

硕 士 学 位 论 文  
开题报告

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

目 录

[1 课题来源和意义 1](#_Toc372986634)

[1.1 课题来源 1](#_Toc372986635)

[1.2 课题背景 1](#_Toc372986636)

[1.3 选题意义 1](#_Toc372986637)

[2 国内外相关研究现状 3](#_Toc372986638)

[2.1 微博用户网络特性研究现状 3](#_Toc372986639)

[2.1.1 基本社会网络特性研究 3](#_Toc372986640)

[2.1.2 用户影响力研究 4](#_Toc372986641)

[2.1.3 用户特征分析与分类研究 5](#_Toc372986642)

[2.2 微博内容的语义信息研究现状 6](#_Toc372986643)

[2.2.1 话题检测 6](#_Toc372986644)

[2.2.2 情感分析 7](#_Toc372986645)

[2.3 有待解决的问题 8](#_Toc372986646)

[3 论文研究方案 10](#_Toc372986647)

[3.1 研究目标 10](#_Toc372986648)

[3.2 研究任务 10](#_Toc372986649)

[3.3 系统总体解决方案 10](#_Toc372986650)

[3.4 拟解决的关键问题 12](#_Toc372986651)

[3.5 拟采取的技术路线 12](#_Toc372986652)

[3.6 可能的创新之处 13](#_Toc372986653)

[4 预期达到的目标和研究成果 14](#_Toc372986654)

[5 论文工作计划 15](#_Toc372986655)

[参考文献 16](#_Toc372986656)

# 课题来源和意义

## 课题来源

本课题来源于网络技术北京市重点实验室。

## 课题背景

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。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。

## 选题意义

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。

在人类社会关系中，个人所从事的行业是一种重要的社会关系。行业是一种反映了生产力三要素（劳动者、劳动对象、劳动资料）不同排列组合的经济活动，能够直接反映人们在社会行业地位及拥有资源等方面的关系，这种结构关系不是体现在个体活动之间，而是体现了社会构成要素之间的关系。在现有研究的基础之上，引入“行业”这一重要的社会关系，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。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。

# 国内外相关研究现状

本课题主要涉及微博用户网络特性研究和内容的语义信息研究，因此对这两方面研究现状进行介绍，并分析得到有待解决的问题。

## 微博用户网络特性研究现状

### 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。

国内的樊鹏翼等人[4]对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。

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[6]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[7]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。

国内的舒琰等人[8]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。例如，Krishnamurthy等人[9]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[10]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[11]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[12]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。

## 微博内容的语义信息研究现状

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等，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。

### 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[13]。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[14]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[15]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[16]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[17]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[18]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[19]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[20]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[21]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[22]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。

### 情感分析

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。

微博中的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[23]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[24]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[25]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[26]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[27]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[28]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。

## 有待解决的问题

根据以上介绍的现状，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，及时搜集并整理该行业的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。

# 论文研究方案

## 研究目标

现有对中文微博的研究仍然处于探索阶段，未考虑用户社会属性对其行为和社会关系的影响。本课题的目标是研究并实现融合行业分析的微博话题检测与情感分析技术，该技术具有以下特点：

1. 能够有效识别用户行业；
2. 从行业这一宏观角度开展微博网络测量研究，全面认识微博网络关系；
3. 在已有的微博相关研究成果的基础上，充分考虑中文自然语言处理技术和中文微博自身的特点，检测行业内话题；
4. 挖掘微博社会网络所隐含的语义信息及情感关联，计算行业内用户情感分布。

## 研究任务

针对上述的研究目标，本课题的主要研究内容如下：

1. 设计并实现获取大量微博数据的方法；
2. 结合微博特点，设计并实现针对微博的行业分类方法；
3. 分析各行业子网络的拓扑结构，包括网络平均路径长度、网络直径、聚集系数、集中趋势（中心度）、节点度分布，比较并分析各子图异同；
4. 分析各行业用户行为特征，包括用户时间分布以及用户博文分布，比较并分析特征异同；
5. 结合行业特点，实现并改进话题检测方法；
6. 分析中文微博自身的特点，实现并改进用户情感分析方法；
7. 研究相关领域的评价指标，设计系统评估方案。

## 系统总体解决方案

根据研究现状和研究目标，拟采取的系统总体方案如图1所示。



图 1 系统总体方案

系统首先通过数据爬取技术获取包括用户个人资料、用户发布的微博、用户关系等原始信息，然后通过行业识别算法获得用户最可能所属的行业，从而得到用户行业信息，再将同一行业内用户信息整合，对行业网络拓扑结构以及行业用户行为特点进行分析，然后进行行业话题检测，最终对行业热点话题进行用户情感分析，得到行业关于此话题的整理倾向。

其中网络拓扑计算中各数据定义如下：

1. 网络平均路径长度：网络中所有节点对之间最短路径的平均值；
2. 网络直径：网络中任意节点对之间最短路径的最大值；
3. 聚集系数：一个节点邻居之间的相互连接的紧密程度，即网络的集团化程度；
4. 集中趋势（中心度）：整个网络中各个点的差异性程度；
5. 节点度分布：在微博中，节点之间的连接关系可表示为两类：“粉丝”和“关注”，其对应的网络关系则为节点的“链入”和“链出”关系。在测量节点度分布时，我们应分别对节点的出度和入度概率分布进行计算，从而得到微博节点度分布的特性。

用户行为分析中各数据定义如下：

1. 用户时间分布：微博用户时间分布是指用户发表博文数目在时间轴上的统计分布，该指标体现了大量用户节点所涌现出来的一种“在线”模式和行为特点，这对于微博用户行为的建模研究具有一定的指导意义。
2. 用户博文分布：微博用户博文分布是指用户发表、转载以及回复博文所呈现出的分布规律，该指标描述了博文信息产生、传播等行为的统计特征，用定量分析的方式描述了微博用户的基本操作行为。

## 拟解决的关键问题

为了研究并实现融合行业的微博话题检测与情感分析技术，必须解决以下三个关键问题：

1. 对普通用户实现行业识别

现有行业分类算法仅对部分实名认证的用户进行了实验，而普通用户可能不具有这样明显的行业特征。要实现对更普遍范围内用户的行业识别，需要将海量微博数据中复杂的，难以理解的，噪声较大的人的特征维转换为一些可计算的值，这将是本课题需要解决的关键问题。

1. 微博短文本表示问题

微博短文本的特点主要体现在：文本表达口语化，不规则字符、谐音词、网络用语多；文本特征词少且稀疏，使得特征之间的相关性难以度量；文本样本数量巨大，分布高度不平衡，少部分的短文本在整体中占有较大比重。如何表示短文本以及是否合理的词加权方法将对行业识别、话题检测以及情感分析这三部分工作有很大的影响，因而是本课题需要解决的关键问题。

1. 行业整体倾向计算问题

现有情感分析方法大多计算单个用户的情感倾向，如何在计算行业整体倾向中体现出不同用户情感所占权重的不同是需要解决的又一关键问题。

## 拟采取的技术路线

针对以上三个关键问题，系统拟采用技术路线如下：

1. 综合微博消息内容和量化后的用户关系作为分类特征，使用支持向量机、贝叶斯等分类方法实现用户行业识别；
2. 通过对比分析，发现最优的特征选择算法和紧凑的文本表示模型，以缓解由于微博文本高维性和高稀疏度而产生的影响；
3. 结合微博的“对话性”，通过引入与用户影响力类似的用户行业影响力，计算得到行业整体倾向。

## 可能的创新之处

本课题针对融合行业的微博话题检测与情感分析技术开展研究，在微博话题检测和情感分析方面均有望取得突破，可能具有的特色与创新之处包括：

1. 结合现有行业词库进行行业识别

在分类特征集的获取上，除考虑通过选取行业领袖所发布的信息中提取出行业特征外，也可以将现有的一些行业词库加入，作为补充信息。在特征值的计算和分类方法的选择上，通过对比分析，得到准确率较高的组合。

1. 利用行业词特征集辅助实现话题检测

在进行行业识别时，需要将用户所发的微博信息和行业词特征集进行匹配，此处可以同步进行词频的记录，为后续话题检测的聚类分析提供一定的数据支持，提高话题聚类的效率与准确率。

1. 通过用户行业影响力计算行业整体倾向

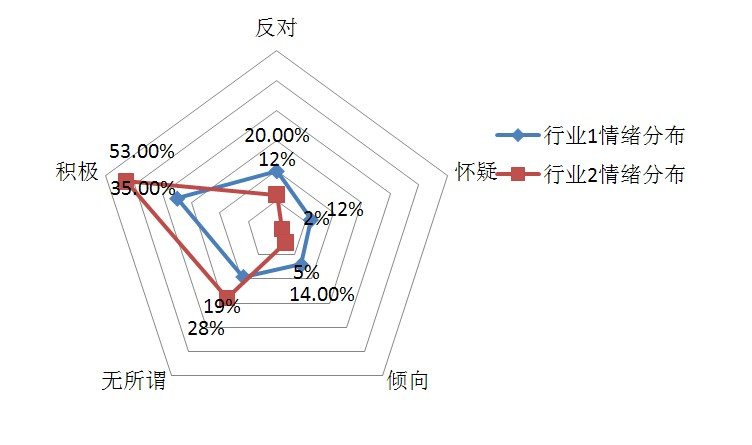
现有情感分析方法大多计算单个用户的情感倾向，在分析得到单个用户的情感后，通过引入与用户影响力类似的用户行业影响力，可以计算得到行业整体倾向。

# 预期达到的目标和研究成果

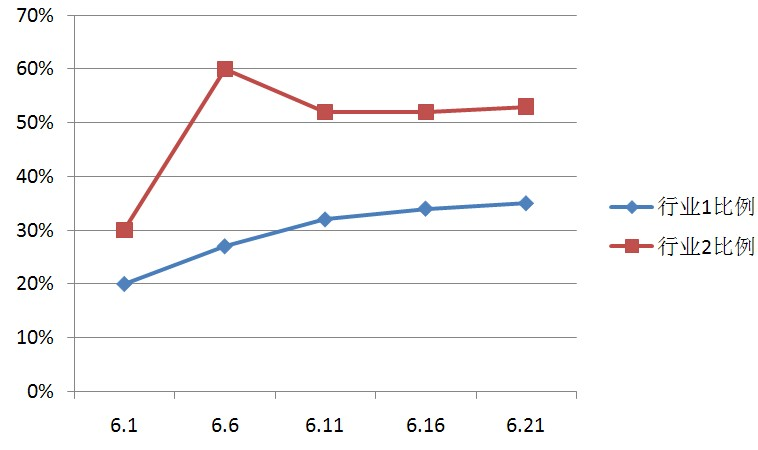
预期形成的最终成果为： 一个原型系统，具体指标如下：

1. 数据获取指标：每天获得用户数据>=1000个；
2. 处理指标：每天处理用户数据>=1000个；
3. 识别准确率指标：用户行业识别准确率>=90%。

该原型系统可以将就特定话题以情绪雷达图以及时间波动图的形式展示最终的行业整体情绪，图2表现的是目前同一行业类对某一具体话题的不同观点和看法的分布情况，图3体现出行业1、2用户在6月1日到6月21日时间段内持支持观点所占比例变化情况。



**图 2 不同行业情绪雷达图**



**图 3 不同行业情绪波动图**

# 论文工作计划

论文的工作计划如表1所示。

**表 1 论文工作计划表**

|  |  |  |  |
| --- | --- | --- | --- |
| **编号** | **起始日期** | **结束日期** | **工作内容** |
| 1 | 2013.11.01 | 2013.11.30 | 文献整理及阅读，撰写开题报告及文献综述 |
| 2 | 2013.12.01 | 2013.12.31 | 设计并实现获取大量微博数据的网络爬虫 |
| 3 | 2014.01.01 | 2014.03.31 | 设计并实现行业分类方法 |
| 4 | 2014.04.01 | 2014.04.30 | 实现话题检测方法 |
| 5 | 2014.05.01 | 2014.05.31 | 实现情感分析方法 |
| 6 | 2014.06.01 | 2014.06.30 | 整体进行系统功能及性能测试 |
| 7 | 2014.07.01 | 2014.07.15 | 整理资料，撰写中期报告 |
| 8 | 2014.07.16 | 2014.09.20 | 资料整理和数据分析 |
| 9 | 2014.09.21 | 2014.12.01 | 论文撰写 |

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

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

[8] 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.

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

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

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

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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)