分类号 TP391.1 学号 10069068

UDC 密级 公开

工学博士学位论文

社交媒体中观点信息分析与应用

博士生姓名 谢松县

学科专业 计算机科学与技术

研究方向 自然语言处理

指导教师 王挺 教授

国防科学技术大学研究生院

二〇一四年十月

Opinion Mining and Application in Social Media

Candidate: Xie Songxian

Supervisor: Professor Wang Ting

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

October 29, 2014

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科学技术大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目：社交媒体中观点信息分析与应用

学位论文作者签名：_____ 日期： 年 月 日

学位论文版权使用授权书

本人完全了解国防科学技术大学有关保留、使用学位论文的规定。本人授权国防科学技术大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅;可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

(保密学位论文在解密后适用本授权书。)

学位论文题目：社交媒体中观点信息分析与应用

学位论文作者签名：_____ 日期： 年 月 日

作者指导教师签名：_____ 日期： 年 月 日

目 录

表 目 录

图 目 录

摘　要

随着社交媒体的日益普及，越来越多的人开始在网上实时地以各种方式表达自己的观点。这些观点覆盖各种话题，并且用户群体庞大，使得网络变成汇集了关于各种主题的大众意见的宝贵资源库。然而社交媒体中的观点通常是在充满噪声的非结构化的文本中表达的，人工就某一话题去阅读所有的和数据并提取总结出其中的观点是不可能的，需要以计算手段自动分析，整合集成，总结出文本中的观点信息。本文主要研究社交媒体用户观点的自动分析与整合集成，对用户在社交媒体上就所关注话题发表的大量观点进行综合建模，并对用户的网络交互行为进行分析。

为了对问题进行系统地研究，我们确定了观点分析的三个主要步骤：文本中情感知识的获取，文本情感倾向性分类，用户观点的集成建模。这三个步骤组成了一个观点集成汇总系统的三个关键组成部分，集成的用户观点信息促进了用户网络行为的分析研究。本文的主要贡献是对四个相互联系协同的观点分析与应用任务提出了新的通用计算方法：

- **中文情感词汇的抽取和情感词典的构建：**目前表示情感知识的词典主要是针对英文构建的情感词典，这些词典在观点信息识别、情感分类特征选择等任务中具有重要作用，是进行观点分析的基础。中文情感词典研究相对较少，还没有形成比较全面可用的情感词典，而靠人工编辑的情感词典费时费力，覆盖度偏低，因此本文首先根据不同语言间情感知识的对映性，借鉴已有的英文情感词典，使用 HowNet 语义知识库的语义关系计算确定了一些情感词汇并计算出情感极性值。为了提高词典的覆盖度以及领域情感知识的适应性，通过实验验证了基于语言规则和统计特征的基于语料库的情感词典的扩展方法，并提出了基于混合特征的扩展方法。

- **基于特征空间划分的情感分类：**情感分类是按照文本中的特征共现规律将文本分类为特定的情感极性类别，是一种特殊的文本分类。用以表达情感的词语特征在情感分类任务中有不同的作用，有的词语就具有通用的情感表达能力，能在不同领域和语境中表示相同的情感极性，而有的词语只有在特定的领域和语境中才能表达特定的情感极性。因此本文提出了将情感分类的特征空间分为领域独立和领域依赖两部分，分别使用两部分特征训练分类器然后组合在一个框架中形成一个更强的情感分类器，这种框架从现成的无需标

注的资源开始，使用自举式的机器学习方法，可以在无监督情况下达到有监督方法的性能。

- **用户观点的整合集成：**社交媒体中用户产生的内容是短小而又分散的，因此用户针对某些话题的观点信息是碎片化在这些非结构化的短文本中。为了能够全面准确的了解用户的观点，本文提出了用户主观模型的概念，将用户产生内容中的话题信息以及用户针对话题的观点信息组合在一起，并将观点按照话题的不同方面进行整合集成，并提出一种通用的观点的表示方法，将同一话题的观点表示为在一个可扩展的情感空间的分布，这种表示能够提供用户更详细和多视角下的观点信息。

- **用户转发行为分析：**作为用户主观模型的直接应用，本文对用户在在社交媒体中信息传播行为的主观动机进行建模分析。针对 Twitter 中用户转发信息的三种常见情形，也就是用户对感兴趣和有吸引力的信息转发，用户基于社交需要对好友的信息转发以及用户对流行度高的信息转发，使用三个主观相似性计算方法进行度量。在转发行为的分析中，三种主观相似性度量与转发行为具有相关性，能够作为转发行为预测的有用特征，并能显著提高现有预测模型的性能。

在对以上四个问题的研究中，我们侧重于使用通用的鲁棒性好的无监督或弱监督方法，因此我们的方法可以适用于话题广泛的大量观点的自动分析，这也是本文区别于一些针对特定领域精心进行特征设计并进行充分训练的其他方法，因为这些方法转换到新领域就会性能下降，领域适应性差。我们尽可能使用现有的无需标注资源，比如一些现成的词典资源，可以为观点分析各种方法提供间接训练指导。基于这种思路使得我们方法显示出良好的通用性和效能，能够在多个领域（比如商业智能和政治学）得到应用。

关键词: **社交媒体**; **情感词典**; **情感分类**; **观点集成**; **信息传播**

ABSTRACT

As Social Media becomes increasingly popular, more and more people express their opinions on the Web in various ways in real time. Such wide coverage of topics and abundance of users make the Web an extremely valuable source for mining people's opinions about all kinds of topics. However, since the opinions are usually expressed as unstructured text scattered in different sources, it is difficult for the users to digest all opinions relevant to a specific topic within a large amount of text pieces, which needs the computational methods to automatically analyze, integrate and summarize the opinions articulated in the text. This thesis focuses on the problem of opinion integration and summarization whose goal is to better support digestion of huge amounts of opinions for an arbitrary topic and model the interaction behavior of users. To systematically study this problem, we have identified three important steps of opinion analysis: extraction of sentiment knowledge, sentiment classification of text, and integration of opinions. These steps form three key components in an integrated opinion summarization system. Accordingly, this thesis makes contributions in proposing novel and general computational techniques for four synergistic tasks:

- **Extraction and construction of Chinese sentiment lexicon:**Current sentiment lexicons are built mainly for English sentiment knowledge, which are basis of opinion analusis and play important roles in such tasks as subjectivity analysis, feature selection of sentiment classification, etc. There are relatively few studies on construction of Chinese sentiment lexicon, and there is no comprehensive lexicon available yet. However the sentiment lexicon compiled by human is time-consuming and laborious, while has a low coverage. Therefore based on the sentiment knowledge mapping between different languages and current English lexicons, we proposed a novel method to identify a number of positive and negative words and calculate their sentiment strength value using semantic relationships of HowNet semantic knowledge dictionary. In order to improve coverage and domain adaptability of sentiment lexicon, we verified language rules based and corpus based statistical extension methods with experiments, and proposed a hybrid features method.

- **Sentiment classification based feature space division:**Sentiment classification classifies the text into predefined categories according to feature co-occurence, and

can be regarded as special text classification. The features of sentiment classification are used in different position: some features represent the same general sentiment polarity across different domains and context, while others represent specific sentiment polarity only in specific domain and context. Therefore, we proposed to divide the feature space of sentiment classification into separate parts, which are domain-dependent part and domain-independent part. Two different classifiers are learned using two feature parts, and then combined together to form a stronger sentiment classifier in a bootstrapping machine learning frame, which started training on an off-the-shelf resources without annotation in an unsupervised bootstrapping way. The bootstrapping method can achieve the performance of supervised methods in an unsupervised situations.

- **Integration of opinions of users**User-generated content of social media is short and dispersed, so that the opinions of users about certain topic are scattered in the unstructured fragmented short text. To be able to understand comprehensively and accurately the users' opinions, we proposes a subjectivity model to combine the topics and the opinions integrated according to the different aspects of the same topic articulated in UGC. We also put forward a general representation of opinion, which defined opinion as sentiment distribution over a scalable sentiment space, and provided a more detailed and informed multiperspective view of the opinions.

- **Retweeting analysis of users:**As a direct application of subjectivity model, we analyzes the subjective motivation of the information dissemination behaviors for the social media users. For three scenarios a Twitter user retweeted a message, that is, the user retweeted for he is interested and attracted by message content, the user retweeted a message of a close friend based on the social needs and the user retweeted because the message is popular, we proposed three subjectivity similarity measurements. For retweeting behavior analysis, the three subjectivity similarities correlated to the behavior, and could serve as a useful features for retweeting behavior prediction, which could significantly improve the performance of existing prediction models.

We focus on general and robust methods which require minimal human supervision so as to make the automated methods applicable to a wide range of topics and scalable

to large amounts of opinions. This focus differentiates this thesis from work that is fine-tuned or well-trained for particular domains but are not easily adaptable to new domains. Our main idea is to exploit many naturally available resources, such as off-the-shelf lexicon, which can serve as indirect signals and guidance for generating opinion summaries. Along this line, our proposed techniques have been shown to be effective and general enough to be applied for potentially many interesting applications in multiple domains, such as business intelligence and political science.

Key Words: Social Media; Sentiment lexicon; Centiment classification; Opinion integration; Information dissemination

第一章　绪论

1.1　研究背景

1.1.1　社交媒体

作为划时代的创新，互联网 20 年以来已深刻影响和改变着我们的生活，思维和行为方式。尤其现在，我们可以通过手机、各种穿戴式智能设备，随时随地保持与互联网不间断联系。根据中国互联网络信息中心的权威报告，截至 2014 年 7 月，我国网民规模达 6.41 亿，手机网民规模已超过 5 亿，互联网普及率为 47.4%[1]。随着互联网技术的迅猛发展，出现了形形色色吸引用户参与的社交媒体（Social Media）平台，并且已经成为人类工作、学习、生活必不可少的重要部分。图 1.1展示了各种国内外的在线社交媒体平台。

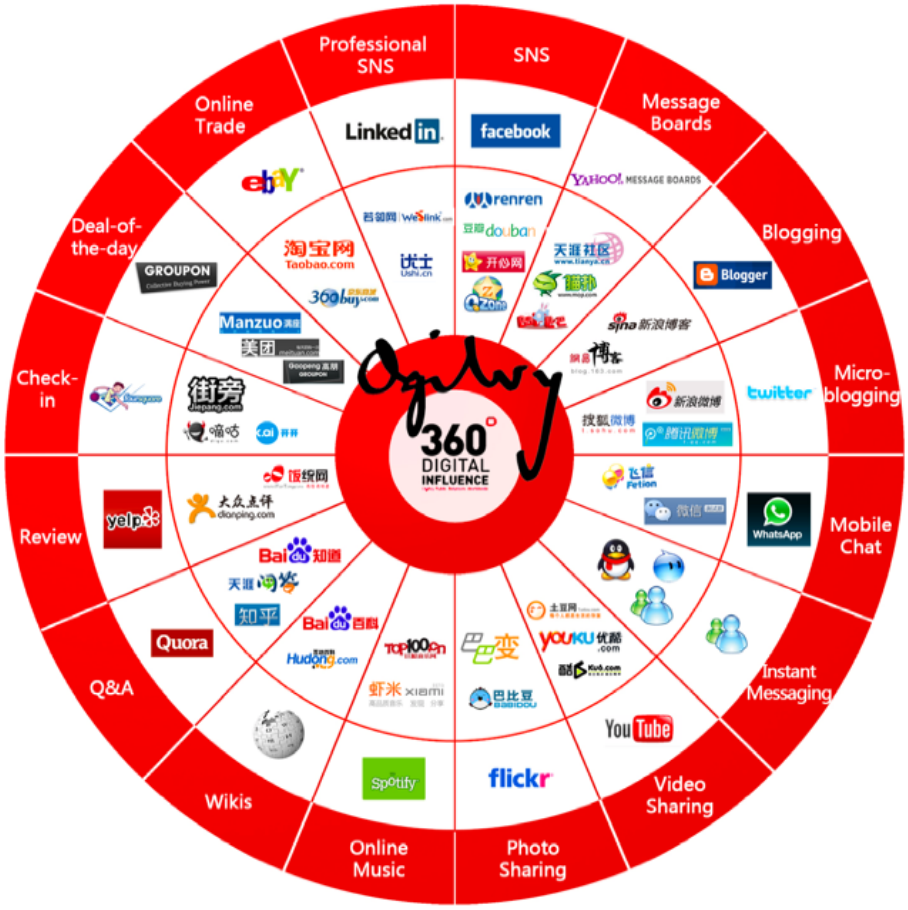


图 1.1　国内外社交媒体

[1]http://www.cnnic.cn/hlwfzyj/hlwfzzx/qwfb/201408/t20140825_47878.htm

社交媒体中的互联网用户不再是单纯的信息接收者，同时也是网络内容的制造者，人们通过社交媒体进行交流而获取和产生信息。以中国为例，目前拥有 12 亿手机用户、5 亿微博用户、5 亿微信用户，每天信息发送量超过 200 亿条，交流无处不在，无时不有。表 1.1是互联网网站信息统计公司 Alexa[2]统计的网络访问统计，从表中可以看出，根据统计，流量前十的互联网网站中社交媒体占了绝大部分。

表 1.1　Alexa 统计访问量前十名网站

排名	网站	排名	网站
1	Google.com	6	**Wikipedia.org**[1]
2	**Facebook.com**	7	**Amazon.com**
3	**Youtube.com**	8	**Twitter.com**
4	Yahoo.com	9	**Qq.com**
5	Baidu.com	10	**Taobao.com**

[1] 表中加黑部分为社交媒体

那么究竟什么是社交媒体呢？社交媒体的典型代表维基百科是这样定义的[3]：

定义 (Social Media): Social media are media for social interaction, using highly accessible and scalable communication techniques. It is the use of web-based and mobile technologies to turn communication into interactive dialogue. ■

从定义中我们可以看出，社交媒体是以互联网的思想和技术为基础的一项应用，用户可以借此进行内容创作、情感交流与信息分享。一般来讲，社交媒体可以分为如表 1.2所示的几种类型。

从表中可以看出，社交媒体有多中不同类型，因此会产生多种不同形式的信息，包括文本、图像以及视频等。社交媒体上的信息由广大的社交媒体使用者产生，因此称为用户产生内容（User-Generated Content，UGC），这些信息依靠用户之间的关系与交互形成相互关联的庞大数据库。Kaplan 和 Haenlein[1] 从数据和信息流动角度定义了社交媒体：首先是作为媒体（media），社交媒体中最突出的特点是它区别于电视、广播和报纸等传统媒体（信息的流动是从少数内容生产者到广大的信息消费者），内容产生的权利扩展到了所有的用户，而且信息流动的方式变得不确定，用户可以在内容消费者和生产者之间多次瞬间改变自己的角色；其次，作为社交工具，我们称这种新媒体是社会化的（social）的媒体，社会化意味着信息内容不只是由个体用户产生，更多是与其他用户的协作产生，信息的内容

[2]网站地址：www.alexa.com，访问时间：2014 年 9 月。

[3]http://en.wikipedia.org/wiki/Socialmedia/

表 1.2　社交媒体的类型

类型	代表性网站
维基（Wiki）	Wikipedia，Scholarpedia，百度百科
博客（Blogging）	Blogger，LiveJournal，WordPress，博客
新闻（Social News）	Digg, Mixx，Slashdot
微博（Micro Blogging）	Twitter，Google Buzz，新浪微博
评论（Opinion & Reviews）	ePinions，Yelp，豆瓣
问答（Question Answering）	Yahoo! Answers，百度知道
媒体分享（Media Sharing）	Flickr，Youtube，优酷
书签（Social Bookmarking）	Delicious，CiteULike
社交网络（Social Networking）	Facebook，LinkedIn，MySpace，人人网

变得更加多样化，因此社交媒体不只是用来产生信息，也成为用户间互相交流通信以及传播信息的便利工具。

虽然社交媒体的出现为我们信息交流提供了便利，但是随着用户数量不断增加，产生的内容达到新的量级，导致我们作为信息消费者遇到了一些新的挑战，使得我们从“信息海洋”中找到有用信息变得更加困难，这通常称为信息超载（information overload）[4]。同时，由于社交媒体发布消息相对廉价和方便, 内容产生门槛降低，任何人都能够参与其中，因此社交媒体中的数据出现了不同于传统媒体数据的新特点。一般来讲，社交媒体中的数据具有以下特点[2]：

- **数量巨大（Big）：**社交媒体中每个用户产生的数据可能不大，但是因为用户群体数量庞大，整体数据规模不可小觑，比如平均每天有超过 300 万条的微博（tweets）发布到 Twitter[5]，每分钟有超过 3000 张照片上传到 Flickr[6]，每年有超过 160 多万的博客（blogs）发表。

- **广泛链接（Linked）：**社交媒体的社会化特性使得用户产生的数据天生就是广泛链接的，最直观的就是用户产生内容往往由于用户之间的各种社交关系链接在一起，是一种新形式的大数据。这种链接的数据显然不是独立同分布的（IID，independent and identically distributed），对于想要使用传统的数据挖掘和机器学习方法研究社交媒体的研究人员是一种挑战[3, 4]。

- **充满噪声（Noisy）：**社交媒体数据产生门槛的降低以及接入手段的增加，使得社交媒体的数据质量参差不齐，充满噪声[5]。不仅于此，社交媒体中的网

[4]信息超载描述了一种状态，就是当一个人在做选择时因为太多的信息而造成决策的困难。
[5]http://www.twitter.com/
[6]http://www.flickr.com/

络结构也充满噪声，一是网络中存在一些传播虚假和垃圾信息的用户[6]，二是用户间建立关系的便捷性使得用户很容易将各种社会关系混杂在一起，无法区分好朋友和陌生人[7]。

- **非结构化（Unstructured）：**社交媒体中用户产生数据，主要是文本数据，是高度非结构化的，尤其是随着移动互联方式的普及，越来越多的用户使用移动设备更新 Facebook 的状态，发送微博，或者回复别人的帖子，这不但导致了文本内容短小，而且错误拼写频繁出现[8]，经常还有一些非自然语言的使用，比如表情符（:)，:() 和缩写（h r u?）等[9]。

- **不完整性（Incomplete）：**为了保护用户的隐私，社交媒体平台一般允许用户将一些个人数据进行隐藏不被他人看到，这些数据包括个人信息，状态更新，朋友列表，发布的视频和照片以及与他人的信息交流等。比如 Facebook 仅有很少部分用户（小于 1%）公开了他们的个人数据[10]。因此社交媒体的数据是极度不完整和稀疏的。

社交媒体的迅速普及与壮大，使得它在政治、经济、教育、社会等多方面发挥着越来越重要的作用。一些互联网公司开始以社交媒体大数据资源为支撑，以 SaaS 形式为用户提供服务。典型的如谷歌和 Facebook 的自助式广告下单服务系统、Twitter 基于实时搜索数据的产品满意度分析以及国内百度推出的大数据营销服务“司南”等。同时，政府也是社交媒体数据的积极使用者，2013 年曝光的棱镜门事件显示出美国国家安全部门在使用社交媒体数据应用的强大实力，其应用范围之广、水平之高、规模之大都远远超过人们的想象。白宫 2014 年 5 月发布的《大数据：抓住机遇，守护价值》报告中重点提及了社交媒体大数据[7]。社交媒体大行其道的今天，自然也会成为品牌营销的手段之一，比如今年世界杯的主要赞助商之一可口可乐就首次挑选了粉丝在 Facebook[8]和 Twitter 上分享的照片，尝试进行 iBeacon 在社交媒体营销中的运用。目前常见的社交媒体的大数据应用有：一是基于用户个人信息、行为、位置、微博等数据而进行的个性化推荐、交叉推荐、品牌监测等营销类大数据应用，被互联网广告、电子商务、微博、视频、相亲等公司普遍采用。第二，公共服务类大数据应用，即不以盈利为目的、侧重于为社会公众提供服务的大数据应用。典型案例如谷歌开发的流感、登革热等流行病预测应用能够比官方机构提前一周发现疫情爆发状况。国内也有搜索引擎公司提供诸如春运客流分析、失踪儿童搜寻的公益大数据服务。三是积极借助外部数据，

[7]来　源：http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[8]http://www.facebook.com/

主要是互联网数据，来实现相关应用。例如，金融机构通过收集互联网用户的微博数据、社交数据、历史交易数据来评估用户的信用等级；证券分析机构通过整合新闻、股票论坛、公司公告、行业研究报告、交易数据、行情数据、报单数据等，试图分析和挖掘各种事件和因素对股市和股票价格走向的影响；监管机构将社交数据、网络新闻数据、网页数据等与监管机构的数据库对接，通过比对结果进行风险提示，提醒监管机构及时采取行动；零售企业通过互联网用户数据分析商品销售趋势、用户偏好等等。

随着社交媒体的迅速发展与参与用户的数目不断增多，社交媒体中可使用的信息也越来越丰富，具有广泛的应用前景。但是社交媒体中信息的庞大规模使得手工分析其内容变得十分困难，因此本文从信息自动化处理的角度对社交媒体的信息，主要是文本信息进行挖掘与分析，发现有用的信息，为社交媒体的相关应用提供帮助，本文特别关注社交媒体中的观点信息。当然社交媒体与传统媒体存在显著的差异，其自身有不同的特点，我们将研究分析其特点为解决观点信息挖掘分析问题找到解决方案。

1.1.2 观点分析

信息分为两种，即客观信息和主观信息。语言学家 Lyons[11] 将语言功能分为描述（descriptive）, 社交（social）的和表达（expressive）三种功能。其中描述功能主要表达客观事实信息（factual information），而社交和表达功能往往表达的是个人的主观信息（subjective information）。主观信息，在语言中主要表现为观点信息，是人们在语言中表达对于谈论的目标事物的态度、情感或者看法[12]。观点常常简化为人对目标的同意或不同意（或者认为目标好或者坏，或持积极（positive）态度还是消极（negative）态度）[13] 等简单表示形式。总结起来，用户在社交媒体中表达的观点信息有三种类型：在评论、论坛、博客以及微博中针对某主题发表的个人体验（experience）和想法（opinion）；关于新闻文章（artitle）、议题（isssues）、话题（topics）发表的评论（comments）；在社交网站，比如 Facebook 上发表的个人状态更新（status）。

以往为了获取用户观点，需要进行问卷调查，而社交媒体的出现，为用户提供了全新的内容共享平台，使大量连接到网络的用户能够在各种社交媒体发表和表达自己观点：比如消费者可以在 Amazon[9], Yelp[10], 以及 TripAdvisor[11]上发表各种商品和服务的评论；用户可以在 Twitter[12]和 Facebook[13]上对最新话题表达自己

[9] www.amazon.com
[10] www.yelp.com
[11] www.tripadvisor.com
[12] www.twitter.com
[13] www.facebook.com

的观点。社交媒体上巨大的用户群以及由他们产生的海量信息成为了分析用户对各种话题所持观点的宝贵资源。这些观点信息无论是对个人还是机构都起到非常重要作用。比如 Horrigan[14] 发现网络中发表的宣传信息对于网络用户在某些话题上观点的形成具有深远影响，用户表达的观点同样也是产品商家[15] 以及政策制定者[16] 不得不考虑的重要因素，有证据显示这种观点的相互作用过程具有显著经济效果[17–19]。此外，大规模的用户看法整合形成的观点可以反映政治倾向[20]，甚至可以提高股票市场的预测[21]。

社交媒体用户产生内容的大容量不可能依靠人工地去发现和总结其中的观点信息，因此需要计算机能够自动对观点信息进行分析和挖掘。观点分析[14]（opnion analysis）[22] 就是对文本中带有情感色彩的主观性信息进行分析、处理、归纳和推理的过程，其目的是自动发现和区分针对目标的情感和观点，目标可以是命名实体、也可以是话题或事件。尽管语言学和自然语言处理已经有很长的研究历史，但是直到 2000 年才开始进行观点挖掘和情感分析等观点分析研究，从此观点分析成为了非常活跃的研究领域。特别是由于社交媒体的出现，研究者第一次拥有了大量的具有主观性信息的数据，正是有了这些数据，规模性的观点分析研究成为可能，可以说观点分析与社会媒体是一起起步和成长的，是社交媒体中数据分析的核心研究。观点分析研究不仅对自然语言理解（Natural Language Understanding）有着重要的影响，而且还对管理学，政治学，经济学和社会科学产生深远影响，因为它们都受到人的主观性的影响。

首先针对观点分析研究应该明确以下几个概念：

定义 (文档（Document）)：文档指的是自然语言中的文本片段，一般文档中至少会讨论一个话题。

定义 (话题（Topic）)：本文中的话题可以是命名实体，事件，或者文档中提及的抽象概念（政治、健康、体育等）。 ■

定义 (情感（Sentiment）)：情感指的是文档作者针对话题表达的态度（attitude）、观点（opinion）或情绪（emotion）。 ■

定义 (情感极性（Polarity）)：情感极性值指的评价观点积极（positive）或消极（negative）程度的度量值，可以是一维的（打分值），二维的（积极值和消极值），也可以是多维的（喜怒哀乐等情感对应值）。 ■

[14]本观点分析包括了情感分析 (sentiment analysis)，观点挖掘（opinion mining）以及主观性分析（subjectivity analysis）等任务，都是对文本中的观点信息进行分析。

针对观点有不同的定义方法，Liu[23] 将观点形式化定义为观点五元组 $(e_i, a_ij, s_ijkl, h_k, t_l)$，其中，$e_i$ 是目标名称，a_{ij} 是目标的不同属性（方面，aspect），h_k 是持有观点的用户，t_l 表示时间，s_{jkl} 是观点的情感值。Kim 和 Hovy 也对观点做了定义[24]，认为观点由四个元素组成：即主题 (Topic)、持有者 (Ho1der)、陈述 (Claim)、情感 (Sentiment)。他们认为观点分析就是发现和确定各个元素的过程。总体来看，比较全面的观点分析可以认为是由三个主要步骤组成：

- **文本中观点信息确定：**需要确定文档中涉及的话题信息，将不同话题对应的文本片段进行对应联系起来，并且需要对文本片段进行主客观分离找到主观性文本，因为观点一般是从主观性的文本中确定的。将主客观文本分离一般需要一些明显带有情感的词语作为标志，这些词语集合在一起形成了能对情感知识进行表示的情感词典。

- **文本情感分类:** 从文本中抽取有用的特征将文本分为不同的情感类别，一般是将文本分为积极或消极极性（或者中性，即客观文本），主要使用机器学习的各种方法，或者基于规则的方法。

- **观点的集成表示:** 经过前面两步，得到了主观文本片段以及文本片段中的具体观点，观点的集成是在更高的层次上的观点分析，是将文本片段中分散的观点整合集成成为一个观点，并根据不同的应用需求以一种合理的方式表示，比如可以见同一作者的所有文本片段中的观点集成起来，称为用户的主观性模型，或者将文本片段按照时间前后串联起来形成观点随时间的演化表示。

观点分析有别于传统的话题分析。话题分析关心的是文本所阐述的客观话题，如文档是属于教育类还是娱乐类的，而观点分析主要用来识别文档中表达的观点、喜好、立场和态度等主观信息，需要分析文档的词语语义、词性，甚至句法和篇章结构等信息。在传统的话题分析中，主题词是最重要的特征，而在观点分析中，情感词是最重要的特征。观点分析涉及语言学领域的诸多问题，由于语言的复杂性和多样性，需要面临以下几个问题：

1. **领域相关性：**某些情感词在不同的领域中具有不同的情感极性，比如：“轻薄” 在通常意义下具有消极极性，如 “举止轻薄”，而在电脑领域，“轻薄” 却表示褒义，具有积极极性。

2. **语境依赖性:** 某些词语具有多个词性，并且不同的词性常常呈现出不同的情感极性。比如 “这款空调经济耐用” 和 “经济呈现快速发展” 在这两句话中，

“经济” 具有不同的词性和情感极性，前者是形容词，具有积极情感极性，后者是名词，具有中性情感极性。

3. **上下文相关性:** 语言中有许多词语本身是不具有情感极性的，但是在特定的上下文环境中，语言描述便具有了情感极性。比如：“小”、“高”、“快” 等词语，在搭配组合 “损失小”、“成绩高”、“进步快” 中，具有积极情感极性，但在搭配组合 “心眼小”、“耗油高”、“耗电快” 中，则具有消极情感极性。

1.2 研究问题

随着以 Twitter，Facebook、新浪微博为代表的社交媒体迅速发展，人们越来越愿意在线分享自己的看法、观点以及体验，他们可以选择博客写作、微博发帖、社交网络状态更新、发表产品评论、或者在论坛中参与讨论发表对于任何话题的看法和观点，因此网络被各种各种意见所充斥，现在的网络可以说是一个“观点网络”，网络已经成为获取观点信息的主要来源。但是站在网络使用者的角度，一方面可以很容易获得大量的带有观点的信息，这些信息远远超出了个人的处理能力，因此用户面临着“信息超载（information overload)”问题；另外一方面，观点的主体是人，而网络中的信息尤其是社交媒体中的信息多是以“碎片化”的形式存在，个人观点分散在许许多多多的信息碎片中，使得用户真正的信息需求（能针对目标主题及时准确从网络中找到公众或个人观点）变得更加困难，因此用户又面临“信息不足（information shortage)”问题。传统的信息检索技术，尤其是搜索引擎，很难解决这样矛盾的信息供求关系。当然目前已经有观点检索系统[25–27]，可以解决如“检索评价某产品的文档，并总结其中的观点”这样的问题，但是还不能满足“大家对某产品的观点或某个朋友对该产品的观点”这样的信息需求。因为这样的问题需要就网络中，尤其是社交媒体中每个用户的观点进行挖掘、分析并整合集成。网络中的信息，一个事实信息可以代表所有的事实信息（One fact stands for all facts）, 但是一个观点不能代表所有的观点（one opinion can not represent all opinions)。从用户的角度来说，一条信息中的观点可能只是他就话题的某个方面表达出的观点，就话题整体的观点需要将所有分散在“信息碎片”中的观点进行集成，并以一种合理的形式表示出来，才能代表一个用户真正的观点。因此本文首先从下面一个科学问题出发，来研究观点分析：

怎么样才能准确的对社交媒体中用户的观点进行表示和分析？

这个问题需要从挖掘用户的信息碎片中的观点出发，是一个观点信息确定、分类以及整合的过程，需要解决文本情感知识表示，情感倾向性的分类以及观点信息的集成等问题。

反过来，因为人是具有主观能动性的，人的行为会受到自己观点和思想的影响。用户在使用社交媒体时会有多种交互行为，比如信息传播行为，人们通过转发分享新闻与观点，加速信息的流动、扩大信息传播的范围。用户的信息传播行为同样会受到自己的主观性的影响，通过观点的分析与与集成可以对用户的主观性进行建模，而用户的主观性模型无疑会对分析用户的一些在线的信息交互行为有帮助。因此本文从以下两个科学问题出发来研究用户的主观性建模，并分析其对用户传播行为的影响：

用户的主观性如何表示和建模？

怎么样使用主观模型分析用户的信息传播行为？

用户在社交媒体中的产生的内容会涉及到多种话题，而且会对话题的不同方面发表观点，因此回答第一个问题需要研究用户产生信息中多样性话题的确定及表示问题，还有用户在不同话题上多种观点的集成及表示问题。在信息传播过程中，用户作为带有自己主观判断的传播主体，会在不同情况下产生传播行为，因此回答第二个问题需要首先确定用户传播行为产生的具体原因，然后研究怎么样从主观动机角度对这些原因度量分析。

1.3 相关研究

本节主要介绍与观点分析与用户传播行为分析相关的一些现有工作，其中观点分析包括观点挖掘，观点集成两个部分相关工作。本文的相关工作分析主要从整体相关工作和局部相关工作进行阐述，本章的相关研究主要介绍的是整体的相关工作，因为这些研究成果可以为本文所研究的具体任务提供思想借鉴和技术支持。以后各个章节中的相关工作则会具体地分析已有的类似工作，以及研究成果。特别需要强调的是，无论是观点分析还是传播行为分析，对社交媒体中文本的处理都是其中一个重要的环节。社交媒体数据的一些特性已经在第 1.1.1节有所介绍，这些特性造成自然语言处理技术在社交媒体上的应用存在着新的挑战，使用自然语言处理技术对社交媒体文本进行处理，主要工作包括文本规范化（Normalization）[2, 28–30]，领域适应化（Domain adaptation）[29–32, 32–41]，预处理（preprocessing）[37, 42, 43] 以及进行一些结构化处理（词性、句法、标注等）[32, 40, 44–53]。

1.3.1 观点挖掘

观点挖掘研究识别文档中针对一主题表达出的观点以及这观点的极性（例如，是积极还是消极）。观点挖掘通过对文档深入分析得到文档中表达的观点信息，是观点分析后续任务的基础，观点挖掘的结果影响着后续分析任务。一般观点挖掘

包含观点识别（identify）和极性分类（classify）两个步骤。观点识别主要是从文档中识别出话题以及与话题相关的带有观点的文本片段。识别带有观点的文本片段(一般是文档中的句子）也称为主客观分析（subjectivity analysis），是将文档中的带有观点的句子与描述客观事实的句子区分开，研究表明将文档中的客观文本过滤有助于提高观点挖掘的准确性[54]，目前主客观分析主要采用机器学习方法进行主客观分类[55–60]。极性分类是将文档就话题表达出的情感倾向进行极性分类，一般是分为积极与消极，也可以是多种类别（当类别为积极、消极以及中性时，与主客观分类一致)。观点挖掘研究方法一般可以分为基于词典、基于统计以及机器学习三类。

1.3.1.1 基于机器学习观点挖掘

在观点挖掘研究早期，机器学习方法和标注数据集的使用加速了研究的进展，目前最常使用的仍然是机器学习方法。机器学习方法是对分类问题的比较成熟的解决方案，一般经过训练和预测两个过程，可以进行如下形式化表示：假设训练数据集是经过极性标注的文档集 D，每个文档都可以用特征（词语，二元组等等）向量表示，文档都被标注了情感极性（在极性空间 S 中的一个值)，机器学习的训练过程可以形式化为，给定训练数据集：$\{(d,s)|d \in D, s \in S\}$，找到映射：

$$g : D \to S, \quad g(d) = \arg max_S f(d,s) \tag{1.1}$$

极性分类也就是找到映射 g，将文档根据打分函数 f 映射到情感极性空间，函数 f 以文档向量和标注的极性作为输入，对未标注的文档给出极性预测的概率值（使用条件概率或联合概率)，训练的过程就是对打分函数 f 的估计过程。一般训练过程有以下几个步骤：（1）首先获取训练数据集，数据集可以是带标注的（有监督)，也可以是无标注的（无监督)；（2）在文档集中发现有用特征，将文档使用特征向量表示；（3）通过分析相关特征共现规律，训练分类器区分文档极性标签；（4）最后使用训练得到的分类器对新文档预测给出极性概率值。

Pang 和 Lee[61] 最先将机器学习方法引入了观点挖掘领域，作者提出了使用三种有监督的分类器（Naive Bayes (NB)，maximum entropy (ME) 和 support vector machines (SVM)）进行电影评论的情感分类，三种分类器都能超过随机选择的基准分类器，平均准确率达到 80%，其中 SVM 表现出最好的性能。Dave 等[57] 扩展了 Pang 的工作，强调使用特征选择对情感表示特征进行过滤，可以将准确率提高到 87%。Pang 和 Lee 使用主客观分析对文档进行预处理，过滤掉其中的描述客观信息的句子，发现可以提高极性分类的准确性。后续的一些工作主要集中于研究如何扩充有用的分类特征[62–65]，训练数据的构建[66, 67] 以及机器学习方法的选择[68]。

1.3.1.2 基于词典的观点挖掘

基于词典的观点挖掘依赖于预先构建好的情感词典，词典里的词语都标注了情感极性值。常用的英文情感词典有 General Inquirer[15]，DAL（Dictionary of Affect of Language）[16]，WordNet-Affect[17]以及 SentiWordNet[69]。基于情感词典的方法一般是用词典确定文本中的带有观情感极性的词语，用以判断文本是否主观文本。也有一些研究使用情感词典词语的情感极性值来计算文本的观点极性[70–72]，一个句子或文档的极性值可以通过将每个词语的极性值取平均来确定，常用的计算方法可以使用公式 1.2来表示：

$$SD = \frac{\sum_{w \in D} S_w * weight(w) * modifier(w)}{\sum weight(w)} \tag{1.2}$$

其中 S_w 是文档中的词语在情感词典中的极性值，$weight(w)$ 是权重函数，可以根据词语相对于话题词的位置进行权重调整，$modifier(w)$ 是专门处理否定、增强或其他改变词语情感值的一些修饰操作。典型工作如 Zhu 等[73] 首先将文档或句子中词语的极性值累积在一起，然后使用简单的基于规则算法计算整个文档或句子的极性值。一些比较成熟的情感分析工具，比如 Sentiment Analyzer[74]，或 Linguistic Approach[75] 针对话题挖掘一些领域相关特征、观点句的模式或词性标签等作为规则加入到文档极性值的计算中，可以得到更精确的极性值，但是需要增加计算复杂性。

1.3.1.3 基于统计的观点挖掘

这种方法是基于语料中表达相似观点的词语经常会出现在一起这一观察基础上的，因此，如果两个词语频繁在同一上下文中同时出现，它们就很有可能有相同的极性。因此一个词语的极性值可以根据它与一个极性恒定的词语（比如“good”）共现的频率来确定。Turney[76, 77] 提出使用点对点互信息（point-wise mutual information (PMI)）[78] 作为统计依据来计算词语的共现：

$$PMI(x, y) = \log_2 \frac{F(x, y)}{F(x)F(y)} \tag{1.3}$$

其中 $F(x, y)$ 表示两个词语的共现频率，$F(x)$ 表示词语的出现频率。词语 x 的极性值可以通过计算该词语与两个相反极性词语的互信息差值来确定：

$$PMI - IR(x) = \sum_{p \in pWords} PMI(x, p) - \sum_{n \in nWords} PMI(x, n) \tag{1.4}$$

[15] http://www.wjh.harvard.edu/~inquirer/
[16] http://www.hdcus.com/
[17] http://wndomains.fbk.eu/wnaffect.html

其中 *pWords* 表示积极极性的基准词集合，*nWords* 表示消极极性基准词集合。为了统计词语出现频率，Turney 使用 AltaVista 搜索引擎检索词语返回的条目数作为词语出现频率。Chaovalit 和 Zhou[79] 扩展了 Turney 方法，通过谷歌搜索引擎确定词语共现频率，提高了准确性。Read 和 Carroll[80] 使用语义空间和分布相似性作为替代方法进一步扩展了 Turney 方法。这种方法更细致的全面的研究是 Taboada 等[81]，他们提出了使用搜索引擎确定共现频率的一些问题。Ben 等[25] 提出使用统计方法构建情感词典与信息检索相结合的方法获取主观性的博客文档。

1.3.2 观点集成

通过观点分析得到的是单个文档中的观点信息，实际使用的时候，我们关注的是更高层次的观点，而不是单独一篇文档的观点，因此需要对文档观点分析结果进行整合集成。这种整合集成可以按照不同的维度进行，比如为了了解一群人的观点分布，需要将每个人发表的所有文档中的观点进行集成。最需要观点集成的研究领域是产品评论，需要从大量用户发表的评论中抽取出产品的特征，并计算不同用户针对相同特征的观点或打分的平均值，以便进行观点集成形成对产品总体的评价。以图 1.2所示产品评论的观点集成为例，观点集成一般包括信息收集，观点识别，观点分类以及推理集成三个步骤[82–84]。

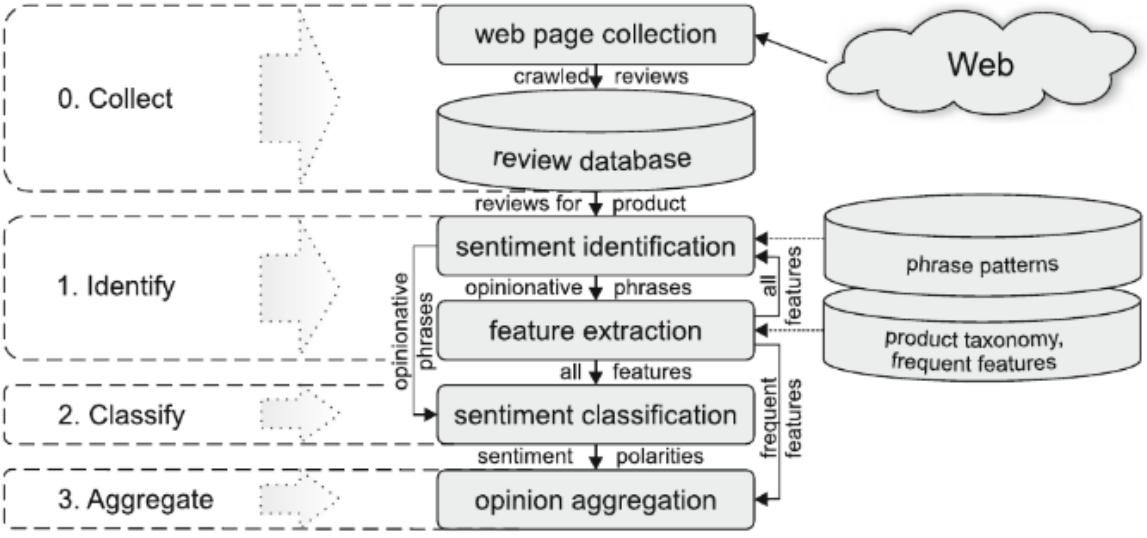


图 1.2　产品评论的观点集成框架

对于文档集 D 中的针对某话题的观点进行集成形式化表示为：

$$\{(f, s_f) | rep(f, D) > \rho_f, s_f = agg(S, f)\} \tag{1.5}$$

其中 f 表示根据某种表示度量方法 $rep(f, D)$ 确定的描述话题的不同方面，s_f 是针对 f 根据集成函数 $agg(S, f)$ 计算出的针对 f 的观点。

观点集成一个主要的挑战就是如何确定描述话题的多个不同方面。Leouski 等[85] 评测了各种文本聚类方法对检索结果中信息的集成效果，发现聚类方法对于文本的交互式检索是有用的。Zeng 等[86] 使用监督学习方法从文本中抽取主要短语并将其聚类表示话题的方面。越来越多的工作使用生成模型发现文档中的隐性方面话题[87, 88]，还有一些工作使用数据挖掘中的联合规则方法对产品的相关方面进行挖掘[83, 89]。

1.3.3 传播行为分析

社交媒体中信息传播具有重要的应用价值，信息传播的主体是人，也就是社交网络的用户，研究人的传播行为是研究信息传播的重要组成部分。社交媒体中最具影响力的是微博上的信息传播，因为用户在微博的转发行为会使得信息在短时间内形成大规模的传播，因此本文主要从微博的转发行为来阐述相关工作。微博的典型代表是 Twitter，Twitter 转发机制，即重新发布其他人发布过的微博，以便于作者的全部粉丝看到转发的信息，使得信息迅速形成病毒式传播（viral propagation）。很多对于转发行为的研究分析涉及影响转发行为的因素，包括 tweet 的文本内容与转发的关系，用户的属性如何决定其他人的转发；Twitter 中信息的一般传播路径与规律等等。

Boyd 等人研究了 Twitter 中转发的各种类型以及转发的原因，他们分析了不同用户，用户属性，用户交流方式对于转发的影响，同时也分析了人们在 Twitter 中喜欢转发的内容[90]。他们发现 18% 的转发 tweet 包含 hashtag，52% 的转发 tweet 包含链接，11% 的转发 tweet 包含连续的转发符号串（如，“RT @user1 RT @user2 ”），另外，9% 的转发 tweet 都包含回复原 tweet 作者的回复字符串（“@reply”）。这说明 tweet 文本中的 hashtag，链接、回复、提交和转发符号都与 tweet 的转发存在着一定的对应关系。

Yang 和 Counts 通过 Twitter 中的提及（“@username”）抽取了用户之间的关系，并在此基础上构造了用户关系的复杂网络。他们研究了信息在这个复杂网络上是如何传播的，包括信息传播的速度，规模，以及范围[91]。他们发现大约只有 25% 的 tweet 是被信息作者的朋友转发，大部分是被粉丝但非朋友转发。这说明 Twitter 中用户形成的复杂网络，影响着人们的转发行为，因此信息在传播路径上具有一定的规律可循。

Macskassy 和 Michelson 分析了一个月用户的 Twitter 数据，他们解释了各种信息传播的方式，尤其是转发的行为模式，他们发现 tweet 的内容是 tweet 被转发的决定因素，因此他们构建了基于内容的转发模型[92]。

Starbird 等人对具体事件在 Twitter 上的传播进行了深入研究，他们分析了 2011 年埃及的政治事件，演示了这个事件的相关信息在 Twitter 上是如何生成，发展，传播的[93]。

Comarela 等人研究了影响用户回复或转发的因素，他们发现以前是否回复，发布信息的频率，信息的时效性，tweet 的长度决定用户是否回复[94]。

除了以上的工作，最新的研究还从不同角度对 Twitter 中的转发行为进行了深入的研究[95–98]。

综上所述，影响用户转发行为的因素主要包括 tweet 文本的内容、tweet 文本的社交媒体属性（如是否包含链接、hashtag、提及等)、tweet 作者的用户属性，tweet 作者的朋友圈子，当然以上的研究都是从宏观上大规模分析 Twitter 转发数据得出的研究结论。从微观的角度则可以考虑给定一个 tweet，未来这个 tweet 是否会被转发，是一个值得研究的问题。

虽然已有的 Twitter 转发研究从许多不同的角度进行了考虑，但是还没有工作从用户的主观动机角度进行研究，本文我们将结合用户观点分析研究的结果对转发行为进行分析。另外，目前的转发大多都是从 tweet 本身进行考虑，并未从受众的角度进行分析，本文将对 tweet、作者、受众三个方面在转发过程中的相互关系进行探讨。

1.4 研究内容与方法

1.4.1 本文研究内容

本文的研究内容主要是围绕社交媒体上的观点信息的分析应用，从两个角度对用户产生的带有观点的内容进行建模：一个角度是从不规范的社交媒体文本中发现观点信息，并在用户层面进行观点集成对用户的主观性进行建模，另外一个角度是利用用户产生内容中分析得到的用户主观模型分析用户在使用社交媒体时的一些在线行为，本文主要分析用户在微博的转发行为。研究框架如图1.3所示。

本文主要研究内容分为四个部分：首先从社交媒体文本中得到观点信息属于观点挖掘研究内容，观点挖掘方法分为基于情感词典和基于机器学习的方法，因此需要进行**情感词典的构建研究**以及判断观点极性的**情感分类研究**；其次从社交媒体文本片段中挖掘到的观点需要进行整合集成，变成具有代表性的观点信息，属于观点集成的研究内容，我们将从用户维度对用户的所有观点进行集成，用于**用户主观性建模**；最后利用用户的主观模型，从行为主观动机角度对用户的信息传播行为进行分析，属于**转发行为分析**研究内容。具体四个研究内容的阐述如下：

1. **情感词典的构建**：使用已有的比较成熟的英文情感词典中的情感知识进行跨语言情感知识转移，构建一个通用的中文情感词典；针对通用情感词典领域

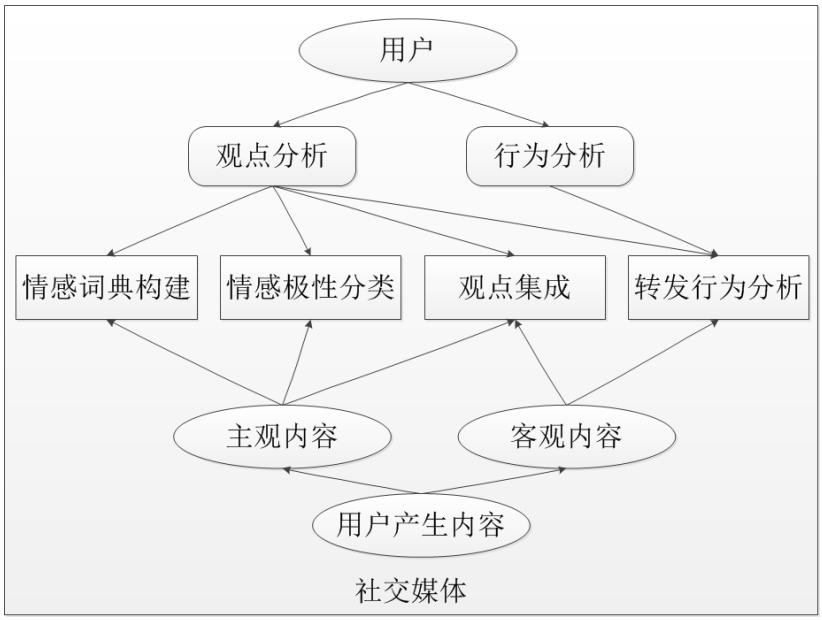


图 1.3　本文研究框架

适应性弱的问题，通过基于语料库情感词典扩展研究，使用语料中的语言特征以及统计特征，对情感词典在领域内进行扩展以增强情感词典的领域适应性。

2. **情感分类**：根据社交媒体情感表达方式的领域依赖性，对情感分类特征空间进行分割，将领域独立的通用特征与领域依赖特征分开，使用两部分特征分别训练分类器，通用分类器使用现成资源训练，领域分类器使用远监督方式训练，最后两个分类器在自举式机器学习框架下组合成性能更强的情感分类器。

3. **用户主观性建模**：提出一个通用的主观模型定义，将用户产生内容中关注的话题和针对这些话题表达的观点组合在一起，对用户在每个话题上发表的所有观点整合集成，并使用一个在情感极性值空间中的分布表示用户在话题上的综合观点，使用一个更简单通用的框架构建主观模型。

4. **转发行为分析**：构建好每个用户的主观模型后，给定微博，发现作者的粉丝中，谁会在未来传播微博，从用户的主观动机角度，分析用户在三种转发情形下的主观动机，即微博内容的吸引力，转发微博的社交需求以及转发微博的认同需求。

总的来说，针对 1.2的第一个问题，本文通过构建情感词典识别社交媒体中带有观点的文本信息，并使用无监督的情感分类方法对观点的极性进行分类；针对 1.2的第二个问题，本文通过将用户关注话题与发表的观点进行组合建模，采用

集成的观点表示方式对用户的主观性进行建模；针对 1.2的第三个问题，本文通过计算主观模型之间的相似性度量用户一些在线行为的主观动机，进行行为的分析。

1.4.2 本文研究方法

社交媒体中的观点分析涉及到信息检索、机器学习、自然语言处理与自然语言理解等多方面的方法和技术，这些方法和技术的综合使用是由社交媒体数据特有的性质以及观点分析应用的特殊需求所决定的。从社交媒体数据特性来看，进行观点分析需要面临以下挑战：

- 社交媒体中的文本篇幅较短而且噪声较多的特点，使得利用标准的机器学习方法进行分析面临数据稀疏问题，也造成自然语言处理技术的困难；
- 庞大的数据容量以及动态的语言特性造成通用的标注数据的匮乏，无法满足机器学习训练要求；
- 社交媒体是一个开放平台，文本涉各种领域，因此各种方法和技术都要满足多领域环境的需求；
- 社交媒体中的数据以数据流的形式不断高速增长，需要能够快速适应新数据并实时处理的技术和方法。

观点分析需要从大量的社交媒体用户产生的内容中发现观点信息，进而进行整合集成并用于分析用户的转发行为，要解决以上问题和挑战，需要达到如下几个主要目标：

1. 使用文本规范化和消除噪音等自然语言处理技术对数据进行预处理，数据的稀疏性需要得到缓解，然后才能进入后面的分析中；
2. 观点极性分类方法应该具有领域独立能力，当领域变化时能够快速适应并且性能不能下降；
3. 采用的所有技术和方法能够以有限的计算能力分析和处理不断增长的数据；
4. 针对训练数据缺乏问题，尽量使用无监督或者弱监督的方法和技术，并且尽量使用已有的资源，减少人工标注。

针对以上挑战和目标，我们确定的研究方法为：

首先在数据和资源选择上，我们首要选择已有的知识资源和标注数据。如果没有对应的知识资源，可以通过资源转化变成我们想要的知识资源，比如情感词

典构建时，我们通过情感知识之间的对应关系，将英文情感词典转化为中文情感词典。如果没有直接标注数据，我们选择采用弱标注的方法得到训练数据，所谓弱标注数据指的是，数据的类别标签是通过启发式从数据中直接确定不需要人工标注，比如在训练情感极性分类器时，我们使用含有明确情感极性的成语微博作为训练数据，是基于微博短小观点表达相对会集中在一些极性相同的词语身这一假设，从而获得大量训练数据。在理想的情况下，用弱标记数据的好处是双重的：首先，我们可以以接近零的代价采集训练数据，因此可以轻松地将我们的应用扩展到其他领域或语言。其次，弱标注语料的规模可以很容易超越常规手动标注的训练语料的数量级。

在学习训练方法的选择上，我们优先选择无监督或半监督的机器学习方法。无监督或办监督学习方法可以减少或无需大量的标注训练数据，而且可以通过迁移方法将学习到的知识进行跨领域或语言转换。比如我们在对微博进行极性分类时，使用自举式机器学习方法将两个弱的分类器结合在一起提高了分类的性能；在构建主观模型时，我们使用 LDA 话题模型识别用户关注的话题，并使用基于规则的情感分析方法获得用户的观点信息。

1.5 本文主要贡献

本文以用户为单位，对用户在社交媒体上产生内容中的观点信息进行识别、分析和集成，并使用得到的观点信息分析用户在社交媒体上的在线交互行为，具体来说本文的主要贡献为：

- 设计了一种中文情感词典的自动构建方法，该方法能够从已有的英文情感词典通过词语之间的对应语义关系转化情感知识，并且能够针对任何领域的语料进行扩充，成为准确性更高的领域情感词典。通过与其他中文情感词典的对比，我们的情感词典完全是自动构建，而且具有更好扩展性和领域适应性。

- 基于词语在表达情感时作用的不同，提出了一种新的无监督情感分类方法，该方法将情感分类的特征空间进行分割，在两个不同的特征空间分别训练分类器，然后以自举式学习框架组成更强的分类器。方法无需人工标注的训练数据，使用现有的成语词典资源和弱标注的远监督方法训练分类器，性能超过了需要大量标注数据训练的有监督分类器。

- 提出了从用户维度进行集成的通用的观点集成方法，该方法将用户感兴趣话题以及在话题上的观点组合对用户主观性建模，并且该模型能够在更细粒度的情感空间中使用分布方式表示观点，使得用户的观点表示更准确，综合反

映了用户在话题每个方面所表达的观点，将该模型应用到观点预测任务时，能显著提高观点预测的准确性。

- 从主观动机角度对用户在 Twitter 上的转发行为进行了分析，利用用户的主观模型设计了一种新的计算主观相似性方法，对用户转发行为的三种情形使用主观相似性进行度量，在真实 Twitter 数据中的实验中验证了三种主观相似性度量与转发行为之间的相关性，并且作为有用特征预测转发行为的准确性超过了目前一些主要方法，通过结合其他影响因素，可以使预测性能得到显著提升。

1.6 本文结构

本文的研究工作主要围绕社交媒体观点信息分析与应用任务展开，我们可以将这两方面的工作分为以下几个主要部分：在观点分析方面，我们首先探讨了如何利用现有资源进行中文情感词典的自动构建，情感词典是观点信息识别和情感分类的基础；然后进一步探讨了如何结合社交媒体的文本特点对文本中表达的观点进行极性分类；最后对社交媒体中的观点信息以用户为维度进行整合集成，形成用户在其产生内容中表达的所有观点信息的主观模型。对分析得到的主观信息应用方面，我们从用户作为信息传播的主体角度，对用户转发行为的主观动机使用主观模型进行度量和分析。上述工作共分为七个章节, 论文主体结构以及章节之间的关系如图 1.4所示。，每个章节内容具体安排如下:

第一章是绪论，首先介绍了本文研究的背景，介绍了社交媒体和观点分析一些基础知识，接着提出研究动机，阐明了本文所涉及的科学问题、研究内容，并给出了研究方法，然后分析了研究问题，确立了依托自然语言处理技术与机器学习方法解决这些问题的基本思路，最后介绍了本文的主要工作和文章的结构。

第二章是应用语义关系自动构建中文情感词典，首先介绍了目前情感词典资源的现状，针对中文情感词典资源缺乏问题，提出了以 HowNet 语义知识库为基础，根据中英文词典语义之间的对应关系将英文情感词典的情感知识转化到中文情感词典中，设计了转化方法以及转化中极性值的计算方法，实验中与现有的几个中文情感词典进行了对比。

第三章是基于语料资源的中文情感词典扩展，是对第一章中构建的通用情感在领域语料中的适应性扩展方法研究，首先介绍了基于语料资源的情感词典构建方法，确定了基于语言特征以及统计特征的扩展方法，并提出了综合使用两种特征的混合特征扩展方法，并分别进行了实验验证。

第四章是无监督的自举式情感分类，本章首先介绍了目前情感分类研究现状，针对领域依赖问题，根据词语在表达情感的不同作用提出了特征空间划分方案，

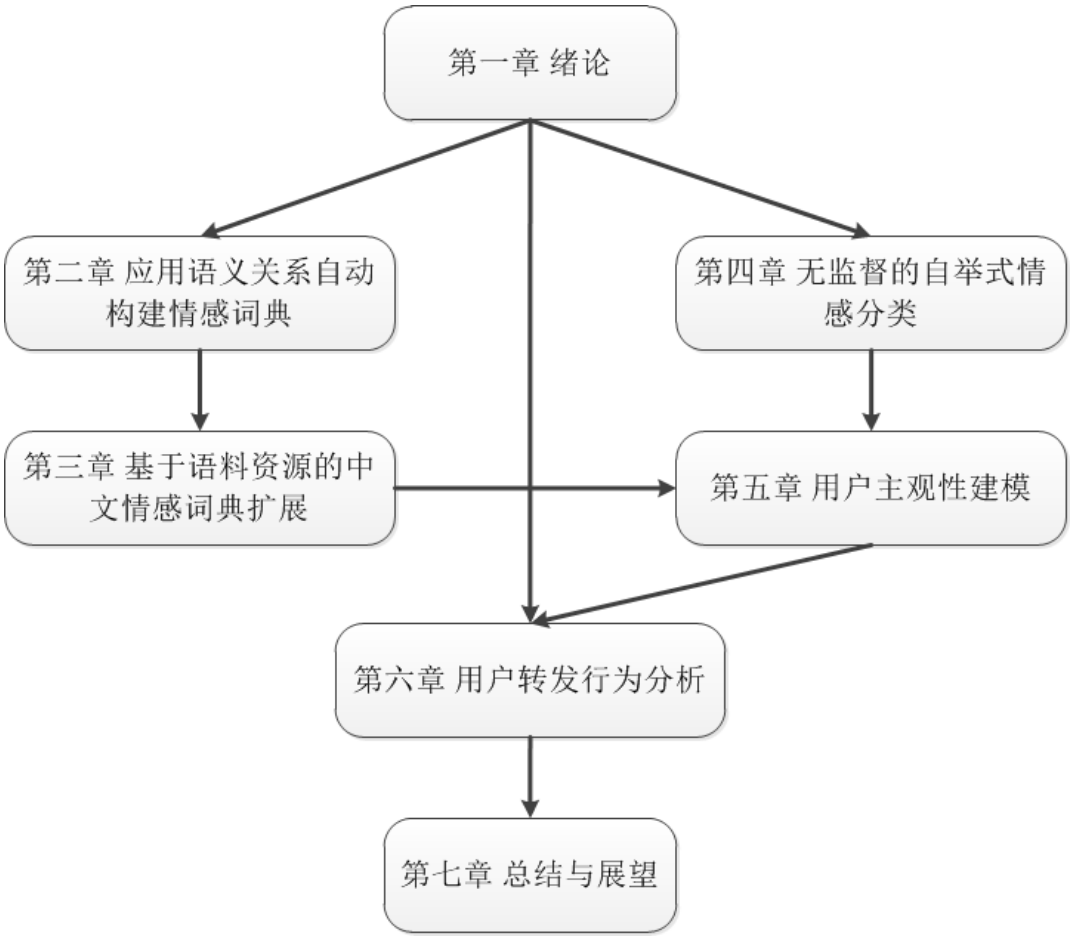


图 1.4　论文整体结构图

并对研究问题进行了形式化，设计了自举式情感分类框架，选用了三种分类器并进行了实验对比分析。

第五章用户主观性建模，首先定义了社交媒体中用户的观点集成问题，然后提出了主观模型的框架，将用户产生内容中的话题和观点组合进行用户观点集成，并设计了通用的模型构建方法，实验中将主观模型应用到观点预测任务，并对模型进行了定性的分析。

第六章用户的转发行为分析，研究的问题是对于给定一个微博，分析微博作者的粉丝中谁会转发该消息，针对该问题，我们使用主观模型从用户的主观动机角度进行分析，设计了主观相似性计算方法，并针对转发行为的三种情形进行度量，最后在实验中定性和定量验证了我们提出方法的有效性。

最后一章是总结部分，我们阐明了本文工作的贡献点，并且指出了工作的一些不足，并对未来社交媒体中观点信息分析与应用的一些问题和方法进行了尝试性地思考。

第二章　应用语义关系自动构建情感词典

2.1　引言

随着互联网的发展, 尤其是社交网络的发展，各种社交媒体的用户发布内容中出现了海量含有用户主观情感色彩的文本数据。针对网络文本的信息处理开始由获得关键词[99]、事件 [100]、话题[101] 等事实信息, 开始向情感观点等主观信息深入，情感分析便是近年来迅速发展的信息处理技术[23]。从数据中提炼出用户的主观信息对于商业情报、舆情分析等具有重要意义。情感分析技术就是对带有情感色彩的主观性文本进行自动推理、分析、归纳的过程，涉及自然语言处理、机器学习、认知科学以及社会心理学等方面的研究[102]。语言的情感表达往往使用具有明确情感色彩的词汇，因此构建带有情感色彩的词典资源是进行情感分析研究的基础。情感分析研究在英文上发展迅速，积累了许多情感词典资源，比如：General Inquirer（GI）[103]，OpinionFinder（OF）[104]，Appraisal Lexicon（AL）[105]，SentiWordNet[69] 以及 Q-WordNet[106]。中文情感分析研究起步较晚，缺乏普遍认可的可靠的中文情感词典[107–109]。目前研究使用主要有 HowNet 情感词典[110]，NTUSD 情感词典[111] 以及大连理工大学的情感词汇本体词库[112]。这些词典主要是以手工或半自动方式编辑而成，覆盖度、可靠性和领域适应性受到限制，并且情感词以主要积极和消极二值区分，缺少情感极性值的细粒度划分。能够将资源丰富的英文词典中的情感知识跨语言向资源相对贫乏的语言进行适应性的转化，以产生其相应情感词典资源，既可以省去耗费大量人力的人工标注过程，又可以克服自动或半自动方法的可靠性和覆盖度问题。

本章提出基于主流的可靠的英文情感词典资源进行转化的中文情感词典的构建方法，可以根据语义关系将英文词语及其情感极性值转化得到中文词语的情感极性值，并且完全是自动的，可靠性和适应性更高。

2.2　词典资源简介

2.2.1　HowNet 语义词典

HowNet 是一个以中英文词语所代表的概念为描述对象，揭示概念与概念之间以及概念的属性与属性之间的关系的知识库。义原是 HowNet 最小语义单元，用于定义和描述概念的属性和概念间的相互关系，义原通过一个树状的层次结构组织构成上下位关系。概念是对词汇语义的一种描述，每一个词可以表达为几个概念[113]。如图 2.1所示，HowNet 采用 KDML（Knowledge Dictionary Mark-up

Language）语言描述概念，其中 W_X 表示词语，G_X 表示词语词性，E_X 表示词语例子，X 为 C 时表示中文，X 为 E 时表示英文。DEF 是对于该概念的定义项，称之为一个语义表达式，其中中英文标注的是义原，“#”等标示符号来对概念属性之间关系进行描述，DEF 中还可以包含概念，概念之间相互交织构成一个网。HowNet 一共有 2234 个义原，收录了近 15 万条概念记录，涵盖了绝大部分中文常用词语，本章将基于 HowNet 的词语进行情感词典的构建。

```
NO.=098818
W_C=医生
G_C=N
E_C=
W_E=doctor
G_E=N
E_E=
DEF=human|人,#occupation|职位,*cure|医治,medical|医
```

图 2.1　HowNet 中概念的定义方式

2.2.2　WordNet 语义词典

WordNet 是由 Princeton 大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英文词典[114]。WordNet 是根据词义而不是词形来组织词汇信息。WordNet 使用同义词集合（Synset）代表概念，词汇关系在词语之间体现，语义关系在概念之间体现。WordNet 将英语的名词、动词、形容词和副词组织为 Synsets，每一个 Synset 表示一个基本的词汇概念，并在这些概念之间建立了包括同义关系（synonymy）、反义关系（antonymy）等多种语义关系。其中，WordNet 最重要的关系就是词的同义反义关系。

2.2.3　SentiWordNet 情感词典

SentimentWordNet 是 Baccianella[69] 等在语义词典 WordNet 基础上使用随机游走的图算法得到的情感词典。词典的每条记录都是一个 WordNet 的 Synset，并且每个 Synset 都计算出了褒义、贬义情感强度值，本文就是利用 SentimentWordNet 的情感强度值以及 HowNet 概念的语义关系进行计算得到中文词语的情感极性值。SentimentWordNet 共有 117,000 多 Synsets，192,493 单词。

2.3　基于语义关系的情感词典构建方法

将英文情感词典的研究成果转化为文资源，可以利用语言之间的语义对应关系减少词典的歧义，使情感词典更加可靠，还可以直接将英文中对情感强度的计

算直接转化为中文词语的情感强度计算，减少了计算开支。本研究正是基于这种动机展开的。HowNet 对义原和概念进行了英汉双语标注，可以作为转化的“桥梁”。但是英文词语和中文词语都存在一词多义现象，不同语义所表达的情感倾向也不同，因此得到的情感极性值也会存在歧义。HowNet 中概念的 DEF 是由义原按语义关系进行描述的，可以利用这种语义关系对词语的情感极性值进行“消歧”。总体来说，解决方案如图 2.2框架所示。

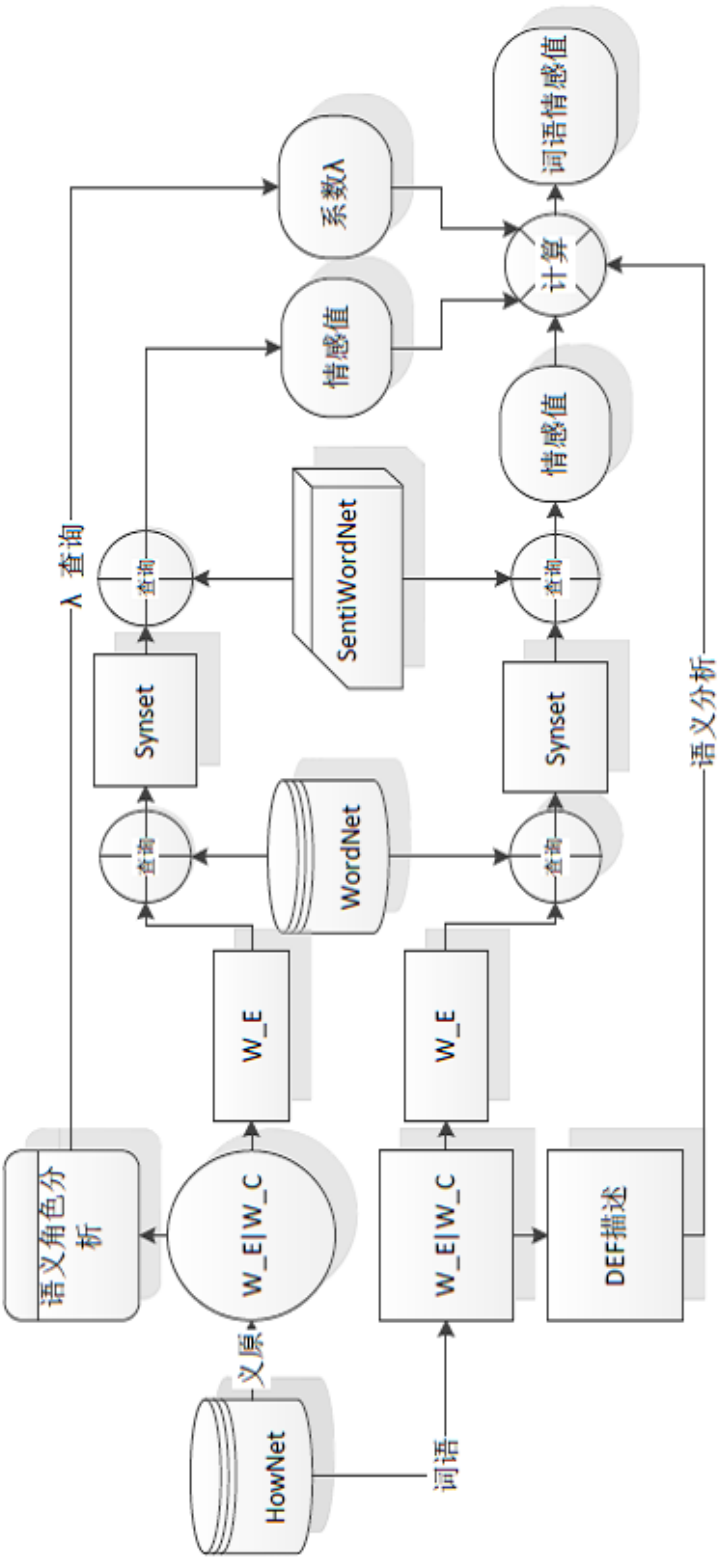


图 2.2　基于语义关系的情感词典解决方案

构建中文情感词典框架可以分为义原和词语抽取及语义分析、义原和词语情感极性值查询与计算以及词语的情感极性值计算三个过程。

2.3.1 词语抽取和义原抽取及语义分析

词语抽取主要是从 HowNet 词典中抽取词语（W_C）和属性定义（DEF）并对 DEF 进行分析。DEF 是由义原和语义关系描述等构成的, 在进行词语倾向计算时，需要根据义原进行词语的语义分析和倾向计算。情感词语抽取处理流程如图 2.3所示。

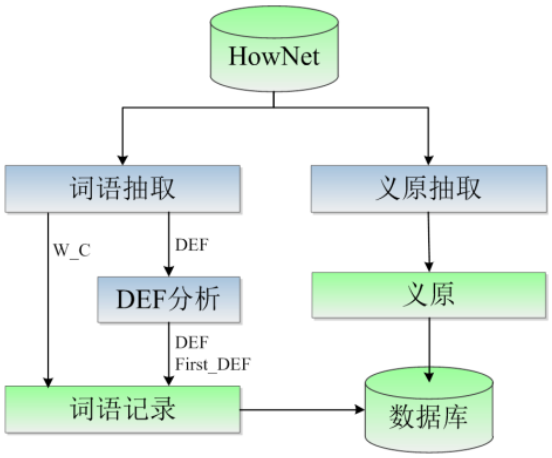


图 2.3　词语和义原抽取处理流程

在抽取得到的词语记录中，主要关注的内容有词语编号（No）、中文词语(W_C)、中文词性（G_C)、英文词语（W_E)、英文词性（G_E)、属性（DEF)、第一属性（First_DEF）等。其中第一属性是指位于属性 DEF 第一位置的义原，通过第一属性可以分析出该词语所属的特征类。

由于 HowNet 中的词语是由义原和语义关系描述等构成的。在进行词语倾向计算时，需要根据义原进行词语的语义分析和倾向计算。在抽取得到的义原的记录中，主要关注的内容有词语编号（No）、特征类别（Category)、中文词语（W_C)、英文词语（W_E)、属性（DEF)、层次（Layer)、父亲节点编号（Father）等。根据记录中的层次（Layer）和父亲节点编号（Father）可以得到义原之间的语义关系，如编号为 33 的义原“依靠”位于“事件类（Event)”的第五层，其父亲节点编号为 32，通过查询编号为 32 的义原，得到其父亲节点义原为“有关（relate)”，表示 DEF 中包含，因此抽取的记录中包含了义原及其在词语中的语义关系。

2.3.2 情感极性值的查询与计算

HowNet 词语是中英双语，因此有的可以直接将抽取到的英文词语（W_E)、英文词性（G_E）直接送入英文情感词典查询其情感极性值。但是大部分词语英文

部分不是一个单词，因此无法直接得到情感极性值，而且由于词语的多义性，也无法获得唯一的情感极性值，因此需要进行“消歧”；HowNet 中词语是由其属性 DEF 定义的，DEF 是由多个义原按照一定的语义关系组合而成的，词语的倾向性可以看作是由义原的倾向性按照一定的规律组合而成的。因此词语的倾向性值可以通过义原的倾向性值根据语义关系计算获得，一方面可以获得直接查询无法获得情感极性值的词语，另外一方面也可以利用 DEF 情感极性值进行修正并消歧。

2.3.2.1 词语倾向性值查询与计算

WordNet 是以词义（sense）来记录的，sense 以同一词义的词集 Synset 表示。通过查询可以得到词语 W_E 所有的 sense，将每个 sense 映射到 SentiWordNet 就可以得到对应的情感极性值。

2.3.2.2 义原倾向性值查询与计算

基于 WordNet 和 SentiWordNet 的义原倾向计算过程如图 2.4所示。在 HowNet 中获取义原后将义原对应英文词语（如“good”）映射到 WordNet 中进行查询，得到该词语所有的 Sense（如“good”的 Sense 共有 27 个）；将这些 Sense 映射到 SentiWordNet 中查询得到对应 Sense 情感极性值；将情感极性值加权根据公式 2.1计算得到义原的情感倾向值（如“good”的倾向值为 PosScore=0.597，NegScore=0.004）。

$$\varphi(s,p)=\frac{\sum_{i=1}\varphi_i(s,p)}{\sum_{p\in P}\sum_{i=1}^{m}\varphi_i(s,p)} \tag{2.1}$$

公式中 P 表示极性类型（积极、消极、中性，“P、N、O”），m 为与义原相对应的 Sense 的总数，s 表示义原，$\varphi(s,p)$ 表示义原的极性值，$\varphi_i(s,p)$ 表示义原在编号为的 Sense 中的类型极性值。

事件类义原有很多在 DEF 描述中可以引起情感极性值的变化，比如“DoNot| 不做，lose| 失去”等会引起情感极性值符号反转，因此我们标注了 819 个事件类义原的在情感极性值计算中的语义角色，并用系数 λ 来表示。

2.3.3 词语情感极性值计算

通过 2.3.2.1部分查询可以获得部分词语的情感倾向值，有些词语由于是多义的，情感极性值可能有几个，因此需要根据词语 DEF 描述中义原情感极性值进行计算修正和消歧。对 HowNet 中词语属性描述 DEF 语义关系的不同提出如下定义：

定义 1 情感倾向值取反：词语 s 的 p 极性值 $\varphi(s,p)$ 取反运算是，将 s 的积极倾向值和消极倾向值互换，过程如公式 2.2：

$$\overline{\varphi(s,p)}=\varphi(s,p),\quad (p,q)\in P\&\&p\neq q \tag{2.2}$$

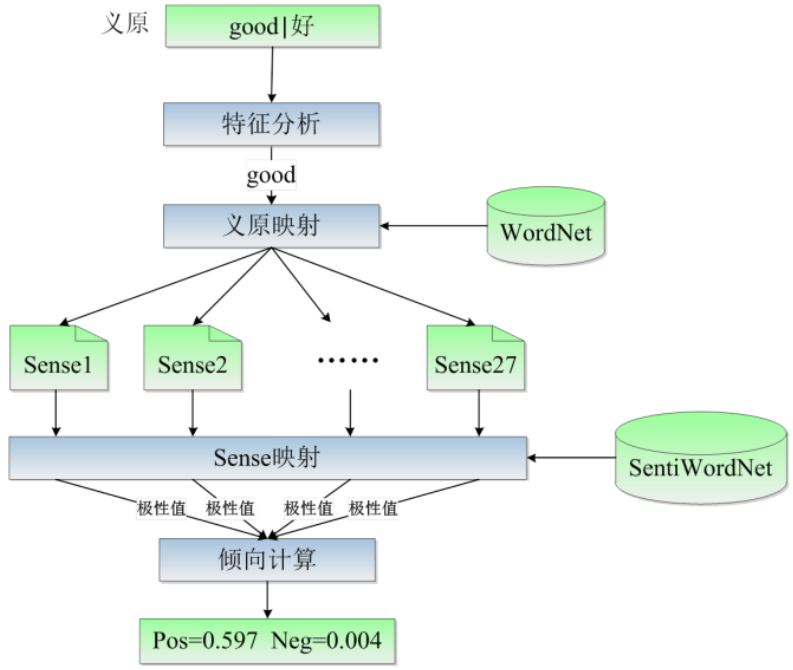


图 2.4　义原情感极性值计算过程

定义 2 因子乘法运算： λ 因子与词语 s 的 p 极性值的乘法运算定义为 λ 乘法运算，过程如公式 2.3：

$$\lambda \times \varphi(s,p) = \begin{cases} \lambda\varphi(s,p), & \lambda > 0 \\ 0, & \lambda = 0 \\ |\lambda|\varphi(s,p), & \lambda < 0 \end{cases} \tag{2.3}$$

λ 取值的确定需要根据义原的类别特征、词语 DEF 的组成特征和义原间的语义关系进行确定，这些都已经在抽取部分和义原情感极性值计算部分记录下来。如词语“好”的 DEF 中每个义原的 λ 可以均取值为 1。词语“扭亏为盈”的 DEF 为“DEF=alter| 改变，StateIni=InDebt| 亏损，StateFin=earn| 赚”，义原“InDebte| 亏损”为初始状态，“earn| 赚”为最终状态，经过分析后，义原“InDebte| 亏损”的 λ 取值为 0，义原“earn| 赚”的 λ 取值为 1。词语倾向计算总结为公式 2.4。其中 $\varphi(s,p)$ 表示词语 s 的 p 极性值，t_i 表示词语 DEF 中第 i 个义原，n 为词语 DEF 中义原总数。

$$\varphi(s,p) = \frac{\sum_{i=1}^{n} \lambda_i \times \varphi(t_i,p)}{\sum_{p\in P}\sum_{i=1}^{n} \lambda_i \times \varphi(t_i,p)} \tag{2.4}$$

其中：$\sum_{p\in P} \varphi(s,p) = 1$。

对于已经通过查询得到情感极性值的词语，可以在多个英文词义 sense 对应的情感极性值 $\varphi(s,p)$ 取最接近 DEF 分析计算得到的情感极性值 $\varphi_m in(s,p)$ 的，然后加和平均，计算公式为：

$$\Psi(s,p)=\frac{\varphi_{min}(s,p)+\varphi(s,p)}{2} \tag{2.5}$$

其中：$\varphi_{min}(s,p)=\min\{|\varphi_s(s,p)-\varphi(s,p)|\}$。

2.4 实验及结果

情感词典的实验评测有两种方法，一是与人工编辑的或者其他可靠性较高的词典进行对比评测，二是将词典应用到情感分析的其他任务上观察性能的提升，本文使用第一种方法。在实验评测时，基准词语由 HowNet 中随机选取了 2000 个词语进行人工判断，人工判断只给出褒贬两种极性。本章生成词典 SentiLex 与 HowNet 情感词典，NTUSD 情感词典以及大连理工大学的情感词汇本体词库 DLLEX 进行对比评价。

2.4.1 评价指标

评价指标采用准确率、召回率以及 F 值作为评测标准。设 a_1 表示褒义判断正确词数；a_2 表示贬义判断正确词数；b_1 表示判断为褒义的词数；b_2 表示判断为贬义词数；c_1 表示基准词典褒义词数；$c_1 2$ 表示基准词典贬义词数。准确率计算公式为：

$$P=\frac{a_1+a_2}{b_1+b_2}\times 100\%$$

召回率计算公式为：

$$R=\frac{a_1+a_2}{c_1+c_2}\times 100\%$$

F 值计算公式为：

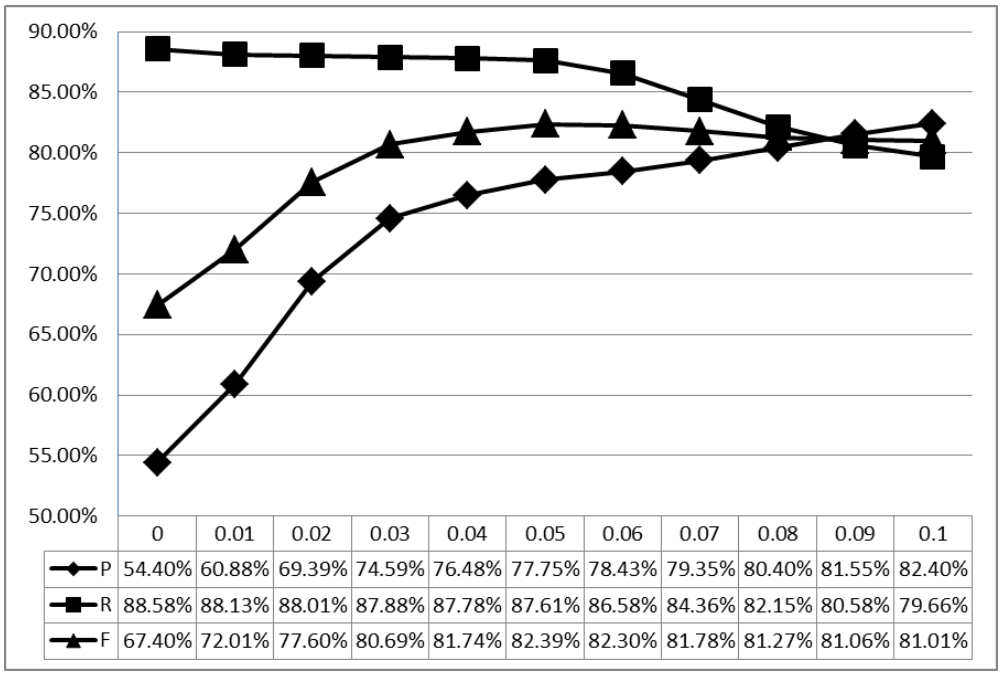
$$F=\frac{2\times P\times R}{P+R}\times 100\%$$

2.4.2 性能评测结果

2.4.2.1 阈值 T 的设定

由于基准词是褒贬二值标注的，因此需要将生成的情感词典连续情感极性值转换为离散褒贬值。将褒义和贬义情感极性值相减得到词语的倾向值来判断词语的极性，为了提高判断的准确性，设定阈值 T，高于 T 为褒义，低于 -T 为贬义。

图2.5为 T 的不同取值对词典性能指标的影响。在 T=0 时，虽然召回率最高达到 88.58%，但准确率最低仅有 54.40%，F 值仅为 67.40%。挡 T=0.05 时，准确率提高到 77.75%，有较大提高，召回率仅下降到 87.61%，下降幅度较小，F 值提高到 82.39%。当 T 提高到 0.05 时性能指标达到最好，因此可以设定 T 为 0.05。



	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
P	54.40%	60.88%	69.39%	74.59%	76.48%	77.75%	78.43%	79.35%	80.40%	81.55%	82.40%
R	88.58%	88.13%	88.01%	87.88%	87.78%	87.61%	86.58%	84.36%	82.15%	80.58%	79.66%
F	67.40%	72.01%	77.60%	80.69%	81.74%	82.39%	82.30%	81.78%	81.27%	81.06%	81.01%

图 2.5　不同 T 值时的性能指标

2.4.2.2　与其他词典性能对比

在 T=0.05 时，SentiLex 与其他词典性能比较如表 2.1所示，SentiLex 准确率为 77.75%，接近最高的 DLLEX 词典 78.40%，而召回率为 87.61%，F 值为 82.39%，均为四个词典中最高。

表 2.1　T=0.05 时的性能对比

	准确率（P）	召回率（R）	F 值
HowNet	74.55%	82.35%	78.26%
NTUSD	64.23%	80.27%	71.36%
DLLEX	**78.40%**	85.58%	81.83%
SentiLex	77.75%	**87.61%**	**82.39%**

2.5　小结

本章对中文情感词典构建相关研究进行了分析，以英文情感词典为基础，设计了基于语义关系的情感词典自动构建方法。方法以 HowNet、WordNet 语义词典

和 SentiWordNet 情感词典为基础，借鉴英文情感词典进行中文情感词典的构建，并且与现有的常用情感词典进行了实验对比。实验结果表明，本文设计的方法取得了较好的评测性能。

第三章　基于语料资源的中文情感词典扩展

3.1　引言

自然语言中，一个词语的语义极性（semantic polarity）表示它对于其语义组（semantic group）或词汇场（lexical field）范式的偏离方向[115]。在自然语言处理领域，情感分析（Sentiment Analysis）能够使用计算手段自动从自然语言中发现观点和情感等主观信息[22, 23]，通常会使用一些标注了极性（积极或消极）的词汇构成的情感词典资源。研究如何能够通过计算方式获得词语的语义极性，自动构建情感词典一直得到计算语言学和自然语言处理研究人员关注。在英文情感词典构建中，Wilson 等[60, 116] 对一些单词进行了人工极性类别的标注形成了 OpinionFinder 词典；Bradley 等[117] 标注了并发布了情感范式的英文词典 ANEW，后来 Nielsen 等[118] 在 Twitter 语言上应用并自动扩展了 ANEW，形成 AFINN 词典。Esuli 和 Sebastiani[119] 以及后来 Baccianella 等[69] 在著名的语义词典 Wordnet 基础上采用自动计算的方式开发出了情感词典 SentiWordnet。Thelwall 等[120] 设计实现了能对词语的情感强度进行估计的方法。情感也可以通过创建情绪词典来进行计算，Plutchik 情绪轮提出了四对对立的情绪状态：joy-trust，sadness-anger，surprise-fear 和 anticipation-disgust[121]。Mohammad 和 Turney[122] 根据 Plutchik 情绪轮分类方法使用情绪分值标注了一些词语形成 NRC 情绪词典。在 2013 和 2014 年举办的 SemEval（Semantic Evaluation）评测中，NRC-Canada 队利用 NRC 词典并扩展出两种新的词典，取得了最好成绩[123, 124]。为了克服以上语法层面建立的词典的上下文语境以及领域适应性问题，一些学者提出了基于概念（concept-based）构建情感词典[125]，其中 SenticNet 是使用常识知识库建立的公开可用的基于概念的情感词典[126]。

中文情感分析研究起步较晚，想对于丰富的英文情感词典资源，缺乏普遍认可的可靠的中文情感词典。目前研究使用主要有 HowNet 情感词典[110]，NTUSD 情感词典[111] 以及大连理工大学的情感词汇本体词库[112]。这些词典主要是以手工或半自动方式编辑而成。我们的前期工作提出了根据语义词典 HowNet 语义关系将英文情感词典跨语言转换为中文情感词典的方法，并构建了带有极性标示以及极性强度值的的情感词典 SentiHowNet[127]。基于语义词典的情感词典构建方法是一种常用的情感词典构建方法。采用这种方法的优势在于可以比较容易获取情感词语，基于词语的语义关系也易于进行情感极性计算。但是，基于语义词典的情感词典构建方法受限于语义词典的规模和语义关系的定义，而且对于专业领域中不断涌现的新词语，对情感词典的覆盖度提出了严峻的挑战。随着互联网应用，

尤其是社交媒体的不断涌现，越来越多的用户在各种网络平台上发布信息，网络上的用户产生内容（User Generated Content，UGC）不断涌现，研究如何利用这些丰富的网络语料对情感词典进行自动扩展具有十分重要的意义。

本章提出的基于语料资源的无监督的情感词典扩展方法，可以使用于无需标注的网络数据语料对中文情感词典进行自动扩充。

3.2 问题描述

基于语料资源的情感词语选择与极性计算，在英文中相关研究通常有两种实现思路：一是基于语言特征的方法，例如，Hatzivassiloglou[115] 等人采用并列或转折连词来判断新的情感词并计算其极性。二是基于统计特征的方法，例如，Turney 等[76] 采用点互信息统计学方法从语料中发现共现度高的情感词并计算其极性。基于以上情感词语选择与极性计算方法的分析，本文将基于中文语料资源扩展情感词典时需要解决的问题描述如下：

1. 研究根据中文语言中的并列、递进以及转折关系对情感词的发现以及极性计算的作用；
2. 根据中文特点，基于统计特征相关知识设计情感词语选择和极性计算方法；
3. 研究采用基于语言特征和统计特征相混合的方式进行情感词语选择和极性计算。

3.3 数据集及预处理

本文使用的数据资源如表 3.1所示，选取的语料资源是谭松波博士提供的酒店、书籍和电子商品评论三个领域的语料文本各 4000 篇[128]。

表 3.1 数据集及词典资源

词典	SentiHowNet	基于前期工作[127]
语料	Hotel	4000 篇
	Book	4000 篇
	NoteBook	4000 篇

其中对语料进行预处理需要将中文文本进行分词并进行词性标注。中文分词处理是对语料进行进一步处理的基础，采用的是中科院设计实现的 ICTCLAS 分词软件[129]；然后将词性标注为形容词（ADJ）和副词（ADV）的，在 SentiHowNet[127] 中出现的进行极性和极性值标注；生成的结构化语料预处理记录格式如图 3.1所

示，主要有词语编号（ID)、词性（Category)、中文词语（W_C)、词语在句子中的编号（Word_Tag)、词语所在语料文件编号（File_Tag)、词语所在句子编号(Sentence_Tag)、极性标注（Senti_Tag)、积极极性值（PosScore）和消极极性值(NegScore)。值得说明的是，极性标注的取值为 Yes 和 No，分别表示已标注和未标注，可以用于在具体的计算过程中直接进行情感词语的选择。

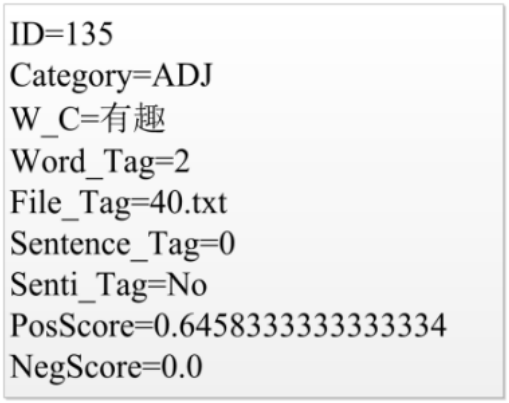


图 3.1　语料预处理记录格式

3.4　基于语言特征的情感词典扩展

早期对于英文语言特征一些研究[115] 发现，由连词（如 and 或 but）连接的两个形容词的极性往往存在一定的关联性，如“and”连接的形容词（如“nice and good”）极性相同，而“but”连接的形容词（如“nice but unnatural”）极性相反。而对于中文来说，基于语言特征的中文情感词是否会遵循想通的规律，非常值得进行研究。

3.4.1　连词选择

连词是用来连接词与词、词组与词组或句子与句子、表示某种逻辑关系的虚词。连词可以表示并列、承接、转折、因果等关系。本文主要研究基于表达并列、转折和递进三种关系的连词如何影响情感词的极性计算，选择的连词为:

- **并列关系连词：**和、跟、与、既、同、及、况、况且、乃至、并、也、又;
- **转折关系连词：**却、虽然、但是、然而、偏偏、只是、不过、至于、致、不料、岂知;
- **递进关系连词：**不但、不仅、何况、并、且、而且。

3.4.2 基于连词的极性计算

基于连词的情感词语极性计算基本思路是，待标注词语的极性值通过SentiHowNet中所有与其在同一句子的词语情感极性值计算获得，然后通过极性值判断其极性。情感极性值计算为:

$$\begin{cases} PosScore(w_t) = \dfrac{\sum_{w\in W_1} PosScore(w) + \sum_{w\in W_2} PosScore(w) + \sum_{w\in W_3} PosScore(w)}{N} \\ NegScore(w_t) = \dfrac{\sum_{w\in W_1} NegScore(w) + \sum_{w\in W_2} NegScore(w) + \sum_{w\in W_3} NegScore(w)}{N} \end{cases} \tag{3.1}$$

其中，$W_1 + W_2 + W_3 = N$，N 表示SentiHowNet与待标注词在同一个句子中情感词语，W_1，W_2 和 W_3 分别表示在SentiHowNet中与待标注词 w_t 在连接词同侧，在并列或递进连接词两侧以及在转折连接词两侧的词语。词语 w_t 极性根据积极与消极极性值大小判定为:

$$Senti\_tag(w_t) = \begin{cases} positive & \text{if } PosScore(w_t) > NegScore(w_t); \\ negative & \text{if } PosScore(w_t) > NegScore(w_t); \\ neutral & \text{if } others \end{cases} \tag{3.2}$$

具体计算过程如算法3.1所示。

3.4.3 实验

实验中用于评测的极性标注标准是基于人工标注和网络注释 (百度百科等) 等多种途径综合获得。评价指标采用正确率、召回率以及F值作为评测标准。针对三个领域的情感词典扩展实验结果如表3.2所示，对于三个语料，其召回率均达到67%以上。其中对于Hotel语料，其正确率最低，为43.69%，而其召回率最高为88.24%。其余语料正确率较高。经分析，Hotel语料中可以用于计算的连词结构的语句所占的比例小于其他语料。从平均值上可以看出，基于连接词的词语极性计算同样适用于中文。

表3.2 各个领域性能评测结果

	正确率（P）	召回率（R）	F值
Hotel语料	43.69%	88.24%	58.44%
Book语料	67.47%	67.47%	67.47%
NoteBook语料	67.21%	67.21%	67.21%
平均值	59.46%	74.31%	64.37%

算法 3.1 基于连词的极性计算

已知:

待标注词语集，$\{w_1\}$;

连词集合，$\{c\}$;

极性已知词语集合，$\{w_2\}$;

for 每一待标注词语 $w_1 \in \{w_1\}$ **do**

  for 每一与 $\{w_1\}$ 在同句子中已标注词 $w_2 \in \{w_2\}$ **do**

    if $\{w_1\}$ 和 $\{w_2\}$ 在 c 同侧 **then**

      $$\begin{cases} PosScore(w_1) += PosScore(w_2) \\ NegScore(w_1) += NegScore(w_2) \end{cases};$$

    else

      if c 为并列或递进连词 **then**

        $$\begin{cases} PosScore(w_1) += PosScore(w_2) \\ NegScore(w_1) += NegScore(w_2) \end{cases};$$

      end if

      if c 为转折连词 **then**

        $$\begin{cases} PosScore(w_1) -= PosScore(w_2) \\ NegScore(w_1) -= NegScore(w_2) \end{cases};$$

      end if

    end if

  end for

  计算极性均值 $\begin{cases} PosScore(w_1) = \dfrac{PosScore(w_1)}{N} \\ NegScore(w_1) = \dfrac{NegScore(w_2)}{N} \end{cases}$;

  根据情感值 $PosScore(w_1)$ 与 $NegScore(w_1)$ 判断极性;

  将 w_1 加入到集合 $\{w_2\}$;

end for

3.5 基于统计特征的情感词典扩展

词语的上下文是词语在实际应用中的语言环境，它在自然语言处理中的价值体现在两个方面：一方面，在自然语言知识获取的过程中，上下文是知识获取的来源；另一方面，在自然语言处理的应用问题解决过程中，上下文扮演着解决所需信息和资源提供者的重要角色。特别是在语料库语言学中，各种机器学习方法的引入使词语的上下文成为计算语言学知识获取和问题求解过程中最为重要的资源，在无监督学习方法中更是如此[130]。本文设计实现的基于统计特征的情感词典扩展方法主要是采用基于上下文的方法进行情感词语极性计算，因为出现在相似上下文环境中的词语具有相似的极性。

上下文的选取时基于核心词左右一定范围进行的，这个固定的范围被称为“窗口”。选择合适的窗口，可以使得上下文的计算提供的信息量足够大，产生的噪声足够小。在英文中，核心词左右 5 个词的范围可以为词语搭配提供 95% 的信息，上下文 ±2 是最好的选择，范围进一步扩大后提供的信息量不会有明显的增加且会带来不必要的计算开销。本章的方法首先是对待标注词语，分析其上下文词语的词性，获取其特征向量；其次，根据其上下文特征向量实现情感词语极性计算。

3.5.1 统计特征选择

定义 3-1 词语 w 的特征向量 $V(w)$ 和窗口 W: 词语 w 的特征向量 $V(w)$ 是指由词语 w 与其相邻上下文词语的词性组成的向量，具体形式为公式:

$$V(w) =< C_{-W}, C_{-W+1}, \cdots C_{-1}, C_0, C_{W-1}, C_W >$$

其中，C_0 表示词语 w 的词性，$C_i(i \neq 0)$ 表示与 w 相邻的词语的词性，i 表示与词语 w 的相对距离，W 表示窗口，即特征向量中与词语 w 相对距离的最大值。

3.5.2 基于上下文的情感词极性计算

基于上下文的情感词极性值计算根据 SentiHowNet 中具有相同的特征向量的词语的极性值进行计算, 通过极性值判断其极性。情感极性值计算为:

$$\begin{cases} PosScore(w_t) = \dfrac{\sum_{V(w)=V(w_t)} \dfrac{|\sum_{w\in W_{positive}} PosScore(w) - \sum_{w\in W_{negative}} PosScore(w)|}{M}}{N} \\ NegScore(w_t) = \dfrac{\sum_{V(w)=V(w_t)} \dfrac{|\sum_{w\in W_{positive}} NegScore(w) - \sum_{w\in W_{negative}} NegScore(w)|}{M}}{N} \end{cases} \tag{3.3}$$

其中 $W_{positive} + W_{negative} = M$，表示与待标注词 w_t 具有同一特征向量的 SentiHowNet 中的情感词，$W_{positive}$ 和 $W_{negative}$ 分别为极性为积极和消极的词语，N 为待标注词 w_t 在不同的上下文环境中的特征向量数。w_t 极性判断依据其积极与消极极性值的大小判断，同公式 3.2。具体计算过程如算法 3.2所示。

3.5.3 实验

对三个领域（Hotel、Book、NoteBook）的情感词典扩展实验结果如图 3.2、图 3.3和图 3.4所示。对于三个语料，当窗口 $W = 1$ 时，准确率最高，分别为 67.65%、72.89% 和 72.13%；当窗口 $W = 2$ 时，召回率有所上升，准确率略有下降；当窗口 $W = 3$ 时，召回率最高，准确率和 F 值下降较多。通过对评测结果进行分析，本文发现在设计基于统计特征的情感词典扩展方法时，采用窗口 $W = 1$ 进行情感词语选择，采用窗口 $W = 12$ 进行情感词语极性计算，可以获得较好的性能。

算法 3.2 基于统计特征的极性计算

已知:

待标注词语集，$\{w_1\}$;

极性已知词语集合，$\{w_2\}$;

每个词特征向量集合，$\{V(w)|w \in \{w_1\} \cup \{w_2\}\}$;

for 每一待标注词语 $w_1 \in \{w_1\}$ **do**

  for w_1 每一特征向量 $V(w_1)$ **do**

    for 每一与 $V(w_1)$ 相同的特征向量 $\{V(w_1) = V(w_2)|w_2 \in \{w_2\}\}$ **do**

      if $SentiTag(w_2) = positive$ **then**

$$\begin{cases} PosScore(w_1)+ = PosScore(w_2) \\ NegScore(w_1)+ = NegScore(w_2) \end{cases};$$

      else

$$\begin{cases} PosScore(w_1)- = PosScore(w_2) \\ NegScore(w_1)- = NegScore(w_2) \end{cases};$$

      end if

    end for

    对各个特征向量下的情感值累加

$$\begin{cases} PosScore(w_1)+ = PosScore(w_1) \\ NegScore(w_1)+ = NegScore(w_1) \end{cases};$$

  end for

  计算极性均值 $\begin{cases} PosScore(w_1) = \dfrac{PosScore(w_1)}{i} \\ NegScore(w_1) = \dfrac{NegScore(w_2)}{i} \end{cases}$;

  根据情感值 $PosScore(w_1)$ 与 $NegScore(w_1)$ 判断极性;

  将 w_1 加入到集合 $\{w_2\}$;

end for

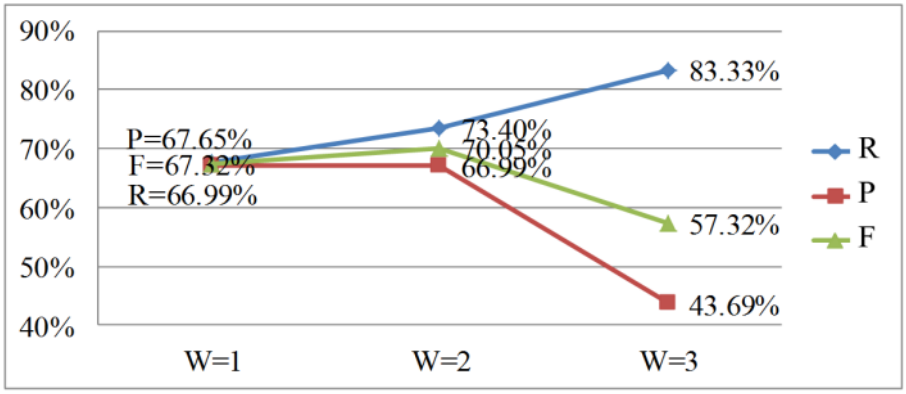


图 3.2　Hotel 语料评测结果

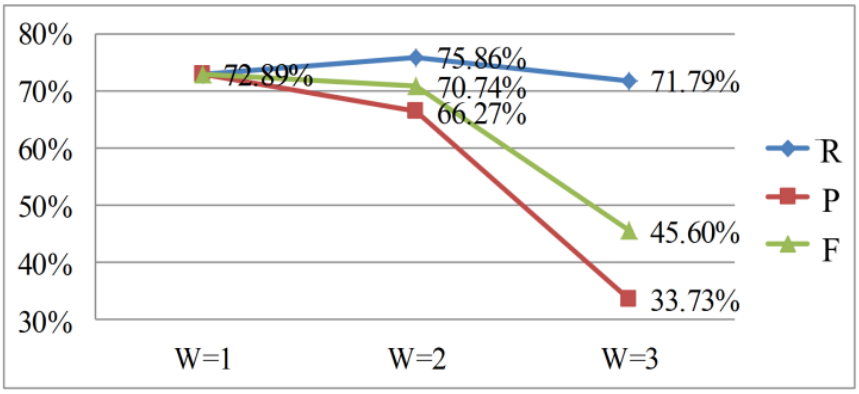


图 3.3　Book 语料评测结果

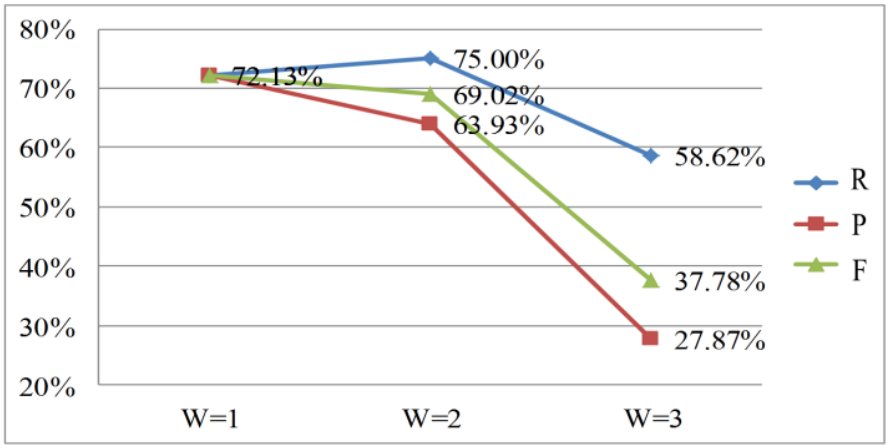


图 3.4　NoteBook 语料评测结果

3.6　基于混合特征的情感词典扩展

对基于语言特征的情感词典扩展方法和基于统计特征的情感词典扩展方法的实验结果进行仔细分析发现，采用语言特征无法进行情感极性计算的词语，可以采用统计特征进行处理；同样的，采用统计特征无法进行情感极性计算的词语，可以采用语言特征进行处理；两种方法可以相互补充。因此本文提出基于混合特征方法。

3.6.1　基于混合特征的情感词极性计算

基于混合特征的情感词语极性计算如算法 3.3。将选取的情感词语集合分别采用两种方法进行极性计算，在将两种方法计算的极性值合成时，遵循以下原则：

1. 优先采用基于统计特征的方法计算出的情感极性值作为待标注词语的情感极性值。

2. 当采用基于统计特征的方法进行计算时，优先设置窗口大小为 2，其次为 1。

3. 当采用基于统计特征的方法无法对待评价词语进行情感计算时，采用基于语言特征的方法进行计算。

算法 3.3 基于混合特征的极性计算

已知:

待标注词语集，$\{w_1\}$;
极性已知词语集合，$\{w_2\}$;
连词集合，$\{c\}$;
每个词特征向量集合，$\{V(w)|w \in \{w_1\} \cup \{w_2\}\}$;

for 每一待标注词语 $w_1 \in \{w_1\}$ **do**
    依据算法 3.2计算情感极性值
    if $\begin{cases} PosScore(w_1) = 0 \\ NegScore(w_1) = 0 \end{cases}$ **then**
        依据算法 3.1计算情感极性值
    end if
    根据情感值 $PosScore(w_1)$ 与 $NegScore(w_1)$ 判断极性;
    将 w_1 加入到集合 $\{w_2\}$;
end for

3.6.2 实验

对三个领域（Hotel、Book、NoteBook）的情感词典扩展实验结果如表 3.3所示。

表 3.3 各个领域性能评测结果

	正确率（P）	召回率（R）	F 值
Hotel 语料	75.49%	74.76%	75.12%
Book 语料	77.11%	77.11%	77.11%
NoteBook 语料	78.69%	78.69%	78.69%

基于语言特征的情感词典扩展、基于统计特征的情感词典扩展和基于混合特征的情感词典扩展的实验评测结果对比情况如图 3.5、图 6 3.6和图 3.7所示，通过分析发现，基于混合特征的情感词典扩展方法的评测性能是在各个领域语料中均是最优的。

3.7 小结

本章详细讨论了基于语料资源的中文情感词典扩展问题描述和方法设计，对基于语言特征的情感词典扩展和基于统计特征的情感词典扩展的关键技术分别进

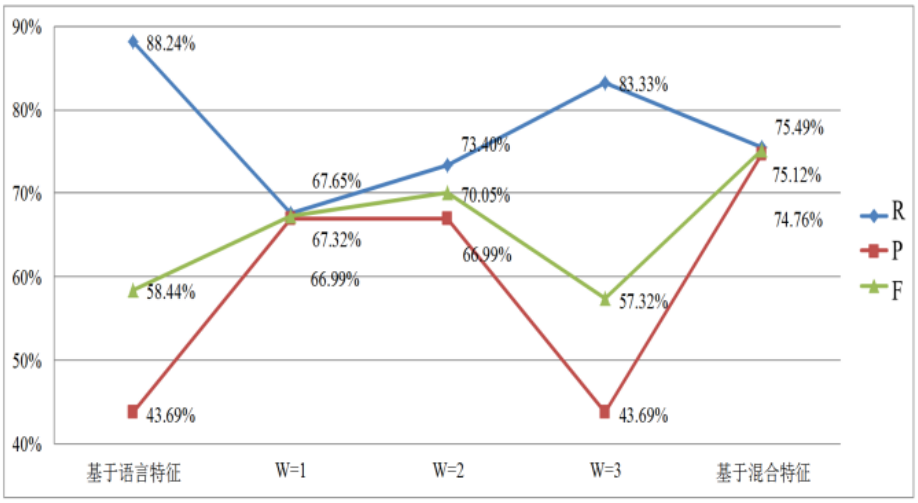


图 3.5　Hotel 语料评测结果综合比较

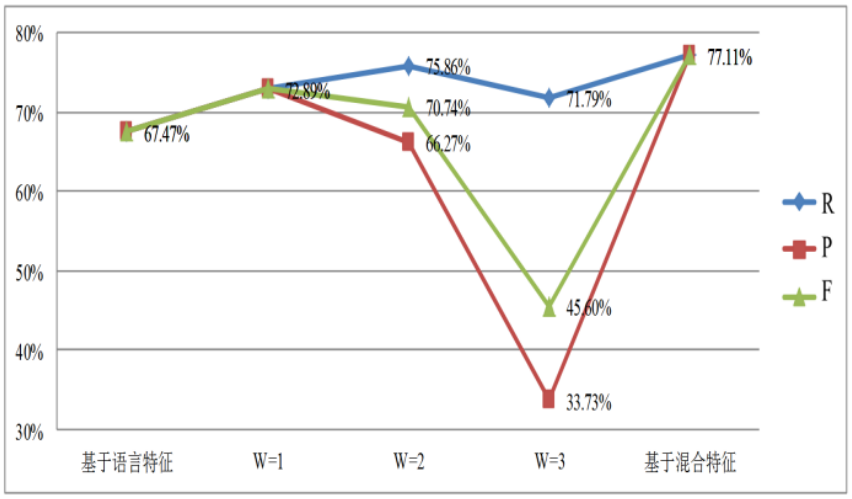


图 3.6　Book 语料评测结果综合比较

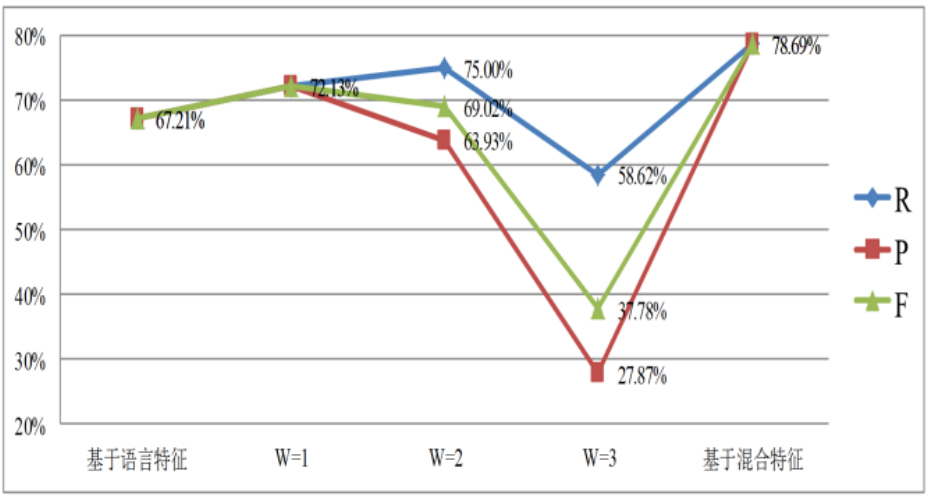


图 3.7　NoteBook 语料评测结果综合比较

行了研究和算法实现，并提出了了基于混合特征的无监督的情感词典扩展方法。通过分析每个方法的实验结果，发现基于混合特征方法能够达到最好性能。

第四章　无监督的自举式情感分类

4.1　引言

文本的情感分析是挖掘文本中主观性信息学的主要手段，是研究文本中观点、态度、情绪和立场等主观性信息是如何表达的。情感分析技术可以从数量庞大的文本数据中抽取并总结主观性信息，为后续的一些应用（商业智能（Business Intelligence），舆情分析（Public Opinion Analysis）或选举预测（Election Prediction）等）提供技术和工具支持[23]。在社交媒体中，一些基于文本的平台比如微博产生了大量针对各种话题或实体的带有主观性信息的数据。这对这些数据的分析，也就是情感分析正逐渐受到各个研究领域（比如推荐系统和搜索引擎）的重视。情感分析中一个主要的方法就是应用各种分类技术，也就是根据作者的主观态度将文本进行分类，一般将这种研究称为情感分类研究。情感分类一般是从一些标注过的训练数据中通过学习得到一个分类模型，学习得到能够将一种情感类型区别于其他类型的一些特征[131]。这种模型的性能主要依赖于其能够学习到的数据中出现的情感的文本表达模式，一般是文本中出现的词语、短语或者词语的各种组合。情感分类也就是情感分析的分类形式，虽然可以被视作文本分类的特殊形式，实际上情感分类是比文本分类更具挑战的任务，因为文本中情感的表达方式严重依赖领域和上下文环境[145]。

随着微博（Twitter、新浪微博等）的出现和广泛使用，用户产生内容（UGC，User-Generated Content）成指数增长，这些内容并且这些内容对于我们来说是很容易获取的，并且这些内容里面有很多用户对于各种话题的观点和情感等主观性信息。因此我们可以很方便的从这些数据中提取出主观性信息，并使其在商业、旅游或者健康领域得到应用。但是对微博进行情感分类特别具有挑战性，因为（1）用户使用微博表达观点的方式是多种多样的，既有正规传统语言的表达方式，又有社交媒体特有的流行的表达方式，比如“cooooooool、OMG、:-(、屌丝、逆袭”等，这些表达方式虽然对于人来说是比较直观和易于理解的，并且更加方便了用户的在线交流，但是对于计算机来说，却是很难准确确定这些表达方式的观点和情感等语义信息。（2）更具挑战性的是，因为用户群体的复杂性，经常会有用户创造出的一些缩写词或者新词，并且会将一些传统的词赋予新的语义在微博中重新使用，这些语言上的变化使得微博上观点的表达方式有别于传统文本的表达方式。综上所述，可以看出微博中的文本噪声、非正式本质以及语言词汇的急剧膨胀使得对微博中表达的主观性信息自动分析需要依赖于微博这种独特的语言环境，因此进行情感分类是困难的。这种情况被称为微博情感分类的领域（或语言环境）

依赖问题，也就是使用其他文本数据集（比如评论或博客）训练出的分类器在微博的情感分类时会出现性能急剧下降，而要获得大量微博训练数据集需要大量的人力，并且微博数据具有时效性，不同时间阶段的微博数据集中观点表达方式也会产生漂移。

本章我们主要关注微博情感分类的领域依赖性问题。为了解决这个问题，基于我们的一些观察，我们提出了一种无监督的自举式（bootstrapping）情感分类框架。该框架首先使用现有的已经有情感标签的语言资源训练得到一个通用的能够跨领域使用的分类器；然后再根据该分类器的跨领域特点使用其作为初始分类器对微博进行分类，获得一些高可信度的微博作为训练集训练得到一个微博分类器；将两个分类器结合迭代使用共同训练（Co-training）过程，逐步在目标数据集扩展并训练微博分类器，直至其分类性能达到最优。

4.2 相关工作

情感分类在观点挖掘研究中越来越受到重视，前期工作主要研究针对评论(商品或电影）进行情感分类。经常使用的方法可以分为基于规则的方法和基于学习的方法，其中基于机器学习方法性能一般比较好因而常被用来作为对表的标准[61]。

现在研究人员逐渐开始注意到微博中用户的主观性信息，并开始结合微博的语言特点进行对微博进行情感分类研究。一些研究显示可以将微博的一些独特的特征结合进情感分类方法中。比如，Barbosa 和 Feng [132] 提出了两阶段支持向量机分类器（Support Vector Machine (SVM) classifier）对 tweet 进行情感分类，证明了该分类其能对 tweet 的类别偏置（biased）和噪声具有很好的鲁棒性；Hu 等 [133] 将社交媒体数据中的情感表达解成情感指征（emotion indication）和情感关联（emotion correlation）两种信号，通过对两类情感信息进行联合建模方式实现了对微博的无监督情感分类；Jiang 等 [145] 主要关注依赖于特定目标的微博情感分类，提出了通过将目标依赖特征（target-dependent features）和相关微博同时进行考虑的监督学习方法，并证明了可以提升情感分类性能；Wang 等 [134] 针对 hashtag 级别的情感分类进行了研究，并提出了一个全新的图模型，然后使用提升（boosting）式分类方法进一步提高了模型的性能；Amir 等 [135] 针对单条微博的情感分类提出了一个分层分类器框架，框架通过抽取对特定目标的微博，将微博按情感类型分开以及分离正负情感类型微博三个层次进行有监督的分类学习；Hu 等 [136] 基于社交理论抽取微博之间的情感关系，提出了一种全新的社会学方法使用这些情感关系以促进情感分类性能，并有效解决了数据中的噪声问题；同样受到社会学理论的启发，Guerra 等 [137] 依据人类通常会持有一致的带有偏执的观点，提出了全新的迁

移学习（transfer learning）方法解决微博基于话题的实时情感分类问题；Thelwall 等 [120, 138] 设计了 SentiStrength 情感分析工具，用于对微博等社交媒体中非正式语言中的情感分析，该工具是基于规则的方法，使用了人工编辑的词典并结合了微博语言中的句法和拼写特点抽取微博中的情感强度，该工具获得了广泛的应用。

以上这些工作通过利用微博的一些网络和语言特点对情感分类方法进行了适应性的改进，以使得这些方法能够适用于微博语言环境，但是没有彻底解决微博情感分类问题的语言环境依赖问题，本章我们提出的方法从一个全新的视角来看情感分类问题，将情感分类的特征空间分各位环境依赖部分（context-dependent part）和环境独立部分（context-independent part）分别进行训练分类器，然后将两种分类器结合进一个自举式（bootstrapping）学习框架中。

4.3 问题的形式化

简单来说，情感分类主要目标就是将文本分类为预先定义的情感极性类别(一般是积极的，positive 或消极的，negative)。形式化上，对于给定的文档语料库 $D = \{d_1, \ldots, d_n\}$，预定义的情感类别 $Y = \{1, -1 \mid \text{positive} = 1, \text{negative} = -1\}$，情感分类的任务就是对每一个文档 d_i 预测一个类别标签 y_i。为了与文本分类问题一致，每个文档可以表示为一个特征向量 $x = R^n$，n 表示特征空间的大小对于情感分类问题来说，对于每一个特征通常将其权重定为二值的，1 表示特征在文档中出现，0 表示没有出现[61]。对于有监督的机器学习，给定训练集 $D = \{x_1, \ldots, x_m\}$，可以学习到分类器：

$$f : D \longrightarrow Y, Y = \{1, -1\} \ . \tag{4.1}$$

对于未来文档 x，同样将其表示为特征向量 $x = (w_1, \ldots, w_v)$（w_i 表示第 i 维权重），就可以使用该分类器去预测其情感类别：$f(x)$。

在以往的情感分类研究中，有一个潜在的假设，就是用于表示文本的特征向量中所有的特征（一般是词语）在表达情感极性时作用是相同的，也就是其出现与否可以在所有的文本中表达相同的情感。实际上这种假设是不成了立的，因为有些词语表达的是客观信息，有些表达主观信息，而且即便是表达主观信息，作用也都不一样。因为有些词语无论用在那种领域或语境下都能表达同样的情感，而有些词语只能在某些具体的语境下表达某种情感。以下面这条微博为例：

tweeet: @Kid_Cloudz: Happy birthday to Yessicaaaa! :D lovee you feggit wish you the best day everrrrr!!!!! @030268.

以词袋模型（bag-of-words）为例，所有的词语都应该抽取出来作为特征加入到特征向量中同等地用于对这条微博的情感倾向进行建模。然而，仔细观察就

会发现，微博中有些词语（@Kid_Cloudz, :D, lovee, everrrrr,!!!!!）实际上只能在微博这种语境中出现并且表达出某种情感倾向，而另外一些词语（Happy, birthday, wish, best, thanks）无论在什么领域或语境下都是正面情感倾向的的标识。基于这样的直观认识，我们可以提出以下特征空间划分的假设：

定理 4.1 (**假设**): 特征空间划分假设：对于微博情感分类问题的特征向量空间，可以将其所有的特征划分为以下两个部分：

- 领域独立部分：也就是通用的特征，该部分特征在任何领域和语言环境下都是某种情感倾向的表达方式。
- 领域依赖部分：也就是具体语言特征，这部分特征只有在微博这种语言环境下才能有具体的语义和表达一定的情感倾向。

这个假设可以更加形式化的表示，对于情感分类问题中一条微博的特征向量 $x = (w_1, \ldots, w_l, w_{l+1}, \ldots, w_v)$，可以划分为两个部分：

$$x = \begin{cases} x_g & : \text{general features} \\ x_c & : \text{context features} \end{cases} . \tag{4.2}$$

其中，$x_g = (w_1, \ldots, w_l)$ 是特征向量空间的通用部分，而 $x_c = (w_{l+1}, \ldots, w_v)$ 是领域依赖部分。

基于以上假设，情感分类问题可以进一步形式化定义为：

定义 (**情感分类**): 根据假设（1），情感分类问题可以表示为 (X_g, X_c, Y)，其中：

- $X_g \subset R^d$ 和 $X_c \subset R^p$ 为两个输入特征空间，$d + p = n$，分别表示两部分空间的维度;
- Y 为输出空间，一般表示为二值空间 $Y = \{1, -1 \mid \text{positive} = 1, \text{negative} = -1\}$;
- 假设有一独立同分布（independently identically distributed）微博实例集合 $D = \{(x_i^g, x_i^c, y_i); i = 1 \ldots m\}$，该集合是从空间 $P = X_g \times X_c \times Y$ 中采样得到，向量 x_i^g 表示实例领域独立部分特征，向量 x_i^c 表示领域依赖部分特征，y 表示实例微博的情感类别；

实际上经过特征空间的划分提供了对于同一微博的两种不同的视角（view），因此可以将数据集 D 看作是 $D_g = \{(x_i^g, y_i); i = 1 \ldots m\} \in (X_g \times Y)^m$ 和 $D_c = \{(x_i^c, y_i); i =$

$1 \cdots m\} \in (X_c \times Y)^m$ 两种不同的集合，因此对于集合 D 的情感分类问题可以视为构建两个分类器通用情感分类器（General Sentiment Classifier）和微博情感分类器（Context Sentiment Classifier）：

$$\begin{cases} GeneralSentimentClassifier: f_g: X_g \mapsto Y \\ ContextSentimentClassifier: f_c: X_C \mapsto Y \end{cases} \tag{4.3}$$

■

当然基于部分特征空间的分类器性能上是否会降低还是一个值得研究的问题，但是本章我们主要研究以下几个问题：

1. 对于从实例中抽取到的同一个特征空间，怎么确定特征空间中领域依赖和领域独立两部分特征？
2. 得到不同的特征空间后，使用什么样的训练数据集来训练得到两个不同的分类器？
3. 两种独立的分类器比同一空间分类器性能上会有什么样的变化，如何将两种分类器结合起来达到更好的性能？

4.4 无监督的情感分类框架

在微博语言中，除了正规的表达方式方式外，一些语言因为比较难以理解而常被视为“火星文”，尤其是对于不长使用微博的人来说对于一条微博中出现的一词语可能不理解其语义。但是整条微博的情感倾向性确能够比较容易读懂，因为微博常常是正规表达方式和“火星文”混合在一起使用的，理解了正规表达部分，也就能理解了整条微博的情感倾向。直观上，这种现象可以通过我们的特征空间分割假设来解释，正规表达部分特征也能从一个不同的视角（view）来阐释整条微博的主观情感。而这些正规表达部分特征 x_g 是不以来于微博语境的，对于任何人（长使用微博的或是很少使用微博的）都是易于理解的。

相似的，对于微博的自动情感分类，基于我们特征空间分割假设，可以认为一条微博的情感倾向性可以通过两部分特征都识别出来。也就是说，如果能够对与一些通用的情感表达知识，在某种程度上也能识别出一条微博的情感极性（根据微博中正规表达方式的比例不同，比例越大就越容易识别）。实际上有很多研究者已经开始研究如何建立各种情感词汇表来对这种通用的情感知识进行建模了，比如我们前面章节的工作中提到的 OpinionFinder 词典[60, 116]、ANEW 词典[117]、AFINN 词典[118]、SentiWordnet[119]、HowNet 情感词典[110]，NTUSD 情感词典[111]、情感词汇本体词库[112] 以及我们的 SentiHowNet[127]。虽然这些词典在尝

试着建立通用的情感表达知识库，但是由于存在一词多义现象，使得一个词语的具体情感倾向性还是需要具体的语言上下文进行“消歧”。因此能够真正找到通用的资源来对跨领域情感知识进行建模不是一件容易的事。但是这样的知识资源却是存在的，比如成语和谚语等具有明确无歧义的情感倾向性，如何能够利用这样的知识资源对通用情感知识进行建模是本章研究的重点。

4.4.1 通用情感分类器

在语言资源中有许多对情感分类研究非常有用的资源，其中成语资源就是其中之一。成语（或谚语，本章中用成语通指这两种语言资源）无论在中文还是英文中都存在，比如中文的“空中楼阁”、英文的 “bring down the house”（搏得满堂喝彩）等。这些成语的情感倾向性是固定不变的，不会随着领域或语境的不同而有歧义。这与我们的通用情感分类器需求十分契合，实际上有很多的专门针对成语编辑的词典资源，为通用情感分类器提供了很好的数据集进行训练。一般的成语词典的条目如下所示：

空中楼阁：贬义词，形容虚构的事物或不现实的理论、方案，脱离实际的理论、计划及空想。

在“空中楼阁:”条目中，有三部分组成：成语本身、情感倾向性（贬义，属消极情感）以及该成语的释义部分。其中释义部分有几个明显表示贬义的词语（虚构的、不现实、脱离实际以及空想)。该词条可以看作是给我们提供了一条带有通用情感知识的标注数据，释义中的词语可以看作通用部分特征 $\{x_i^g\}$, 情感标签 y_i 就是成语的情感标签。由于成语的情感倾向是不依赖于任何领域和语境的，因此我们可以认为存在如下假设：

定义 (假设): 每条成语条目可以看作是一条不依赖于任何领域的情感标注数据。

在假设 4.2基础上，我们可以根据现存的成语词典构建一个训练数据集用于训练通用情感分类器 f_g，该分类器用于对通用情感知识的建模。

4.4.2 微博情感分类器

由于通用特征只是全特征空间的一部分，在识别情感倾向时仅代表跨领域或语境的情感表达方式。在微博这种特殊的语言环境下，情感的表达通常有其独特的方式，比如表情符、简写、以及故意不规范的拼写等等。为了能够更好的识别出微博中的细腻的情感倾向，必须要考虑微博中领域依赖部分的特征。

为了对微博情感特征的领域依赖部分进行建模，有两个问题必须考虑。首先是如何界定微博中抽取的特征中那些是领域依赖的特征。随着用户发布微博数量的急剧膨胀以及用户在语言使用上的自主性，一些微博特有的主管观点或情感的新的表达方式不断出现，并且同样使用一些通用词语，在特定的语境下也会出现不同于其固有的语义信息。不断涌现的新词和词语在为上的语义变异使得界定领域依赖部分特征变得非常困难。但是众所周知，微博文本属于短文本，每条微博都有字数上的限制（一般是要求 140 字以内），因此用户在一条微博中表达就某件事情表达某种情感倾向时，除了描述事情所用词语外，只能够用很少的词语描述情感倾向。因此我们可以假设，如果一条微博中含有某个成语或谚语，如果没有否定词，整条微博的情感倾向可以看作是和成语或谚语的情感倾向一致，并且除成语外的其他词语形成的情感特征可以视为领域依赖特征。第二个问题是如何找到足够的标注数据来训练得到依赖于微博特征的微博情感分类器。之前有些研究提出了远监督（distant supervision）方法来解决微博标注数据缺失的问题[139, 140]，主要是基于微博中含有明显的情感倾向的一些表达方式（表如表情符）为基准来发现一些含有噪声的微博作为训练数据。我们也是利用这样的思想，但是我们利用成语资源作为我们的情感表达基准，找到包含成语的微博（过滤掉含有否定词的部分）作为微博依赖的情感分类器的训练数据。

4.4.3 分类器的组合

我们的一个基本假设就是认为用户在表达一种主观情感时可能会使用不同的表达方式，一是可以使用通用的情感词语，另外也可以使用微博特有的一些表达方式，更有可能会混合使用通用词语和微博上流行的特有的表达。因此我们可以将其情感表达的特征空间划分为通用特征和领域依赖特征，主要目的是将相同的信息从相互补充的两种视角（view）来分析，以训练不同的分类器达到更好的效果。虽然理论上两个分类器都能对微博达到情感分类的效果，但是性能会受到训练数据的数量和质量的制约。很明显的，无论是成语的释义还是微博文本，都是比较短小的，而且微博常常会有噪声，因此从这样的数据抽取的特征向量会比较稀疏，对分类器的性能造成影响。为了克服这些困难，我们提出了一个自举式（bootstrapping）的学习框架将两种分类器组合在一起，相互补充，通过优化，发挥二者的最大效能。框架如图 4.1所示。

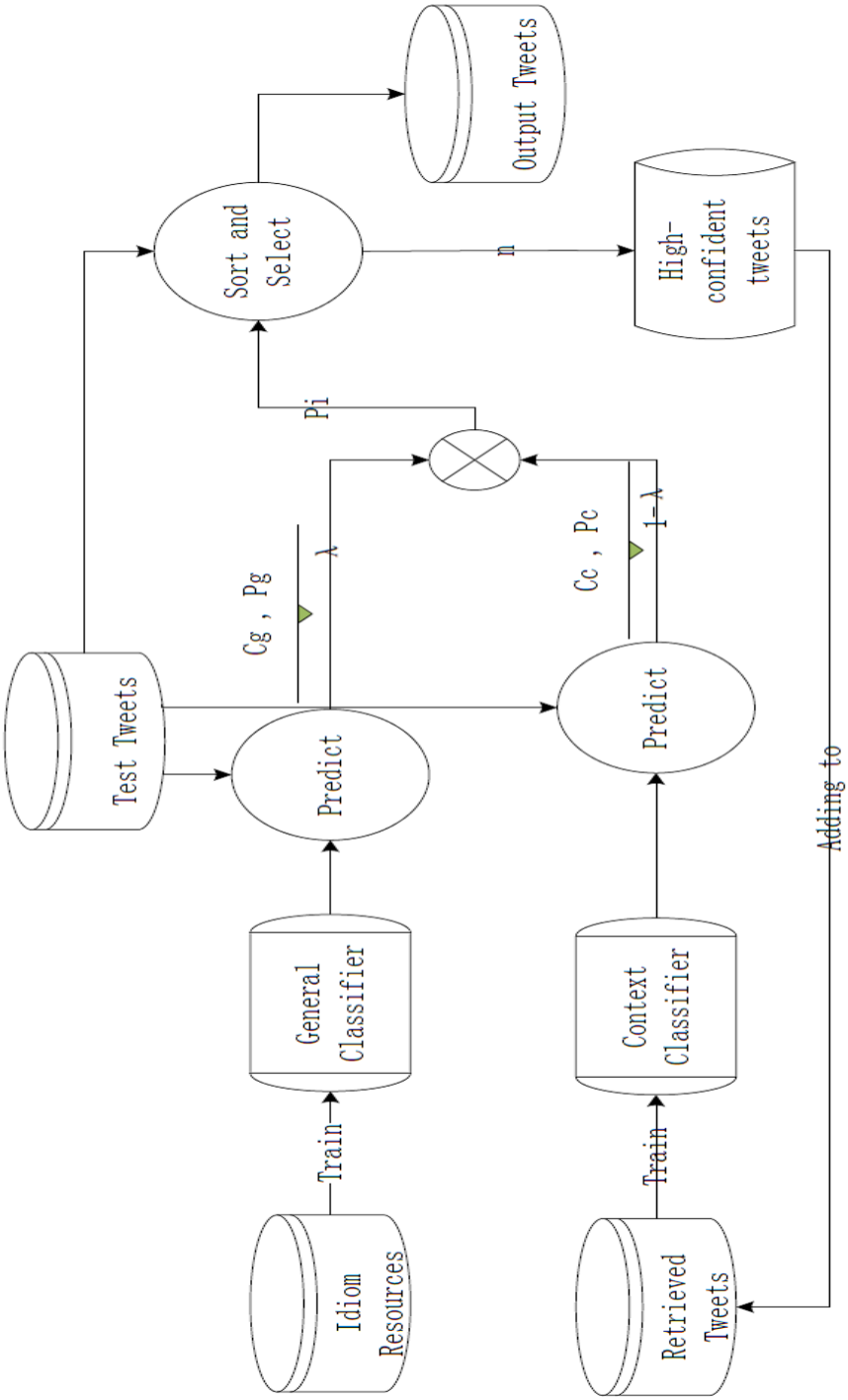

Output Tweets
Sort and Select
n
High-confident tweets
Pi
Cg , Pg
λ
Cc , Pc
1-λ
Test Tweets
Predict
Predict
General Classifier
Context Classifier
Train
Train
Idiom Resources
Retrieved Tweets
Adding to

图 4.1　自举式学习框架

该框架中，我们要通过不断迭代训练学习到通用情感分类器 P_g 和微博情感分类器 P_c，使得这两个分类器不但从单个视角达到分类性能的最优，也需要在对相同的测试数据上分类结果一致，组合性能能够提高。根据使用的分类器的不同，我们假设可以将分类器的输出用 $\{p_g, p_c\}$ 来表示（例如 SVM 的距离输出或生成模型的概率输出）来表示分类结果的可信度。对于每一个待分类测试数据，首先使用两个情感分类器对其进行分类，预测其情感倾向性标签为 $c_i = \{c_g, c_c\}$，并输出可信度 $p_i = \{p_g, p_c\}$，然后将可信度按照公式～4.4合成为一个：

$$p_i = \begin{cases} \lambda * p_g + (1-\lambda) * p_c & \text{if} \quad c_g = c_c; \\ 0 \qquad if \quad c_g \neq c_c; \end{cases} . \tag{4.4}$$

其中 λ 是控制不同部分特征影响权重的系数，首先将其初始化为 $\lambda = 0.5$，然后随着迭代的进行逐步增加 λ 以使得组合起来的分类器逐步适应针对微博的情感分类。根据两个分类器对每个测试数据预测情感标签 c_i ($c_i \in \{1, -1\}$)，将测试数据分为两组，并分别按照预测输出的的组合可信度的降序排列。在排序的两组数据中分别取其前 n 条可信度最高的微博数据作为新的依赖于微博语境的微博情感分类器的训练数据加入到训练集中，以逐步提高该分类器对微博情感分类的适应性。这样的过程循环多次进行迭代，直至所有数据的情感分类组合可信度的变化因为小于某个指标而收敛。

总体来说，整个框架可以被视作是一个自举式（bootstrapping）共同训练（Co-training）[140] 机器学习算法过程，所不同的是该框架并没有使用标注好的训练数据，而是从现成的成语词典资源作为训练的起始点，是一个无监督的学习框架，因此节省了人工或自动标注微博数据的过程，对于数量庞大的微博数据来说，该框架更加实用。

4.4.4 分类器算法

对于两个分类器，我们采用跟 Pang 等 [61] 文章中一样的三种机器学习算法：朴素贝页斯（Naïve Bayes）算法，最大熵（Maximum Entropy）算法以及支持向量机（Support Vector Machine）算法。这三种算法的有效性已经得到 Pang 等 [61] 的验证，其中支持向量机取得的性能是最好的（准确率达到 82.9%）。

4.4.4.1 Naïve Bayes 分类器.

Naïve Bayes 在文本分类任务中是最常用的分类器。对与情感分类问题，为了确定一篇文档 d_i 的情感倾向性类别 c_j，需要计算后验概率 $P(c_j \mid d_i)$。根据贝页斯法则和多项式分布，基于每一维特征概率的独立性假设，可以得到：

$$P\left(c_j \mid d_i\right) = \frac{P\left(c_j\right)\prod_{k=1}^{|d_i|} P\left(w_{d_i,k} \mid c_j\right)}{\sum_{r=1}^{|C|} P\left(c_r\right)\prod_{k=1}^{|d_i|} P\left(w_{d_i,k} \mid c_r\right)} \quad . \tag{4.5}$$

通过计算每一情感类别的后验概率，概率最大的可以视为文档 d_i 的情感类别。

4.4.4.2 最大熵分类器.

最大熵（Maximum Entropy）分类器与 Naïve Bayes 分类器一样也是通过计算后眼概率来判断文档的情感类别，所不同的是最大熵分类器是计算条件概率:

$$P\left(c_j \mid d_i, \vec{\theta}\right) = \frac{1}{Z} exp\left(\vec{\theta}, \vec{f}\left(d_i, c_j\right)\right) \quad . \tag{4.6}$$

其中 $\vec{\theta}$ 表示特征向量，$\vec{f}\left(d_i, c_j\right)$ 表示将训练实例 $\left(d_i, c_j\right)$ 映射到特征向量空间的特征函数，Z 是归一化因子。最大熵分类器用训练数据集 D 的训练学习过程就是解一个最优化问题:

$$\vec{\theta^*} = argmax_{\vec{\theta}} \prod_{i=1}^{|D|} P\left(c_j \mid d_i, \vec{\theta}\right) \quad . \tag{4.7}$$

4.4.4.3 支持向量机分类器.

支持向量机（Support Vector Machines）分类器是一种判别式的机器学习方法。支持向量机分类器的训练过程发现一个支持向量确定的决策平面将在训练数据能够分为两类，然后使用支持向量确定测试数据的类别。训练过程是接一个受限的最优化问题:

$$\begin{aligned} \vec{\alpha^*} = argmin\left(-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha j x_i x_j < \vec{x_i}, \vec{x_j} >\right) \\ \text{Subject to: } \sum_{i=1}^{n}\alpha_i y_i = 0, 0 \leqslant \alpha_i \leqslant 1 \end{aligned} \quad . \tag{4.8}$$

情感分类问题通常使用线性支持向量机分类器。

4.5 实验

为了验证所提出框架的性能，我们使用一部现成的在线成语词典和从腾讯微博[1]中抓取的数据进行了一系列的实验。

[1] http://t.qq.com/

4.5.1 实验描述

4.5.1.1 数据集

我们从成语覆盖比较全的中国教育在线网[2]上抓取了在线的成语词典，经过整理的到了有 8,160 个条目的成语词典，其中褒义的（正面情感倾向）的成语有 3,648 条，贬义的（负面情感倾向）的成语有 4,512 条，我们使用这些数据训练通用情感分类器。微博情感分类器的训练数据是通过腾讯微博公开 API 抓取的数据，从 2013 年四月 15 日开始到五月 15 日一个月的时间我们监控腾讯微博的实时数据流，查询抓取了至少含有一条成语的微博数据，形成 120,346 条微博数据的数据集。经过筛选过滤掉噪声和过短的数据，最后得到 91,268 条微博数据集用于训练微博情感分类器。为了测试我们所提出的两种分类器组合形成的自举式分类器的性能，我们使用了中国计算机学会（CCF）举办的第一届自然语言处理与中文计算会议（Natural Language Processing and Chinese Computing）中的微博情感分析与语义关系抽取评测（the First Chinese tweet Sentiment Analysis and Semantic Relationship Extraction Evaluation）[3]的标注数据集作为测试数据。

4.5.1.2 实验配置

为了能够多角度测量分类器的性能，有各种评测指标，但是我们的实验不是为了比较这些评价指标的不同，因此我们选择了简单直观的准确率作为分类器性能的评价指标。对于分类器，我们选择了自然语言处理的工具 NLTK（Natural Language ToolKits）[141] 中的 Naïve Bayes 分类器和最大熵分类器，以及常用的 Libsvm[142] 工具包的支持向量机分类器。分类中所有的参数设置都经过交叉验证进行了优化。

4.5.1.3 评价基准

为了客观评价我们的方法的性能，我们设置了三个评价基准用于对比评价。一个是 50% 的基准，因为我们所用的测试数据集是平衡数据集，所以即便是随机判断的分类器的准确率可以达到这样的准确率；第二个是用一个基于情感词典的情感分类器的准确率作为基准，我们使用的是前面两章构建的 SentiHowNet 情感词典，通过计算每条微博中的包含的情感词语情感值叠加来计算综合情感值，然后判断微博的情感倾向性；第三个基准是有监督的机器学习方法构建的情感分类器，我们按照 Pang 等 [61] 文章中的方法使用测试数据通过 5 倍交叉验证方式训练了朴素贝页斯（Naïve Bayes）、最大熵（Maximum Entropy）以及支持向量机（Support Vector Machine）分类器，把三种分类器的在测试数据集上的准确性作为基准。

[2]China Education Network：http://chengyu.teachercn.com

[3]http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html

4.5.1.4 数据预处理

中文文本信息不像英文那样靠空格自然形成了词语结构，因此需要对中文进行分词预处理才能进行词袋特征的抽取。我们使用常用的中科院 ICTCLAS[4]分词软件上述所有数据进行分词处理，并进行了停用词过滤。

4.5.2 实验结果

为了确定公式 4.4中的 λ 值，我们从 0 到 1 对 λ 值进行了遍历实验，每一遍历步 λ 值增加 0.1，实验结果如图 4.2所示。

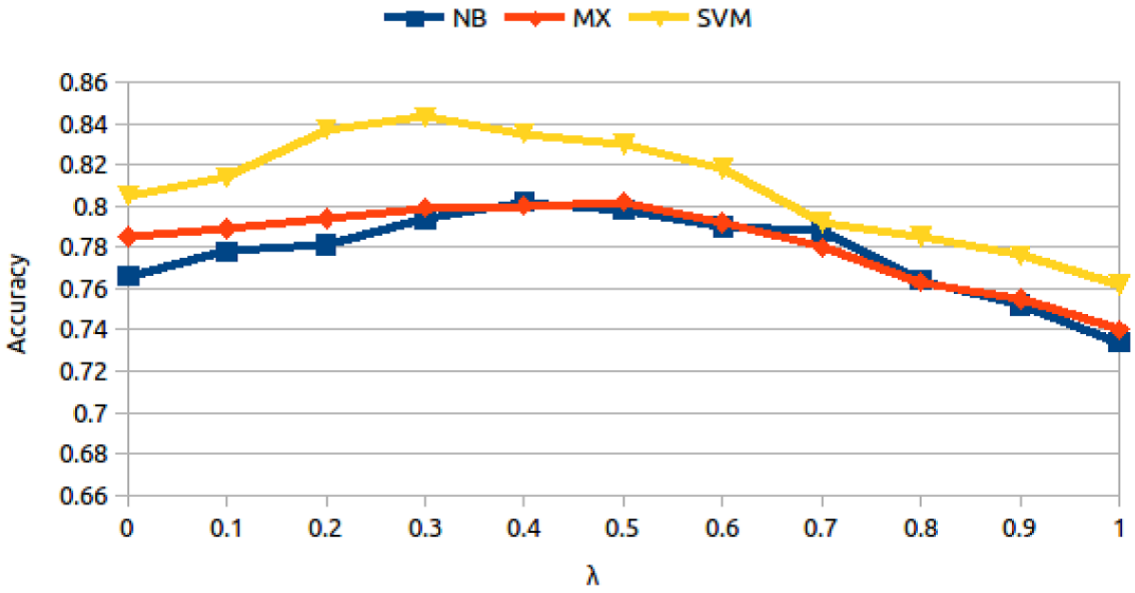


图 4.2 λ 值的确定。

从图中可以确定对于三种分类器所确定的 λ 值：对于 Naïve Bayes 分类器，$\lambda = 0.4$；对于最大熵分类器，$\lambda = 0.5$；对于支持向量机分类器，$\lambda = 0.3$。

确定了 λ 后，经过自举式的学习训练，并在测试集上评价，最后的结果如表 4.1所示。

从表中可以看出，首先无论是通用情感分类器还是微博情感分类器，性能上都超过了随机基准的 50% 准确率，这证明了无论是从那种视角进行分类，两种分类器都是有效的，胜过随机猜测。因此在没有任何标注数据来训练有监督或半监督分类器的情况下，我们的特征分割假设可以作为情感分类的一种有效的方法。

其次，通用情感分类器的准确率比基于情感词典的分类器要稍微好些，这是因为尽管通用情感分类器和基于词典的分类器都使用领域独立的词语来对情感知识进行建模，但是通用情感分类器是经过成语等知识资源所抽取的特征空间进行

[4]http://ictclas.nlpir.org/

表 4.1　结果对比表

Classifier	**NB**	**MX**	**SVM**
Lexicon Classifier	0.725	0.725	0.725
Supervised Classifier	0.785	**0.806**	0.826
General Classifier	0.734	0.740	0.762
Context Classifier	0.766	0.785	0.805
Combined Classifier	**0.802**	0.802	**0.843**

训练的，而情感词典中的词语都是独立使用并且还是具有一定的歧义；而对于微博情感分类器，准确率都比基于词典的分类器和通用情感分类器要好，因为它是使用微博依赖部分的特征训练得到的，更能使用微博语言环境，测试数据中出现的微博的“火星语言”越多越能体现出微博情感分类器的性能优越性。

最后，使用自举式学习框架的组合分类器显示出了最好的性能，因为它结合了通用分类器和微博分类器的综合性能，其准确率也超过了准确率比较高的有监督的分类器，这说明我们提出的方法既能很好的利用通用情感表达知识把握微博的总体情感倾向，也能照顾到微博特有的情感表达方式，准确掌握微博中细致的情感倾向。

4.6　小结

本章中我们针对情感分类的领域依赖性问题提出了无监督的自举式学习框架，并在微博中进行了验证。通过将情感分类问题的特征空间进行多视角分割，将整个特征向量特征空间分为领域独立的通用特征部分和领域依赖的微博特征空间，因此可以在两个特征空间分别训练得到通用情感分类器和微博情感分类器。然后我们使用了自举式的机器学习框架将两种分类器组合起来，达到更好的分类效果。实验证明我们所提出的方法性能上超过了现有的主要一些情感分类方法。

第五章　用户主观性建模

5.1　引言

随着基于内容的社交媒体的兴起，越来越多的用户开始愿意在这些社交媒体比如 Twitter 上针对各种话题发表短的文本信息表达意见和观点。因此这些社交媒体的文本数据中蕴含的广泛的话题、各种讨论以及丰富的主观性信息成为研究社交媒体用户的主观性珍贵的资源。在本章中我们所说的主观性是指用户感兴趣的话题（产品、政治人物和事件等等）以及用户对这些话题所持观点。使用社交媒体数据研究用户的主观性反过来也会有利于社交媒体的应用，比如用户观点查询、观点追踪或者用户行为的预测等。然而社交媒体产生的数据是巨大的，并且用户的主观性信息散播在“碎片化的信息”中，使得从这些数据中挖掘和消化各种不同用户所有的的观点变得极具挑战性。例如，如果在 Twitter 中查询“iphone”（由于 Twitter 数据的实时流动性，不同时间查询会有不同的效果，我们的查询日期为 2014 年 2 月 14 日），会返回大概 231,233 用户的 830,879 条微博（tweet，本文中我们统称为微博），意味着很多用户发表了不止一次微博来表达对“iphone”的观点。因此为了能够更好的了解到不同用户各种不同的观点，需要能够自动从用户发表的所有内容中（UGC）挖掘出“碎片化的观点”，将这些主观性信息进行集成（integrate），然后呈现出用户对于某些感兴趣话题的主要观点[143]。实际上用户感兴趣的话题会有很多，因此发表的内容也是多种多样的，因此如何从一条条独立的“信息碎片”中找到用户感兴趣的话题以及观点对主观性信息研究来说是很有意义且极具挑战的。

本章中我们针对这一用户的主观性建模问题提出了主观性模型，该模型分为两部分，使用一个框架将话题和观点结合起来。其中一部分是用户的兴趣话题分布，用于对用户对各种话题的兴趣读建模；另一部分是用户在每个话题上的观点的分布。具体来讲，图 5.3展示了框架的总体结构，该框架通过三步来解决用户话题观点集成问题：（1）首先使用用户层次的话题模型（user-level topic model）从用户发表的微博（我们以 Twitter 平台为例，当然我们的框架也可以适用于其他平台）中抽取出用户感兴趣的话题；（2）使用得到的话题模型和情感分析技术对用户每条微博进行话题和观点的分析；（3）综合并集成用户所有微博的话题与观点信息形成用户的主观模型。我们可以通过

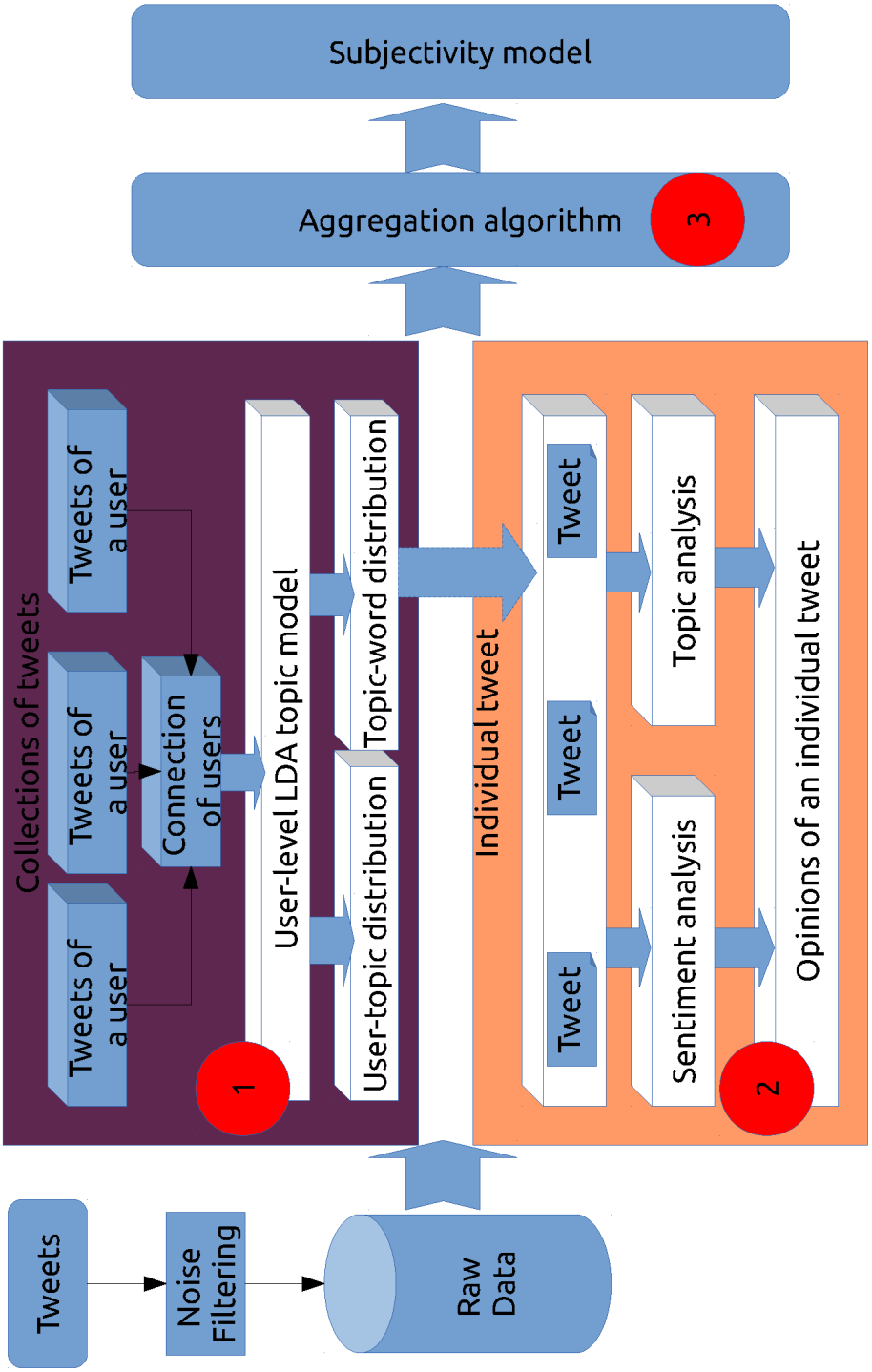


图 5.1 主观模型总体框架.

5.2 相关工作

虽然观点挖掘（opinion mining）最先是针对产品评论（review）和新闻评论（news comment）[22, 23]，近年来越来越多的研究工作开始关注于 Twitter 等短文本社交媒体，主要是针对单个短文本进行情感分析[132, 144–147]，往往忽视了用户之间与数据之间的关系。也有一些前期工作研究自动挖掘用户层次的主观性信息[148, 149]，通常是着眼于识别用户发表观点的目标[150] 或是针对特定目标确定用户的情感倾向性[151]，而没有考虑到话题的每个方面（aspect）。自从 Blei 等[152] 发表了潜语义话题模型（Latent dirichlet allocation，LDA），已经有了各种扩展的 LDA 模型用于从大规模语料中抽取用户的话题[153, 154]，也有很多模型将情感分析与话题模型想结合形成话题 -情感（topic-sentiment model）模型，这种模型将情感倾向与话题关联起来，与我们提出的主观模型很相似，主要有 Mei 等的 TSM 模型[155] 和 Lin 等的 JST 模型[156]。

随着用户在社交媒体上公开信息的增多，研究者能够获得越来越多的数据对用户建模，通过这些用户模型使得研究用户行为等研究更加容易。Hannon 等[157] 首先提出使用微博内容以及 Twitter 的社交网络关系对 Twitter 用户进行建模。Macskassy 和 Michelson[92] 使用 Wikipedia 作为外部知识库确定用户产生内容中的实体来对用户兴趣进行建模。Ramage 等[158] 使用 4S 维度利用话题模型对用户的微博及你想那个分析建模，得到的模型在信息过滤和朋友推荐等应用中显示出了很好的效果。Xu 等[159] 提出了一个混合模型用于分析用户的发帖行为，混合模型将突发新闻、朋友发帖以及用户兴趣三个重要因素结合在一起。Pennacchiotti 和 Popescu [160] 提出了一个全面的方法对用户建模用于用户分类任务，确认了从用户产生内容中挖掘深入的特征的价值，方法反映了对用户及其网络结构的深入理解。

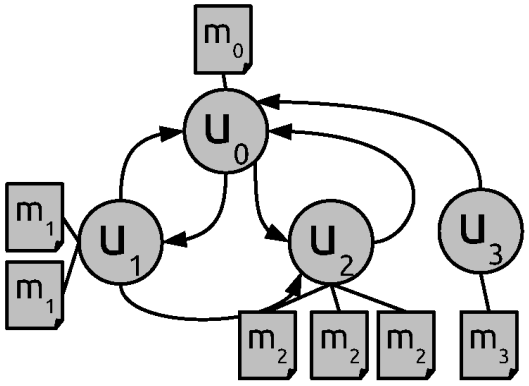
上述这些工作都确认从用户自己发布的内容中挖掘关键信息的重要性，并且开始从四方面信息进行建模，即基本信息（“Who you are”），发帖行为（“How you tweet”），发帖内容（“What you tweet”）以及网络关系（“Who you tweet”），但是很少能够对用户的兴趣和观点进行综合建模，也就是全面反映用户的主观性，本章中我们将提出主观模型对用户的主观性进行综合建模。

5.3 观点集成问题

正如我们在引言部分所介绍，用户在使用社交媒体平台的时候通常会就感兴趣的多个话题的多个方面多次发表自己的观点。因此要确定一个用户在某个话题上观点不能只看他的一条微博，应该将他所有与特定话题相关的微博中的主观信息进行综合才能得出用户的真正观点。在本章我们提出观点集成问题（**Opinion**

Integration Problem (OIP)）来对用户的主观性进行建模，我们主要关注用户层次（user-level）观点信息，而不是单个微博层次（tweet-level）的观点信息，因为观点挖掘的最终目标是找到人的主观想法而不只是单条微博中的主观信息，而对单条微博中观点的挖掘只是对用户主观性建模的一个中间步骤。此外，很多情况下用户的单条微博中的观点信息因为受到长度限制以及上下文语境的缺失是不明确的，但是通过看用户的所有微博就可以知道他的明确的观点[147]。

我们所提出的话题相关的观点集成问题（OIP）可以使用图 5.2进行说明。



	u_0	u_1	u_2	u_3
Topic0	$O_{0,0}$	$O_{1,0}$	$O_{2,0}$	$O_{3,0}$
Topic1	$O_{0,1}$	$O_{1,1}$	$O_{2,1}$	$O_{3,1}$
Topic2	$O_{0,2}$	$O_{1,2}$	$O_{2,2}$	$O_{3,2}$
Topic3	$O_{0,3}$	$O_{1,3}$	$O_{2,3}$	$O_{3,3}$

图 5.2　观点集成问题示例。

我们给出观点集成问题的正式定义：

定义 (观点集成问题)：如图 5.2所示，假设 Twitter 上的一个异构网络（heterogeneous network）由用户集 $V=\{u_i\}$，用户关系集合 $E=\left\{(u_i,u_j)|u_i,u_j\in V\right\}$ 以及用户发表的相应微博集合 $M_i=\{m_i\}$ 构成，其中用户所关注的话题 $T=\left\{Topic_j\right\}$ 以及用户对话题的观点可以从用户自己发表的微博中确定和抽取出来。对于一个特定用户 u_i，其对特定话题 $Topic_j$ 的观点 $O_{i,j}$ 不是他某条微博 m_i 中包含的观点，而应该是从他所有微博中与话题 $Topic_j$ 相关的所有微博 $M_i=\{m_i\}$ 中集成得出的。

对观点集成问题有两点因素必须考虑：首先为了观点针对目标的一致性，异构网络中无论是用户还是微博谈及的话题必须是在同一个话题空间，以使得无论话题的表示形式（比如概念（concept）表示或是话题模型的词袋向量空间的多项式分布表示）还是粒度都能够标准规范；其次，也是最重要的，就是集成的观点的表示形式问题，由于观点是与话题紧密相连的，一个用户针对某话题所发表的所有微博会覆盖与话题相关的所有方面，并且对于不同的方面会有不一样的喜好，比如对于手机“iphone”，用户可能喜欢它好看的外观和智能化操作系统，却不喜欢电池的待机时间过短，因此采用什么样的形式表示集成后的观点能准确表达出用户的总体主观性是一个十分重要的问题。本章中我们提出一个全新的通用主观模型来满足以上两个因素。

5.4 主观模型

心理学已经对主观性进行了广泛的研究，并基于用户的历史行为和言论中体现出的主观性来表示其独特个性[161]。在语言学上，语言中的主观性定义为作者在其发表的文本中表现出自己的立场、态度和情感[162]。社交媒体的出现为用户提供了能够及时自主针对感兴趣话题表达自己意见以展现自己独特主观性的发帖平台，因此在社交媒体平台上，用户的”**主观性**“，我们定义为用户产生内容中话题和针对话题的观点，因为主观性不但涉及到观点信息，也应该包含观点的目标。

在这一节，我们首先给出主观模型的形式化定义以满足提出的观点集成问题的需求。一般来讲，用户层面的观点分析是将用户针对某话题的情感倾向性分为“正面的（positive)”或是“负面的（negative)”。“正面的“表示该用户对话题支持或者喜欢该，而“负面的”表示不支持或不喜欢。我们所提出的主观模型中采用了更加通用（general）的“观点”定义，也就是观点针对某话题观点是在一个带有情感强度的更细粒度的情感表达空间里的情感分布，这个细粒度的情感表达空间可以更好的区分细微的观点差别，比如对话题持支持度为 8 的观点比支持度为 5 的观点更加“正面（positive）“。其实对观点的定义还没有统一的标准，我们采用这种比较广义的定义是为了能使得我们的模型能够更加通用。本章中为了具体化，我们统一在 Twitter 平台对主观模型进行定义和讨论，其实该模型可以适用于其他的平台或媒体。之所以将模型命名为”**主观模型**“是因为它是对社交媒体中用户产生数据中的主观性信息进行建模。模型的定义如下:

5.4.1 模型定义

假设 $G = (V, E)$ 表示 Twitter 上某个异构社交网络，其中 V 是网络中的用户，$E \subset V \times V$ 是用户之间的关注关系（follow relationship)。对于每一个用户 $u \in V$，对应的微博集合 M_u 表示其发帖的历史。可以认为在这个社交网络中存在一个话题空间 T 包含了 V 中所有用户谈论的所有话题，以及某个情感强度值空间 S 用于表示用户对某话题的观点。对于用户 $u \in V$ 的”**主观性（subjectivity）**“，指的是用户所发表的微博 M_u 中所涉及的所有话题以及观点。

定义 (主观模型): *用户 u 的主观模型 $P(u)$ 是用户在话题空间 T 中所谈论话题 $\{t\}$ 以及他对每个话题所持有的观点 $\{O_t\}$，观点用情感强度空间 S 的情感分布表示。*

$$P(u) = \{(t, w_u(t), \{d_{u,t}(s) \mid s \in S\}) \mid t \in T\} \tag{5.1}$$

其中：

- 对于用户 u，权重 $w_u(t)$ 表示其在话题空间中每个话题 $t \in T$ 的兴趣强度，并且 $\sum_{t=1}^{|T|} w_u(t) = 1$。

- 用户 u 对每个话题 t 的观点 O_t 指其对话题所持情感在情感强度空间 S 的分布 $O_t = \{d_{u,t}(s) | s \in S\}$，并且 $\sum_{s=1}^{|S|} d_{u,t}(s) = 1$。 ■

主观模型通过将用户感兴趣话题与观点同时考虑对用户的主观性进行建模，用户兴趣使用一个话题分布表示，对话题的观点用一个情感强度分布表示，主要目标是为了研究用户层面的观点信息，获得用户兴趣和主观思想的比较全面理解。

5.4.2 主观模型的构建

根据主观模型的定义，我们使用了两个分布对用户的主观性进行建模：一个是话题分布，一个是针对每个话题的观点分布，二者都需要从用户发布的历史数据中经过计算推断得出。然而对 Twitter 数据进行内容分析面临一些挑战：Twitter 上微博数量十分巨大，但是每条微博由于受限于 140 字的限制而很短小，并且各种不规范的语言被广泛使用，这些都很容易使得机器学习方法和自然语言处理技术失效[163]。因此能够有效的对 Twitter 的数据内容进行建模处理需要一些能适应这些挑战并且尽量不使用有监督的方法技术。本章中，我们主要使用一些无监督的方法从用户产生数据中挖掘话题和观点信息来构建用户的主观模型。我们提出了一个通用的框架来构建主观模型，该框架的主要优势就是利用 Twitter 的社交网络结构来帮助客服微博短文本造成的稀疏问题以及微博中标注数据的不足[164]。

5.4.2.1 话题分析

微博所涉及的话题一般是隐含的需要从其内容中推导得出。先前的研究主要是通过找到关键词（key words）[165]，抽取实体（extracting entities）[166]，链接到外部知识库（external knowledge categories）[92]，或者使用语义框架（semantic framework）[167]。对于这些方法来说，一个主要的问题是稀疏问题，因为同样涉及到同一个话题用户人然会使用各种不同的词汇来表达。在话题模型出现后，对于 LDA[152] 及其扩展的研究工作[168] 显示出对微博语料更加有效。LDA 的话题在概念上更加宽泛，因为 LDA 的每一个话题都是由所有相关的词语所构成。因此我们采用了用户层面（user-level LDA）话题模型来从用户所有的微博中发现隐含的话题，该模型的生成过程可以使用图 5.3来示意。

生成过程如下：

- 对每个用户 u, 获取话题分布 $\theta_u \sim Dir(\alpha)$;

- 对用户微博中的每个词语 $w_{u,n}$，$n \in \{1, \cdots, N\}$:

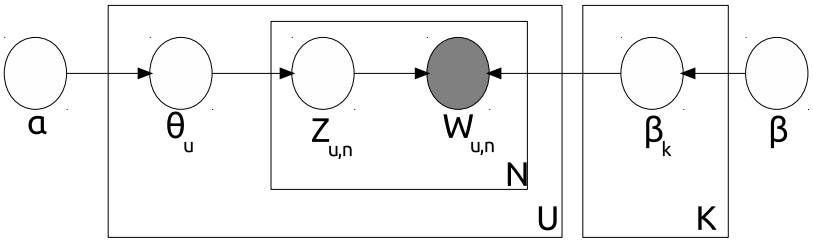


图 5.3　用户层面 LDA 话题模型

– 获取一个话题 $z_{u,n} \sim Multinomial(\theta_u)$；
– 基于话题 $z_{u,n}$，从话题的多项分布中获取词语 $w_{u,n}$：$p(w_{u,n}|z_{u,n},\beta_k)$。

为了从用户产生的内容中提取出讨论的话题，用户所发的微博应该和 LDA 模型的文档自然对应起来。由于我们的目标是了解用户感兴趣的话题而不是单条微博谈论的话题，所以我们将一个用户所有的微博连接起来组成一篇大的微博文档作为 LDA 模型的文档。因此 LDA 模型中的一篇文档就对应于一个用户，因此一个用户感兴趣的话题就可以使用在话题空间的一个多项式分布来表示，正好可以和我们的主观模型的话题权重相对应。在用户层面的 LDA 模型中，给定用户集合 V 以及话题数目 K，一个用户 $u \in V$ 的所有微博文档可使用话题上一个多项分布 θ_u 来表示，该分布具有参数为 α 的 Dirichlet 先验分布；一个话题 $k \in K$ 可以用所有词汇上的一个多项分布 β_k 来表示，该分布具有参数为 η 的 Dirichlet 先验分布。模型中的两个分布能够使用 Gibbs 采样或变分推理（variational inference）进行估计。本章中我们使用的是基于变分推理的话题模型工具 Gensim[169]，该工具使用的是在线批处理模式的变分推理，而且能够对新文档进行话题推理。

5.4.2.2　观点分析

微博用户经常会通过发表一些跟自己感兴趣话题相关的微博来表达自己的观点，因此为了挖掘微博用户的主观信息，我们需要了解用户每条微博蕴含的情感倾向，这需要情感分析技术。情感分析主要有基于规则以及机器学习两种方法。机器学习的训练过程需要大量标注数据，因为 Twitter 的庞大的数据量以及语言的动态性不太可能达到这样的要求，因此我们采用基于规则的方法，基于规则的方法具有很好的灵活性，可以将 Twitter 语言的一些特点转换成为规则而更适用于 Twitter[133, 138]。

SentiStrength 是专门针对社交媒体中的不正规短文本进行情感分析的工具包[138]。SentiStrength 将对应于社交媒体语言的规则结合进了基于词典的方法，非

常适用于我们能够对微博进行情感分析的要求。SentiStrength 对每条微博情感分析后赋予两个情感值：一个正面情感倾向强度值（在 $[1,5]$ 范围内）和一个负面情感倾向强度值（在 $[-5,-1]$ 范围内）。SentiStrength 情感分析的输出不是简单的正负倾向二值结果，而是细粒度的情感强度值，正好符合主观模型能够获得细粒度情感空间上情感分布的需求。为了计算的方便，我们将 SentiStrength 的两个输出结果映射为一个值，使用 $[0,8]$ 的离散整数值表示情感的强度，映射函数为：

$$o=\begin{cases} p+3 & if\ |p|>|n| \\ n+5 & if\ |n|>|p| \\ 4 & if\ |p|=|n| \end{cases} \tag{5.2}$$

其中 p 代表 SentiStrength 输出的正面情感倾向值，n 代表负面情感倾向值。在 $[0,8]$ 情感空间中，强度值 4 和 5 表示中性（neutral）情感，强度值大于 5 表示正面情感倾向，强度值小于 4 表示负面情感倾向。使用 SentiStrength 对用户的每条微博进行情感分析后，就可以 $[0,8]$ 情感空间内在进行观点的集成了。

5.4.2.3 构建主观模型

对用户微博进行话题分析以及情感分析后，我们就可以为用户开始构建主观模型了。对于一社交网络的用户集合 V，用 $M_u=\{m_i\}$ 表示每一用户 $u\in V$ 所发布的所有微博。按照用户层面 LDA 话题模型要求，将 M_u 中所有微博连接在一起形成一片长的微博文档 d_u，然后可以用这些微博文档 $\{d_u|u\in V\}$ 使用 LDA 话题模型进行训练获得话题个数为 K 的话题空间。训练得到的话题模型用参数 θ 表示每个用户在话题空间 T 中感兴趣的话题的分布，参数 β 表示每个话题在所有微博词汇上的分布。使用 SentiStrength 对每个用户的每条微博 m 进行情感分析得到每条微博大的情感强度 s_m。于是对用户 u 构建主观模型 $P(u)$ 的过程如算法 5.1所示：

算法中，由于微博的短小，我们简单假设微博 m 的情感 s_m 是针对微博所讨论的所有话题 Z_m 的，并没有区分针对不同话题的不同观点。

5.4.3 与生成模型比较

前期研究中提出了一些基于话题模型的生成模型，能够扩展基本的话题模型将文档中表达的情感与话题结合统一建模[155, 156]。TSM 模型（Topic Sentiment Mixture model）[155] 认为文档中表示情感等主观信息的语言与描述话题的语言是独立的，因此可以将表示情感的语言模型跟表示话题的语言模型分开，在文档的生成过程中词语要么从话题中要么从情感中采样获得。JST 模型（Joint Sentiment/Topic model）[156] 提出了一种新的方式来分析文档中的情感信息，使用话题模型抽取话

算法 5.1 主观模型的构建过程

已知:

The users set of a local network, V;

The tweets set published by each user u, M_u;

求:

The subjectivity model for each user u, $P(u)$;

Topic analysis with a user-level LDA, getting a topic model $P(\theta,\beta|M_u,V)$;

for all tweet $m \in M_u$ **do**

Sentiment analysis, outputting sentiment of m, s_m;

end for

for user $u \in V$ **do**

the topic distribution is the corresponding component of parameter θ, θ_u;

the topics he tweets about are $Z_u = \{t|p(t|\theta_u) > 0, t \in T\}$;

end for

for tweet $m \in M_u$ **do**

topics of m can be identified by the topic model:

$$Z_m = \{t|p(t|\theta,\beta,Z_u) > 0, t \in T\}. \tag{5.3}$$

end for

for each topic $t \in Z_u$ **do**

for sentiment value $s \in S$ **do**

count the number of tweets which talk about topic t with sentiment value s: $N_s = \sum_{m \in M_u} I(s_m), if\, s_m = s \& t \in Z_m$;

end for

calculating opinion towards topic t: $O_t = \left\{\frac{N_s}{\sum_{s \in S} N_s} | s \in [0,8]\right\}$.

end for

establishing subjectivity model of user u:

$$P(u) = \left\{\left(t, p(t|\theta_u), \left\{\frac{N_s}{\sum_{s \in S} N_s}\right\}\right) | t \in Z_u, s \in S\right\}. \tag{5.4}$$

return $P(u)$;

题过程中将话题和情感关联起来，因此可以同时对话题和情感信息建模。这些模型在发现话题相关的观点信息时跟我们提出的主观模型是很相似的，都能同时对用户的感兴趣话题以及话题相关的观点建模。

但是他们通常需要通过学习获得一个通用词语 -情感分布来对文档中的情感建模，这对短小和不正规的社交媒体语言尤其是 Twitter 来说是困难的。相较于话题的表达，情感等主观信息更难发现，u 因为情感信息常常隐含在一些微妙的语

言表达方式中（比如反讽），还有一些具体的领域和语境中也会具有独特的情感表达方式。微博中的情感除了一些正规的表达语言外，还有很多微博特有的语言特色来表示，比如表情符、字母大小写、字母重复以及惊叹号的使用等等。微博上的这些语言特色很难用概率分布来表示用于对情感建模。可是基于规则的情感分析方法可以很容易通过将这些语言特色转化为规则来抓住微博语言中微妙的情感表达方式。因此，我们的模型采用了基于规则的情感分析工具发现微博中的情感信息，更适合于 Twitter 等短文本社交媒体上用户主观信息的建模。

5.4.4 主观模型的应用

从用户微博中学习得到的主观模型能够在用户观点挖掘分析以及行为分析预测（转发、关注等行为）得到应用。本章我们以用户观点的预测为例来验证主观模型的作用，也就是学习到的主观模型能否有效的对用户将来针对某话题的观点进行预测。根据据用户观点的一致性，我们认为用户不会就某一话题表达随机正面或负面的观点，例如按照常理一个支持某候选人的用户更趋向于针对该候选人发表正面观点的微博。社会学上称这种现象为人的主观偏执（bias），也就是人的主观性[170]。因此得到了用户的主观模型，就可以预测用户就某一话题发表的微博所持观点。

我们将观点预测问题形式化为三元组 $< author, m, t >$，其中 *author* 微博 *m* 作者，微博涉及话题为 *t*。观点预测的目标就是计算得出用户 *author* 的微博 *m* 针对话题 *t* 表达出的观点极性 $p = \{positive, negative\}$。情感分析领域针对这一问题的主要方法是从微博中抽取出文本的表达模式，然后探索这些模式来预测观点的极性。单条短微博经常会由于缺乏上下文信息而使得观点模糊不清，但是用户的主观模型是从用户所有的微博中建立的，因此具有足够的信息，并且由于用户主观性的一致性，主观模型的观点信息比一条短文本更加可靠。因此我们提出使用主观模型来提高观点预测效果。具体来说，对微博 *m*，其作者的主观模型算法 5.1构建，微博 *m* 的情感值通过某种方法比如 SentiStrength 得出为 s_m。微博 *m* 所谈论的话题可以使用的公式 5.3推导得出：

$$\widehat{t} = argmax(\widehat{P}(t|\theta, \beta, Z_u)|t). \tag{5.5}$$

用户 *author* 的观点分布可以从主观模型 $P(author)$ 中确定，$O_{author,\widehat{t}}$，为在情感值上的分布，我们可以计算出用户在话题 $\widehat{t}$ 上归一化情感值：

$$\widehat{s_m} = \sum_{i \in T} d_i * v_i \tag{5.6}$$

表 5.1　Twitter 数据集统计

Total users	139,180	Friends per user	14.8
Total edges	4,175,405	Followers per user	14.9
Total tweets	76,409,820	Tweets per user	549

其中 v_i 表示情感值，d_i 表示情感值对应的分布。于是可以使用微博的情感值 s_m 和归一化的主观模型情感值 $\widehat{s_m}$ 进行观点极性 p 预测:

$$p = \begin{cases} positive & if \ \widehat{s_m + s_m} 2 \frac{|S|}{2} + 1 \\ negative & if \ \widehat{s_m + s_m} 2 \frac{|S|}{2} \\ neutral & otherwise \end{cases} \tag{5.7}$$

其中 S 是情感强度空间。

5.5　实验

5.5.1　数据集及设置

我们使用的是通过 Twitter 公开的 API 抓取的现成的数据集[171]。数据集的具体细节在表 5.1中列出。

由于 LDA 模型的计算复杂度，直接从 139,180 个用户的所有微博中构建主观模型需要耗费大量的时间。根据社交网络的同质性[172]，也就是“物以类聚（birds of a feather flock together）”的原则[173]，社交网络中连接紧密的用户更趋向于讨论形同话题持有相似观点[174]。在 Twitter 上，用户之间的连接关系对应着认可或关注，或者意味着有相似话题或观点。因此我们利用社交网络的社区结构（community）将 139,180 个用户划分为不同的社区，在社区内部为用户构建主观模型。社区的划分我们使用的是 Igraph[1]工具包，最后将用户划分为 106 个社区。对于用户数少于 15 的小社区，使用 LDA 进行话题分析效果比较差，因此我们将这些社区和用户过滤掉。同时我们也过滤掉了发微博少于 5 的、每次发微博字数少于 3 个的或微博中只有网络连接的用户 15,756 个。最后形成的数据集中有 122,329 个用户，分部于 33 个社区中。我们为每个用户在其社区内局部网络中按照算法 5.1构建主观模型。

除了主观模型，我们也使用生成模型 JST 和 TSM 进行了对比试验。其中模型的 Dirichlet 先验参数设为 $\alpha = 50/T$（T 为话题数目），$\beta = 0.01$。JST 的不对称情

[1] http://igraph.org/

感先验参数 γ 依照经验设为 $(0.01, 1.8)$。两个模型推导都进行 2,000 次迭代 Gibbs 采样。

5.5.2 样例分析

为了定性展示我们所提出的主观模型的有效性，我们给出了一个学习得到的用户的主观模型如图所示。该用户发表了 533 条微博，所有微博用词云图 5.4来显示。

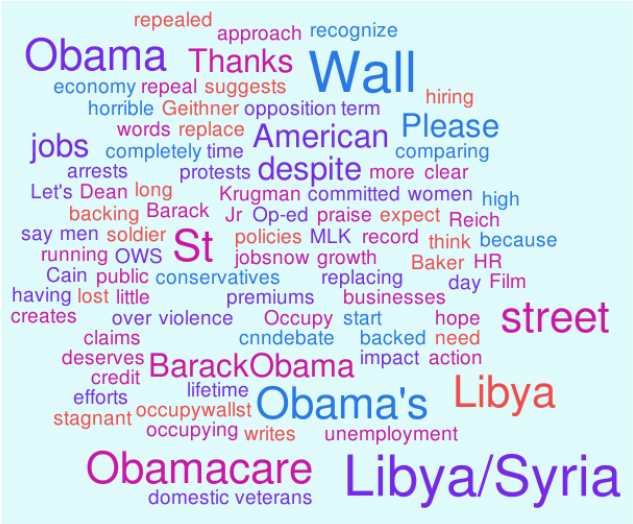


图 5.4 微博词云

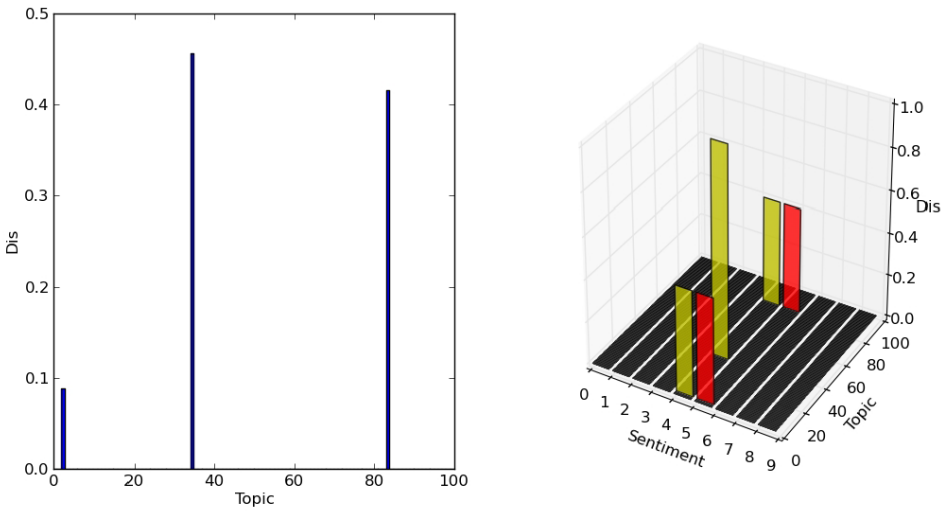


图 5.5 主观模型样例

图 5.5是用户在 [0, 100] 话题空间和 [0, 8] 情感强度空间的可视化的主观模型。很明显，从词云图上可以看到该用户讨论了三个话题（话题 2：“#Obamacare”，话题 32：“#libya” 以及话题 83：“#occupywallst”），子图 5.5左侧展示了用户在每个话题上的兴趣权重。子图 5.5右侧是用户对这三个话题所持观点分布，总体来看，对于话题 “#libya” 情感分布 100% 在强度 4 上，属于中性，对话题 “#Obamacare” 和 “#occupywallst” 情感分部都是 50% 在强度 4 以及 50% 在强度 5 上，属于中性偏正面倾向。从这个样例可以看出，我们的主观模型对用户的主观性进行了详细的建模，不但有用户的兴趣分布，也有细粒度的观点分部信息。

5.5.3 观点预测性能

为了定量评价主观模型的有效性，我们将主观模型在观点预测任务上与两个生成模型（TSM and JST）和一些主流的情感分析方法进行了对比实验。由于缺少可用的标注数据，我们主要跟几个无监督的情感分析方法进行了对比，这些方法主要有：

- **OF:OpinionFinder** 是一个公开可用的情感分析软件包，主要是用于句子层面的主观性分析[116]。
- **S140:Sentiment140** 使用远距离监督（distant supervision）方式（使用表情符获取训练数据）进行微博的情感分类。
- **STR:SentiStrength** 将微博中的一些语言特点转化成规则结合基于词典的方法专门针对微博等短文本社交媒体进行情感分析[138]。

我们从数据集中随机选择了 1,000 个至少有 80 条微博的用户条，然后每个用户随机选择一条微博组成了 1,000 条微博的数据集用于性能评测。为了能更容易识别微博的话题，我们优先选择带 hashtag 的微博。所有的 1,000 条微博进行人工标注作为评测标准。话题模型的话题数分别设置为 50, 100, 150 以及 200，评价指标使用的是准确率，结果如表 5.2所示。

从表中可以看出：

首先，OpinioFinder 的准确率是最低的 65.85%，主要原因是 OpinioFinder 主要是针对评论而设计的主观性信息分析工具，不适用于 Twitter 这种语言环境；

第二点，两个情感分析方法的准确率都显著的好于 OpinioFinder（Sentiment140：70.45%，SentiStrength：69.98%）；

[1]主观模型中, 左侧图代表话题分部: $(w_u(2) = 0.08, w_u(32) = 0.48, w_u(83) = 0.44)$；右侧图代表观点的分部: $O_2 = (d_{u,2}(4) = 0.5, d_{u,2}(5) = 0.5)$, $O_{32} = (d_{u,32}(4) = 1.0)$, $O_{83} = (d_{u,83}(4) = 0.5, d_{u,83}(5) = 0.5)$。

表 5.2　评价结果。相对于 OF 显著的性能提升使用 * 标记。

Method	50	100	150	200
OF	65.85%			
S140	70.45%			
STR	69.98%			
TSM	63.46%	72.94% *	67.83%	66.65%
JST	61.25%	68.57% *	75.88% *	67.03%
SUB	71.53% *	81.05% *	78.32%	74.54%

第三点，总体上两个生成模型的准确率都好于 OpinioFinder，证明了将情感信息与话题进行关联的重要性，它们的准确率都比 Sentiment140 和 SentiStrength 稍好，但是不显著；

最后，我们的模型（SUB）准确率在四种话题设置下都显著地超过了三个情感分析方法，并且将主观模型的信息加入到 SentiStrength 的分析结果后，显著提高了 SentiStrength 的性能，与两个生成模型相比较，我们的模型性能明显比 TSM 要好，稍好于 JST，这是因为我们的模型构建所用的情感分析方法更适合与 Twitter 语言，能够更准确的分析微博中的情感信息。

5.6　小结

本章中，我们定义并研究了社交媒体中用户的观点集成问题，提出了主观模型作为解决方案并设计了新的算法来从用户的历史数据中构建主观模型。使用主观模型可以自动总结用户在各个话题上的综合观点。实验证明，主观模型能有效的对用户的主观性进行建模，并且在观点预测任务中基于主观模型的方法性能显著比现有的几个情感分析方法要好，而且比 TSM 和 JST 两个生成模型更适于与 Twitter。

第六章　转发分析

6.1　引言

信息传播在市场营销以及选举策略等应用场景发挥着很大的作用，因为通过信息传播可以触发大量的人参与并逐步层叠扩展到更多的人。信息传播的研究引起了很多研究者的关注，尤其是在线的社交网络的研究者，他们对信息传播提出了一些通用的模型，用于帮助模拟信息流动（information flow）[175, 176] 以及探测信息级联（information cascades）的爆发[177]。可是他们常常都是将用户看作是网络中单一的节点，忽略了用户在信息传播过程中的自主性。作为 Web2.0 时代的信息消费者和生产者，每个用户都可以在社交媒体上发帖表明自己的兴趣和观点，选择信息阅读和传播。在社交网络上一条信息能否获得广泛传播依赖于“口碑相传（word of mouth）”效应，该效应会引发用户的传播行为。随着自然语言处理（Natural Language Processing）和数据挖掘（data mining）技术的发展，社交媒体上用户的意图可以使用他们产生的数据对用户进行建模来分析。本章中，我们关注一个有趣的问题：“口碑相传”效应机制问题，也就是给定某个特定用户的一条新微博，我们想要预测在所有收到该微博的所有用户中谁最有可能参与到该条信息的后续传播中去。作为一个典型的场景，我们用在 Twitter 中的一个异构网络来说明这个问题，如图 6.1所示，网络中 Tony 和他的网络中的朋友讨论了两个话题：“苹果手机（Iphone）”以及电影”冰雪奇缘（Frozen）”，现在 Tony 发了一条有关电影“冰雪奇缘”的新微博，我们想去发现他的朋友中谁会转发这条微博进行传播。

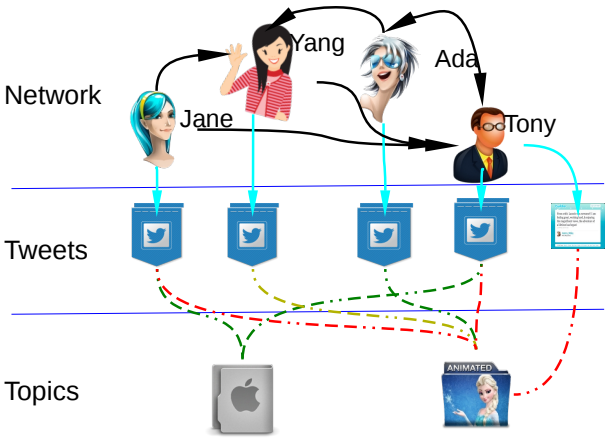


图 6.1　问题图示

不同社交网络平台上的信息传播行为是不一样的，本章我们主要研究 Twitter 上的转发行为。因为庞大的用户群以及信息的爆炸式增长，Twitter 在互联网上信息传播中扮演着重要角色，尽管微博长度上受到限制，但是 Twitter 为用户提供的转发按钮为信息的快速传播提供了前所未有的机制。据统计 Twitter 上的微博超过四分之一是转发别人的[91]，因此如果能够理解了转发行为是如何发生的也就能很好的解释 Twitter 上的信息传播。

作为信息传播的参与者，用户很自然地会在相互通信交互中表达出自己感兴趣的话题并发表自己的观点。在心理学的研究中证实人的主观能动性（subjective initiative）本质决定了人的主观性会影响其行为[178]，同样根据偏颇吸收（Biased Assimilation）理论，人总是趋向于选择和传播跟自己偏执观点（biased opinions）相似的信息[179]。因此能够全面掌握用户的主观性是研究用户参与信息传播意图的一个很重要的方面。针对转发行为的前期研究已经发展出了一些方法和模型用于找到影响转发的因素[92, 180]。然而据我们所知，还没与研究关注到用户转发行为背后的主观动机。就 Twitter 的口碑效应来说，转发意味着包括接受消息，评估内容以及衡量时候转发三个环节的一个完整过程，其中最重要的就是评估消息中是否有价值的信息值得和朋友分享。因此对用户的主观动机进行建模会为转发行为分析提供重要的研究视角。直觉上，依据“物以类聚（like attracts like）”原则，具有主观性的用户更趋向于转发那些能够迎合他口味（观点）的信息。就我们在图 6.1中的例子来说，用户对两个话题所持观点已经在他们之前发布的微博中表达了，Tony 和 Jane 对电影“冰雪奇缘”持正面肯定观点，而 Ada 持负面观点，Yang 持中性观点。如果 Tony 新发布的是对冰雪奇缘正面肯定的微博，Jane 是最有可能转发这条微博的就很容易理解了。因此用户的主观性是如何影响其转发行为的是本章的关注点。

为了研究主观性和转发行为的关系，有两个问题需要回答：（1）怎样准确对用户的主观性进行建模？（2）怎么样从主观性上有效度量用户认为微博传播的值得性？回答这两个问题不是一件很容易的事，其中上一章我们已经提出了使用主观模型来解决社交媒体上用户观点集成问题，本章我们将继续使用主管模型的概念，提出一个更加通用的框架，并就主管模型提出一个全新的相似性计算方法来度量传播的值得性，而且就影响转发行为的因素，我们会结合主观相似性度量来分析三个最有可能引起转发行为的因素。

[0]每个用户的观点用不同颜色表示，“红色”当标正面观点，“绿色”代表负面观点，“黄色”代表中性观点。

6.2 相关工作

在微博转发行为分析方面，很多工作已经在转发行为特征、查找提高微博转发性的因素以及设计能估计转发概率的模型方面展开研究。Suh 等 [181] 通过研究发现带有网络连接 URL 以及 hashtag 标记的微博更有可能被转发。Macskassy 和 Michelson [92] 发现从微博内容中推到得出的模型能够解释大多数的转发行为。Comarela 等[94] 发现对微博作者的先前反应，微博作者发帖频率，微博内容的新鲜程度以及微博的长度会影响关注者的转发可能性。Starbird 和 Palen [93] 特别针对危机发生时的微博信息转发机制进行了研究，发现带有跟带有危机话题关键词的微博更有可能被转发。Osborne 和 Lavrenko [182] 通过引入一些有用特征，比如微博的新颖性和作者被加入朋友列表的次数，来使用被动主动算法（passive aggressive algorithm）训练模型预测转发行为。Jenders 等 [96] 从微博及其作者的网络结构、信息内容以及情感方面分析了一些“显式”和“隐式”的影响转发的特征。Naveed 等 [183, 184] 引入了微博的趣味性指标，并使用一些比如表情符、情感以及话题等特征对其进行量化来预测一条微博被转发的可能性。Feng 和 Wang [185] 构建了一个图模型并将微博以及用户的所有信息源结合进图的节点和边，并提出了一个特征敏感的因子分解模型（factorization model）对微博依据被转发的可能概率进行重排序。Pfitzner 等 [186] 提出了一种叫做情感分歧（emotional divergence）指标来评价微博被转发的可能性，并研究证实了高情感分歧值的微博会有更高五倍的机会被转发。

总体来说，上述所有工作主要是回答“某条微博会否被某些用户转发”这样一个问题，但是他们还不能回答“从一个信息消费者角度来看的某条微博是否值得用户采用转发行为进行信息传播”这样的问题，本章中我们会结合上章提出的主观模型从用户的兴趣和观点角度就用户的主观动机来分析转发行为。

6.3 基于主观模型的转发分析

为了研究用户转发行为的主观动机，首先需要了解用户的主观性，也就是能清楚用户喜欢什么和不喜欢什么，也就是用户感兴趣的话题和用户对话题所持的观点，这就是上章我们为用户所建立主观模型的作用。随着社交媒体普及率越来越高，社交媒体上带有用户主观性信息的用户产生内容（UGC）也越来越多，自然语言处理领域的观点挖掘（opinion mining）[23] 研究开始通过计算方式自动对用户的观点信息进行建模。并且也出现了一些基于方面（aspect-based）情感分析或话题情感模型（topic-sentiment model）[155, 187] 将针对话题的观点投射为二值极性，评价等级或情绪类型等某一个单一的情感值，但是这些模型因为这种观点简单的

表示形式而使得其作用受到限制。因此在主观模型中，我们通过将话题和观点结合为一个模型并将话题和观点使用新的表示形式（在话题空间和情感空间的分布）来对用户的主观性进行建模。

6.3.1 主观模型

在 Twitter 上，用户一般是对多个不同话题感兴趣，比如图 6.1中的例子中 Tony 和 Jane 都对苹果手机话题和冰雪奇缘电影刚兴趣并表达了自己的观点。一般来讲，用户对话题的兴趣度会随着话题的不同而不同，而且，即便对同一个感兴趣话题，当一个用户其其不同的方面（aspects）发表微博表达的观点也不是完全一样的，因此我们认为对用户的主观性进行建模时：

- 每一个用户 u 都对应着一个在话题空间 T 的 $|T|$ 维的话题分布 $W_u = \{w_u(t)\}|W_u \in R^{|T|}, \sum_t w_u(t) = 1$，其中 $w_u(t)$ 表示用户在话题 t 上的兴趣度。
- 用户 u 对话题 t 的观点应该表示为一个在情感空间 S 的 $|S|$ 维的情感分布 $O_t = \{d_{u,t}(s)|O_t \in R^{|S|}, \sum_s d_{u,t}(s) = 1\}$，其中 $d_{u,t}(s)$ 表示用户观点中带有情感值 s 的可能性。

从技术角度来讲，我们提出主观模型的目标就是设计一个通用的框架能够从社交媒体用户产生的数据中同时学习到用户兴趣（对应的话题分布）和观点（对应的观点分布），以此为基础进行用户一些在线行为的分析，本章中我们主要研究用户的转发行为。之所以说我们提出的主观模型是通用的，因为它不但将用户的兴趣和观点融进一个整体框架，更重要的是，在主观模型中观点表示为一个在可扩展的情感空间上的概率分布。这个情感空间既可以是表示情感正负极性的二值空间，又可以是连续值表示的情感强度空间，或是离散值表示的情绪类型空间，因此可以覆盖所有的观点表示形式。图 6.2是一个在 $[0, 100]$ 话题空间和 $[0, 8]$ 情感强度空间主观模型的可视化示例。

左侧图中表示用户在话题 2，32 和 83 上的兴趣度：

$$(w_{u,2} = 0.08, w_{u,32} = 0.48, w_{u,83} = 0.44)$$

右侧图中表示在每个话题上的观点分布：

$$O_2 = (d_{u,2,4} = 0.5, d_{u,2,5} = 0.5)$$

$$O_{32} = (d_{u,32,4} = 1.0)$$

$$O_{83} = (d_{u,83,4} = 0.5, d_{u,83,5} = 0.5)$$

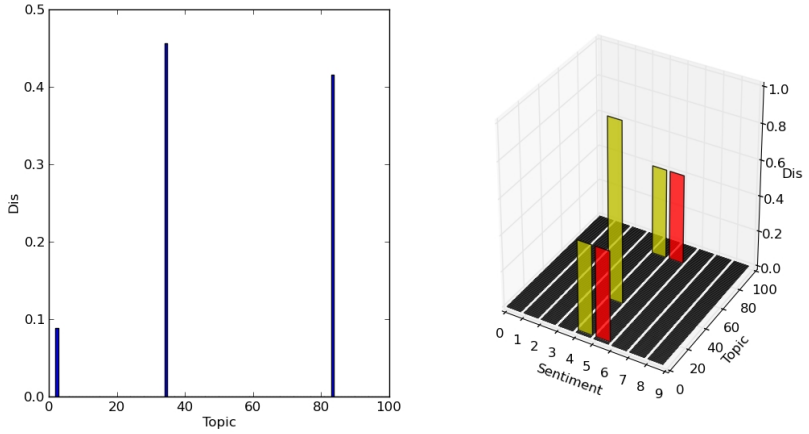


图 6.2　可视化主观模型示例

在构建主观模型时，我们的框架是将话题分析和观点分析分开进行的。具体点来讲，首先使用的是用户层面（user-level）的 LDA 话题模型从用户所有的微博 M_u 中训练一个全局话题模型 $TM = (\theta, \phi)$, 其中 θ 表示用户在话题空间 T 的兴趣度分布，ϕ 表示话题在词表上的分布。由于微博比较短小，通常认为每条微博谈论的是一个话题，因此我们可以根据全局话题模型计算微博 t 从话题模型中产生的概率值为 t 指定一个最有可能的话题：

$$z_t = \arg\max_k \prod_{w \in t} P(w|\phi_k) \tag{6.1}$$

然后就可以将用户 u 所有谈论同一话题的微博数进行归一化后获得用户 u 在话题上的兴趣度：

$$w_{u,k} = \frac{|\{t : t \in M_u \wedge z_t = k\}|}{|M_u|} \tag{6.2}$$

关于观点分布，正如我们在图 6.1中的例子看到，Tony 和 Jane 总体上都是对电影“冰雪奇缘”持正面观点，但是他们有可能是因为不同的原因而喜欢这部电影的。Jane 可能非常喜欢电影浪漫的故事情节，但是对它的动画画面稍微有点失望；而 Tony 喜欢这部电影可能是因为被这部电影的动画技术所折服，却不喜欢它略显幼稚的公主王子题材。之前的情感分析研究主要是将观点通过计算得出某一个单一值，尤其是正负极性二值为主，并不区分针对话题的每个观点在不同方面的具体情感，也无法计算观点的大小顺序，比如那个用户更喜欢电影些。因此在主观模型中，我们将观点表示为在情感空间 S 的一个分布用以更精确的表示和区分观点。假设微博 t 通过情感分析得出其在情感空间的情感值为 s_t，用户在某一

话题 k 上的观点分布可以通过将其所有谈论该话题微博在每一个情感值上的数目归一化处理后获得：

$$\begin{aligned} O_k &= \{d_{u,k,s} | s \in S\} \\ &= \{\frac{|t : t \in M_u \wedge z_t = k \wedge s_t = s|}{|M_u|} | s \in S\} \end{aligned} \tag{6.3}$$

6.3.2 主观相似性

得到用户的主观模型后，需要定义一个相似性度量方法来计算用户之间或用户与微博之间主观性上的距离，以量化“物以类据（like attracts like)”效应。首先我们定义在同一话题上观点的相似性计算方法。

6.3.2.1 观点相似性

在主观模型中观点是定义在情感空间上的分布，其每一维都代表着在对应情感值上的情感比重。实际上，为了区分观点，情感空间中的情感值并不是独立的，情感值之间有一定的顺序和大小来表示情感的强度。比如情感值为 8 的观点比情感值为 5 的观点持更正面的观点。因此常用的一些计算分布相似性的方法，比余弦相似性（cosine similarity）以及 KL 距离（KL-divergence)，对于主观模型中观点分布相似性的计算就不适合了。如表 6.1所示，假设在一个 $S = [0, 8]$ 情感空间中表示的三个观点：观点 O_k^1 是最负面（100% 分布在情感值 0 上。)，观点 O_k^2 是正面的（50% 分布在情感值 6，50% 分布在情感值 7 上。)，观点 O_k^3 最正面（100% 分布在情感值 8 上)。

表 6.1 观点相似性示例

	0	1	2	3	4	5	6	7	8
O_k^1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
O_k^2	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.0
O_k^3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

假设使用常规的分布的相似性计算方法余弦相似性，就会发现三个观点之间的相似性都是 0，因而出现了相似性计算方法失效，这是与事实不相符的，因为观点 O_k^2 与观点 O_k^3 比观点 O_k^1 与观点 O_k^3 更相似，它们都是持正面观点。因此观点的相似性计算不能简单将观点视为一般的概率分布来计算，或者只是观点空间的一个距离值。为了准确计算观点之间的相似性，我们将观点在情感空间的距离和分布上的相似性结合起来，提出了如下的计算观点 O_k^u, O_k^v 之间相似性方法：

$$Sim(O_k^u, O_k^v) = \frac{|S| - |\sum_{i=0}^{|S|} d_i^u v_i - \sum_{i=0}^{|S|} d_i^v v_i|}{|S|} \tag{6.4}$$

其中 d_i 是第 i^{th} 维的情感分布，v_i 是相应的情感值。

使用方法 6.4计算表 6.1中观点之间相似性为:

$$Sim(O_k^1, O_k^3) = 0$$

$$Sim(O_k^2, O_k^3) = 6/8$$

$$Sim(O_k^1, O_k^2) = 2/8$$

。这与我们对观点相似性的直觉理解是一致的。

6.3.2.2 主观相似性

在主观模型中，用户感兴趣的话题表示为在话题空间 T 上不同话题的兴趣度分布，因此两个主观模型 SM_u 和 SM_v 之间的主观相似性可以将话题上的权重与对应的观点分布相似性结合起来进行集成计算:

$$Sim(SM_u, SM_v) = \sum_{k=1}^{|T_{u,v}|} \theta_u(k) Sim(O_k^u, O_k^v) \tag{6.5}$$

其中 $T_{u,v}$ 表示两个用户之间的共同话题，可以通过对他们之间感兴趣话题的交集来获得；$\theta_u(k)$ 代表用户 u 在话题 k 上的兴趣度权重。

值得注意的是，当我们测量用户 u 在主观性上与用户 v 有多相似性时，话题权重我们使用的是用户 u 的话题权重，因此这个主观相似性度量方法是不对称的。之所以这么做，我们的直觉是因为用户的主观性是个人的内在感觉，因此对于另一个用户与自己主观想法上有多相似也是一个单向的感知，无需对称性的考虑。因此在度量主观相似性时，$Sim(SM_u, SM_v) \neq Sim(SM_v, SM_u)$。

6.3.3 转发行为分析

用户的转发行为收到多种因素的影响，但是从用户的角度来讲，三种情形下会引发用户的转发:

1. 微博的内容对用户具有吸引力，因此用户的转发行为是根据自己的主观判断引发的;
2. 微博是由关系密切的好朋友发出的，因此用户的转发行为是因为社交需要;

3. 微薄内容是突发新闻或有趣段子而具有流行性，因此用户的转发行为是趋同需求（conformity needs）[188] 的结果。

这些情形是用户转发行为的不同原因，从主观动机角度分析，我们使用三个主观相似性来量化这三个因素进行转发行为的分析。

在以下的分析中，对于一条微博 t，假设 F 表示该微博作者 u_a 的关注者，当作者 u_a 发布微博 t 后，所有用户 F 都会看到 t，至于哪个用户会转发 t，我们需要分析其主观动机。对于每一个关注者 $f \in F$，可以定义一个四元组 $< f, u_a, t, r_f >$，其中 r_f 是一个二值标签用以表示微博 t 是否会被用户 f 转发，需要我们通过分析进行预测。

6.3.3.1 吸引力度量

一般来讲，用户根据自己的主观判断，看到一个有吸引力的微博就会转发。因此我们可以通过计算微博 t 与微博关注者 f 之间的主观相似性来定量地度量这种吸引力。对于一条微博，它所讨论的话题 z_t 可以使用公式 6.1指定，对其进行情感分析可以得到情感值 s_t，因此微博也能够使用主观模型进行建模，它的话题分布和观点分布都是一个 100% 的单一分布。于是微博 t 对于用户 f 的吸引力就可以使用我们定义的主观相似性计算方法 6.5进行度量：

$$Sim(f,t) = \theta_f(z_t) Sim(O^f_{z_t}, O^t_{z_t}) \tag{6.6}$$

6.3.3.2 社交性度量

这种情形下，转发行为是基于用户的社交交互需要。由于微博是由志同道合（like-minded）的好朋友发的，转发行为是因为友谊触发而不一定是微博内容。这种情况下可以通过计算用户 f 与微博作者 u_a 之间的主观相似性来度量二者之间友谊的亲密程度：

$$Sim(f,u_a) = \sum_{k=1}^{|T_{u,v}|} \theta_f(k) Sim(O^f_k, O^{u_a}_k) \tag{6.7}$$

同时也应该考虑到，不同类型的朋友对用户 f 的影响力（influence）是不同的，比如用户 f 可能会关注很多人，但是可能只会与少数几个互动频繁（转发等互动）。而且用户 f 并不是对朋友的每条微博都感兴趣，例如在图 6.1中的例子中，Jane 可能会对 Tony 所发的关于电影“冰雪奇缘”的微博，但是对他的关于苹果手机微博不感兴趣。因此我们对用户之间的主观相似性 $Sim(f,u_a)$ 附加一个权重以反映不同类型朋友对用户 f 的影响力，该权重有四部分因子组合而成。

专家指数因子（Expert Factor）$w_E(u_a)$： 该因子代表着微博作者 u_a 在微博接收朋友圈中相对的专家指数，专家指数越高的用户就会对其他用户有更多的影响力。本章中我们简单的根据用户 u_a 的发帖数量在所有朋友圈中发帖总数的比例来计算专家指数。

$$w_E(u_a) = |M_{u_a}|/|\{M_u|u \in u_a \cup F\}| \tag{6.8}$$

领导力因子（Leadership Factor）$w_L(u_a)$: 我们将用户的领导力为该用户拥有的粉丝（followers）数。因此领导力因子可以通过归一化计算为:

$$w_L(u_a) = \log(|F|)/\log(\max) \tag{6.9}$$

其中 max 是 Twitter 中用户的最大流行度（maximum popularity）[1]。

相似性因子（Similarity Factor）$w_S(u_a, f)$: 用户 u_a 和 f 之间的兴趣的相似性可以通过他们主观模型中话题分布之间的反 KL 距离（inverse KL-divergence）来度量:

$$w_S(u_a, f) = 1/KL(\theta_{u_a}, \theta_f) \tag{6.10}$$

交互因子（Interaction Factor）$w_I(u_a, f)$: 用户 u_a 和 f 之间的交互数量 $Interation_{u_a,f}$ 包括他们之间的对话，相互之间的提及以及相互之间的转发等。该因子可以通过将 $Interation_{u_a,f}$ 用用户 u_a 和 f 所有的微博数目归一化计算获得:

$$w_I(u_a, f) = |Interation_{u_a,f}|/|\{M_{u_a}, M_f\}| \tag{6.11}$$

总体上，将以上四个因子组合后可以得到影响权重:

$$w_{u_a,f} = \lambda_1 * w_E(u_a) + \lambda_2 * w_L(u_a) + \lambda_3 * w_S(u_a, f) + \lambda_4 * w_I(u_a, f) \tag{6.12}$$

其中 λ_i 是一个可选权重向量以反映不同因子的影响，并且 $\sum_{i=1}^{4} \lambda_i = 1$。本章中我们将其均衡设为 $\lambda_i = 0.25$。

6.3.3.3 流行性度量

用户在使用 Twitter 时，如果发现一条微博是非常流行的（具有新颖性或传染性），在趋同（conformity）效应的作用下，用户就会很有可能对其进行转发。这种情形下，微博 t 的内容一般在话题和观点上与其作者 u_a 的主观性不太一致，因此微博 t 与其作者 u_a 之间的主观相似性 $Sim(u_a, t)$ 会相对较低:

$$Sim(u_a, t) = \theta_{u_a}(z_t) Sim(O^{u_a}_{z_t}, O^{t}_{z_t}) \tag{6.13}$$

[1] http://twittercounter.com/pages/100

用户的转发行为是由于微博 t 的流行性而不是其因为内容具有吸引力或者是好朋友发布的，为了度量其流行性影响，我们对 $Sim(u_a, t)$ 增加一个流行性系数，该系数可以通过计算接收微博 t 的用户 f 所关注朋友中转发微博 t 的比例来确定。

6.4 实验

6.4.1 数据集与实验设置

实验中我们使用了论文[189]提供的 Twitter 数据集[2]，在构建数据集时，他们使用 Twitter Streaming API 随机选取了 500 条目标微博，每条微博至少被其作者的粉丝转发过一次，对这 500 条微博进行连续几个小时的监控找到转发微博的那些用户。同时以这 500 条微博为入口，收集了微博作者及其粉丝的最近发布的 200 条微博。获得的数据集总共有 45,531 个用户，共 6,277,736 条微博，在监控期间有 5,214 个用户转发了 500 条微博中的至少一条。为了避免数据不平衡带来的偏置，我们从数据集中采样抽取了 5,214 个没有转发的用户作为反例，与转发者一起构成平衡测试数据集。数据集的统计如表 6.2所示：

表 6.2 数据集统计分析

Total tweets which have been monitored	500
Average number of followers per tweet	89
All followers	45,531
All historical tweets	6,277,736
Total retweeters	5,214
Total non-retweeters	40,317

在构建主观模型时与上一章一样，使用了 Gensim[169] 进行话题模型训练，话题数目设为 50,100,150 和 200；使用 SentiStrength[138] 对每条微博进行情感分析，并且为了更好的适应于微博情感表达方式，我们使用了 Nielsen 等[123] 为 Twitter 构建的情感词典。

6.4.2 相关性检验

首先，为了验证主观相似性会影响转发行为，我们采用了一个 ANOVA（Analysis of Variance）[190] 假设性检验方法对我们提出的用主观相似性表示的三个因素与转发行为之间的相关性进行分析，使用该检验方法我们对“**转发者（retweeters）和非转发者（non-retweeters）具有相同的主观相似性均值**”这一零

[2]下载地址：https://sourceforge.net/projects/retweeter/

假设（null hypothesis）进行检验。结果如表 6.3所示，表中加黑部分表示 *p-value* 低于显著性水平。

表 6.3　ANOVA 检验结果

Similarity		$Sim(f,t)$	$Sim(f,u_a)$	$Sim(u_a,t)$
50	*F*	**12.182**	2.212	4.236
	p	**4.44e^{-06}**	0.140	0.272
100	*F*	**43.892**	**31.145**	**28.466**
	p	**8.65e^{-11}**	**3.55e^{-08}**	**1.32e^{-09}**
150	*F*	**22.356**	**12.240**	**14.664**
	p	**2.43e^{-08}**	**6.25e^{-06}**	**8.46e^{-07}**
200	*F*	**31.675**	**20.616**	6.145
	p	**4.22e^{-06}**	**2.92e^{-05}**	0.26

表中如果平均值差异是偶然，*F-ratio*=1.00，否则
F-ratio >1.00 (*p-value* <0.01)。

从表中可以看出，当话题数是 100 和 150 时，所有的主观相似性检验都是 *F-ratio* 大于 1.00，且 *p-values* 低于显著性水平。这表示所有的主观相似性与转发行为具有相关性，能够作为转发行为的有用特征。后续实验我们将话题数目固定为 100 来进行讨论。

6.4.3　样例分析

在本节，我们用一个实际样例来说明主观模型以及主观模型在解释转发行为中的作用。我们从 500 条目标微博中选取了其中的一条，其内容为：

Sometimes the right person for you was there all along. You just didn’t see it because the wrong one was blocking the sight.

微博作者以及两个关注者（一个是转发者，一个是未转发者）构建的主观模型如图 6.3所示。

图 6.4显示了 14^{th} 号话题、微博作者与两个关注者的所有微博的词云图（word cloud diagrams）[3]。

该微博谈论的是 14^{th} 号话题，话题是关于“love between people”，且作者对该话题的观点偏中性，这与图 6.3中微博的主观模型以及图 6.4中 14^{th} 号话题词云图是一致的；微博作者有 188 条微博，主要集中在 14^{th} 号话题，观点分布为 $O_{u_a}^{14} = (0, 0.04, 0.05, 0.25, 0.35, 0.25, 0.05, 0.01)$，偏中性；至于两个关注者，转发者有 196 条

[3]我们使用 TagCrowd (`http://tagcrowd.com/`) 生成词云图。

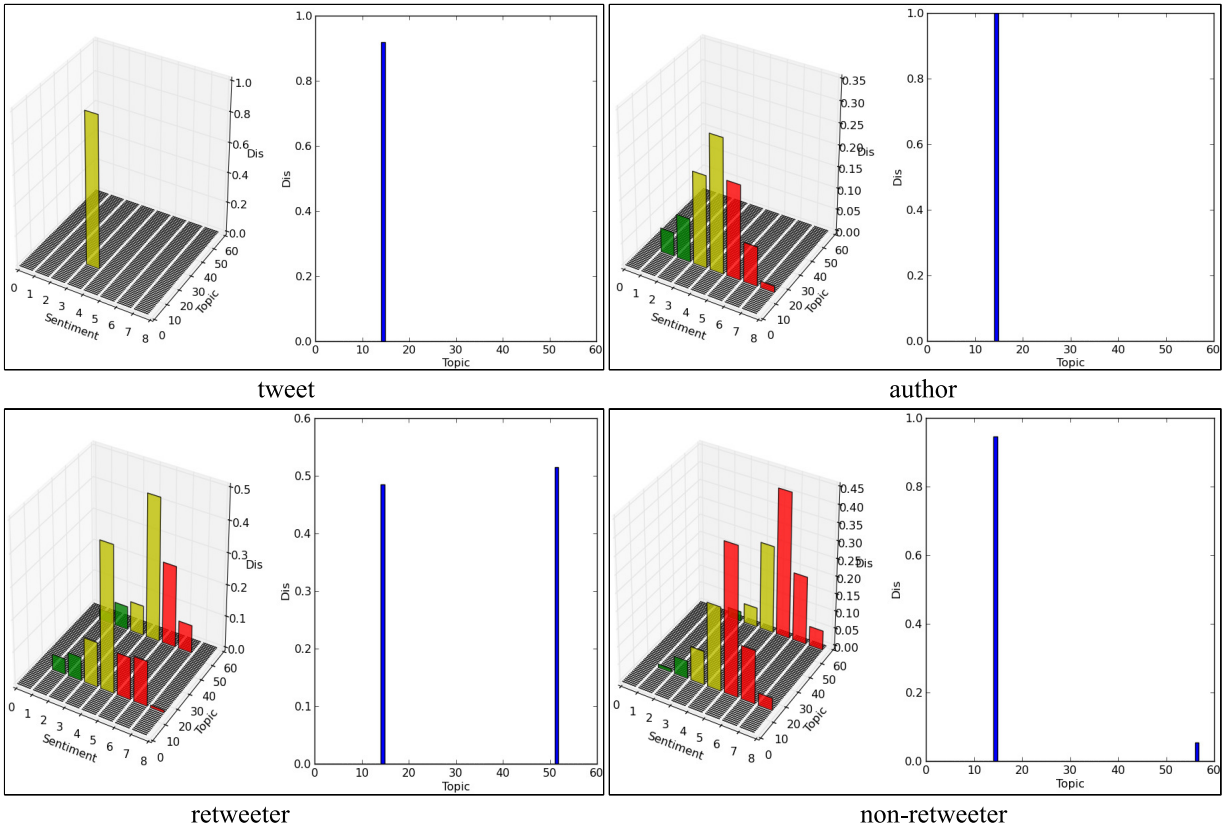


图 6.3　主观模型示意图

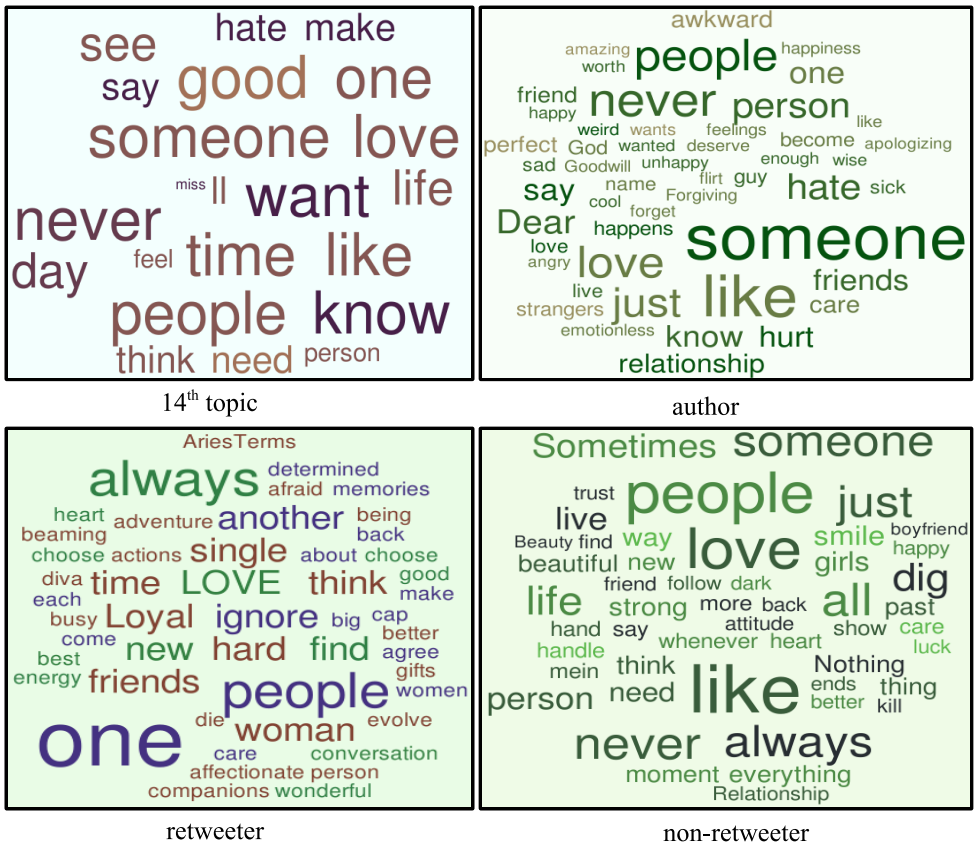
微博谈论到了 14^{th} 号话题和 52^{nd} 号话题，话题分布比较均匀（其中 $w_{u_r}(14)=0.48$），对 14^{th} 号话题，观点分布为 $O_{u_r}^{14}=(0,0.02,0.04,0.15,0.50,0.13,0.15,0.01)$，偏中性；未转发者有 156 条微博谈论到了 14^{th} 号话题和 56^{th} 号话题，主要谈论了 14^{th} 号话题（其中 $w_{u_n}(14)=0.98$），观点分布为 $O_{u_n}^{14}=(0,0.01,0.04,0.10,0.25,0.45,0.13,0.02)$，偏正面。

表 6.4是针对转发者和未转发者计算的三个主观相似性，可以看出对于转发者来说除了微博与其作者的主观相似性外，其他两个主观相似性都明显高于未转发者。

表 6.4　主观相似性比较

Similarity	$Sim(f,t)$	$Sim(f,u_a)$	$Sim(u_a,t)$
Retweeter	0.854	0.967	0.886
Non-retweeter	0.805	0.919	0.886

从以上分析中看得出，对于两个微博的关注者，单单感兴趣话题来说，未转发者与微博以及微博作者更为相似（在共同话题 14^{th} 话题的兴趣度：作者 $w_{u_a}(14)=1.0$，未转发者 $w_{u_n}(14)=0.98$，转发者 $w_{u_r}(14)=0.48$），但是考虑到主观



14th topic　　author

retweeter　　non-retweeter

图 6.4　14^{th} 号话题、微博作者与两个关注者词云图

相似性，转发者因为与微博以及作者有更相似的观点分布而主观相似性更高（与微博主观相似性 $Sim(f,t)$:0.854 vs 0.805，与作者主观相似性 $Sim(f,u_a)$：0.967 vs 0.919），因此主观相似性的不同引发了他们不同的行为，从这个例子可以看出我们所提出的主观模型以及考虑三个因素的主观相似性在解释用户转发行为上的作用。

6.4.4　转发预测

为了进一步定量的评价所提出的方法的有效性，我们分三阶段进行了转发预测实验。首先，将我们的模型与其他基于话题的模型进行对比实验。这些模型包括使用词袋模型对用户兴趣建模的 TF-IDF 模型、从用户产生内容中抽取实体对用户兴趣建模的基于实体模型（entity）以及使用用户微博中的 hashtag 的 hashtag 模型[166]，对这些模型计算相似性的时候使用的是余弦相似性。

第二个阶段，将我们的模型与两个生成式的话题情感模型（generative topic-sentiment models）TSM 模型[155] 以及 JST 模型[156] 进行了对比实验。虽然 TSM 模型和 JST 模型也能同时对话题和话题相关的情感建模，但是他们的输出为正负二值极性。在训练这两个模型时，我们同样也是将用户所有微博作为一篇文档作为

表 6.5　LUO 方法使用特征

转发历史特征（RH）	**取值范围**	**Description**
用户转发数目（Num_fRu）	$N = \{0, 1, 2, ...\}$	粉丝转发作者 tweet 的数目
用户提及数目（Num_fMu）	$N = \{0, 1, 2, ...\}$	粉丝提及作者 tweet 的数目
用户被转发数目（Num_uRf)	$N = \{0, 1, 2, ...\}$	作者转发粉丝 tweet 的数目
用户被提及数目（Num_uMf)	$N = \{0, 1, 2, ...\}$	作者提及粉丝 tweet 的数目
用户转发比例（Ratio_retweet）	$[0, 1]$	粉丝的 tweet 中转发 tweet 的比例
用户提及比例（Ratio_mention）	$[0, 1]$	粉丝的 tweet 中提及 tweet 的比例
用户特征（FS）	**取值范围**	**Description**
发布 tweet 数目（Posts）	$N^+ = \{1, 2, 3, ...\}$	作者以往发布 tweet 的数目
粉丝数目（Followers）	$N = \{0, 1, 2, ...\}$	作者的粉丝数目
朋友数目（Friends）	$N = \{0, 1, 2, ...\}$	作者的朋友数目
分组数目（Listed）	$N = \{0, 1, 2, ...\}$	作者的分组数目
验证用户（Verified）	0 or 1	作者是否被官方验证
用户活跃时间特征（FAT）	**取值范围**	**Description**
时区时间（Timezone）	0 or 1	粉丝是否与作者在同一个时区
用户户活跃时间（PostTimeConsis）	$[0, 1]$	粉丝发布 tweet 不同时间的数目比例
用户兴趣特征（FI）	**取值范围**	**Description**
相似兴趣（SimInterest）	$(-1, 1)$	tweet 与粉丝以往发布 tweet 的相似度

输入，并且使用我们定义的主观相似性计算方法 6.5来计算三个主观相似性，只不过我们见三个值组合起来作为特征同时加入到分类器中使用评测其预测性能。

第三个阶段，考虑到影响转发行为的其他因素，比如网络结构或用户使用习惯等元数据，我们也将模型和综合其他因素的方法进行了对比。主要是和 Luo[189] 的方法（标记为“LUO”）进行了对比，表 6.5是其使用的一些特征。其中对用户兴趣建模 LUO 方法使用的是简单的词袋模型，为了验证我们方法从用户产生数据中提炼出的主观性信息的作用，我们将 LUO 方法中的用户兴趣特征替换为我们的主观相似性特征进行组合实验（使用“LUO+”前缀表示)。

实验中使用的是逻辑回归分类器（logistic regression classifier)，用 5 倍交叉验证方式（5-fold cross-validation）训练测试，评价指标采用准确率。关于基准

（baseline）设置，我们认为如果一个粉丝在曾经转发某用户的某条微博，那么他很会继续转发，因此将其预测为目标微博的转发者。评测结果如表 6.6所示。

表 6.6　准确率评测结果

Feature	Accuracy(%)	Feature	Accuracy(%)
baseline	60.85		
TF-IDF	62.85 $*$	LUO	71.76 $*$
entity	68.76 $*$	LUO+entity	72.15 $*$
hashtag	59.12	LUO+hashtag	68.44 $*$
TSM	67.44 $*$	LUO+TSM	68.23 $*$
JST	68.13 $*$	LUO+JST	70.53 $*$
$Sim(f,t)$	73.88 $*$ ‡	LUO+$Sim(f,t)$	74.04 $*$ ‡
$Sim(f,u_a)$	70.04 $*$	LUO+$Sim(f,u_a)$	70.27 $*$
$Sim(u_a,t)$	69.64 $*$	LUO+$Sim(u_a,t)$	71.86 $*$
sim_{all}	**75.64** $*$ ‡	LUO+sim_{all}	**78.15** $*$ ‡

显著性水平 ($p < 0.05$)，使用 $*$ 标记性能显著超过基准，用 ‡ 标记性能显著超过 LUO。

首先，除了 hashtag 模型外，其他模型的准确率都显著超过了基准准确率（60.85%），hashtag 模型准确率为 59.12%，主要原因是微博中 hashtag 的使用率过低而造成的数据稀疏。

第二，对比实验中，两个主观相似性指标 $Sim(f,t)$ 和 sim_{all} 准确率显著超过了 LUO 方法（71.76%），其中最高准确率为 sim_{all}（75.64%），是将三个主观相似性组合作为特征加入到分类器中，TF-IDF 模型（62.85%）仅仅比基准准确率稍好，entity 模型（68.76%）准确性接近 $Sim(f,u_a)$（70.04%）和 $Sim(u_a,t)$（69.64%），差别并不显著。

第三，两个生成模型（TSM: 67.44%，JST: 68.13%）准确率不如我们的模型，主要原因在于他们的情感表示形式是二值极性表示，不能够很好的区分不同的观点，而我们的模型采用新的在情感空间的分布表示，可以区分用户细致的观点差别，从而可以对用户转发行为的主观动机建模。

最后，在组合实验中，$Sim(f,t)$ 指标对准确率的提高显著（LUO+$Sim(f,t)$，准确率提高 2.12%），但是其他两个主观相似性以及 entity 模型加入后准确率提高不明显，加入 hashtag 和两个生成模型后准确率反而会降低，值得注意的是将三个主观相似性同时加入到 LUO 方法中准确率提高最多（LUO+sim_{all}，准确率提高 6.39%）。

总结以上分析，转发预测结果显示主观模型以及考虑三个因素的主观相似性可以很好的预测用户的转发行为，能够作为分析转发行为的有效途径。

6.5 小结

本章在前面主观模型的基础上从主观动机角度进行了转发行为的分析，提出了新的主观相似性计算方法，并通过考虑影响用户转发行为的吸引力、社交性以及流行性三个不同因素，对用户的转发行为进行分析并量化为三个主观相似性。实验结果证明了主观相似性与转发行为存在相关性，可以很好的预测用户的转发行为，对于理解用户的转发动机有帮助。

第七章　总结与展望

社交媒体已经逐步发展完善，随着用户使用社交媒体的普及，带有用户观点信息的文本数据正爆炸式增长，本文主要围绕社交媒体中观点信息的挖掘、分析以及在行为分析中的应用展开研究。通过情感词典资源的建设、情感极性分类以及利用情感分析结果对社交媒体用户的主观性建模和应用等任务，我们充分利用了社交媒体作为媒体所产生的文本特点和社交媒体用户之间的社交功能来帮助解决这些问题。

对社交媒体中文本数据中的观点信息分析研究能够从社交媒体海量数据中发现有借鉴意义的信息，无论对于其他研究还是商业应用都有价值。为了能够得到观点信息需要从文本中抽取分离出能够识别用户看法、态度、立场以及情感的表达方式，我们特别针对社交媒体的文本进行了情感知识辞典的构建和对社交媒体非规范化文本的情感分类研究，因此可以从数据中挖掘分析观点信息。

在得到文本中的观点信息后，可以帮助我们来认识作为社交媒体使用主体的用户，对用户在社交媒体上表达的观点进行集成分析，对用户的主观性进行建模。得到的主观模型对于我们理解用户的在线行为分析提供帮助。

7.1　工作总结

本文的主要工作可以从以下五个方面来总结：

首先，针对现有中文情感词典相对较少并且缺乏标准性问题，我们提出了借鉴现有的丰富的英文情感词典资源进行跨语言的情感知识转化研究。为了更准确的反映词语的情感情感极性，我们结合中文语义知识库 HowNet，将知识库中的语义关系融合进词语的情感值计算过程中，利用 HowNet 的义原与词语的中英文对应关系见英文词典 SentiWordnet 的情感知识转化为中文词语的情感知识，形成中文情感词典 SentiHowNet。

第二，仅仅依靠从词典资源标注或转化的情感知识是不够的，社交媒体语言的动态性决定需要一种能够及时从社交媒体语料数据中发现新的情感知识词典的方法，我们通过研究中文的语言规则和统计特征，实验验证从语料中抽取词语并计算情感值形成新的情感词典方法，使得情感词典可以适应社交媒体语言不断增长与变化的特点得以及时补充情感知识。

第三，从社交媒体文本尤其是一些不规范的短文本中发现观点信息需要对文本的情感倾向性进行分类，通过将情感分类问题形式化为特殊的文本分类问题，我们提出了特征空间划分的假设，并使用现成的无须标注成语资源在不同的特征

空间训练通用的分类器，并使用自举式机器学习方法解决了社交媒体的文本缺少大规模的标注文本而无法训练分类器的问题。

第四，用户在使用社交媒体时发表的信息是短小的、碎片化的，因此用户的观点信息散布在这些碎片化的文本中。以往的情感分析只是针对文本片段分析得出文本片段的观点信息，无法完整呈现出一个用户整体的观点，我们首先提出用户观点集成问题，并就这一问题提出了主观模型框架。在主观模型中，我们将用户在社交媒体中感兴趣的话题以及针对这些话题的观点进行了集成，并提出了观点的一种通用的表示形式，将观点表示为在情感空间的一种分布。主观模型可以对用户在社交媒体中的观点信息集成表示，解决了用户信息的碎片化而造成的表示不全面准确问题。

最后，针对信息传播研究中忽略的人的传播主观动机问题，结合主观模型对用户的主观性的建模分析，我们研究了 Twitter 中用户的转发行为，以此帮助理解 Twitter 中信息是如何传播的。我们将用户转发行为的主观动机量化为主观相似性，通过分析影响用户转发行为的三个因素，也就是内容上的吸引力，朋友间的社交需要以及微博信息的流行性，转化为三个主观相似性度量并分析研究了它们与转发行为之间的关系，实验证明了主观相似性度量与转发行为的相关性，以及在预测转发行为的有效性。

7.2 工作展望

展望未来，社交媒体中的观点信息的分析研究及其相关方向还有很多工作需要完成。这里总结以下亟待探索的研究方向和路线:

1. 以 Twitter 为代表的社交媒体一个重要特点就是消息的实时性，许多研究工作都围绕在 Twitter 中发现实时信息展开，包括新事件发现[191–197]、实时灾害报道（如地震、疾病、火灾等）[198–202]，另外，TREC 的 Twitter 检索[203–209]也将实时性作为一个重要指标。本文的研究中，我们并未对观点挖掘与分析受到实时性的影响进行讨论研究[?]。

2. 本文的社交媒体中观点信息的研究还是对比较常用的几个类型（比如评论和微博) 进行的研究，实际上社交媒体还有很多类型，如 Facebook[1]、YouTube[2]、Flickr[3]等等。这些社交媒体肯定有自己独特的特点，在其数据上进行观点的挖掘与分析需要研究其独特的情感表达方式；另外，多种社交媒体综合和跨媒体的信息交互与传播也会对观点信息的分析提出新课题，这就需要研究者

[1] https://www.facebook.com/
[2] https://www.youtube.com/
[3] https://www.flickr.com/

在充分理解各种社交媒体的特点和人们对各种社交媒体不同使用习惯上，提出方法解决问题。

3. 目前将观点分析结合进其他应用和任务研究是一个新的研究方向，主要是基于目前越来越多的应用和任务需要以社交媒体用户的观点作为有用的特征来使用，比如在股市指数的预测中人的情感指标会影响者投资意愿，未来我们见结合更多的和实际任务或应用有针对性的进行观点信息的分析。

总之，针对社交媒体中的观点信息的研究还有许多问题等待着去解决，我们将继续深入研究相关问题。

致　谢

“Learning is more imortant than knowing”，尤其是对于工作过一段时间的人来说，在感知到自己所知甚少时候能有机会重新学习，进行课题研究，需要感谢的人实在是太多。

首先感谢我的导师王挺教授，从一开始能够接受一名在职的考博学生，您就以开放而又严谨的治学态度给予我最大的支持，感谢您在过去的五年中精细的学术指导和研究建议，您在学术领域的专业深度和开阔视野激发了我在文本信息处理研究的巨大兴趣，感谢您让我拥有充分的研究自由，培养了我深入思考和独立解决问题的能力，这些都对于我顺利完成博士课题研究都是必不可少的。

感谢课题组的唐晋韬、周云、李岩、麻大顺、刘培磊、岳大鹏、刘海池、汝承森、张文文、姜仁会、胡长龙、李欣奕，和大家一起亦师亦友共同探讨自然语言处理领域最前沿的问题让我获益非浅，在艰辛的求学到路上大家互相帮助，苦中作乐的日子让我重新体会到了无私的同学友谊。

感谢计算机学院的学院领导在工作、考博和学习期间给我的关怀和指导；感谢学员大队的同事，在我求学阶段给我指导、鼓励还有协助；感谢博士队队领导和各位博士战友，在一起“共同战斗”的日子永远值得回味。

最后，也是做重要的，感谢我的家人，没有家人的支持就没有我顺利的博士学习研究：感谢我的妻子黄丽达，感谢你牺牲自己的工作学习对我的支持；感谢我五岁的女儿，你的出生给我生活带来无尽的乐趣；感谢我的岳父母，在我读博期间对我们这个小家的生活上无微不至的照顾；感谢在山东的父母，远在千里你们的关爱依然！

参考文献

[1] Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media [J]. Business horizons. 2010, 53 (1): 59–68.

[2] Eisenstein J. What to do about bad language on the internet [C]. In Proceedings of NAACL-HLT. 2013: 359–369.

[3] Jensen D, Neville J. Linkage and autocorrelation cause feature selection bias in relational learning [C]. In ICML. 2002: 259–266.

[4] Taskar B, Abbeel P, Wong M-F, et al. Label and link prediction in relational data [C]. In Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data. 2003.

[5] Agichtein E, Castillo C, Donato D, et al. Finding high-quality content in social media [C]. In Proceedings of the international conference on Web search and web data mining. 2008: 183–194.

[6] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks [C]. In Proceedings of the 26th Annual Computer Security Applications Conference. 2010: 1–9.

[7] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks [C]. In Proceedings of the 19th international conference on World wide web. 2010: 981–990.

[8] Rossion B, Delvenne J-F, Debatisse D, et al. Spatio-temporal localization of the face inversion effect: an event-related potentials study [J]. Biological psychology. 1999, 50 (3): 173–189.

[9] Speriosu M, Sudan N, Upadhyay S, et al. Twitter polarity classification with label propagation over lexical links and the follower graph [C]. In Proceedings of the First workshop on Unsupervised Learning in NLP. 2011: 53–63.

[10] Mislove A, Viswanath B, Gummadi K P, et al. You are who you know: inferring user profiles in online social networks [C]. In Proceedings of the third ACM international conference on Web search and data mining. 2010: 251–260.

[11] Lyons J. Semantics. 2 vols. 1977.

[12] Wiebe J, Wilson T, Bruce R, et al. Learning subjective language [J]. Computational linguistics. 2004, 30 (3): 277–308.

[13] Rachels J, Rachels S. The elements of moral philosophy [M]. Random House New York, 1986.

[14] Hoffman T. Online reputation management is hot – but is it ethical? [EB/OL]. 2008. http://www.computerworld.com/article/2537007/networking/online-reputation-management-is-hot----but-is-it-ethical-.html.

[15] HORRIGAN J. Online Shopping [EB/OL]. 2008. http://www.pewinternet.org/2008/02/13/online-shopping/.

[16] Mullen T, Malouf R. A preliminary investigation into sentiment analysis of informal political discourse [C/OL]. In AAAI symposium on computational approaches to analysing weblogs (AAAI-CAAW). 2006: 159–162. http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-031.pdf.

[17] Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards [J]. The Journal of Finance. 2004, 59 (3): 1259–1294.

[18] Archak N, Ghose A, Ipeirotis P G. Show me the money!: deriving the pricing power of product features by mining consumer reviews [C]. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007: 56–65.

[19] Chevalier J, Mayzlin D. The effect of word of mouth on sales: Online book reviews [J]. J. Marketing Res. 2006: 345–354.

[20] Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. [J]. ICWSM. 2010, 10: 178–185.

[21] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market [J]. Journal of Computational Science. 2011, 2 (1): 1–8.

[22] Pang B, Lee L. Opinion Mining and Sentiment Analysis [J]. Found. Trends Inf. Retr. 2008, 2 (1-2): 1–135.

[23] Liu B. Sentiment analysis and opinion mining [J]. Synthesis Lectures on Human Language Technologies. 2012, 5 (1): 1–167.

[24] Kim S-M, Hovy E. Determining the sentiment of opinions [C]. In Proceedings of the 20th international conference on Computational Linguistics. 2004: 1367.

[25] He B, Macdonald C, He J, et al. An effective statistical approach to blog post opinion retrieval [C]. In Proceedings of the 17th ACM conference on Information and knowledge management. 2008: 1063–1072.

[26] Macdonald C, Ounis I, Soboroff I. Overview of the TREC 2007 Blog Track. [C]. In TREC. 2007: 31–43.

[27] Ounis I, Macdonald C, Lin J, et al. Overview of the trec-2011 microblog track [C]. In Proceeddings of the 20th Text REtrieval Conference (TREC 2011). 2011.

[28] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network [C]. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. 2003: 173–180.

[29] Gimpel K, Schneider N, O'Connor B, et al. Part-of-speech tagging for twitter: Annotation, features, and experiments [R]. 2010.

[30] Owoputi O, O'Connor B, Dyer C, et al. Improved part-of-speech tagging for online conversational text with word clusters [C]. In Proceedings of NAACL-HLT. 2013: 380–390.

[31] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling [C]. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005: 363–370.

[32] Ritter A, Clark S, Etzioni O, et al. Named entity recognition in tweets: an experimental study [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 1524–1534.

[33] Foster J, Cetinoglu O, Wagner J, et al. From news to comment: Resources and benchmarks for parsing the language of web 2.0 [J]. 2011.

[34] Han B, Baldwin T. Lexical Normalisation of Short Text Messages: Makn Sens a# twitter. [C]. In ACL. 2011: 368–378.

[35] Han B, Cook P, Baldwin T. Automatically constructing a normalisation dictionary for microblogs [C]. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 421–432.

[36] Han B, Cook P, Baldwin T. Lexical normalization for social media text [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, 4 (1): 5.

[37] Liu X, Zhou M, Wei F, et al. Joint inference of named entity recognition and normalization for tweets [C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012: 526–535.

[38] Liu F, Weng F, Jiang X. A broad-coverage normalization system for social media language [C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012: 1035–1044.

[39] Hassan H, Menezes A. Social text normalization using contextual graph random walks [C]. In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics. 2013.

[40] Finin T, Murnane W, Karandikar A, et al. Annotating named entities in Twitter data with crowdsourcing [C]. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010: 80–88.

[41] Liu X, Zhang S, Wei F, et al. Recognizing Named Entities in Tweets. [C]. In ACL. 2011: 359–367.

[42] Li C, Weng J, He Q, et al. TwiNER: named entity recognition in targeted twitter stream [C]. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012: 721–730.

[43] Liu X, Wei F, Zhang S, et al. Named entity recognition for tweets [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, 4 (1): 3.

[44] Liu X, Zhou M. Two-stage NER for tweets with clustering [J]. Information Processing & Management. 2012.

[45] Sharifi B, Hutton M-A, Kalita J. Summarizing microblogs automatically [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 685–688.

[46] Chakrabarti D, Punera K. Event Summarization Using Tweets. [C]. In ICWSM. 2011.

[47] Takamura H, Yokono H, Okumura M. Summarizing a document stream [C]. In Proceedings of the 33rd European conference on Advances in information retrieval. 2011: 177–188.

[48] Weng J-Y, Yang C-L, Chen B-N, et al. IMASS: an intelligent microblog analysis and summarization system [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. 2011: 133–138.

[49] Harabagiu S M, Hickl A. Relevance Modeling for Microblog Summarization. [C]. In ICWSM. 2011.

[50] Ren Z, Liang S, Meij E, et al. Personalized time-aware tweets summarization [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, 2013: 513–522.

[51] Shen C, Liu F, Weng F, et al. A Participant-based Approach for Event Summarization Using Twitter Streams [C]. In Proceedings of NAACL-HLT. 2013: 1152–1162.

[52] Judd J, Kalita J. Better Twitter Summaries? [C]. In Proceedings of NAACL-HLT. 2013: 445–449.

[53] Chang Y, Wang X, Mei Q, et al. Towards Twitter context summarization with user influence models [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 527–536.

[54] Schler J, Schler J. The importance of neutral examples for learning sentiment [C]. In In Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations (FINEXIN. 2005.

[55] Wiebe J, Wilson T, Bell M. Identifying Collocations for Recognizing Opinions [C]. In In Proc. ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. 2001: 24–31.

[56] Wiebe J, Riloff E. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts [C/OL]. In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg, 2005: 486–497. http://dx.doi.org/10.1007/978-3-540-30586-6_53.

[57] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C]. In Proceedings of the 12th international conference on World Wide Web. 2003: 519–528.

[58] Pang B, Lee L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts [C/OL]. In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA, 2004. http://dx.doi.org/10.3115/1218955.1218990.

[59] Riloff E, Wiebe J, Phillips W. Exploiting subjectivity classification to improve information extraction [C]. In Proceedings of the National Conference On Artificial Intelligence. 2005: 1106.

[60] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis [J]. Computational linguistics. 2009, 35 (3): 399–433.

[61] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [C/OL]. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. Stroudsburg, PA, USA, 2002: 79–86. http://dx.doi.org/10.3115/1118693.1118704.

[62] Melville P, Gryc W, Bldg W, et al. Sentiment analysis of blogs by combining lexical knowledge with text classification [C]. In In KDD. 2009: 1275–1284.

[63] Vegnaduzzo S. Acquisition of subjective adjectives with limited resources [C]. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. 2004.

[64] Devitt A, Ahmad K. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach [C]. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007.

[65] Osherenko A, André E. Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect? [C/OL]. In Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction. Berlin, Heidelberg, 2007: 230–241. http://dx.doi.org/10.1007/978-3-540-74889-2_21.

[66] Goldberg A B, Zhu X. Seeing Stars when There Aren'T Many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization [C/OL]. In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. Stroudsburg, PA, USA, 2006: 45–52. http://dl.acm.org/citation.cfm?id=1654758.1654769.

[67] Täckström O, McDonald R. Discovering Fine-grained Sentiment with Latent Variable Structured Prediction Models [C/OL]. In Proceedings of the 33rd European Conference on Advances in Information Retrieval. Berlin, Heidelberg, 2011: 368–374. http://dl.acm.org/citation.cfm?id=1996889.1996937.

[68] Mcdonald R, Hannan K, Neylon T, et al. Structured Models for Fine-to-Coarse Sentiment Analysis [C]. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007.

[69] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. [C]. 2010: 2200–2204.

[70] Angela. Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives [J/OL]. http://www.aisb.org.uk/convention/aisb08/proc/proceedings/02%20Affective%20Language/11.pdf.

[71] Tsytsarau M, Palpanas T, Denecke K. Scalable Discovery of Contradictions on the Web [C/OL]. In Proceedings of the 19th International Conference on World Wide Web. New York, NY, USA, 2010: 1195–1196. http://doi.acm.org/10.1145/1772690.1772871.

[72] Missen M，Boughanem M．Using WordNet’s Semantic Relations for Opinion Detection in Blogs [M] // Boughanem M，Berrut C，Mothe J，等．Advances in Information Retrieval：第 5478 卷．Springer Berlin Heidelberg，2009：2009: 729–733.

[73] Zhu J, Zhu M, Wang H, et al. Aspect-based Sentence Segmentation for Sentiment Summarization [C/OL]. In Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion. New York, NY, USA, 2009: 65–72. http://doi.acm.org/10.1145/1651461.1651474.

[74] Yi J, Nasukawa T, Bunescu R, et al. Sentiment Analyzer: Extracting Sentiments About a Given Topic Using Natural Language Processing Techniques [C/OL]. In Proceedings of the Third IEEE International Conference on Data Mining. Washington, DC, USA, 2003: 427–. http://dl.acm.org/citation.cfm?id=951949.952133.

[75] Thet T T, Na J-C, Khoo C S, et al. Sentiment Analysis of Movie Reviews on Discussion Boards Using a Linguistic Approach [C/OL]. In Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion. New York, NY, USA, 2009: 81–84. http://doi.acm.org/10.1145/1651461.1651476.

[76] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]. 2002: 417–424.

[77] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems (TOIS). 2003, 21 (4): 315–346.

[78] Church K W, Hanks P. Word Association Norms, Mutual Information, and Lexicography [J/OL]. Comput. Linguist. 1990, 16 (1): 22–29. http://dl.acm.org/citation.cfm?id=89086.89095.

[79] Chaovalit P, Zhou L. Movie review mining: A comparison between supervised and unsupervised classification approaches [C]. In Proceedings of the Hawaii International Conference on System Sciences (HICSS). 2005.

[80] Read J, Carroll J. Weakly Supervised Techniques for Domain-independent Sentiment Classification [C/OL]. In Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion. New York, NY, USA, 2009: 45–52. http://doi.acm.org/10.1145/1651461.1651470.

[81] Taboada M, Anthony C, Voll K. Methods for Creating Semantic Orientation Dictionaries [C]. In Conference on Language Resources and Evaluation (LREC). 2006: 427–432.

[82] Tang H, Tan S, Cheng X. A Survey on Sentiment Detection of Reviews [J/OL]. Expert Syst. Appl. 2009, 36 (7): 10760–10773. http://dx.doi.org/10.1016/j.eswa.2009.02.063.

[83] Liu B, Hu M, Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web [C/OL]. In Proceedings of the 14th International Conference on World Wide Web. New York, NY, USA, 2005: 342–351. http://doi.acm.org/10.1145/1060745.1060797.

[84] Ng R, Pauls A. Multi-document summarization of evaluative text [C]. In In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics. 2006: 3–7.

[85] Leouski A, Croft W. An evaluation of techniques for clustering search results [J]. 1996.

[86] Zeng H J, He Q C, Chen Z, et al. Learning to cluster web search results [C]. In Proceedings of the 27th annual international conference on Research and development in information retrieval. Sheffield, United Kingdom, 2004: 210–217.

[87] Su H, Mei Q, Zhai C. A probabilistic approach to spatiotemporal theme pattern mining on Weblogs. [C]. In Proceedings of the 15th International Conference on World Wide Web. 2006.

[88] Titov I, McDonald R. Modeling Online Reviews with Multi-grain Topic Models [C/OL]. In Proceedings of the 17th International Conference on World Wide Web. New York, NY, USA, 2008: 111–120. http://doi.acm.org/10.1145/1367497.1367513.

[89] Popescu A-M，Etzioni O．Extracting Product Features and Opinions from Reviews [M] // Kao A，Poteet S．Natural Language Processing and Text Mining．Springer London，2007：2007: 9–28．

[90] boyd d, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter [C]. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on. 2010: 1–10.

[91] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks [C]. In Proceedings of the 19th ACM international conference on Information and knowledge management. 2010: 1633–1636.

[92] Macskassy S A, Michelson M. Why do people retweet? anti-homophily wins the day! [C]. In ICWSM. 2011.

[93] Starbird K, Palen L. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising [C]. In Proceedings of the acm 2012 conference on computer supported cooperative work. 2012: 7–16.

[94] Comarela G, Crovella M, Almeida V, et al. Understanding factors that affect response rates in twitter [C]. In Proceedings of the 23rd ACM conference on Hypertext and social media. 2012: 123–132.

[95] Kupavskii A, Umnov A, Gusev G, et al. Predicting the Audience Size of a Tweet [C]. In Seventh International AAAI Conference on Weblogs and Social Media. 2013.

[96] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 657–664.

[97] Ahmed M, Spagna S, Huici F, et al. A peek into the future: predicting the evolution of popularity in user generated content [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 607–616.

[98] Bao P, Shen H-W, Huang J, et al. Popularity prediction in microblogging network: a case study on sina weibo [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 177–178.

[99] Yuan S, Wang J, van der Meer M. Adaptive Keywords Extraction with Contextual Bandits for Advertising on Parked Domains [J]. Computing Research Repository. 2013, abs/1307.3573.

[100] 张辉, 李国辉, 贾立, et al. 一种基于 TF· IEF 模型的在线新闻事件探测方法 [J]. 国防科技大学学报. 2013, 35 (3): 55–60.

[101] 刘健, 李绮, 刘宝宏, et al. 基于话题模型的专家发现方法 [J]. 国防科技大学学报. 2013, 35 (2): 127–131.

[102] 黄萱菁, 张奇, 吴苑斌. 文本情感倾向分析 [J]. 中文信息学报. 2011, 25 (6): 118–126.

[103] Stone P J, Dunphy D C, Smith M S. The General Inquirer: A Computer Approach to Content Analysis. [M]. Cambridge, Massachusetts: MIT press, 1966.

[104] Wilson T, Hoffmann P, Somasundaran S, et al. OpinionFinder: A system for subjectivity analysis [C]. In Proceedings of hlt/emnlp on interactive demonstrations. 2005: 34–35.

[105] Taboada M, Grieve J. Analyzing Appraisal Automatically [C]. Stanford University, Stanford California, 2004.

[106] Agerri R, Garc I A-Serrano A. Q-WordNet: Extracting Polarity from WordNet Senses. [C]. Valletta, Malta, 2010.

[107] 朱嫣岚, 闵锦, 周雅倩, et al. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报. 2006, 20 (1): 14–20.

[108] 朱征宇, 孙俊华. 改进的基于《知网》的词汇语义相似度计算 [J]. 计算机应用. 2013 (08): 2276–2279+2288. 页数: 5.

[109] 黄硕, 周延泉. 基于知网和同义词词林的词汇语义倾向计算 [J]. 软件. 2013, 34 (2): 73–74,94.

[110] 知网 HowNet 评价词词典. 2013.

[111] Ku L W, Chen H H. Mining opinions from the Web: Beyond relevance retrieval [J]. Journal of the American Society for Information Science and Technology. 2007, 58 (12): 1838–1850.

[112] 情感词汇本体库. 2013.

[113] 刘群, 李素建. 基于《知网》的词汇语义相似度计算 [C]. 中国台北, 2002.

[114] Fellbaum C. WordNet: An Electronic Lexical Database [M]. Cambridge, Massachusetts: MIT Press, 1998.

[115] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives [C]. 1997: 174–181.

[116] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis [C]. 2005: 347–354.

[117] Bradley M M, Lang P J. Affective norms for English words (ANEW): Instruction manual and affective ratings [R]. 1999.

[118] Nielsen F A R. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs [J]. arXiv preprint arXiv:1103.2903. 2011.

[119] Esuli A, Sebastiani F. Sentiwordnet: A publicly available lexical resource for opinion mining [C]. 2006: 417–422.

[120] Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web [J]. Journal of the American Society for Information Science and Technology. 2012, 63 (1): 163–173.

[121] Plutchik R. The nature of emotions [J]. American Scientist. 2001, 89 (4): 344–350.

[122] Mohammad S M, Turney P D. Crowdsourcing a word–emotion association lexicon [J]. Computational Intelligence. 2013, 29 (3): 436–465.

[123] Mohammad S M, Kiritchenko S, Zhu X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets [C]. 2013.

[124] Kiritchenko S, Zhu X, Cherry C, et al. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews [C]. 2014.

[125] Tsai A C-R, Wu C-E, Tsai R T-H, et al. Building a concept-level sentiment dictionary based on commonsense knowledge [J]. IEEE Intelligent Systems. 2013, 28 (2): 22–30.

[126] Cambria E, Olsher D, Rajagopal D. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis [C]. In Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[127] 谢松县, 刘博, 王挺. 应用语义关系自动构建情感词典 [J]. 国防科技大学学报. 2014, 36 (3): 111–115.

[128] Lin Z, Tan S, Cheng X, et al. Effective and efficient?: bilingual sentiment lexicon extraction using collocation alignment [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. New York, NY, USA, 2012: 1542–1546.

[129] 张华平. NLPIR/ICTCLAS2014 分词系统. 08-01 2014.

[130] 鲁松, 白硕. 自然语言处理中词语上下文有效范围的定量描述 [J]. 计算机学报. 2001, 24 (7): 742–747.

[131] Lourenco Jr R, Veloso A, Pereira A, et al. Economically-efficient sentiment stream analysis [C]. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014: 637–646.

[132] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data [C]. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. 2010: 36–44.

[133] Hu X, Tang J, Gao H, et al. Unsupervised sentiment analysis with emotional signals [C]. In Proc. of the 22nd WWW. 2013: 607–618.

[134] Wang X, Wei F, Liu X, et al. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach [C]. In Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 1031–1040.

[135] Asiaee T A, Tepper M, Banerjee A, et al. If you are happy and you know it... tweet [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 1602–1606.

[136] Hu X, Tang L, Tang J, et al. Exploiting social relations for sentiment analysis in microblogging [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 537–546.

[137] Calais Guerra P H, Veloso A, Meira Jr W, et al. From bias to opinion: a transfer-learning approach to real-time sentiment analysis [C]. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 150–158.

[138] Thelwall M, Buckley K, Paltoglou G, et al. Sentiment strength detection in short informal text [J]. Journal of the American Society for Information Science and Technology. 2010, 61 (12): 2544–2558.

[139] Go A, Bhayani R, Huang L. Twitter Sentiment Classification using Distant Supervision [J]. Processing. 2009: 1–6.

[140] Marchetti-Bowick M, Chambers N. Learning for microblogs with distant supervision: Political forecasting with twitter [C]. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012: 603–612.

[141] Loper E, Bird S. NLTK: The Natural Language Toolkit [EB/OL]. 2002. http://arxiv.org/abs/cs/0205028.

[142] Chang C-C, Lin C-J. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology. 2011, 2: 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[143] Lu Y, Zhai C. Opinion integration through semi-supervised topic modeling [C]. In Proceedings of the 17th international conference on World Wide Web. 2008: 121–130.

[144] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys [C]. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. 2010: 241–249.

[145] Jiang L, Yu M, Zhou M, et al. Target-dependent twitter sentiment classification [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. 2011: 151–160.

[146] Li G, Hoi S C, Chang K, et al. Micro-blogging sentiment detection by collaborative online learning [C]. In Data Mining (ICDM), 2010 IEEE 10th International Conference on. 2010: 893–898.

[147] Tan C, Lee L, Tang J, et al. User-level sentiment analysis incorporating social networks [C]. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 1397–1405.

[148] Mostafa M M. More than words: Social networks’ text mining for consumer brand sentiments [J]. Expert Systems with Applications. 2013, 40 (10): 4241–4251.

[149] Malouf R, Mullen T. Taking sides: User classification for informal online political discourse [J]. Internet Research. 2008, 18 (2): 177–190.

[150] Liu H, Zhao Y, Qin B, et al. Comment target extraction and sentiment classification [J]. Journal of Chinese Information Processing. 2010, 24 (1): 84–89.

[151] Zhai Z, Liu B, Xu H, et al. Constrained LDA for grouping product features in opinion mining [M] // Zhai Z, Liu B, Xu H, et al. Advances in knowledge discovery and data mining. Springer, 2011: 2011: 448–459.

[152] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. the Journal of machine Learning research. 2003, 3: 993–1022.

[153] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents [C]. In Proceedings of the 20th conference on Uncertainty in artificial intelligence. 2004: 487–494.

[154] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora [C]. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. 2009: 248–256.

[155] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs [C]. In Proceedings of the 16th international conference on World Wide Web. 2007: 171–180.

[156] Lin C, He Y. Joint sentiment/topic model for sentiment analysis [C]. In Proceedings of the 18th ACM conference on Information and knowledge management. 2009: 375–384.

[157] Hannon J, Bennett M, Smyth B. Recommending twitter users to follow using content and collaborative filtering approaches [C]. In Proc. of the 4th ACM ReSys. 2010: 199–206.

[158] Ramage D, Dumais S, Liebling D. Characterizing Microblogs with Topic Models [C]. In ICWSM. 2010.

[159] Xu Z, Zhang Y, Wu Y, et al. Modeling user posting behavior on social media [C]. In Proc. of the 35th ACM SIGIR. 2012: 545–554.

[160] Pennacchiotti M, Popescu A-M. A Machine Learning Approach to Twitter User Classification. [C]. In ICWSM. 2011.

[161] Engbert K, Wohlschläger A, Thomas R, et al. Agency, subjective time, and other minds. [J]. Journal of Experimental Psychology: Human Perception and Performance. 2007, 33 (6): 1261.

[162] Stein D, Wright S. Subjectivity and Subjectivisation: Linguistic Perspectives [M/OL]. Cambridge University Press, 2005. http://books.google.com.hk/books?id=mWlS5Q8uBYcC.

[163] Cambria E, White B. Jumping NLP curves: A review of natural language processing research [J]. IEEE Computational Intelligence Magazine. 2014, 9 (2): 48–57.

[164] Lin C, He Y, Everson R. A comparative study of Bayesian models for unsupervised sentiment detection [C]. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning. 2010: 144–152.

[165] Chen J, Nairn R, Nelson L, et al. Short and tweet: experiments on recommending content from information streams [C]. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems. 2010: 1185–1194.

[166] Abel F, Gao Q, Houben G-J, et al. Analyzing user modeling on twitter for personalized news recommendations [M] // Abel F, Gao Q, Houben G-J, et al. UMAP. Springer, 2011: 2011: 1–12.

[167] Gangemi A, Presutti V, Reforgiato Recupero D. Frame-Based Detection of Opinion Holders and Topics: A Model and a Tool [J]. Computational Intelligence Magazine, IEEE. 2014, 9 (1): 20–30.

[168] Weng J, Lim E-P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers [C]. In Proc. of the third ACM WSDM. 2010: 261–270.

[169] Řehůřek R，Sojka P. Software Framework for Topic Modelling with Large Corpora [C]. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta，May 2010: 45–50. http://is.muni.cz/publication/884893/en.

[170] Walton D N. Bias, critical doubt and fallacies [J]. Argumentation and Advocacy. 1991, 28: 1–22.

[171] Li R, Wang S, Deng H, et al. Towards social user profiling: unified and discriminative influence model for inferring home locations [C]. In KDD. 2012: 1023–1031.

[172] Lazarsfeld P F, Merton R K. Friendship as a social process: A substantive and methodological analysis [M] // Berger M, Abel T. Freedom and control in modern society. New York: Van Nostrand, 1954, 1954:.

[173] McPherson M, Smith-Lovin L, Cook J M. Birds of a feather: Homophily in social networks [J]. Annual review of sociology. 2001: 415–444.

[174] Thelwall M. Emotion homophily in social network site messages [J]. First Monday. 2010, 15 (4).

[175] 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.

[176] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network [C]. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003: 137–146.

[177] Cheng J, Adamic L, Dow P A, et al. Can cascades be predicted? [C]. In Proceedings of the 23rd international conference on World wide web. 2014: 925–936.

[178] Moore J, Haggard P. Awareness of action: Inference and prediction [J]. Consciousness and cognition. 2008, 17 (1): 136–144.

[179] Hyman J. Three Fallacies about Action [J]. Behavioral and Brain Sciences. 2000, 23: 665–666.

[180] Feng S, Zhang L, Li B, et al. Is Twitter A Better Corpus for Measuring Sentiment Similarity? [C]. In EMNLP'13. 2013: 897–902.

[181] Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network [C]. In Social Computing (SocialCom), 2010 IEEE Second International Conference on. 2010: 177–184.

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

[183] Naveed N, Gottron T, Kunegis J, et al. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter [C]. In WebSci '11: Proceedings of the 3rd International Conference on WebScience. 2011.

[184] Naveed N, Gottron T, Kunegis J, et al. Searching microblogs: coping with sparsity and document quality [C]. In Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 183–188.

[185] Feng W, Wang J. Retweet or not?: personalized tweet re-ranking [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 577–586.

[186] Pfitzner R, Garas A, Schweitzer F. Emotional Divergence Influences Information Spreading in Twitter. [C]. In ICWSM. 2012.

[187] Lek H H, Poo D C. Aspect-Based Twitter Sentiment Classification [C]. In Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference. 2013: 366–373.

[188] Cialdini R B, Goldstein N J. Social influence: Compliance and conformity [J]. Annu. Rev. Psychol. 2004, 55: 591–621.

[189] Luo Z, Osborne M, Tang J, et al. Who will retweet me?: finding retweeters in twitter [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 869–872.

[190] Fisher S R A, Genetiker S, Fisher R A, et al. Statistical methods for research workers [M]. Oliver and Boyd Edinburgh, 1970.

[191] Petrović S, Osborne M, Lavrenko V. Streaming first story detection with application to twitter [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 181–189.

[192] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter. [C]. In ICWSM. 2011.

[193] Weng J, Lee B-S. Event Detection in Twitter. [C]. In ICWSM. 2011.

[194] Naaman M, Becker H, Gravano L. Hip and trendy: Characterizing emerging trends on Twitter [J]. Journal of the American Society for Information Science and Technology. 2011, 62 (5): 902–918.

[195] Benson E, Haghighi A, Barzilay R. Event discovery in social media feeds [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. 2011: 389–398.

[196] Petrović S, Osborne M, Lavrenko V. Using paraphrases for improving first story detection in news and Twitter [C]. In Proceedings of The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 338–346.

[197] Kanhabua N, Nejdl W. Understanding the Diversity of Tweets in the Time of Outbreaks [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 1335–1342.

[198] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors [C]. In Proceedings of the 19th international conference on World wide web. 2010: 851–860.

[199] Paul M J, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. [C]. In ICWSM. 2011.

[200] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: Detecting influenza epidemics using twitter [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 1568–1576.

[201] Abel F, Hauff C, Houben G-J, et al. Twitcident: fighting fire with information from social web streams [C]. In Proceedings of the 21st international conference companion on World Wide Web. 2012: 305–308.

[202] Yin J, Karimi S, Robinson B, et al. ESA: emergency situation awareness via microbloggers [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2701–2703.

[203] Efron M, Golovchinsky G. Estimation methods for ranking recent information [C]. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 495–504.

[204] Metzler D, Cai C, Hovy E. Structured event retrieval over microblog archives [C]. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 646–655.

[205] Zhang X, He B, Luo T, et al. Query-biased learning to rank for real-time twitter search [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 1915–1919.

[206] Soboroff I, McCullough D, Lin J, et al. Evaluating real-time search over tweets [J]. Proc. ICWSM. 2012: 943–961.

[207] Choi J, Croft W B. Temporal models for microblogs [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2491–2494.

[208] Amati G, Amodeo G, Gaibisso C. Survival analysis for freshness in microblogging search [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2483–2486.

[209] Miyanishi T, Seki K, Uehara K. Combining recency and topic-dependent temporal variation for microblog search [M] // Miyanishi T, Seki K, Uehara K. Advances in Information Retrieval. Springer, 2013: 2013: 331–343.

作者在学期间取得的学术成果

发表的学术论文

[1] Songchen Xie and Ting Wang. Construction of Unsupervised Sentiment Classifier on Idioms Resources. In *Journal of Central South University*, (2014) 21: 1376−1384, Springer. （**SCI 期刊，影响因子 0.496**）

[2] Songxian Xie, Jintao Tang, Ting Wang. Resonance Elicits Diffusion: Modeling Subjectivity for Retweeting Behavior Analysis. In *Cognitive Computation*, Published online July-14 2014, Springer.（**SCI 期刊，影响因子 1.100**）

[3] Songchen Xie and Ting Wang. Dividing for Combination: A Bootstrapping Sentiment Classification Framework for Microblogs. In *Proceedings of the 2013 International Conference on Information Science and Cloud Computing* (**ISCC2013**), Guangzhou, China, Dec 2013.

[4] Songchen Xie, Jintao Tang and Ting Wang. Topic Related Opinion Integration for Users of Social Media. In *Proceedings of the 3rd National Conference of Social Media Processing* (**SMP 2014**), Beijing, China, Nov 2014.（**CCF C 类会议，社交媒体领域重要会议**）

[5] 谢松县, 刘博, 王挺. 应用语义关系自动构建情感词典. 国防科技大学学报.2014, 36 (3): 111–115.