

硕 士 学 位 论 文  
文献综述

融合行业的微博话题检测与情感分析技术的研究与实现

作者姓名 杨 颖

作 者 学 号 ZY1221217

学 科 专 业 软件工程

专 业 方 向 软件工程与管理

学院指导教师 王海泉 讲师

企业实习导师 周礼元

企业实习导师单位 北京三源合众科技有限公司

培 养 院 系 软件学院

论文起止时间 2013年11月-2014年12月

开题报告提交时间 2013年11月

**摘 要**

微博即微型博客，是Web2.0时代下衍生出的一种新型社会网络，已逐渐成为人们日常交流、通信、娱乐的基本工具。通过微博能够将人的社会行为以数据的形式加以记录和保存，对这些数据的分析挖掘，有可能发现其中蕴含着的关于个人行为、群体行为的特征和规律。如果在现有研究的基础之上，引入“行业”这一重要的社会关系，将能够更为准确的分析个体和群体的行为，特别是对社会热点事件所涉及的专业性，从而更好的引导舆论。本文首先给出微博的概念，然后从微博社会网络特性研究和微博内容的语义信息研究两方面综述了基本研究现状与发展趋势，分析了现有针对微博进行的行业分类方法，最后指出了当前研究中存在的问题。

关键词：微博，行业分析，话题检测，情感分析

Abstract

Microblog is a new social network developed in the Web2.0 era and already becomes a basic tool for people to communicate and entertain themselves. Through the tools like microblog, human social behavior can be recorded and saved in the form of data. By these data analyzed and data mining, characteristics and behavioral law among personal behavior, group behavior and social interaction between groups may be founded. However, existing studies tend to ignore the social attributes of microblog users of their behavior and social relations. If "industry" this important social relations can be added to the existing research, the behavior of individuals and groups will be more accurately analyzed. Especially on hot social events involving professional, public opinion will be better guided. This paper firstly gives the concept of microblog. Secondly it shows an overview of basic research status and development trend of the existing social networks characteristics and semantic information of microblog. Thirdly the exciting methods to classify user occupation are introduced and finally the problems in the current study are pointed out.

**Key words**: Microblog, Occupation classification, Topic detection, Sentiment analysis

目 录

[第一章 微博概述 1](#_Toc372980866)

[第二章 微博社会网络特性研究 2](#_Toc372980867)

[2.1 微博用户网络特性分析 2](#_Toc372980868)

[2.1.1 基本社会网络特性研究 2](#_Toc372980869)

[2.1.2 用户影响力研究 2](#_Toc372980870)

[2.1.3 用户特征分析与分类研究 4](#_Toc372980871)

[2.2 微博消息网络的特性分析 5](#_Toc372980872)

[第三章 微博内容的语义信息研究 6](#_Toc372980873)

[3.1 微博内容的基本语义分析 6](#_Toc372980874)

[3.1.1 自动摘要 6](#_Toc372980875)

[3.1.2 话题检测 7](#_Toc372980876)

[3.1.3 标签推荐 8](#_Toc372980877)

[3.1.4 可信度研究 9](#_Toc372980878)

[3.2 微博内容的情感语义分析 9](#_Toc372980879)

[第四章 当前存在的问题 11](#_Toc372980880)

[参考文献 12](#_Toc372980881)

# 微博概述

微博(Microblog)即微型博客，是一种集成化、开放化的互联网社交服务，用户可通过Web、即时通信、电子邮件和手机等方式用很短的文字描述其当前状态。为方便与移动终端的接入，其每一篇“微博”限定在140字左右，同时也可上传音视频、图片。用户与用户之间信息的传递通过“关注—被关注”(Following-Follower)来实现，用户之间在微博平台上通过转发的方式对其他用户的微博进行传播。随着移动通信网络和Web技术的不断发展，微博逐渐成为人们日常交流、通信、娱乐的基本工具。

微博的出现以2006年推特(Twitter)的创办为标志，从2008年开始Twitter得到了广泛的应用。自2009年中国各主流门户推出了各自的微博产品，微博在中国也得到了迅猛发展，以新浪微博为代表，包括腾讯、搜狐、网易、凤凰等其他门户纷纷加入微博阵营，用户数目与日剧增，截至2013年6月底，我国微博用户规模达到3.31亿[1]。通过微博这样的工具能够将人的社会行为以数据的形式加以记录和保存，这就为研究者提供了一个大范围的、长时间的、连续的人群的各种行为和互动的可靠数据来源，通过对这些数据的分析挖掘，有可能发现其中蕴含着的关于个人行为、群体行为和群体间社会交互行为的特征和规律。

目前，对微博的研究主要从两方面展开：一是分析微博的社会网络特性，微博是一种新兴的社会网络，因此也具有社会网络的各种特性，微博社会网络可以划分为两类，微博用户形成的社会网络和微博消息在传播过程中形成的社会网络，很多研究都是围绕这两种类型的社会网络展开特性分析；二是分析微博内容中蕴含的语义信息，微博信息呈现文本碎片化、海量等特点，信息利用成本高，无法直接获取微博中蕴含的语义信息，因此很多研究试图从微博内容中挖掘语义信息，特别是情感语义，用于发现用户对于热点事件的观点。

# 微博社会网络特性研究

## 微博用户网络特性分析

### 基本社会网络特性研究

自2006年Twitter获得广泛应用后，微博这一新兴社会网络逐渐引起学术领域的关注，研究人员对其社会网络特性进行了相关分析。Java等人[2]对Twitter的基本功能及特点进行了详细介绍，并对其社会网络特性进行了初步分析，数据集包括76,177个用户和1,348,543条微博信息，结果表明Twitter表现出一定的幂律分布和小世界等特性；同时还研究了Twitter用户社会网络的拓扑结构和地理位置等特征，并从个人和社区两个不同层次对用户使用Twitter的意图进行了分析，结果表明用户一般通过Twitter讨论日常事件或共享信息。Kwak等人[3]对整个Twitter进行了定量分析，数据集包括Twitter上的4,170万用户、14.7亿用户社会关系、4,262个热点话题和1.06亿微博等大量数据信息，通过分析Twitter用户间“关注—被关注”的拓扑结构，对Twitter用户社会网络统计特性进行了分析，统计结果表明Twitter在一定程度上表现出用户间的互惠性，但其社会网络特性较一般社会网络存在一定的偏差，例如，用户的follow数并不呈现幂律分布以及分割度更小等不同于一般社会网络的基本特性。

也有国内学者对国内的微博社交网络展开了相关研究。樊鹏翼等人[4]对国内最大的微博系统——新浪微博——开展了主动测量，并结合已有的在线社会网络测量结果，对新浪微博的网络拓扑和用户行为特征进行了分析和比较。主要发现包括：1）新浪微博网络具有小世界特性；2）新浪微博网络的入度分布属于幂次分布，而出度分布表现为某种分段幂律函数；3）与类似社会网络相比，新浪微博网络的出入度不具有相关性；4）新浪微博网络属于同配网络。夏虎[5]计算了新浪微博、腾讯微博以及Twitter三个数据集的网络基本特征，通过对比找出不同移动社交网络的相同点和不同点，经过对比分析发现腾讯微博度分布的中间部分是由朋友关系占主体，因此比较适合于进行基于朋友的推荐；而在新浪微博中，度分布的中段以后是由大度节点占主体，比较适合于进行基于关键节点的推荐。

### 用户影响力研究

在用户网络中发现用户影响力不仅有助于用户推荐，对于微博网络中的商业运营模式也有着重要的意义，如利用用户影响力实现广告推送等。因此，用户影响力探测也是微博用户网络特性分析中的一个研究热点。微博用户影响力探测的方法可分为两类。一种方法是利用用户关系网络图的整体拓扑结构探测用户影响力；另外一种方法则是通过用户发布微博的网络传播影响力间接探测用户影响力。

基于用户关系网络图的方法从两个不同的角度去度量用户影响力。最简单直接的方法是利用用户的关注数大小，即网络图中节点度的大小，来评定用户影响力的大小。这种方法计算简单但效果不佳。另一种方法则是将用户这种“关注—被关注”关系看作是Web网页间的超链接关系，利用Web网页排名中常用的PageRank和HITS等算法进行用户影响力评定。例如，Java等人[2]利用HITS算法对Twitter用户网络图中用户影响力进行探测；Kwak等人[3]则利用PageRank算法对Twitter中的用户影响力进行探测，并通过用户的跟随者数和用户发布微博的转发数等不同方法与之进行对比；Weng等人[6]提出了一种TwitterRank算法，在PageRank算法基础上，考虑用户所关注话题间的相似度和用户关系拓扑结构，从而发现Twitter中与话题相关且具有一定影响力的用户。

基于微博在整个用户网络中的传播覆盖度，即用户发布微博的被转发次数或其他用户在微博中提到该用户的次数，来度量用户的影响力大小。如Cha等人[7]对比分析了3种不同的用户影响力度量方法：用户的跟随者数、用户的微博转发数和用户在微博中通过“@”被关联的次数，认为用户的跟随者数越多，并不能真正说明该用户在用户群中的认可度越高，而用户的微博转发数以及用户在微博中通过“@”被关联的次数则能更准确地度量用户的实际影响力。Ye等人[8]将社交影响力划分为3种影响力和5种排序准则，分别是粉丝数量影响力、回复影响力、转推影响力，按粉丝数、消息数（回复和转推）、响应用户数（回复者和转推者）的排序准则。通过采用斯皮尔曼等级相关系数和肯德尔等级相关系数(Kendall Tau Rank Correlation Coefficient)，对5种排序准则进行了计算比较，认为拥有最多回复者的用户集合是最稳定的，并按照回复者数量进行社交影响力排序作为标准。前文提到的Kwak[3]等人在分析Twitter用户网络的基本特性时，提出了微博转发树的概念，但并未用于度量用户影响力，而Bakshy等人[9]则利用转发树的概念作为用户影响力的度量标准，认为在Twitter中用户的影响力是通过用户发布微博的转发规模所决定的，即消息传播的广度和深度。

国内的舒琰等人[10]结合网页PageRank算法，针对微博单纯通过对比粉丝人数的排名，提出了一种新的根据粉丝重要性的排名算法。首先通过抓取大量的用户数据，建立PageRank矩阵，结合传统网页PageRank的计算方法来计算微博用户的PageRank值。该算法能有效减少垃圾用户对微博排名的影响，来提高排名的公平性与准确性。

### 用户特征分析与分类研究

现实社会网络中，用户间各种不同的关联关系是不尽相同的。例如，用户A与用户B是基于朋友关系建立的关联关系，而用户A与用户C可能存在一种敌对的关联关系。微博作为现实社会网络在虚拟互联网中的具体展现，相应地，微博用户社会网络中不同类型的链接关系也必然存在差异，研究用户间不同的链接关系对于更深入的理解微博社会网络特性有着重要的作用[11]。通过微博用户社会网络分析用户特征，并根据这些特征进行用户分类也是重要的研究内容之一。例如，Krishnamurthy等人[12]通过分析Twitter用户关注和被关注数之间的关系分析了用户的特征，将用户分为三类：广播人(broadcaster )、一般人(acquaintance)和垃圾虫(miscreant)。Pal等人[13]收集同一主题中的微博，然后提取该主题下所有微博发布者的特征，并根据其特征将用户聚成两类，将所聚类别中的作者进行排序，并找出最具权威的用户，实验结果对权威用户的发现提供了许多有用特征。Welch等人[14]认为在Twitter网络结构图中不同节点之间的边代表用户间不同的链接关系，分别针对用户间的follow 关系和微博转发关系进行了相应的分析，并指出利用这种链接关系对用户排名算法有较好的改进。而前文提到的国内研究者樊鹏翼等人[4]发现新浪微博用户发博时间具有明显的日分布和周分布模式以及新浪微博用户博文数目分布表现为威布尔分布的特性。

目前将微博内容挖掘和用户行业分析结合的研究比较少。Zhou等人[47]在利用微博消息内容、用户行为以及用户关系这三种数据进行用户行业分类问题上进行了探索，使用序列最小优化算法对从房地产、娱乐、体育、IT这4个行业各获取500名实名认证用户的数据进行训练，得到21,974个用于分类的特征词，实验表明该方法能够较为准确的对用户行业进行分类。然而文章中并没有考虑用户所发微博的评论数和转发数，以及用户间共同好友数目这些重要的关系衡量指标。加利福尼亚大学伯克利分校的一个项目[[1]](#footnote-1)中提到了针对twitter所做的一个行业分析的工具，它是首先通过Klout scores得到行业领袖的人选，再由行业领袖所发布的信息得到行业的特征，最后根据K-means聚类进行用户行业分析得出用户所属的行业。它的目的主要是帮助求职者快速建立一个行业专业人士的关注圈。

以上两种方法都与行业领袖的选取有很大的关系，这种方法的难点和缺陷在于：

1. 行业领袖的选择是否足够和全面对最后的结果有很大影响。
2. 从行业领袖所发布的信息中提取出行业特征也有一定的难度和偏差，因为行业领袖发布的信息不一定与行业有很大关联。

## 微博消息网络的特性分析

与一般在线社会网络相同，微博也允许用户在线、实时发布文本信息，然而，不同的是微博在信息长度上限制在140个字符之内，同时语法结构自由，支持手机等移动设备实时发布信息，这使得微博消息传播网络无论是在传播范围上还是速度上都具有更大的优势。因此，微博消息网络的特性分析及消息在网络中的传播机制也是最近的研究热点。

Yang等人[15]从用户贡献模式（即用户每月发布微博数目的分布情况）、Web导航（即用户发布微博中含有超链接的目的指向）和用户社会网络整体结构模式等三个方面对比分析了Twitter与传统博客在信息传播结构上的区别。Kwak等人[3]认为转发方式是Twitter消息传播中最有效的方式之一，基于微博转发关系，针对不同的热点话题，构建了微博转发树，并对微博转发机制进行了研究。通过对微博转发树的广度进行分析发现，Twitter用户并非通过直接接收的方式获得信息，即大部分用户并不是该消息发布者的直接关注者，而是通过用户与用户间转发微博而间接收到消息，且微博一经转发，不管用户关注者有多少，该微博总会被传播到一定数量的用户。对微博转发树的深度进行分析发现，微博转发树中约占97.6%其转发深度小于6。这体现了Twitter消息网络中信息传播范围广且速度快的特点，即病毒式传播特点。

国内的张赛等人[16]提出了一种三角和算法用于探测用户粉丝数阈值。该算法根据散点分布的统计规律来估计使微博热度达到某一值的粉丝数的临界值，发现为使微博热度大于10，用户粉丝数应大于150。文章中其他测量分析结果表明：新浪微博具有很强的“名人效应”，用户频繁地发帖并不能引起较大的关注，热门微博的热度几乎都以激增方式增长。

# 微博内容的语义信息研究

微博不仅具有社会网络的结构性特征，微博内容本身也包含了丰富的语义信息。基于微博内容的语义分析，其研究工作主要是从用户发布的微博内容中挖掘出有价值的信息，可分为面向事实(Fact-Oriented)的文本挖掘和面向观点(Opinion-Oriented)的文本挖掘两类。其中面向事实的文本挖掘主要包括热点话题检测、主题抽取、垃圾信息处理、自动摘要等，在本文中归纳为微博内容的基本语义分析。而面向观点的文本挖掘即情感分析或观点挖掘是指从用户发布的信息中挖掘出其对讨论主题的潜在情感信息。因此，基于微博的情感语义分析研究工作主要是指对微博内容进行情感分析和观点挖掘。

## 微博内容的基本语义分析

### 自动摘要

自动摘要/文摘即利用计算机自动编制摘要/文摘，最早的自动文摘和信息抽取模型是Sager等人[17]提出的。Sager等人采用子语语法，从医学文章中抽取信息。该方法依赖子语领域的语义知识。我国研究者针对汉语特点进行了不懈的努力，取得了一些成就。哈尔滨工业大学王开铸教授等人[18]提出了偏重于篇章物理结构的“篇章计算模型”，并研制了一个基于篇章理解的军事领域自动文摘实用系统MATAS(Military Area Text Automatic Abstract System)。该系统考虑了句子之间的语义联系，但是系统不能自动判断段落的文体，需要人工干预。由于文体将直接影响到该系统中篇章的形式化表示，所以该系统提出的方法难以实现。

然而微博具有海量数据、短文本性、文本缺失（受到发帖字数的限制，用户只能用精炼的语言发表微博，文本的规范性和完整性较差，为了更好地阐述自己的观点，用户在发帖时嵌入超文本，如图片、视频、网页链接等）、实时性、丰富的社交信息（关注关系、转发、评论操作）等特性。因此利用传统的自动摘要方法并不能很好的进行微博摘要分析。Zhao等人[19]针对微博的特点提出了基于上下文话题相关的PageRank算法，对微博进行关键词提取和排序，然后利用基于概率的得分函数计算关键词短语间的相关度和兴趣度，最后利用这些关键词对某话题特定时间段内的所有微博进行自动摘要生成。针对新浪微博，童薇等人[20]综合考虑了微博数据的文本特征(转帖、评论、内嵌链接、用户标签hashtag、命名实体等)、语义特征、时序特性和社交关系特性，提出了一种通过提取事件关键要素，即关键词、命名实体、发帖时间和用户情感倾向性，构成事件摘要的方法。

### 话题检测

话题检测与跟踪(Topic Detection and Tracking, TDT)是文本挖掘的一个方向，旨在帮助人们应对信息过载问题。话题检测有一些广为人知的方法，例如CMU、UMass和Dragon等[21]。这些算法各有特点，但是主要的处理对象是新闻媒体的长篇报道，不能有效地适用于微博消息。Zhao等人[22]提出了非监督LDA话题模型的改进形式Twitter-LDA模型，对Twitter与纽约时报在信息传播力(包括内容和速度两方面)进行了对比，并认为Twitter传播力更强。Hong等人[23]利用微博的转发次数作为度量微博流行程度的度量标准，并利用机器学习的方法，通过分析微博的内容、微博的时间特性、消息和用户的元数据以及用户社会网络图作为特征，预测新的微博发布后在多长时间内会被转发。赵文清等人[24]给出一种基于词共现图的方法来识别微博中的新闻话题。该方法首先在微博数据预处理之后，综合相对词频和词频增加率两个因素抽取微博数据中的主题词，然后根据主题词间的共现度构建词共现图，把词共现图中每个不连通的簇集看成一个新闻话题，并使用每个簇集中包含信息量较大的几个主题词来表示微博新闻话题。马彬等人[25]提出了基于线索树的双层聚类的话题检测方法，通过利用融合了时序特征和作者信息的话题模型(Temporal-Author-Topic, TAT)进行线索树内的局部聚类，借以实现垃圾微博的过滤，最后利用整合后的线索树进行全局话题检测。郑斐然等人[26]综合考虑短文本中的词频和增长速度而构造复合权值，用以量化词语是新闻词汇的程度；在话题构造中使用了上下文的相关度模型来支撑增量式聚类算法，相比于语义相似度模型，其更能适应该问题的特点。另一类方法是通过情感变化来发现热点话题。杨亮等人[27]依据热点事件的出现会使用户所发微博中情感词数量增多，情感发生变化，提出了情感分布语言模型，通过分析相邻时段间情感分布语言模型间的差异，实现对热点事件的发现。张鲁民等人[28]通过构建情感向量，采用改进的Kleinberg方法对情感状态进行监测，发现突发事件情感特征及突发期，并采用谱聚类方法对处于突发期的博文进行聚类分析，抽取突发事件。田野[29]提出了结合了MACD算法(Moving Average Convergence and Divergence，指数平滑异同移动平均线)和LDA算法(Latent Dirichlet Allocation，潜在狄利克雷分布)的话题检测方法，分别进行突发事件内容的获取和已知事件的相关词扩展。利用 MACD 算法，计算微博文本中单位时间片的词频变化，利用短周期移动平均线和长周期移动平均线之间的聚合及分离情况，识别平台文本流中讨论量突增内容，以此抽取有可能成为讨论热点的事件。利用LDA算法，计算事件相关的“词袋”内容及各相关词在该事件中的关联权重，以若干词组合的方式作为关键词查询的补充，以此扩展事件相关内容的抽取结果。周刚等人[30]基于微博中联系人存在的关注和粉丝等结构化信息、帖子之间转发评论等内在关联关系，提出了针对微博的MB-SinglePass话题检测算法。该算法针对微博短文本特征稀疏的问题，利用同义词典，引入了微博特征扩展技术，丰富了特征信息。同时，针对单一使用余弦相似度、雅各比相似度和语义相似度的不足，采用了组合相似度策略。

### 标签推荐

微博为用户提供了更加便捷的日志工具，用户可通过微博发布大量的日常信息，而这些信息中通常隐含着用户的兴趣爱好，因此，与基于Web网页内容的自动标注类似，可利用微博内容自动为用户生成标签，如Wu等人[31]利用TF-IDF (Term Frequency-Inverse Document Frequency)与Text-Rank两种不同的算法来自动提取用户发布微博中的关键词，从而标注用户的兴趣爱好，其中Text-Rank算法的效果明显好于TF-IDF算法。王晟等人[32]提出了一种基于贝叶斯个性化排序的微博推荐算法，对用户进行个性化微博推荐。该基于贝叶斯个性化排序的微博推荐算法，以微博对的形式提取微博系统中的隐式信息，对这些微博对进行学习，从而得到用户对不同微博的兴趣值。根据每条微博发出的时间，估计每条微博对的可信度。发出时间越接近的微博对，它的可信度就越高，并且对用户的兴趣值影响就越大。赵岩露等人[33]在综合兴趣模型研究现状的基础上，结合微博数据集对微博用户的特征进行分析，建立微博用户兴趣模型，并提出基于微博用户兴趣模型的发现算法。同时采用从系统内明星中进行聚类发掘，并根据用户反馈修正更新推荐结果的方法解决冷启动问题。吴铭[34]基于图数据挖掘中链接预测的基本思想，通过计算尚未建立链接的两个人发生链接的概率，实现社会网络中人与人之间的关系推荐。文章对传统的链接预测方法进行扩展，验证了一种同时考虑网络结构信息和节点属性信息的数学模型——指数随机图模型ERGM/P\*(Exponential Random Graph Models, ERGM)在关系推荐中的应用。

### 可信度研究

微博信息的发布取决于用户的自律，可信度受到质疑。不完整的信息经过用户不断转发后，难以找到信息源。由于每个用户都可使用微博发布信息使得在微博网络中信息泛滥，最终导致信息的平均可靠度也随之下降。Castillo 等人[35]分析了微博的可信度，利用四个特征来度量微博的可信度：基于消息的特征，如消息的长度、是否存在“#”符号、是否存在问号或感叹号以及情感词汇的数目等；基于用户的特征，如用户注册时间、关注人数、被关注人数、过去发布微博的数量等；基于话题的特征，如有多少微博包含URL；基于消息传播的特征，如微博转发树的深度和广度等。结果表明：可信度高的微博被转发次数也较多；微博的原始发布者一般集中在少数用户中；转发微博的用户往往具有转发的习惯。曹鹏等人[36]提出了一种Twitter中近似重复消息的判定方法，统计字符种类和最短编辑距离两种字符串距离以判定Twitter中近似重复的消息。该方法可在一定程度上提高微博的信息利用率。谈磊等人[37]提出一种采用复合分类模型对用户进行分类的方法，并开发了一个对微博类社交网络用户进行分类的系统。通过研究用户的属性和行为特点，比较属性间的相关性，从两方面兼顾了分类的准确性和效率。

## 微博内容的情感语义分析

情感分析也可称为观点挖掘，随着Web2.0的发展，越来越多的用户在网络上发布具有不同情感趋向的信息，研究这些用户信息中潜在的情感信息，挖掘用户的潜在观点一直是研究的热点，但已有的观点挖掘研究主要集中于在线产品评论或传统博客上，较少有针对微博的观点挖掘研究。

微博正日益成为一个普遍流行的实时性交流工具，大量网络用户每天都会发布并传播高达几千万的微博，在这些微博中包含着不同用户的日常生活记录，因此微博为情感分析与观点挖掘提供了丰富的数据来源，从中挖掘出相关用户对某个特定主题或事件的观点，如对使用过的产品或服务的满意程度以及用户的政治或宗教观点等。同时，由于微博内容简短、结构自由、实时性高且数据量大也为进行用户情感分析和观点挖掘提出了挑战。

微博用户和用户之间存在社会网络关系，用户通过微博所体现的观点集之间也存在语义上的关联关系，而事件特征之间同样存在隐含的关系。这些隐含的关联对情感分析会产生潜在的影响。Bermingham等人[38]研究发现针对微博进行情感分析相对传统博客的效果将会更好，微博已经成为情感分析与观点挖掘的有效文本领域。Go等人[39]利用机器学习的方法对微博消息进行情感分类，即判断一条微博消息的情感倾向是正面还是负面。在训练集的选择上，利用微博中的表情作为类别标记，然后利用朴素贝叶斯和支持向量机等不同的分类算法训练分类器，从而实现微博的情感分类。在应用方面，基于Twitter的实时搜索引擎TweetFeel[[2]](#footnote-2)能够判断信息是负面情绪还是正面，然后用一个计数器进行对比。国内研究方面，谢丽星等人[40]对于三种情感分析的方法进行了深入研究，包括表情符号的规则方法、情感词典的规则方法、基于SVM的层次结构的多策略方法，实验表明基于SVM的层次结构多策略方法效果最好。其次，针对层次结构的多策略方法的特征选择进行了详细分析，包括主题无关、主题相关的特征。实验表明引入主题相关的特征的准确率高于使用主题无关的特征时获得的准确率。刘志明等人[41]使用三种机器学习算法、三种特征选取算法以及三种特征项权重计算方法对微博进行了情感分类的实证研究。实验结果表明，针对不同的特征权重计算方法，SVM和贝叶斯分类算法各有优势，信息增益特征选取方法相比于其他的方法效果明显要好。综合考虑三种因素，采用SVM和信息增益，以及TF-IDF作为特征项权重，三者结合对微博的情感分类效果最好。庞磊等人[42]提出一种基于情绪知识的非监督情感分类方法，利用情绪词和表情图片这两种情绪知识对大规模微博非标注语料进行筛选并自动标注，用自动标注好的语料作为训练集构建微博情感文本分类器，对微博文本进行情感极性自动分类。同时，王昊等人[43]利用基于情感的HITS算法对日本地震发生后一周内爬取的新浪微博进行事件分析，同时采用基于规则的情感分类的方法分析人们在特定主题词下表现的情感。

# 当前存在的问题

总的来说，现有对微博及中文微博的研究仍然处于探索阶段，仍然有大量关键问题还需做深入细致的研究。如何在已有的微博相关研究成果和观点挖掘方法的基础上，结合中文自然语言处理技术和中文微博自身的特点，提出新的模型和方法，挖掘中文微博中蕴含的语义信息及用户观点，是需要重点解决的问题。根据前三章的现状，可以发现现有研究中往往忽略了微博用户的社会属性对其行为和社会关系的影响，具体问题总结如下：

1. 在针对微博网络开展测量研究时大多从社会个体层次的结构，即微观结构入手，了解微博系统的拓扑结构特征、用户行为特征等信息，缺乏从宏观结构（[组织](http://wiki.mbalib.com/wiki/%E7%BB%84%E7%BB%87)或群体层次的结构）对微博网络的研究。
2. 在进行话题检测时缺乏对行业的关注，娱乐、体育等关注度较高的行业出现的话题容易被发现，而医生、航空航天等关注度较低的行业人士热烈讨论的话题不易被发现。
3. 在进行情感分析时缺乏对行业的区分，在社会热点事件涉及一定专业性时，例如在昆明、大连、厦门各地爆发的反对“PX”化工厂的社会群体性事件中，涉及到相当专业的化工、生物和医学的内容，在这样的背景下，如果能够及时搜集并整理化工行业人士对此问题的看法，并形成相对全面的，具有代表性的行业观点，将大大有助于实现微博舆论的监测、引导以及控制，有效避免由谣言的滋生和蔓延对我国社会稳定和经济发展产生的负面影响[4, 42]。

# 参考文献

[1] 中国互联网络信息中心. 第32次中国互联网络发展状况调查统计报告, 2013.

[2] Java A, Song X, Finin T, et al. Why we twitter: understanding microblogging usage and communities[C]. San Jose, California: ACM, 2007.

[3] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]. Raleigh, North Carolina, USA: ACM, 2010.

[4] 樊鹏翼，王晖，姜志宏，等. 微博网络测量研究[J]. 计算机研究与发展, 2012, v.49(04): 691-699.

[5] 夏虎. 移动社交网络结构和行为研究及其应用[D]. 电子科技大学, 2012.

[6] Weng J, Lim E, Jiang J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]. New York, New York, USA: ACM, 2010.

[7] Cha M, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[J]. ICWSM, 2010, 10: 10-17.

[8] Ye S, Wu S F. Measuring message propagation and social influence on Twitter. com[M]. Social informatics, Springer, 2010, 216-231.

[9] Bakshy E, Hofman J M, Mason W A, et al. Everyone's an influencer: quantifying influence on twitter[C]. Hong Kong, China: ACM, 2011.

[10] 舒琰，向阳，张骐，等. 基于PageRank的微博排名MapReduce算法研究[J]. 计算机技术与发展, 2013, v.23;No.190(02): 73-76.

[11] 文坤梅，徐帅，李瑞轩，等. 微博及中文微博信息处理研究综述[J]. 中文信息学报, 2012, v.26(06): 27-37.

[12] Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter[C]. New York, NY, USA: ACM, 2008.

[13] Pal A, Counts S. Identifying topical authorities in microblogs[C]. 2011.

[14] Welch M J, Schonfeld U, He D, et al. Topical semantics of twitter links[C]. 2011.

[15] Zhou M, Xu Y, Zhao X. Study of Feature Extract on Microblog User Occupation Classification[C]. 2012.

[16] Yang J, Counts S. Comparing Information Diffusion Structure in Weblogs and Microblogs.[C]. 2010.

[17] 张赛，徐恪，李海涛. 微博类社交网络中信息传播的测量与分析[J]. 西安交通大学学报, 2013, v.47(02): 124-130.

[18] Sager N. Sublanguage grammers in science information processing[J]. Journal of the American Society for Information Science, 1975, 26(1): 10-16.

[19] 刘挺，吴岩，王开铸. 基于信息抽取和文本生成的自动文摘系统设计[J]. 情报学报, 1997, 16(1): 24-29.

[20] Zhao X, Jiang J, He J, et al. Topical keyphrase extraction from Twitter[J]. 2011.

[21] 童薇，陈威，孟小峰. EDM:高效的微博事件检测算法[J]. 计算机科学与探索, 2012, v.6;No.51(12): 1076-1086.

[22] Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study final report[J]. 1998.

[23] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[M]. Advances in Information Retrieval, Springer, 2011, 338-349.

[24] Hong L, Dan O, Davison B D. Predicting popular messages in twitter[C]. 2011.

[25] 赵文清，侯小可. 基于词共现图的中文微博新闻话题识别[J]. 智能系统学报, 2012, v.7;No.37(05): 444-449.

[26] 马彬，洪宇，陆剑江，等. 基于线索树双层聚类的微博话题检测[J]. 中文信息学报, 2012, v.26(06): 121-128.

[27] 郑斐然，苗夺谦，张志飞，等. 一种中文微博新闻话题检测的方法[J]. 计算机科学, 2012, v.39(01): 138-141.

[28] 杨亮，林原，林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报, 2012, v.26(01): 84-90.

[29] 张鲁民，贾焰，周斌. 基于情感计算的微博突发事件检测方法研究[J]. 信息网络安全, 2012, No.140(08): 143-145.

[30] 田野. 基于微博平台的事件趋势分析及预测研究[D]. 武汉大学, 2012.

[31] 周刚，邹鸿程，熊小兵，等. MB-SinglePass:基于组合相似度的微博话题检测[J]. 计算机科学, 2012, v.39(10): 198-202.

[32] Wu W, Zhang B, Ostendorf M. Automatic generation of personalized annotation tags for twitter users[C]. 2010.

[33] 王晟，王子琪，张铭. 个性化微博推荐算法[J]. 计算机科学与探索, 2012, v.6;No.49(10): 895-902.

[34] 赵岩露，王晶，沈奇威. 基于特征分析的微博用户兴趣发现算法[J]. 电信工程技术与标准化, 2012, v.25;No.182(11): 79-83.

[35] 吴铭. 基于链接预测的关系推荐系统研究[D]. 北京邮电大学, 2012.

[36] Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]. 2011.

[37] 曹鹏，李静远，满彤，等. Twitter 中近似重复消息的判定方法研究[J]. 中文信息学报, 2011, 25(1): 20-27.

[38] 谈磊，连一峰，陈恺. 基于复合分类模型的社交网络恶意用户识别方法[J]. 计算机应用与软件, 2012, v.29(12): 1-5.

[39] Bermingham A, Smeaton A F. Classifying sentiment in microblogs: is brevity an advantage?[C]. 2010.

[40] Go A, Huang L, Bhayani R. Twitter sentiment analysis[J]. Entropy, 2009, 17.

[41] 谢丽星，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报, 2012, v.26(01): 73-83.

[42] 刘志明，刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, v.48;No.740(01): 1-4.

[43] 庞磊，李寿山，周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程, 2012, v.38;No.408(13): 156-158.

[44] 王昊，杨亮，林鸿飞. 日本地震的微博热点事件分析[J]. 中文信息学报, 2012, v.26(05): 7-13.

1. http://blogs.ischool.berkeley.edu/i290-abdt-s12/files/2012/11/Twitter-Professional-Analysis-Tool1.pdf [↑](#footnote-ref-1)
2. http://www.tweetfeel.com [↑](#footnote-ref-2)