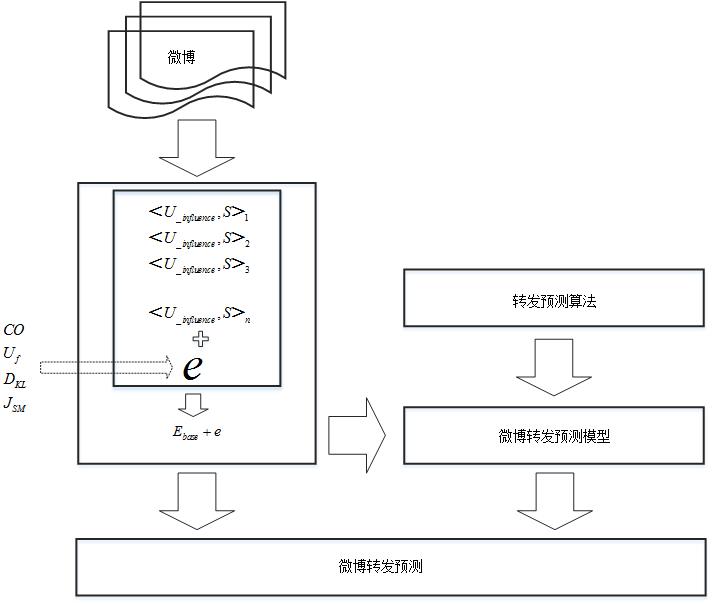


图 17 多特征融合的转发预测实验流程图

## 实验与结果分析

1. **基于基本微博转发特征的实验及结果分析**

基于基本微博转发特征的实验结果如表 3所示，从实验结果来看，将用户影响力、用户活跃度以及用户兴趣与微博相似度等特征用于微博转发预测能够取得较好的预测效果。

表 3 基于用户特征和内容特征转发预测实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| Classifiers | Precision | Recall | F-measure |
| BayesNet | 0.877 | 0.867 | 0.869 |
| NaiveBayes | 0.802 | 0.736 | 0.74 |
| C4.5 | 0.886 | 0.883 | 0.884 |
| LibSvm | 0.839 | 0.841 | 0.84 |

1. **基于背景热点话题特征的微博转发预测实验及结果分析**

基于背景热点话题特征的微博转发预测实验如表 4所示：

表 4 基于背景热点的微博转发预测实验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| classifiers | Precision | | Recall | | F-measure | |
| baseline | BHT-based | baseline | BHT-based | baseline | BHT-based |
| BayesNet | 0.877 | 0.938 | 0.867 | 0.937 | 0.869 | 0.936 |
| NaiveBayes | 0.802 | 0.756 | 0.736 | 0.714 | 0.74 | 0.643 |
| C4.5 | 0.886 | 0.951 | 0.883 | 0.948 | 0.884 | 0.947 |
| LibSvm | 0.839 | 0.942 | 0.841 | 0.938 | 0.84 | 0.936 |

表 4中baseline所对应的实验结果是4.3.1所描述特征对应的实验结果，BHT-based是本节所选特征，即4.3.2所描述基于背景热点的特征用于微博转发预测对应的实验结果。从实验结果可以看出，选取融合背景热点的特征用于微博转发预测，能取得很好的预测效果，而且相对于基于用户特征和内容特征的方法预测效果提升明显；从准确率（Precision）来看，预测效果最好的是决策树（C4.5）算法，达到了95.1%，从综合评价指标（F-Measure）来看，预测效果最好的也是决策树（C4.5）算法，达到了94.7%，相对于baseline的实验这两个指标的提高分别为6.5%和6.3%，提升效果明显。根据实验对比，发现并不是对于所有的方法，基于背景热点特征的预测效果都更好，朴素贝叶斯方法（NaiveBayes）相对于baseline的实验预测效果没有提升，反而降低了，这是因为本文所选的基于背景热点的微博转发特征并不是条件独立的，它们之间通过背景热点相互关联，而朴素贝叶斯方法要求所选特征条件独立，在实验的过程中将不满足条件独立的特征当作条件独立的，这影响了朴素贝叶斯算法分类的准确性，所以预测效果并没有提升。

1. **多特征融合的微博转发预测实验及结果分析**

多特征融合的微博转发预测实验结果如以下表格所示。

表 5 增加单个特征变量对比实验结果(Precision)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifiers | Precision | | | | |
| baseline | B+f1 | B+f2+f3 | B+f4 | B+f5 |
| BayesNet | 0.877 | 0.928 | 0.935 | 0.93 | 0.92 |
| NaiveBayes | 0.802 | 0.735 | 0.817 | 0.714 | 0.725 |
| C4.5 | 0.886 | 0.931 | 0.938 | 0.931 | 0.933 |
| LibSvm | 0.839 | 0.929 | 0.936 | 0.928 | 0.931 |

表 6 增加单个特征变量对比实验结果(Recall)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifiers | Recall | | | | |
| baseline | B+f1 | B+f2+f3 | B+f4 | B+f5 |
| BayesNet | 0.867 | 0.978 | 0.976 | 0.978 | 0.973 |
| NaiveBayes | 0.736 | 0.959 | 0.919 | 0.962 | 0.965 |
| C4.5 | 0.883 | 0.998 | 0.994 | 0.997 | 0.986 |
| LibSvm | 0.841 | 0.987 | 0.981 | 0.985 | 0.983 |

表 7 增加单个特征变量对比实验结果(F-Measure)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifiers | F-Measure | | | | |
| baseline | B+f1 | B+f2+f3 | B+f4 | B+f5 |
| BayesNet | 0.869 | 0.953 | 0.955 | 0.953 | 0.946 |
| NaiveBayes | 0.74 | 0.832 | 0.865 | 0.82 | 0.828 |
| C4.5 | 0.884 | 0.865 | 0.965 | 0.963 | 0.958 |
| LibSvm | 0.84 | 0.957 | 0.957 | 0.956 | 0.956 |

其中，Base列对应4.3.1节所选特征的实验结果，f1对应基于背景热点话题的用户转发兴趣特征；f2+f3对应基于背景热点的用户活跃度特征（包括基于背景热点的发表活跃度和转发活跃度）；f4对应微博用户转发行为与背景热点发展趋势一致性特征；f5对应基于背景热点的微博内容特征。

根据对比实验结果，可以看出，对于本文提出的基于背景热点的几个特征，在基本实验的基础上加上其中任意一个特征，都使得微博转发预测效果得到了明显提升，这说明基于背景热点的特征对于微博转发预测是有效的。预测模型与预测模型之间进行对比，可以发现，决策树算法（C4.5）预测效果要稍优于其他预测模型；单个特征之间作对比，发现一般情况下，基本特征加上基于背景热点的微博内容特征（f5），预测效果比基本特征加上其他特征的差，可以认为基于背景热点的微博内容特征对转发预测效果的提升不如其他特征。

表 8 逐个增加特征变量对比实验结果(Precision)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifiers | Precision | | | | |
| baseline | B+f1 | B+(f1~f3) | B+(f1~f4) | B+(f1~f5) |
| BayesNet | 0.877 | 0.928 | 0.932 | 0.932 | 0.94 |
| NaiveBayes | 0.802 | 0.735 | 0.903 | 0.903 | 0.813 |
| C4.5 | 0.886 | 0.931 | 0.931 | 0.932 | 0.953 |
| LibSvm | 0.839 | 0.988 | 0.935 | 0.94 | 0.947 |

表 9 逐个增加特征变量对比实验结果(Recall)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifiers | Recall | | | | |
| baseline | B+f1 | B+(f1~f3) | B+(f1~f4) | B+(f1~f5) |
| BayesNet | 0.867 | 0.978 | 0.978 | 0.979 | 0.939 |
| NaiveBayes | 0.736 | 0.959 | 0.812 | 0.824 | 0.784 |
| C4.5 | 0.883 | 0.998 | 0.998 | 0.998 | 0.951 |
| LibSvm | 0.841 | 0.986 | 0.987 | 0.988 | 0.943 |

表 10 逐个增加特征变量对比实验结果(F-Measure)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifiers | F-Measure | | | | |
| baseline | B+f1 | B+(f1~f3) | B+(f1~f4) | B+(f1~f5) |
| BayesNet | 0.869 | 0.953 | 0.955 | 0.955 | 0.938 |
| NaiveBayes | 0.74 | 0.832 | 0.882 | 0.862 | 0.789 |
| C4.5 | 0.884 | 0.865 | 0.963 | 0.963 | 0.966 |
| LibSvm | 0.84 | 0.958 | 0.96 | 0.963 | 0.942 |

在上一步的实验中，本文通过在基本特征的基础上每次增加一个基于背景热点的特征，经过实验和对比分析，发现基于背景热点的微博内容特征对预测效果的提升最小；同时，将上一步的实验结果与综合所有基于背景热点特征（B+(f1~f5)）的实验进行对比，发现上一步的预测效果比综合所有基于背景热点特征的预测效果要好。为此，本文逐个增加基于背景热点的转发特征，进行实验，实验结果准确率（precision）、召回率（recall）、综合评价指标（F-Measure）分别如表 8、表 9、表 10所示。经过对比分析，发现分别增加特征f1、f2+f3、f4时预测效果基本都有所提升，但当增加特征f5时，预测效果有所下降。由此，可以认为基于背景热点的微博内容特征在多特征融合的微博转发预测中不利于预测效果的提升。

## 本章小结

本章在第三章研究的基础上，构建微博转发预测特征，基于机器学习方法训练转发预测模型，预测微博转发行为。首先，作为对比实验，基于用户特征和内容特征进行了微博转发预测实验；然后，选择基于背景热点的微博转发预测特征，进行微博转发预测实验，研究本文所研究的微博转发预测特征对微博转发预测效果是否有提升以及提升效果；最后，综合用户特征、内容特征和基于背景热点话题的特征进行微博转发预测实验，并从多个角度、不同层面做了对比实验。通过本章的实验，验证了本文所研究的基于背景热点话题的微博转发预测特征与用户转发行为具有密切关系，且将这些特征用于微博转发预测能够明显提升微博转发预测效果。

# 基于背景热点微博转发系统设计与实现

## 概述

经过第三章和第四章的分析和实验，本文基于背景热点提出的微博转发特征从多方面体现了微博用户转发行为受背景热点事件的影响；在将这些特征用于微博转发预测的实验中，这些特征对微博转发预测的提升效果比较明显。本章将在已有实验的基础上，将各个功能模块集中到一个系统中。根据已有的实验设计可知，系统将主要包括数据采集、特征提取、微博转发预测等功能模块。

## 总体系统设计

本节为基于背景热点的微博转发预测系统提供总体的系统设计。基于背景热点微博转发预测系统包括数据采集、数据存储、微博转发预测模型、转发预测共四个基本模块，系统设计图如图 18所示。

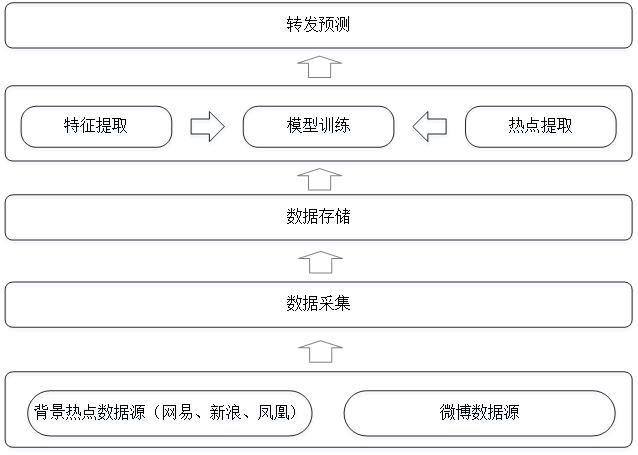


图 18 基于背景热点微博转发预测总体系统设计图

其中，数据采集模块负责从新浪微博网采集微博数据，以及从网易资讯排行榜、凤凰新闻排行榜和新浪新闻排行榜采集背景热点相关原始数据，采集到的数据存储在MySQL数据库中；数据存储模块采用MySQL数据库，默认使用InnoDB数据库引擎；微博转发预测模型模块包括特征提取、热点提取和模型训练三个子模块，该模块从数据库读取背景热点数据和微博数据，输出基于背景热点的微博转发预测模型；转发预测模块基于已有的微博转发预测模型，给定一条微博，判断其是否会得到转发。

## 数据采集

数据采集模块流程图如图 19所示，数据采集模块主要实现两个功能：采集微博数据和采集背景热点数据。微博数据从新浪微博获取，背景热点数据的数据源包括网易新闻排行榜、新浪新闻排行榜和凤凰新闻资讯排行榜。

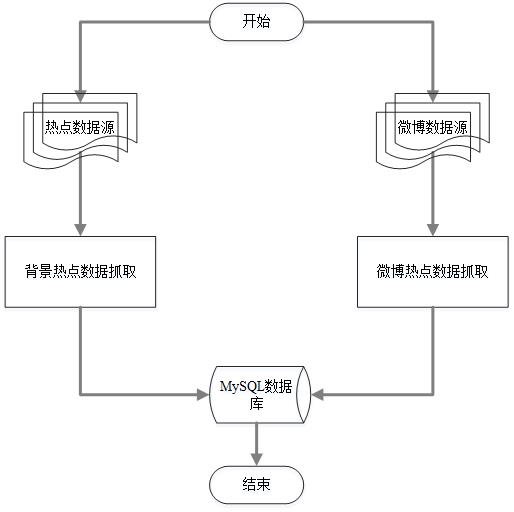


图 19 数据采集模块流程图

数据采集的实现通过开源工具WebCollector[[[50]](#endnote-48)]实现。三个不同背景热点数据源的数据采集具有相同的采集过程，不同内容抽取处理方式。由于热点数据更新不是很频繁，所以采用周期轮询的方式对网站进行数据采集。本文主要采集热点新闻相关的新闻标题、新闻标签、新闻时间、新闻来源等信息。

微博数据的采集分为两个阶段，认证登录和数据采集。

微博的认证登录采用OAuth 2.0[[[51]](#endnote-49)]身份鉴权的方式。OAuth 2.0认证和授权过程涉及服务提供方、用户和客户端三方，服务提供方存储待访问资源如网页数据、视频、照片等的服务端，用户指受保护的待访问资源的拥有者，客户端是指需要访问服务方所提供的资源的第三方应用。通过OAuth 2.0认证方式获取登录所需的Access Token，然后再利用获得的Access Token登录新浪微博，登录之后便可类似网页浏览新浪微博一样获取微博数据。当Access Token过期时只要重新获取即可。

微博数据的采集仍然通过开源工具WebCollector实现。首先，利用新浪微博的高级检索功能对背景热点相关的关键词进行检索，然后采集所返回的结果集中的微博数据。本文主要采集微博和用户相关的用户ID、用户昵称、用户粉丝数、用户关注数、用户粉丝列表、用户关注列表、用户标签、微博ID、微博内容、微博转发数、微博点赞数、原博主用户ID、微博微博发表时间等信息。

## 数据存储

本文数据存储采用MySQL数据库，数据库引擎采用默认的InnoDB引擎，这是因为相比于ISAM和MyISAM等数据库引擎，InnoDB更加灵活且支持事务。数据存储主要包括背景热点数据的存储和微博数据的存储。

背景热点数据的存储包含背景热点新闻表、背景热点表两张表。背景热点新闻表主要存储新闻标题、新闻URL、新闻发布时间、新闻标签、新闻内容、新闻来源等内容。背景热点新闻数据库表设计如表 11所示。背景热点表主要保存热点名称、表示热点的关键词等内容，背景热点数据库表设计如表 12所示。

表 11 背景热点新闻表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 说明 | 长度 | 备注 |
| ID | N |  | 38 | PK、NN |
| URL | VAR | 新闻原文链接地址 | 1024 |  |
| TITLE | VAR | 新闻标题 | 1024 |  |
| CONTENT | CLOB | 新闻内容 |  |  |
| SOURCE | VAR | 数据源（网易或新浪或凤凰网） | 38 |  |
| CREATE\_TIME | TS | 新闻发布时间 |  |  |
| TAG | VAR | 新闻标签 | 1024 |  |

表 12 背景热点表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 说明 | 长度 | 备注 |
| ID | N |  | 38 | PK、NN |
| HOT\_NAME | VAR | 热点名称 | 1024 |  |
| HOT\_KEYWORDS | VAR | 表示热点的关键词 | 1024 |  |

微博消息的存储主要包括微博ID、用户ID、用户名、用户昵称、微博内容、微博链接、图片URL、转发数、评论数、赞数、微博时间等。微博消息存储表如表 13所示。

表 13 微博消息存储表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 说明 | 长度 | 备注 |
| MESSAGE\_ID | N | 微博消息ID | 38 | PK、NN |
| USER\_ID | N | 用户ID | 1024 |  |
| USER\_NAME | VAR | 用户名 | 1024 |  |
| NICK\_NAME | VAR | 用户昵称 | 1024 |  |
| MESSAGE\_CONTENT | VAR | 微博内容 | 4000 |  |
| MESSAGE\_URL | VAR | 微博链接 | 1024 |  |
| IMAGE\_URL | VAR | 图片URL | 1024 |  |
| RETWEET\_COUNT | N | 转发数 | 11 |  |
| COMMENT\_COUNT | N | 评论数 | 11 |  |
| UP\_CLICK\_COUNT | N | 赞数 | 11 |  |

微博用户信息的存储主要包括用户ID、用户名、性别、微博数、关注数、粉丝数、兴趣爱好、标签、关注列表、粉丝列表等。用户信息存储表如表 14所示。

表 14 用户信息存储表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 说明 | 长度 | 备注 |
| UID | N | 用户ID | 38 | PK、NN |
| NAME | VAR | 用户名 | 100 |  |
| SEX | N | 用户性别 | 1 |  |
| STATUSNUM | N | 用户微博数 | 11 |  |
| FRIENDNUM | N | 用户关注数 | 11 |  |
| FOLLOWERNUM | N | 用户粉丝数 | 11 |  |
| INTEREST | VAR | 用户兴趣 | 1024 |  |
| TAG | VAR | 用户标签 | 1024 |  |
| FRIENDLIST | VAR | 关注列表 | 4000 |  |
| FOLLOWLIST | VAR | 粉丝列表 | 4000 |  |

## 微博转发预测模型

该模块主要完成根据背景热点信息和微博数据提取微博转发特征，然后采用贝叶斯信念网络、朴素贝叶斯、决策树和支持向量机等机器学习算法，训练微博转发预测模型，用于微博转发预测。该模块流程图如图 20所示，首先，从采集到的背景热点数据中提取背景热点表示，然后基于背景热点信息和微博数据，提取融合背景热点的微博转发预测特征，最后以背景热点事件前期对应数据作为训练集，训练微博转发预测模型，将训练出来的模型序列化保存，用作微博转发预测的备用。该部分的主要工作在于提取背景热点信息、提取微博转发特征、训练微博转发预测模型及预测微博转发行为。

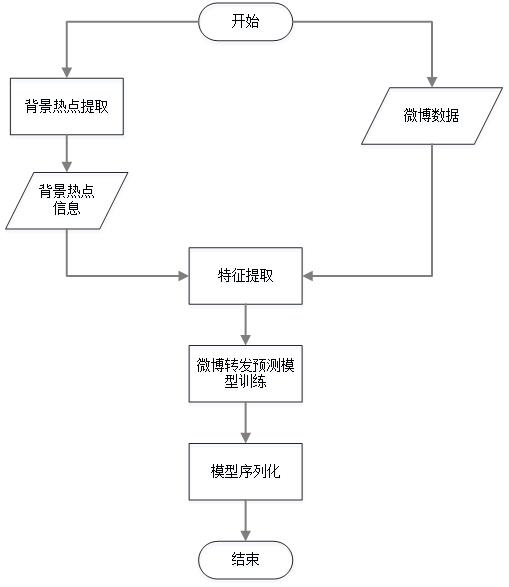


图 20 微博转发预测模型流程图

### 背景热点提取

提取背景热点信息部分主要工作是提取表示背景热点事件的关键词集合。在5.3节所采集的数据的基础上，每篇文档可以通过向量空间模型表示成一个向量： ，其中，表示所抓取到的背景热点新闻文档集合。从中随机选取个初始质心：，其中。非中心文档到质心的距离定义为：



其中，**，。**根据误差平方和（SSE）最小原则将分配到距离最近的质心，然后重新计算质心，重复此过程直到质心不再变化。计算方式如下：



经过聚类学习之后得到个背景热点话题。话题包含个新闻文档，即，对文档进行分词、去除停用词以及去除噪音数据等处理，然后计算词的TFIDF值，取分值高的词作为话题的表示，即将话题表示成词集合：

### 提取微博转发特征

本文提取的微博转发特征主要有：发布用户影响力、用户转发活跃度、用户发布活跃度、用户兴趣与微博相似度、基于背景热点话题的用户转发兴趣、基于热点话题的用户活跃度、用户转发行为与热点话题发展趋势一致性特征和基于背景热点话题的微博内容特征。本文以用户粉丝数作为发布用户影响力，以用户发表微博的频繁程度和用户转发微博的活跃度作为用户发布活跃度和用户转发活跃度。

**用户兴趣与微博相似度：**

记用户历史微博表示为,其中()表示用户的一条历史微博。由于微博具有文本短，词稀疏的特点，所以本文选择以词集合的形式表示用户转发兴趣。对用户微博进行分词，去除停用词后，形成用户转发兴趣的词语级别的表示：。对于一条待预测微博，同样进行分词、去除停用词等处理，得到一条微博的词语级别表示为：。

本文以Jaccard相似性系数来计算用户兴趣与微博的相似度，计算方式如下：



**基于背景热点话题的用户转发兴趣：**

记背景热点话题文档级别的表示，其中()表示背景热点话题文档集合中的一篇文档。根据表示背景热点话题的文档集合，提取关键词，获得背景热点话题词语级别的表示。用户转发兴趣通过用户历史微博来表示：。

本文定义基于背景热点话题的用户转发兴趣为用户转发兴趣与背景热点话题的匹配程度，其计算方法如下：



**基于热点话题的用户活跃度：**

融合背景热点话题的用户转发活跃度通过热点话题期间与热点话题相关的微博的累积转发量表示，该特征能够表明用户对热点话题的感兴趣程度。计算公式如下：

用户在一段时间t内转发的与热点话题相关的微博的频繁程度：



其中：表示用户在时间t内发布的微博i，表示对应热点话题的词语级表示，是微博是否与热点话题相关的阈值。

**用户转发行为与热点话题发展趋势一致性特征：**

用户转发行为可以用一段时间内用户转发微博的数量变化来表示，一段时间内用户转发微博数量服从概率分布；背景热点话题的发展趋势可以通过报道与该热点话题相关的新闻数量的变化表示，设热点话题发展趋势服从概率分布。本文通过计算两个分布之间的相似度来计算用户行为与热点话题传播趋势的一致性特征。KL距离是一种很好的能够描述两个概率分布之间差异性的方法，其具有非对成性。KL距离越小表示两个分布越相似，KL距离等于0时表示两个分布完全一样。表示用户转发行为的概率分布和表示热点话题发展趋势的概率分布的KL距离计算如下：





考虑到KL距离的非对称性，本文以



来表示用户转发行为与背景热点发展趋势的差异性，即用户转发行为与背景热点发展趋势的一致性。

**基于背景热点话题的微博内容特征：**

背景热点话题内容可以表示为。本文对微博进行分词、去除停用词等预处理之后，将微博表示成一个词语级别的集合：。由于微博内容很短且都是特征词语，本文用Jaccard相似系数来表示微博内容与背景热点话题之间的相似性，即融合热点话题的微博内容特征：



## 微博转发预测

完成特征提取之后，微博转发特征可以表示为,其中，表示原博主的用户影响力，表示微博内容与微博用户兴趣相似度；表示基于背景热点的用户转发兴趣特征，表示用户转发与热点话题相关的微博的频繁程度，表示用户转发行为与背景热点发展趋势的一致性特征，表示融合背景热点的微博内容特征。以背景热点事件前期的微博作为训练数据，选择贝叶斯信念网络、朴素贝叶斯、决策树和支持向量机等机器学习算法，训练微博转发预测模型，并序列化保存模型。对新出现的待预测的微博数据，分别用四个微博转发预测模型，预测微博是否会被转发，分别给出各个模型的预测结果。

## 本章小结

本章内容主要是基于背景热点的微博转发预测系统的设计与实现。本文微博转发预测系统包括数据采集、背景热点提取、微博转发预测三个主要功能模块。

# 总结与展望

现有微博转发预测相关研究提取用户或微博有关转发特征，进行微博转发预测研究，能取得较好的预测效果，但现有研究工作没有考虑背景热点事件这一用户转发行为的影响因素。本文总结了现有关于微博转发预测研究现状，首先通过实验验证了用户转发行为受背景热点事件的影响；然后分析背景热点相关因素对用户转发行为的影响情况，并依此为依据提取背景热点相关的转发特征；最后基于背景热点研究微博转发预测问题。本文的主要工作包括以下几方面：

1. 本文通过以下几个方面研究了背景热点话题与用户转发行为之间的关系，以此来研究基于背景热点的微博用户转发行为。首先，本文研究了基于背景热点话题的用户转发兴趣对用户转发行为的影响，本文以具体事例研究发现背景热点话题与用户转发兴趣越相近，用户转发与其相关的微博的可能性越大，并且即使用户对热点话题不感兴趣，由于话题的热点效应，用户也可能转发；其次，本文研究了用户转发行为与背景热点话题发展趋势之间的一致性对用户转发的影响，发现它们之间的一致性程度越高，用户转发与热点话题相关微博的可能性越大；最后，本文研究了基于背景热点微博内容特征对用户转发的影响，发现基于背景热点的微博内容特征也能提升微博转发预测效果，只是提升幅度不如其他两个特征。

2. 根据背景热点话题对微博转发行为的影响，提取基于背景热点话题的微博转发特征，用于微博转发预测实验。本文提取的背景热点话题相关的转发特征有：基于背景热点话题的用户转发兴趣特征、基于背景热点话题的用户活跃度特征、基于背景热点话题的微博内容特征以及用户转发行为与背景热点话题发展趋势的一致性特征。

3. 基于背景热点话题研究微博转发预测。本文在所提取基于背景热点话题的微博转发特征的基础上，综合基本的用户特征和微博内容特征，从多个维度进行微博转发预测对比实验。通过实验，发现将基于背景热点的微博转发特征用于微博用户转发行为预测，能够取得很好的预测效果。并且，从对比实验中发现，不增加基于背景热点的微博内容特征在微博转发预测时能够取得更好的效果。

最后，本文以上两部分工作，实现了一个简单的微博转发预测原型系统，系统包括数据采集、背景热点提取、微博转发预测三个主要功能模块。

虽然本文的研究工作对微博转发预测的效果有比较明显的提升，但是本文的相关工作仍然有需要完善的地方，主要包括以下几方面：

首先，微博文本是短文本，词稀疏，所以基于微博的短文本分析具有极大的挑战性。虽然本文将用户转发前后的微博内容都作为微博内容参与计算，然后进行分词、去除停用词等处理，用词集合表示微博，在实验中也取得了较好的效果。但是，如果能对微博做一些语义扩展，比如将WordNet[[[52]](#endnote-50),[[53]](#endnote-51)]用于微博短文本的语义扩展；或者利用Word2Vec[[[54]](#endnote-52),[[55]](#endnote-53)]等工具丰富词语所携带的信息，或者基于神经网络或深度学习方法使词语包含更多的上下文信息。

其次，本文虽然综合考虑了微博用户特征、微博内容特征以及基于背景热点话题的特征，并将它们用于微博转发预测实验，也取得了较好的实验结果；但是，如果时间和硬件条件允许，能够获得更多的微博数据和微博用户数据，提取更加丰富的有效转发特征，或许能取得更好的实验效果，应用场景更广泛。

再次，从本文的研究工作可以看出，有些用户对背景热点话题并不感兴趣，有时他们也会转发相关的微博。因此，可以更进一步地研究在背景热点事件发展过程中，用户兴趣的变化。

# 参考文献

1. Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter[C]//ICWSM. 2011.
2. 李英乐, 于洪涛, 刘力雄. 基SVM的微博转发规模预测方法[J]. 计算机应用研究, 2013, 30(9): 2594-2597.
3. 曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779-790.
4. Kanavos A, Perikos I, Vikatos P, et al. Modeling ReTweet Diffusion Using Emotional Content[M]//Artificial Intelligence Applications and Innovations. Springer Berlin Heidelberg, 2014: 101-110.
5. Perikos I, Hatzilygeroudis I. Recognizing emotion presence in natural language sentences[M]//Engineering Applications of Neural Networks. Springer Berlin Heidelberg, 2013: 30-39.
6. Ekman P. Basic Emotions. Handbook of Cognition and Emotion Psychology Department[J]. 1999.
7. Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1633-1636.
8. Kschischang F R, Frey B J, Loeliger H A. Factor graphs and the sum-product algorithm[J]. Information Theory, IEEE Transactions on, 2001, 47(2): 498-519.
9. Loeliger H A. An introduction to factor graphs[J]. Signal Processing Magazine, IEEE, 2004, 21(1): 28-41.
10. Yang Z, Tang J, Li J, et al. Social community analysis via a factor graph model[J]. IEEE Intelligent Systems, 2010 (3): 58-65.
11. Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter[C]//System Sciences (HICSS), 2010 43rd Hawaii International Conference on. IEEE, 2010: 1-10.
12. Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[C]//Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010: 177-184.
13. 张旸, 路荣, 杨青. 微博客中转发行为的预测研究[J]. 中文信息学报, 2012, 26(4): 109-114.
14. Hong L, Dan O, Davison B D. Predicting popular messages in twitter[C]//Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 57-58.
15. Bandari R, Asur S, Huberman B A. The Pulse of News in Social Media: Forecasting Popularity[C]//ICWSM. 2012: 26-33.
16. Ma H, Qian W, Xia F, et al. Towards modeling popularity of microblogs[J]. Frontiers of Computer Science, 2013, 7(2): 171-184.
17. Sanli C, Lambiotte R. Temporal Pattern of Online Communication Spike Trains in Spreading a Scientific Rumor: How Often, Who Interacts with Whom?[J]. Frontiers in Physics, 2015, 3: 79.
18. Zhao Q, Erdogdu M A, He H Y, et al. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1513-1522.
19. Gao S, Ma J, Chen Z. Modeling and Predicting Retweeting Dynamics on Microblogging Platforms[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 107-116.
20. Ma Z, Sun A, Cong G. On predicting the popularity of newly emerging hashtags in twitter[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1399-1410.
21. McCallum A, Nigam K. A comparison of event models for naive bayes text classification[C]//AAAI-98 workshop on learning for text categorization. 1998, 752: 41-48.
22. Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.
23. Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data[J]. Remote sensing of environment, 1997, 61(3): 399-409.
24. Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[J]. 2003.
25. Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.
26. Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks[J]. Physical Review E, 2001, 63(6): 066117.
27. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks[J]. Physical review letters, 2001, 86(14): 3200.
28. Galuba W, Aberer K, Chakraborty D, et al. Outtweeting the twitterers-predicting information cascades in microblogs[C]//Proceedings of the 3rd conference on Online social networks. 2010, 39(12): 3aAS3.
29. Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358.s
30. Cheng J, Adamic L, Dow P A, et al. Can cascades be predicted?[C]//Proceedings of the 23rd international conference on World wide web. ACM, 2014: 925-936.
31. Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth[J]. Marketing letters, 2001, 12(3): 211-223.
32. Kempe D, Kleinberg J, Tardos é. Maximizing the spread of influence through a social network[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003: 137-146.
33. Bakshy E, Hofman J M, Mason W A, et al. Everyone's an influencer: quantifying influence on twitter[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 65-74.
34. Jin F, Dougherty E, Saraf P, et al. Epidemiological modeling of news and rumors on twitter[C]//Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013: 8.
35. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 199-208.
36. Greene D, O'Callaghan D, Cunningham P. Identifying topical twitter communities via user list aggregation[J]. arXiv preprint arXiv:1207.0017, 2012.
37. Page, L., et al. ThePagerank citation ranking: Bring order to the web. Technical report, 1998.
38. Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46.5 (1999): 604-632.
39. Romero, Daniel M., et al. "Influence and passivity in social media." Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2011. 18-33.
40. 伍倩. 基于协同过滤的用户兴趣发现[D]. 上海交通大学, 2012.
41. 仇钧. 基于微博社会网络的用户兴趣模型研究[D]. 上海: 上海交通大学, 2013.
42. Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53.
43. Ali K, Van Stam W. TiVo: making show recommendations using a distributed collaborative filtering architecture[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 394-401.
44. Takács G, Pilászy I, Németh B, et al. Scalable collaborative filtering approaches for large recommender systems[J]. The Journal of Machine Learning Research, 2009, 10: 623-656.
45. http://baike.baidu.com/subview/45281/8045345.htm#viewPageContent
46. Tatler B W, Baddeley R J, Gilchrist I D. Visual correlates of fixation selection: Effects of scale and time[J]. Vision research, 2005, 45(5): 643-659.
47. https://github.com/CrawlScript/WebCollector
48. Hardt D. The OAuth 2.0 authorization framework[J]. 2012.
49. Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
50. Cilibrasi R L, Vitanyi P. The google similarity distance[J]. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(3): 370-383.
51. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
52. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

# 攻读学位期间取得的成果

[1] 陈江, 刘玮, 巢文涵, 王丽宏. 融合热点话题的微博转发预测研究[A].全国第十四届计算语言学学术会议（CCL2015）[C].(已录用)

[2] 刘玮, 王丽宏, 陈江, 张同虎, 马宏远, 王博. 网络转发行为预测方法及装置:中国, 国家计算机网络与信息安全管理中心. CN104933475A.2015/09/23.(专利)

# 致谢

值此论文完成之际，谨向给予我无私帮助的老师和同学们致以最诚挚的谢意！

在校内导师巢文涵、校外导师王丽宏的精心指导下，我顺利完成了本论文的撰写。王老师作为国家互联网应急中心的总工程师，工作繁忙，但她仍然坚持不计回报地投入精力对我的研究工作进行指导，王老师对我学术上和生活上的指导和关照，让我很是钦佩和感激。巢老师作为我的校内导师，就算是在我去国家互联网应急处理中心实习不在学校的日子里，依然时刻关注我研究工作的进展，每次遇到瓶颈问题时，老师的指导和意见都会起到关键催化作用，除了学术研究，生活上巢老师也给了我很多的指导和关照。研究生生活只有短短的两年半，但是两位导师在学术研究上和生活中树立的榜样，是我在以后的工作和生活当中都应该学习的。

除了两位导师的直接指导，实验室的师兄、师姐、同窗、班里的同学也给了很多的帮助。在这里我要感谢国家互联网应急中心刘玮博士的精心指导，除了两位导师，刘玮博士是对我研究工作指导最多的了，真诚地感谢她在繁忙工作之余对我工作的细心指导。同时也感谢实验室徐庆星、周庆、王文声、张永强、李哲豪、王耀华、张全龙、黄彦、魏萍等同窗以及其他所有帮助过我的同学，大家的宽容、无私、互帮、互学丰富了我的研究生生活。同时，我要感谢我的室友高峰斌、张红刚、李宁等人对我晚归或早起等不良生活习惯的包容。

两年半的研究生生活，离不开家人的理解、关心和支持，在这里，也要特别感谢我的家人对我研究生生活物质上的和精神上的支持，对我生活的关心，以及对不能陪在他们身边的理解。是他们的爱和理解，我才能全心地投入到研究工作当中，顺利地完成学业。

最后，衷心地感谢各位评审老师百忙之中抽出宝贵的时间评阅我的论文，谢谢！

再次感谢所有帮助、支持和理解我的人，我将在以后的工作和生活中更加努力，活到老学到老，一步一步走向更美好的生活。

1. Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter[C]//ICWSM. 2011. [↑](#endnote-ref-2)
2. 李英乐, 于洪涛, 刘力雄. 基于 SVM 的微博转发规模预测方法[J]. 计算机应用研究, 2013, 30(9): 2594-2597. [↑](#endnote-ref-3)
3. 曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779-790. [↑](#endnote-ref-4)
4. Kanavos A, Perikos I, Vikatos P, et al. Modeling ReTweet Diffusion Using Emotional Content[M]//Artificial Intelligence Applications and Innovations. Springer Berlin Heidelberg, 2014: 101-110. [↑](#endnote-ref-5)
5. Perikos I, Hatzilygeroudis I. Recognizing emotion presence in natural language sentences[M]//Engineering Applications of Neural Networks. Springer Berlin Heidelberg, 2013: 30-39. [↑](#endnote-ref-6)
6. Ekman P. Basic Emotions. Handbook of Cognition and Emotion Psychology Department[J]. 1999. [↑](#endnote-ref-7)
7. Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1633-1636. [↑](#endnote-ref-8)
8. Kschischang F R, Frey B J, Loeliger H A. Factor graphs and the sum-product algorithm[J]. Information Theory, IEEE Transactions on, 2001, 47(2): 498-519. [↑](#endnote-ref-9)
9. Loeliger H A. An introduction to factor graphs[J]. Signal Processing Magazine, IEEE, 2004, 21(1): 28-41. [↑](#endnote-ref-10)
10. Yang Z, Tang J, Li J, et al. Social community analysis via a factor graph model[J]. IEEE Intelligent Systems, 2010 (3): 58-65. [↑](#endnote-ref-11)
11. Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter[C]//System Sciences (HICSS), 2010 43rd Hawaii International Conference on. IEEE, 2010: 1-10. [↑](#endnote-ref-12)
12. Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[C]//Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010: 177-184. [↑](#endnote-ref-13)
13. 张旸, 路荣, 杨青. 微博客中转发行为的预测研究[J]. 中文信息学报, 2012, 26(4): 109-114. [↑](#endnote-ref-14)
14. Hong L, Dan O, Davison B D. Predicting popular messages in twitter[C]//Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 57-58. [↑](#endnote-ref-15)
15. Bandari R, Asur S, Huberman B A. The Pulse of News in Social Media: Forecasting Popularity[C]//ICWSM. 2012: 26-33. [↑](#endnote-ref-16)
16. Ma H, Qian W, Xia F, et al. Towards modeling popularity of microblogs[J]. Frontiers of Computer Science, 2013, 7(2): 171-184. [↑](#endnote-ref-17)
17. Sanli C, Lambiotte R. Temporal Pattern of Online Communication Spike Trains in Spreading a Scientific Rumor: How Often, Who Interacts with Whom?[J]. Frontiers in Physics, 2015, 3: 79. [↑](#endnote-ref-18)
18. Zhao Q, Erdogdu M A, He H Y, et al. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discover and Data Mining. ACM, 2015: 1513-1522. [↑](#endnote-ref-19)
19. Gao S, Ma J, Chen Z. Modeling and Predicting Retweeting Dynamics on Microblogging Platforms[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 107-116. [↑](#endnote-ref-20)
20. Ma Z, Sun A, Cong G. On predicting the popularity of newly emerging hashtags in twitter[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1399-1410. [↑](#endnote-ref-21)
21. McCallum A, Nigam K. A comparison of event models for naive bayes text classification[C]//AAAI-98 workshop on learning for text categorization. 1998, 752: 41-48. [↑](#endnote-ref-22)
22. Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185. [↑](#endnote-ref-23)
23. Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data[J]. Remote sensing of environment, 1997, 61(3): 399-409. [↑](#endnote-ref-24)
24. Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[J]. 2003. [↑](#endnote-ref-25)
25. Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27. [↑](#endnote-ref-26)
26. Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks[J]. Physical Review E, 2001, 63(6): 066117. [↑](#endnote-ref-27)
27. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks[J]. Physical review letters, 2001, 86(14): 3200. [↑](#endnote-ref-28)
28. Galuba W, Aberer K, Chakraborty D, et al. Outtweeting the twitterers-predicting information cascades in microblogs[C]//Proceedings of the 3rd conference on Online social networks. 2010, 39(12): 3âAS3. [↑](#endnote-ref-29)
29. Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358.s [↑](#endnote-ref-30)
30. Cheng J, Adamic L, Dow P A, et al. Can cascades be predicted?[C]//Proceedings of the 23rd international conference on World wide web. ACM, 2014: 925-936. [↑](#endnote-ref-31)
31. Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth[J]. Marketing letters, 2001, 12(3): 211-223. [↑](#endnote-ref-32)
32. Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003: 137-146. [↑](#endnote-ref-33)
33. Bakshy E, Hofman J M, Mason W A, et al. Everyone's an influencer: quantifying influence on twitter[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 65-74. [↑](#endnote-ref-34)
34. Jin F, Dougherty E, Saraf P, et al. Epidemiological modeling of news and rumors on twitter[C]//Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013: 8. [↑](#endnote-ref-35)
35. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 199-208. [↑](#endnote-ref-36)
36. Greene D, O'Callaghan D, Cunningham P. Identifying topical twitter communities via user list aggregation[J]. arXiv preprint arXiv:1207.0017, 2012. [↑](#endnote-ref-37)
37. Page, L., et al. *ThePagerank citation ranking: Bring order to the web*. Technical report, 1998. [↑](#endnote-ref-38)
38. Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." *Journal of the ACM (JACM)* 46.5 (1999): 604-632. [↑](#endnote-ref-39)
39. Romero, Daniel M., et al. "Influence and passivity in social media." *Machine learning and knowledge discovery in databases*. Springer Berlin Heidelberg, 2011. 18-33. [↑](#endnote-ref-40)
40. 伍倩. 基于协同过滤的用户兴趣发现[D]. 上海交通大学, 2012. [↑](#endnote-ref-41)
41. 仇钧. 基于微博社会网络的用户兴趣模型研究 [D][D]. 上海: 上海交通大学, 2013. [↑](#endnote-ref-42)
42. Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53. [↑](#endnote-ref-43)
43. Ali K, Van Stam W. TiVo: making show recommendations using a distributed collaborative filtering architecture[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 394-401. [↑](#endnote-ref-44)
44. Takács G, Pilászy I, Németh B, et al. Scalable collaborative filtering approaches for large recommender systems[J]. The Journal of Machine Learning Research, 2009, 10: 623-656. [↑](#endnote-ref-45)
45. 注1 http://news.sina.com.cn/hotnews/ [↑](#footnote-ref-2)
46. 注2 http://news.163.com/rank/ [↑](#footnote-ref-3)
47. 注3 http://news.ifeng.com/hotnews/ [↑](#footnote-ref-4)
48. http://baike.baidu.com/subview/45281/8045345.htm#viewPageContent [↑](#endnote-ref-46)
49. Tatler B W, Baddeley R J, Gilchrist I D. Visual correlates of fixation selection: Effects of scale and time[J]. Vision research, 2005, 45(5): 643-659. [↑](#endnote-ref-47)
50. https://github.com/CrawlScript/WebCollector [↑](#endnote-ref-48)
51. Hardt D. The OAuth 2.0 authorization framework[J]. 2012. [↑](#endnote-ref-49)
52. Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41. [↑](#endnote-ref-50)
53. Cilibrasi R L, Vitanyi P. The google similarity distance[J]. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(3): 370-383. [↑](#endnote-ref-51)
54. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013. [↑](#endnote-ref-52)
55. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119. [↑](#endnote-ref-53)