融合热点话题的微博转发预测研究*

陈江[1]，刘玮*[2,3,4]，巢文涵[1]，王丽宏[2]

(1.北京航空航天大学，北京市，100191；

2.国家计算机网络应急技术处理协调中心，北京市，100029；

3. 中国科学院计算技术研究所，北京 100190；

4.中国科学院大学，北京 100049)

摘要：微博转发行为是实现信息传播的重要方式，微博转发预测对微博影响力分析、微博话题分析具有重要价值。现有微博转发预测研究大多围绕消息属性、用户属性等微博自身特征，本文提出融合热点话题的微博转发预测方法，对背景热点话题内容和传播趋势对用户转发行为的影响进行量化分析，提出融合背景热点信息的转发兴趣、转发活跃度、行为模式等特征，并基于分类算法建立了面向热点话题相关微博的转发预测模型，在真实数据上的实验结果表明，本文方法的预测准确性达到 96.6%，提升幅度最高达到 12.14%。

关键词：转发行为；转发预测；热点话题

中图分类号：TP391　　**文献标识码：**A

Research on Microblog Forwarding Prediction based on Hot Topics

Jiang Chen1, Wei Liu* 2,3,4, Wenhan Chao1, Lihong Wang2

(1.Beihang University, Beijing 100191;

2. National Computer network Emergency Response technical Team/Coordination Center of China, Beijing 100029;

3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190;

4. University of Chinese Academy of Sciences, Beijing, 100049)

Abstract: Microblog forwarding is an important way to achieve the information dissemination, Microblog forwarding prediction is of great value in the analysis of microblog influence and microblog topic analysis. Existing microblog forward prediction methods mostly focus on microblog attributes and user attributes. In this paper, a microblog forwarding prediction method based on hot topics is proposed. We quantitatively analyzed the impact of hot topics content and transmission tendency on users' forwarding behavior, and then introduced features that fused with hot topics such as forwarding interest, forwarding activity and behavior pattern. Finally, we established the hot topic oriented microblog forwarding prediction model based on the classification algorithm. Our experimental results on real data show that the accuracy of this method is 96.6%, and the max improvement of the proposed method is 12.14%.

Keywords: Microblog Forward, Forwarding Prediction, Hot Topic

1. 引言

微博是一个基于用户关系的信息分享、传播以及获取平台[1]。微博从 2009 年发布至今，迅速以其内容简洁、交互简便和快速传播等特点，发展成为人们表达观点、抒发情绪、传递信息的重要社会媒体。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，我国微博用户规模为 2.04 亿，其中手机微博用户数为 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，使得某些微博能够在短时间内形成极大的关注和影响。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

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目（61170230）；国家科技支撑计划项目（2012BAH46B01）；国家高技术研究发展计划（SS2014AA012303）；国家高技术研究发展计划（863 计划）（2014AA015105）

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。

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，社会上发生的热点话题信息是其中重要的影响因素。

本文以此为出发点，研究融合背景热点话题的用户转发行为预测方法。本文提出融合热点话题的微博转发预测方法，对背景热点话题内容和传播趋势对用户转发行为的影响进行量化分析，提出了融合背景热点信息的转发兴趣、转发活跃度、行为模式等特征，并基于分类算法建立了面向热点话题相关微博的转发预测模型。在真实数据上的实验结果表明，本文方法的预测准确性达到 96.6%，提升幅度最高达到 12.14%。

本文的组织结构如下：引言部分介绍问题背景和研究现状；第 2 节介绍相关工作；第 3 节介绍问题描述；第 4 节介绍热点话题对用户转发的相关性问题；第 5 节介绍融合热点话题的特征分析方法；第 6 节实验结果和分析；第 7 节是总结和下一步工作。

2. 相关工作

微博转发研究工作主要集中在提取转发和非转发行为区分度高的特征。Petrovic S[2]等人研究 Twitter 平台的转发预测问题，考虑了 tweet 用户相关特征，如粉丝数、关注数、tweet 发布量等，以及 tweet 本身特征，如标签、URL、tweet 长短等，基于机器学习方法构建转发预测模型。Galuba W 等人 [3]研究了用户 URL 提及频繁程度，通过追踪 URL 传播的方式，研究 Twitter 平台中 URL 的传播规律，构建基于用户提及 URL 的预测模型。李英乐[4]和曹玖新[5]等人通过微博客中用户特征和微博内容特征来研究微博转发及其预测问题，但方法过于依赖微博内容对微博转发的影响。Kanavos A[6]等人构建微博情绪模型，基于 tweet 内容特征的情感倾向研究 tweet 传播的广度和深度。论文[7][8][16]将微博是否会被转发转化成一个二分类问题，基于机器学习方法构建分类模型进行微博转发预测。

转发行为是促使微博在微博网络中病毒式传播[9][10]的关键问题，转发特征的研究主要分布在用户转发行为的研究[11][12][13]和微博转发规模预测的研究中 [14][15]，Zhang Y等人[16]研究不同特征对转发行为影响的差异性，从而构建基于特征加权的转发预测模型。Petrovic S等人[2]基于passive-aggressive算法预测微博是否会被转发，他们的研究发现微博博主是否认证及其粉丝数等会影响微博是否会被转发。Bandari R等人[17]将微博转发数量按不同等级划分（1-20,20-100,100-2400），构建多分类模型来预测微博转发规模。Ma Z[18]等人提取tweet的特征，基于机器学习方法，构建了标签的流行度预测模型对转发规模进行预测。

社交网络具有复杂网络特性，也有学者基于社交网络结构研究微博传播规律。这方面的研究主要基于社交网络的结构特征，构建社交网络拓扑图，在此基础上研究信息传播规律。Szabo G[19]等人研究在线内容的流行度问题，并构建流行度预测模型，但他们的研究具有平台局限性，可推广性差。Yang J[20]等基于传播关系网络，通过信息已经流过的节点，构建线性影响模型，预测信息传播的实时动态。

综上所述，现有的研究者主要基于微博特征或基于网络结构特征，研究微博转发预测问题，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[13]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，而用户对微博内容的综合认知程度会受到多种复杂因素影响。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。

针对上述问题，本文研究融合背景热点话题的用户转发行为预测方法。首先，提出话题背景知识获取和特征向量计算方法，用于表示用户对热点话题的综合认知程度，并将直接判断用户对微博的感兴趣程度问题转换成判断用户对微博所属热点话题的感兴趣程度问题，能够避免因用户历史微博内容局限性而导致的预测准确性低问题。其次，提出利用用户历史转发行为趋势特征及其与热点话题传播趋势一致性的计算方法，以此代表用户对热点话题的关注程度，进而表示该用户对热点话题微博的感兴趣程度，避免直接通过计算单条微博与用户兴趣相似程度所带来的不准确性问题。最后，基于分类算法建立面向热点话题相关微博的转发预测模型，在真实数据上开展实验验证。

3. 问题描述

融合背景热点话题的微博转发预测问题可以描述为$F = f(U, W, H)$，其中：U表示用户特征，W表示微博特征，H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，F表示用户行为，即用户u对微博w的动作，$F \in \{1, -1\}$，$F = 1$表示用户u转发了微博w，$F = -1$表示用户u没有转发微博w，用户u是否转发微博w的转发预测问题可以转化为二分类问题。

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。本文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，研究背景热点话题对微博转发行为的影响。在传统分类模型基础上，引入热点话题特征扩展特征空间，提高预测准确性。背景热点话题对转发行为的影响主要考虑热点话题内容和传播趋势两方面因素，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。

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。

4. 热点话题对用户转发的影响研究

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。热点话题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。

以“世界杯”热点话题为例，我们爬取新浪微博 2014 年 4 月 12 日至 9 月 13 日期间数据，统计微博总量变化趋势和话题相关微博总量的变化趋势。如图 1 所示，热点话题期间微博空间的微博总量和话题相关的微博总量都呈现出明显的增长趋势，表明用户转发行为会受到热点话题的影响。

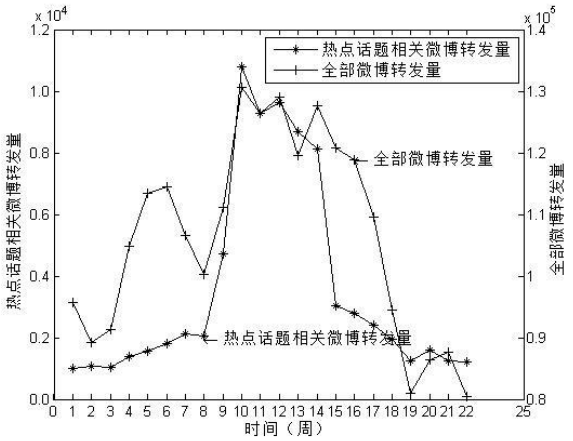

图 1 热点话题期间微博转发量变化趋势

进一步的，我们研究热点话题对微博用户转发行为的影响。我们针对 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。结果如图 2 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。我们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，表明仅基于用户历史微博计算的用户兴趣难以有效预测在新的热点话题下的用户转发行为，用户的转发行为会受到当前社会热点话题的影响。

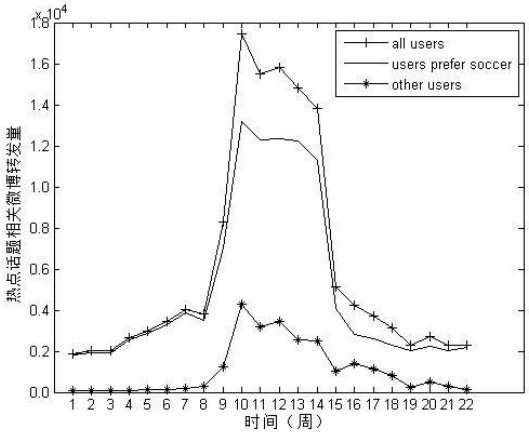

图 2 热点话题对各类用户微博转发量的影响

5. 特征分析

5.1 融合热点话题的用户转发兴趣特征

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[21]。微博用户的兴趣部分地通过用户的转发行为体现出来，我们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。

从第 4 节的分析，我们可以看出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，我们提出融合热点话题的用户转发兴趣特征，来计算用户转发兴趣与热点话题的匹配程度。

热点话题内容表示：背景热点话题文档级别的表示$D_topic = \{d_1, d_2, \dots, d_n\}$，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$S_topic = \{w_1, w_2, \dots, w_m\}$。

用户转发兴趣表示：通过用户历史转发微博内容来表示用户转发兴趣。用户历史转发微博表示为$D_user = \{d_1, d_2, \dots, d_n\}$, 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：$I\_user = \{w_1, w_2, \dots, w_m\}$。

融合热点话题的用户转发兴趣特征计算：定义为用户转发兴趣与背景热点的匹配程度，计算方法如下：

$$CO = |I\_user \cap S\_topic|$$

我们计算了转发微博数量对应用户转发兴趣特征的变化关系。如图 3 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（CO越大），则用户所转发与背景热点话题相关微博的数量越多，表明融合背景热点话题的用户转发兴趣能够有效预测用户的转发行为。

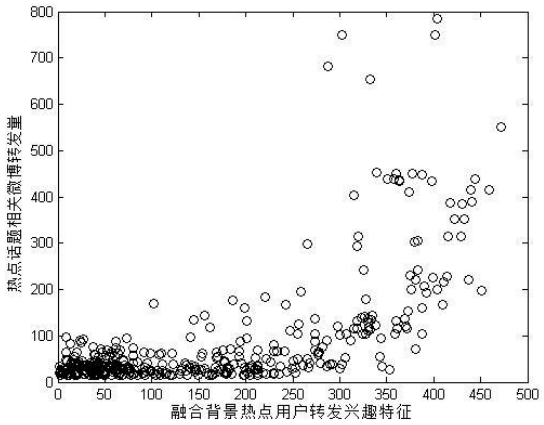

图 3 微博转发量对应融合背景热点话题的用户转发兴趣的关系图

5.2 融合热点话题的用户转发行为特征

1、融合背景热点话题的用户活跃度

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：

$$F = \left\{ \left| m_i^t \right| \left| \left| m_i^t \cap S \right| > \tau \right. \right\}$$

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

2、融合背景热点话题的用户行为一致性

第 4 节从内容上考虑影响用户转发行为的因素，从图 3 我们还可以看出，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。

一段时间内用户转发微博数量的变化趋势可以看做是时间轴上的一个概率分布P_user；我们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P_topic。我们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。

计算分布相似度，我们采用 *KL*（Kullback-Leiber divergence）散度又称相对熵（relative entropy）方法，该方法是用来描述两个概率分布之间差异性的一种方法[22]，*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。我们用 *KL* 距离来反映概率分布P_user和概率分布P_topic之间的关系。

$$D_{KL}(P\_user||P\_topic) = \sum P\_user\log(\frac{P\_user}{P\_topic})$$

$$D_{KL}(P\_topic||P\_user) = \sum P\_topic\log(\frac{P\_topic}{P\_user})$$

考虑到 KL 距离的非对称性，我们以

$$D_{KL} = \frac{1}{2}(D_{KL}(P\_user||P\_topic) + D_{KL}(P\_topic||P\_user))$$

来计算概率分布之间的关系。通过上述计算方法我们计算了转发微博与行为一致性之间的对应关系。

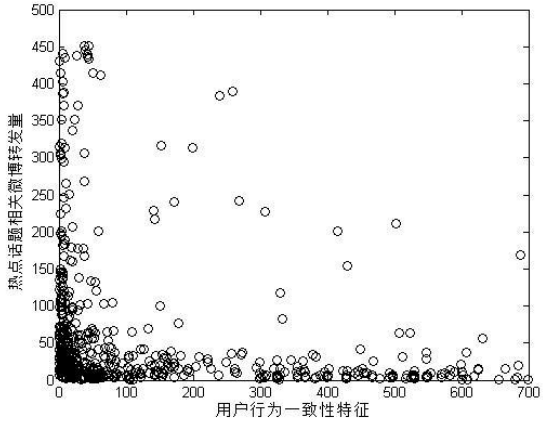

图 4 微博转发量对应行为一致性特征的关系图

如图 4 所示，当 *KL* 距离增大时，分布之间差异变大，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小，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，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。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。

5.3 融合热点话题的微博内容特征

本节针对待预测微博内容来分析热点话题对微博转发的影响。考虑到微博内容与热点话题越相关，得到转发的概率就越大，我们提出融合热点话题的微博内容特征，同样，我们用词集合$S_topic = \{w_1, w_2, \dots, w_m\}$来表示背景热点话题内容。我们对微博进行分词、去除停用词的预处理之后，将微博表示成一个词语级别的集合：$M\_mes = \{w_1, w_2, \dots, w_m\}$。由于微博内容很短且都是特征词语，我们用 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：

$$J_{SM} = \frac{|S\_topic \cap M\_mes|}{|S\_topic \cup M\_mes|}$$

我们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 5 所示，横坐标为J_{SM}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绿色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蓝色曲线表明，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。

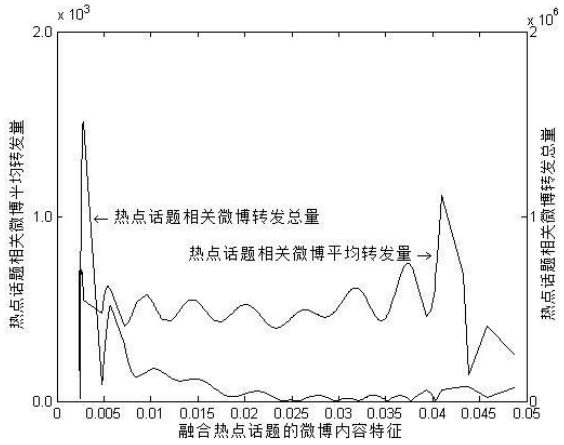

图 5 微博转发对应融合热点话题的微博内容特征对的关系图

6. 实验结果与分析

6.1 数据集构建

本文研究如何引入热点话题来提高微博转发预测准确性，本文首先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，利用新浪微博检索功能，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。我们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：

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

其中：n 表示用户个数，m 表示微博个数，$a_{ij} \in F$。标记后的数据集元素为一个三元组 $< u_i, m_j, a_{ij} >$，当$a_{ij} = 0$ 或 1时表示用户u_i转发微博m_j，当$a_{ij} = -1$时表示用户u_i没有转发微博m_j。

6.2 特征选取与对比方法

分类特征选择本文提出了融合热点话题的用户转发兴趣特征、用户活跃度、用户行为一致性、微博内容特征，如表 1 所示特征 1 到 4。

本文采用李英乐[4]等人的方法作为对比试验，该方法使用特征较全面且预测准确性较高，在特征可计算条件下，选择用户影响力、用户转发活跃度、用户发布活跃度、用户兴趣与微博相似度作为对比特征，如表 1 所示特征 5 到 8。

表 1 特征选取

序号	特征含义
1	融合热点话题的用户转发兴趣特征
2	融合热点话题的用户活跃度
3	融合热点话题的用户行为一致性
4	融合热点话题的微博内容特征
5	发布用户影响力（用户粉丝数）
6	用户转发活跃度
7	用户发布活跃度
8	用户兴趣与微博相似度

6.3 转发预测及评价指标

本文采用 SVM、朴素贝叶斯、贝叶斯信念网络、决策树等分类算法，来测试所选分类特征应用于转发预测时的效果。

评价方法采用准确率（Precision）、召回率（Recall）和综合评价指标(F-Measure)。

6.4 实验结果及分析

表 2 实验结果对比

classifiers	Precision			Recall			F-Measure		
	baseline	Ours	Combine	baseline	Ours	Combine	baseline	Ours	Combine
BayesNet	0.877	0.938	**0.94**	0.867	0.937	**0.939**	0.869	0.936	**0.938**
NaiveBayes	0.802	0.756	**0.813**	0.736	0.714	**0.784**	0.74	0.643	**0.789**
C4.5	0.886	0.951	**0.953**	0.883	0.948	**0.951**	0.884	0.947	**0.966**
LibSvm	0.839	0.942	**0.947**	0.841	0.938	**0.943**	0.84	0.936	**0.942**

我们将特征分为三组，分别在四种分类器上进行了对比试验。Baseline 方法是对比的基准方法，仅采用了用户和微博自身特征，Ours 表示本文所提特征，即融合了热点话题的转发特征，Combine 表示将用户和微博自身特征与融合热点话题的转发特征相结合，进行了综

合测试。如表 2 所示，在多个分类器上的测试结果表明本文所提出的融合热点话题的用户转发兴趣特征、用户活跃度、用户行为一致性、微博内容特征能够有效提升转发预测准确性，与传统用户和微博自身特征相结合后，能够进一步提升效果，其中，采用 C4.5 分类器时的预测效果最好，达到 96.6%，对基于 SVM 分类器的预测模型提升效果最高，达到 14.12%，采用 NaiveBayes 分类器的实验中，ours 的实验效果略差于 baseline，这是因为我们选取的特征不完全满足朴素贝叶斯的条件独立性假设，导致其在分类准确率上有一定的牺牲，但是从数据上可以看出，本文所提特征在与 baseline 特征结合后能够提升分类准确率。引入外部热点话题并融合其内容和传播趋势对用户转发行为的影响因素，能够有效提升转发行为的预测准确性。本文的训练集和测试集按照时间先后相互独立，预测准确性的提高也表明了本文所提特征能够很好的刻画用户转发行为模式，具有较好的长期预测效果。

7. 总结与展望

微博转发行为是实现信息传播的重要方式，微博转发预测对微博影响力分析、微博话题分析具有重要价值。现有微博转发预测研究大多围绕消息属性、用户属性等微博自身特征。本文融合背景热点话题研究了外部热点话题对用户转发行为的影响，并对影响因素进行量化分析，提出了融合背景热点信息的转发兴趣、转发活跃度、行为模式等特征。根据热点话题前期用户的转发行为，预测用户是否会转发热点话题相关的微博。并基于分类算法建立了面向热点话题相关微博的转发预测模型，在真实数据上的实验结果表明，本文方法的预测准确性达到 96.6%，提升幅度最高达到 12.14%。通过引入背景热点话题内容和传播趋势特征，能够有效提升用户转发行为预测准确性。在未来工作中，可以进一步改进热点话题内容表示方法，以及热点话题内容和用户兴趣相似性度量方法，进一步的提高预测效果。

参考文献

[1]http://baike.baidu.com/link?url=Qsdt8nZWb5Q_iTpNaS41WlK2ZxMJeaUC8g9cuHWpK2V01Grlj6wiUx7C4170CTm2988GAfKuQoMHuWdmq1V65C0zVgKyuU1qMYlZ44yMBe_

[2] Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter[C]//ICWSM. 2011.

[3] 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.

[4]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.

[5]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.

[6] 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.

[7] Ma H, Qian W, Xia F, et al. Towards modeling popularity of microblogs[J]. Frontiers of Computer Science Selected Publications from Chinese Universities, 2013, 7(2):171-184.

[8] Ying-Le L I, Hong-Tao Y U, Liu L X. Predict algorithm of micro-blog retweet scale based on SVM[J]. Application Research of Computers, 2013, 30(9):2594-2597.

[9] Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks[J]. Phys.rev.e, 2001, 63(6):138-158.

[10] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks.[J]. Physical Review Letters, 2001, 86(14):3200-3203.

[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] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[J]. Cikm ', 2010:1633-1636.

[14] Jiang Y, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. Fourth International Aaai Conference on Weblogs & Social Media, 2010.

[15] Hong L, Dan O, Davison B D. Predicting Popular Messages in Twitter[J]. Www68, 2011.

[16] Zhang Y, Rong L U, Yang Q. Predicting Retweeting in Microblogs[J]. Journal of Chinese Information Processing, 2012, 26(4):109-108.

[17] Bandari R, Asur S, Huberman B A. The Pulse of News in Social Media: Forecasting Popularity[J]. Sixth International Aaai Conference on Weblogs & Social Media, 2012.

[18]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.

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

[20] Yang J, Leskovec J. Modeling information diffusion in implicit networks[C]//Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 599-608.

[21] http://baike.baidu.com/subview/45281/8045345.htm#viewPageContent

[22]http://baike.baidu.com/link?url=uSpPDwuklxXEAOvRxbkQCpT_pGlLfjTt1dbrHfKS5Iz4Kq8UfNXRmuujk1A77KpPpSLfj8fVbwf5qYcUD2i1a

作者简介：

第一作者：陈江（1990–），男、硕士研究生，研究方向为社交网络数据挖掘、网络信息安全。E-mail: jiangchencj@163.com

通讯作者：刘玮（1984–），女、湖北、工程师，博士研究生，研究方向为社交网络数据挖掘、信息过滤。E-mail: liuwei@isc.org.cn

第三作者：巢文涵：博士，北京航空航天大学计算机学院讲师。主要研究兴趣为：机器翻译、自然语言处理、智能检索、数据挖掘等。